Tellus
SERIES A DYNAMIC METEOROLOGY AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Tellus (2011), 63A, 415–428
Printed in Singapore.
TELLUS

EuroTEPS — a targeted version of ECMWF EPS for the European area

By INGER-LISE FROGNER[1,*] and TROND IVERSEN[1,2] [1]*Norwegian Meteorological Institute, P.O. Box 43 Blindern, NO-0313 Oslo, Norway* [2]*University of Oslo, dep. Geosciences, P.O. Box 1072 Blindern, NO-0316 Oslo, Norway*

(Manuscript received 30 April 2010; in final form 10 December 2010)

ABSTRACT

A global ensemble prediction system (EPS) needs initial perturbations to account for the uncertainties in the analysis. In this study singular vectors (SVs) are used to create these perturbations. In the operational EPS at the European Centre for Medium-Range Weather Forecasts (ECMWF) the SVs are computed over the whole globe. The aim of this study is to evaluate the possible benefit of complementing the regular SVs with higher resolution SVs that are targeted to three European areas after 24 h. The system thus adds perturbations that puts Europe in focus, and produces ensemble spread, in our area of interest for short-range forecasts without spending computational resources on ensemble spread elsewhere. This system is called European Targeted Ensemble Prediction System (EuroTEPS), and is evaluated and compared to ECMWF's operational EPS over Europe for 9 weeks in 2008. The spread/skill relationship for mean sea level pressure is better for EuroTEPS than for EPS, and there is a considerable reduction in the ensemble mean forecast error. For most parameters and scores, EuroTEPS is comparable or better than ECMWF EPS over the European region for the first two forecast days. Further developments of the system are suggested.

1. Introduction

It was formally established almost half a century ago that the atmosphere is a chaotic system (Lorenz, 1963), in the sense that small errors in the initial conditions for predictions will amplify as the forecast evolves. These errors can only be approximately estimated, and there exists an infinite range of plausible initial conditions that are consistent with the uncertainty due to in situ and remote observations and their limited spatial coverage. A single model forecast from a 'best guess' initial state gives only one of the similar infinite ranges of possible future atmospheric developments. This knowledge has over the years led to development of ensemble prediction systems (EPSs) where carefully selected initial perturbations are added to the analysis in order to span the possible future states of the atmosphere (e.g. Buizza et al., 1993; Toth and Kalnay, 1993). For this several model runs are needed, but as running the model several times is expensive, much effort is dedicated to the generation of initial perturbations which will give the best system possible with as few members as possible. Uncertainty also originates from approximations to the governing equations and their boundary conditions. Numerical truncation errors are inevitable, and since there are no known lower limits for the scale truncation beyond which there is negligible upscale cascade due to non-linearity (Fjørtoft, 1953), the governing equations behaves semi-stochastically in practice despite their basically deterministic nature (e.g. Palmer and Williams, 2010). Model uncertainty is therefore also accounted for in weather prediction in different ways, for example by adding random tendencies to the equations (e.g. Buizza et al., 1999), or by combining several model versions (e.g. Palmer et al., 2004).

Taking into account the chaotic nature of weather by running EPS does not only provide us with different possible solutions of the future atmospheric states, but also gives us information about the predictability of the atmosphere and provides probabilities of the occurrence or non-occurrence of weather events. Furthermore, by combining information from all the ensemble members, for example, into an ensemble mean forecast, unpredictable components can be filtered from the forecasts. An EPS therefore contains a range of potentially important information compared to single deterministic forecast.

One such EPS is developed at the European Centre for Medium Range Weather Forecasts (ECMWF). It has proven to be a very useful tool in weather forecasting for almost two decades (see e.g. Vitart et al., 2008). This system uses singular vectors (SVs) that maximize the total perturbation energy over 48 h and the extratropical northern hemisphere for computing the initial perturbations (Buizza et al., 1993). Over the years the

*Corresponding author.
e-mail: inger.lise.frogner@met.no
DOI: 10.1111/j.1600-0870.2010.00504.x

system has improved substantially, including improved data assimilation techniques, increased number of observations going into the assimilation, model upgrades, increased horizontal and vertical resolution, and improved methods for constructing the ensemble system itself. As a global system EPS seeks to be a valuable probabilistic system in all parts of the globe. Its forecast target is mainly the medium range (day 3–10). ECMWF has, for example, also emphasis on predicting tropical cyclones (Anwender et al., 2010).

In this paper, we investigate the properties and scores of a special version of EPS (EuroTEPS) that is optimal for Europe and for the shorter forecast lead times (1–2 d). Following the method proposed by Leutbecher (2007), we combine several different sets of SVs to compute the initial perturbations. For this we use a targeting procedure giving perturbations with maximum total energy amplitudes over Europe at the end of the optimization period chosen (24 h; Buizza, 1994). The intention is to obtain perturbations, and hence ensemble spread, which are optimal for our area of interest with less computational resources spent to obtain ensemble spread elsewhere. Furthermore, fewer ensemble members may be needed to produce the same spread in Europe than if the operational EPS is used. By forcing the targeted singular vectors (TSVs) to be orthogonal to the regular operational SVs optimized over 48 h and combining the TSVs and the SVs, ensemble spread is obtained beyond the 24 h optimization interval for the TSVs. In this paper, we evaluate the predictive properties of EuroTEPS relative to the operational ECMWF EPS as it was before the latest resolution increase in January 2010. Further work is underway to extend EuroTEPS in full consistency with the latest (and upcoming) EPS, which will be evaluated and published later.

Reliable predictions of severe, high-impact weather events are increasingly emphasized in numerical weather prediction. However, although the leading global EPSs now have considerable resolution, many subsynoptic and mesoscale processes are still not well represented in these models. Since the spatial and temporal extension of severe weather phenomena often is very limited, several EPSs with limited-area models have been developed and become operational over the last years (Marsigli et al., 2005; Frogner et al., 2006; Bowler et al., 2008; Aspelien et al., 2011), and even further downscaling to approximately convection-permitting scales are under development (e.g. Gebhardt et al., 2008; Schwartz et al., 2010; Kristiansen et al., 2011). This is also the rationale behind the invention of Grand Limited Area Ensemble Prediction System (GLAMEPS) (Iversen et al., 2011), for which EuroTEPS basically is developed. GLAMEPS aims at an operational pan-European probabilistic forecasting system for the short range on subsynoptic scales, with emphasis on reliable and valuable forecast probabilities for weather events. The purpose of EuroTEPS is to provide a basis for sufficient ensemble spread over Europe for the first 2 d of prediction, with as few ensemble members as possible. EuroTEPS will provide initial and lateral boundary conditions for ensembles run with both the HIRLAM model and the ALADIN model, and in addition it will provide separate members to the GLAMEPS ensemble.

With the extension of Leutbecher (2007), EuroTEPS is based on experience with the Norwegian targeted ensemble prediction system (TEPS) that has been running operationally at ECMWF for The Norwegian Meteorological Institute (met.no) since February 2005 (Frogner et al., 2006, Aspelien et al., 2011). A weakness of the method is its use of total energy as a norm for the TSV-calculations. In the short range, actual initial state errors may not grow as fast as the total energy TSVs would indicate. Unless other types of perturbations that reflect actual analysis errors are added, it is quite probable that EuroTEPS will show underspread during the first 12–24 h of the forecasts in order to produce correct ensemble spread up to 2 d. Alternatively, if the initial amplitudes are chosen to match the error during the first 12–24 h, the spread will become exaggerated during the second day of the forecasts. In GLAMEPS this is accounted for (at least partly) by using multi-analysis (Iversen et al., 2011). Later it is planned to extend GLAMEPS with ensemble transform Kalman filtering (ETKF, Bishop et al., 2001) which will further account for actual initial state uncertainty. For the pure EuroTEPS, however, it is expected that the ensemble spread will underestimate the root mean square error of the ensemble mean forecasts for many variables over the first 12–24 h.

In Section 2, the set-up of the experiments is described. The evaluation of EuroTEPS with different number of ensemble members is compared to operational EPS from ECMWF in Sections 3 and 4. The discussion is given in Section 5, and conclusions are drawn in the same section.

2. Description of EuroTEPS and experiments

The Norwegian operational TEPS system focuses on Northern Europe, and therefore has a relatively small target area in that region. This target area is not suitable for EuroTEPS, which is a system for Europe as a whole. One option could then be to use one, pan-European target domain for the SVs that would cover this big region. However, constructed in this way one would encounter situations when small parts of the big target region are much more active than the rest. This would result in ample perturbations in the active part and very little in the remaining, possibly major part, of the area. To enable a more even distribution of perturbations across Europe, we use three target areas instead of one single. The target areas together cover Europe and they have considerable overlaps to counteract the possibility of developments being excluded because it occurs on the boundary between two target areas. The target areas are displayed in Fig. 1 together with the GLAMEPS integration domain. The latitude/longitude area that completely covers the three target areas is used as the verification area in this study.

The TSVs are targeted at final optimization time (24 h) to these three overlapping areas. For reference, the SVs used for

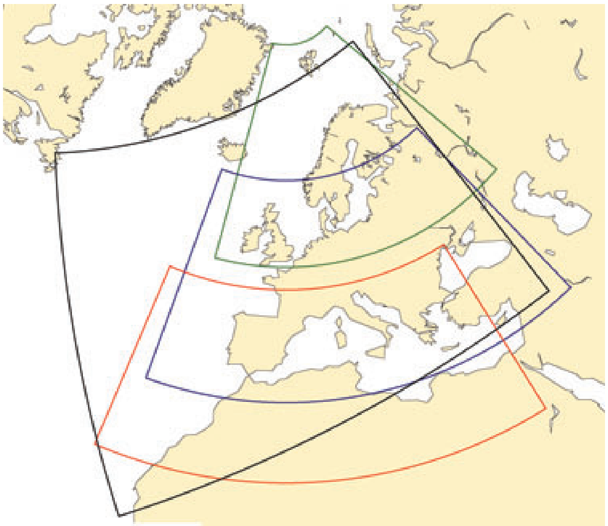

Fig. 1. The three overlapping target areas together with the GLAMEPS integration domain (HIRLAM version).

operational EPS at ECMWF are targeted at final optimization time (48 h) to the northern hemisphere (NH) north of 30°N. The TSVs are calculated for each of the three target areas using the total energy norm and a resolution of T159, whilst the operational NH SVs have resolution T42. For each target area 10 TSVs are calculated to be orthogonal to the operational SVs with respect to the total energy inner product (Leutbecher, 2007). Finally, they are also mutually orthogonalized. EuroTEPS is a system for which additional ensemble spread for the first 24 h are sought over Europe, hence the choice of 24 h optimization time for the TSVs. However, perturbations solely generated from the TSVs will for a great part leave the area of interest before the end of the desired forecast range of 48 h. Therefore, we also employ the already operationally calculated 50 operational NH SVs optimized over 48 h.

The initial ensemble perturbations for EuroTEPS are computed using the same method as ECMWF uses for the operational EPS. The 30 orthogonalized TSVs are combined with the 30 TSVs calculated 24 h earlier and linearly evolved over the 24 h up to present time. These are the TSVEVOs. Furthermore, the operational NH SVs and their 50 linearly evolved counterparts (SVEVOs) over the recent 48 h are used. All these 160 normalized vectors (using total energy norm) are linearly combined to unique perturbations using weights drawn randomly from a gaussian distribution. The number of perturbations computed is dependent of the number of ensemble members of the system, for example, 25 perturbations are computed for a system with 50 members plus control. Each unique perturbation is added and subtracted to the initial state analysis, and the entire set of initial state perturbations add to zero. The total amplitude of the perturbations is finally decided in order to match the ensemble mean error as closely as possible. The tuning parameter for the amplitude of the TSVs can be different from that used for the amplitude of the NH SVs. Different amplitudes have been tried, and for the present experiments we use a factor of 0.0105 for the operational NH SVs and SVEVOs, which is a reduction of 25% compared to the operational EPS for the cycle used in these experiments (CY32R3). The factor for the TSVs and TSVEVOs is also set to 0.0105, hence giving a ratio of 1 between operational and TSVs.

EuroTEPS has been produced with ECMWF IFS CY32R3 with resolution T399L62 (corresponding to ca. 55-km-grid mesh width and 62 levels). Different numbers of ensemble members are investigated: 51, 21, 13 and 11 members. The perturbations for all the different sets have all been made from the same 160 SVs, but the SVs have been combined differently to yield the desired number of perturbations for the four different sets. We compare the different ensemble sets for EuroTEPS to the operational EPS from ECMWF with 51 members, as well as with subsets of EPS. We call the different EuroTEPS runs EuroTEPS <NN>, where <NN> = 51, 21, 13 or 11. The subsets of EPS are labelled in the same way, so EPS51 is the full operational EPS from ECMWF, whilst for example, EPS11 is a subset of EPS with only the first 10 members in addition to the control.

The GLAMEPS experiment period is 2 weeks in summer 2007 (12 August 2007 to 25 August 2007, 00 and 12 UTC) and 7 winter weeks in 2008 (17 January 2008 to 5 March 2008, 00 and 12 UTC). The two periods include several extreme events, both in terms of wind and precipitation. Three-hourly data in full model levels suitable for input to HIRLAM and ALADIN models are provided for the GLAMEPS project.

3. The spread-skill relationship

The purpose of designing the system with three overlapping target areas is to make sure that there will be sufficient ensemble spread in the whole area of interest for the two first forecast days with as few ensemble members as possible. In Fig. 2 the normalized difference in spread (=root mean square difference between ensemble members and ensemble mean) between EuroTEPS and operational EPS ([spread EuroTEPS11 – spread EPS11]/spread EPS11) is shown for mean sea level pressure (MSLP) for +12, +24, +36 and +48 h, over the 7 winter weeks. The results for the sets with 13, 21 and 51 ensemble members are similar (not shown), but the signal increases with decreasing number of ensemble members. As seen from Fig. 2, EuroTEPS produces additional spread in the entire area of interest for all forecast lead times up to +48 h compared to the operational EPS. It is also evident that the targeted perturbations start to leave the area in the later part of the forecasts. The same conclusion holds when looking at Z500 (not shown). Hence, in this respect EuroTEPS produces as desired.

The spread-skill relationship for MSLP for the same 7-week period in winter of 2008 is shown in Fig. 3. The top panel shows results from the systems with 11 members, and the lower panel

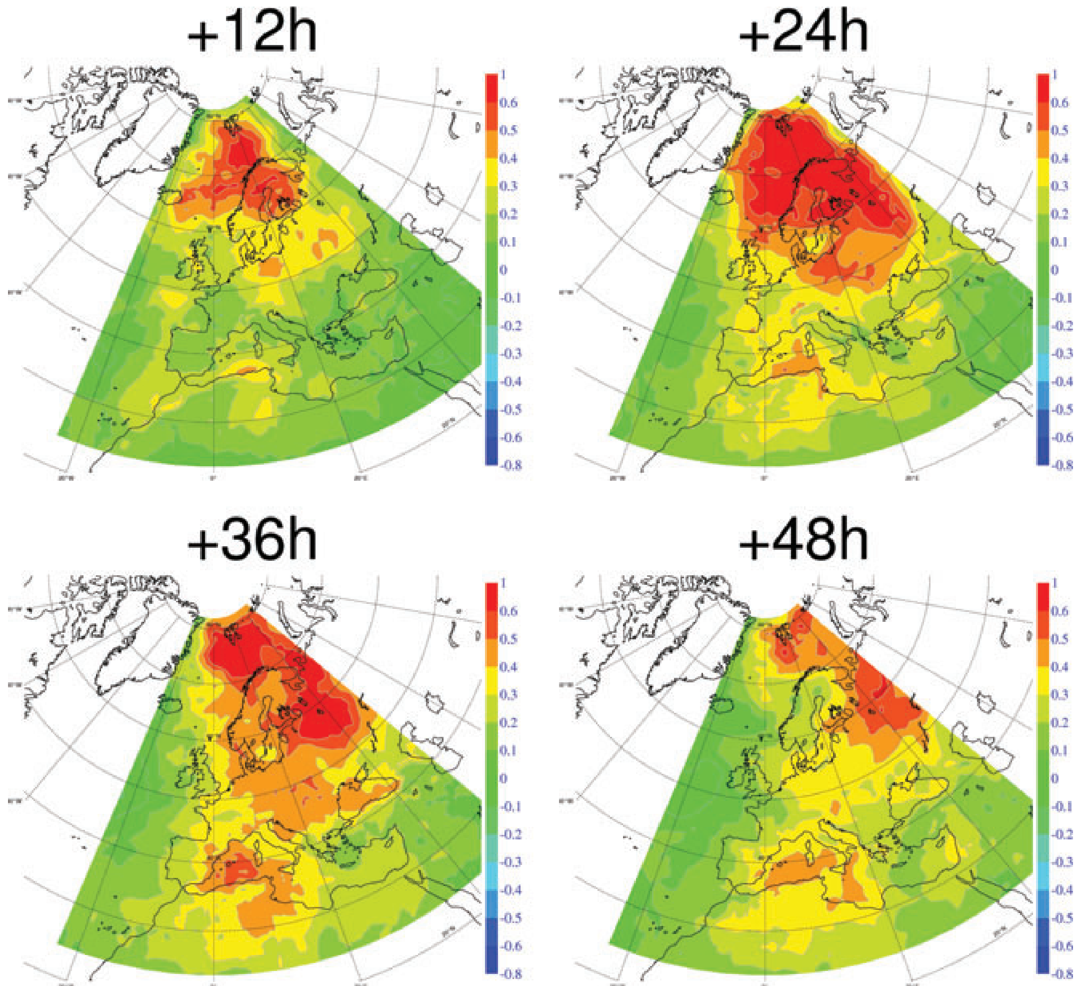

Fig. 2. Normalized spread between EuroTEPS11 and EPS11, for the 7 winter weeks in 2008. Forecast time +12 h upper left, forecast time +24 h upper right, forecast time +36 h lower left and forecast time +48 h lower right.

shows results from the systems with 51 members. For a well-calibrated system the spread and the skill should match so that the two curves should be on top of each other. As can been seen from the top figure, the spread/skill relationship is better for EuroTEPS than for EPS, as the spread for EPS is too small. The same conclusion holds for EuroTEPS and EPS with 13 and 21 members (not shown). The fit between spread and skill is good for EuroTEPS with 11 members after the first day and up to the end of the forecast period. For EuroTEPS with 51 members there is overspread from day 1 and onwards, while EPS has too low spread. For the first 24 h, the spread for all systems is too low. The plausible reason for this is that the SVs only include modes or patterns whose total energy grow very fast over the optimization time, and no specific estimates of actual analysis errors. The forecast errors for the first hours will normally be dominated by patterns that may not grow particularly fast, and some may even decay. In general it will take 12–24 h before the fast growing patterns may dominate the forecast errors in reality.

When comparing the two 51 member ensembles (lower panel in Fig. 3) it is worth noting that the error of the ensemble mean for forecast range +12 h to +48 h is noticeably lower for EuroTEPS than for EPS. This is only seen in the comparison between the 51-member ensembles. Both for EPS and for EuroTEPS the error of the ensemble mean is lower when the number of ensemble members is increased (see the two panels of Fig. 3), but the reduction is largest for EuroTEPS, indicating that with more members dedicated to the verification area EuroTEPS is able to produce a better ensemble mean forecast than EPS. The spread-skill relationship for Z500 (not shown here) shows too large spread for EuroTEPS, whilst the relationship for EPS is better. This is as expected and is further discussed in Section 5. However, as for MSLP the error of the ensemble mean is noticeably lower for EuroTEPS for the case with 51 member ensembles.

The rank histograms for the four EuroTEPS systems are shown in Fig. 4. All four systems show similar characteristics,

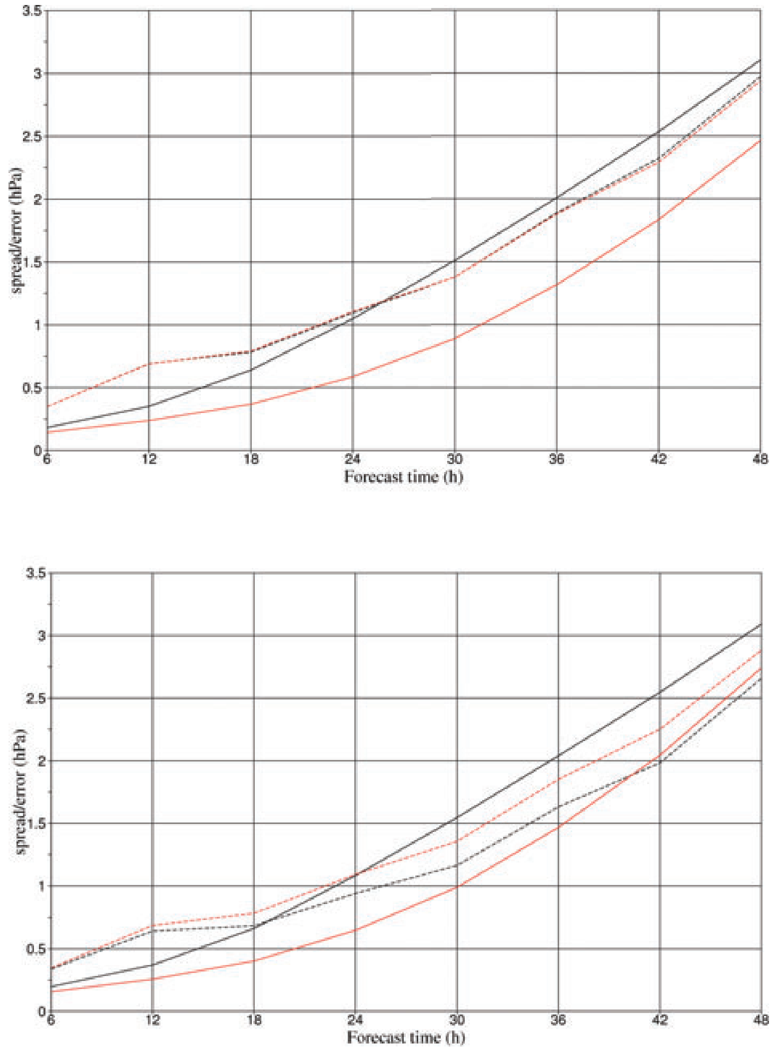

Fig. 3. The spread/skill relationship for EuroTEPS in black and EPS in red, for MSLP. Solid lines: spread (standard deviation of ensemble members). Dashed lines: RMS error of ensemble mean. Upper figure: EuroTEPS11 and EPS11. Lower figure: EuroTEPS51 and EPS51.

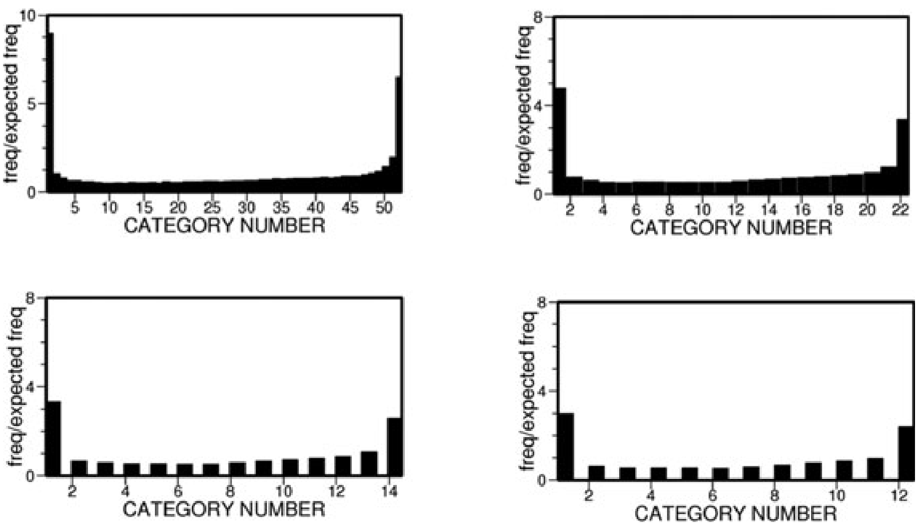

Fig. 4. Rank histograms for MSLP for forecast time +24 h for EuroTEPS51 (upper left), EuroTEPS21 (upper right), EuroTEPS13 (lower left) and EuroTEPS11 (lower right).

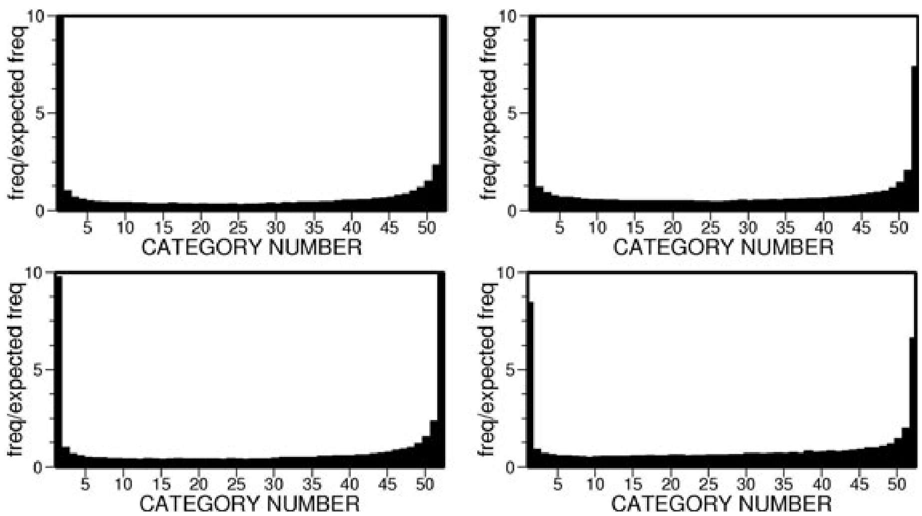

Fig. 5. Rank histograms for EPS51 (top panel) and for EuroTEPS51 (lower panel) for MSLP. Forecast time +12 h (left) and +30 h (right).

with a typical U-shape indicative of a slightly underestimated variability in the modelled ensemble. A slight U-shape should, however, be expected as we not explicitly account for observational uncertainty when producing the diagrams (Sætra et al., 2004; Bowler, 2006). A skew-shaped rank histogram indicates conditional biases (Hamill, 2001). The rank histograms for EPS51 and EuroTEPS51 are in Fig. 5, for forecast time +12 and +30 h. Both systems have the typical U-shape indicating under estimated spread, but the frequency of the smallest and highest categories are smaller for EuroTEPS, and in particular for +30 h. Hence EuroTEPS performs better in terms of rank histograms.

4. Validation of EuroTEPS in terms of probabilistic scores

MSLP, two-metre temperature (T2m) and 10-metre wind speed (ff10m) have been evaluated in terms of Brier Skill Scores (BSS), relative operating characteristics (ROC) area, ROC curves, expected value versus cost-loss ratios (*c*/*L*), reliability diagrams and rank histograms. The verification is done by comparing model calculations to synop observations inside the verification area. For scores that are calculated for predicted events defined relative to thresholds, the thresholds are defined relative to a climatological average computed from 5 yr of monthly mean fields from the years 2002 to 2006, inclusive. For BSS and the expected value as a function of the *c*/*L*, estimates of the climatological frequency of occurrence of an event are needed. This frequency is determined from the observations in the verification period (so-called sample climatology).

In Fig. 6 EuroTEPS with different ensemble sizes are compared using the area under the ROC curve (top panel) and BSS (lower panel) for MSLP more than 5 hPa higher than the climatological mean. The differences between the different con-

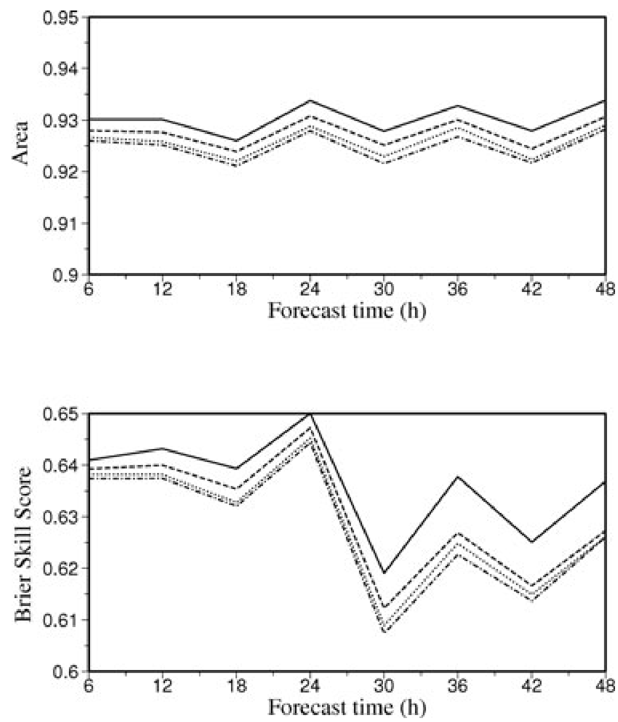

Fig. 6. The area under the ROC-curve (top panel) and BSS (lower panel) for EuroTEPS with different number of members, for MSLP. Solid line: 51 members, dashed line: 21 members, dotted line: 13 members and chain-dashed line: 11 members. Event threshold is 5 hPa higher than climatological mean.

figurations of EuroTEPS are small for these two measures; note the scale on the ordinate. As expected the system with 51 members performs best, then the performance decreases with decreasing number of members. This is also the case for other thresholds. In Fig. 7 the cost/loss curve for this event for

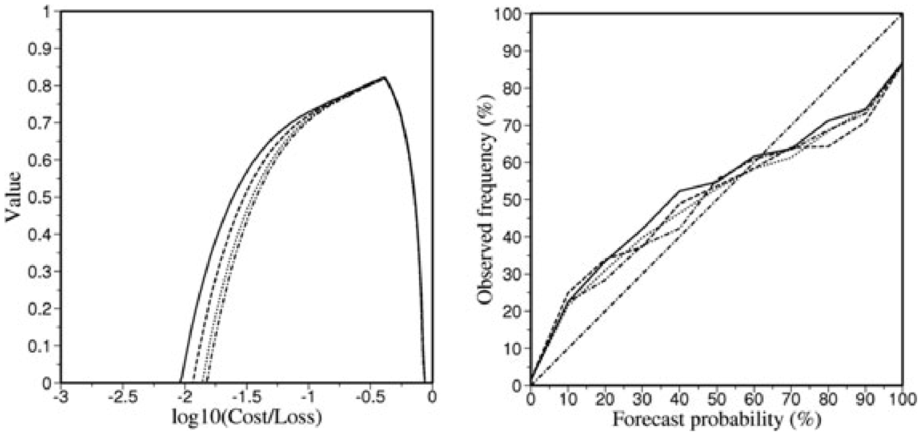

Fig. 7. Estimated value of prediction vs. cost/loss fractions (left-hand panel), and reliability curves (right-hand panel) for EuroTEPS with different number of members, for MSLP. Solid line: 51 members, dashed line: 21 members, dotted line: 13 members and chain-dashed: 11 members. Event threshold is 5 hPa higher than climatological mean. Forecast lead time is 24 h.

forecast time +24 h is shown together with the reliability curve for the same forecast time. Here the differences between the systems are more evident. The estimated value of EuroTEPS51 is clearly better than the other systems for small cost-loss ratios. As one would normally expect, the second best is EuroTEPS21, then follows EuroTEPS13, and finally EuroTEPS11. With respect to reliability of predicted probabilities (Fig. 7, right-hand panel), the picture is less clear. There is a tendency when looking at several thresholds and forecast times that EuroTEPS51 is best (e.g. closest to the diagonal) for the higher probabilities, but for the lower probabilities the differences are too small to draw any conclusions. All EuroTEPS systems are however underforecasting the lower probabilities and overforecasting the higher probabilities. For ff10m and T2m the overall picture

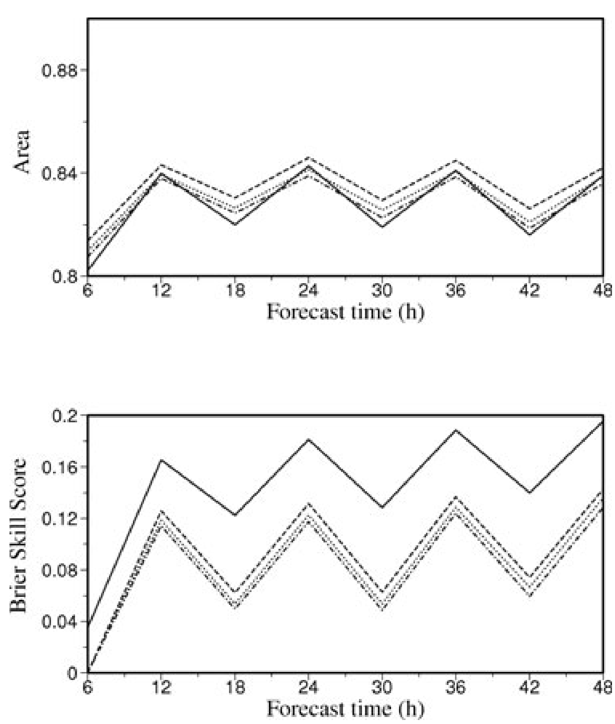

Fig. 8. Area under ROC curve (top panel) and BSS (lower panel) for 10 m wind speed for a threshold of 5 m s^{-1} higher than climatological mean. Solid: 51 members, dashed: 21 members, dotted: 13 members and chain-dashed: 11 members.

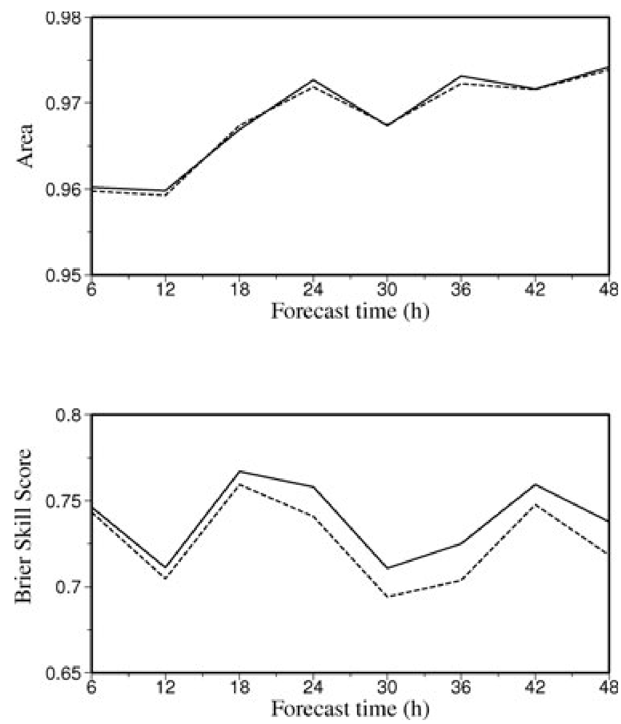

Fig. 9. Area under ROC curve (top panel) and Brier Skill Score (lower panel) for mean sea level pressure for a threshold of more than 15 hPa compared to climatology. Solid: EuroTEPS51, dashed: operational EPS51 from ECMWF.

is the same, but for ff10m, EuroTEPS51 is clearly better in terms of BSS (Fig. 8, note the scale on the ordinate in this figure).

A comparison of EuroTEPS51 and operational EPS51 for MSLP for the event threshold of 15 hPa above climatological mean is shown in Fig. 9. The lower panel shows that EuroTEPS is able to beat operational EPS51 in terms of BSS, and also has higher value of the predictions as a function of cost/loss Fig. 10, lower panel). The ROC curves show that the two

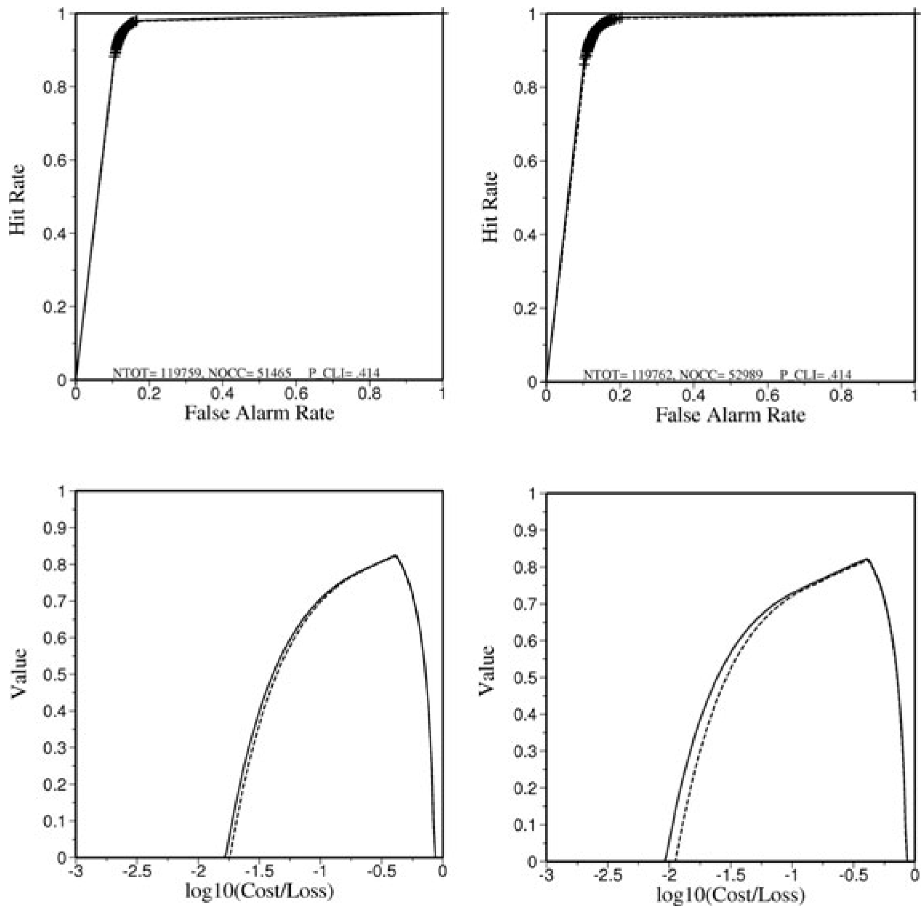

Fig. 10. ROC curves (upper panel) and cost/loss analysis (lower panel) for forecast time +12 h (left) and +24 h (right) for MSLP for a threshold of 5 hPa compared to climatology. Solid: EuroTEPS51, dashed: EPS51. NTOT is the total number of times the event can possibly occur (the number of observations in the verification area multiplied with the number of cases), and NOCC is the number of times the event actually occurred in the sample. P_CLI is the climatological frequency of the event for the experiment period.

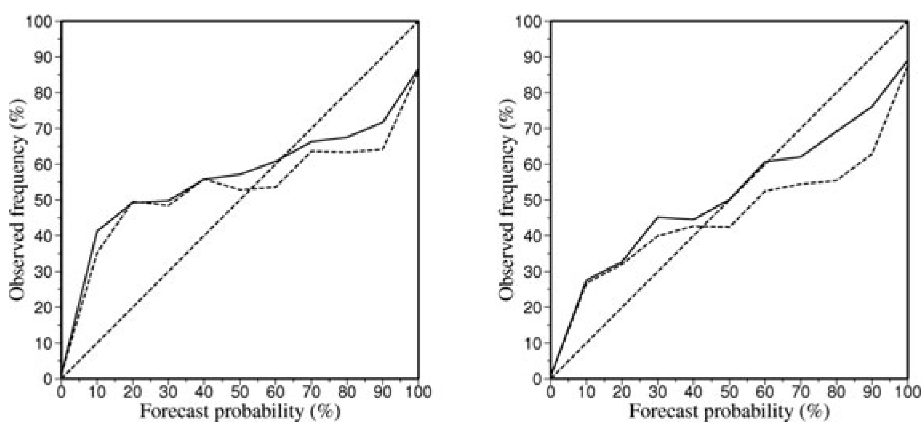

Fig. 11. Reliability for EuroTEPS51 (solid) and EPS51 (dashed) for MSLP with a threshold of 15 hPa compared to climatology, for +12 h (left-hand panel) and +24 h (right-hand panel).

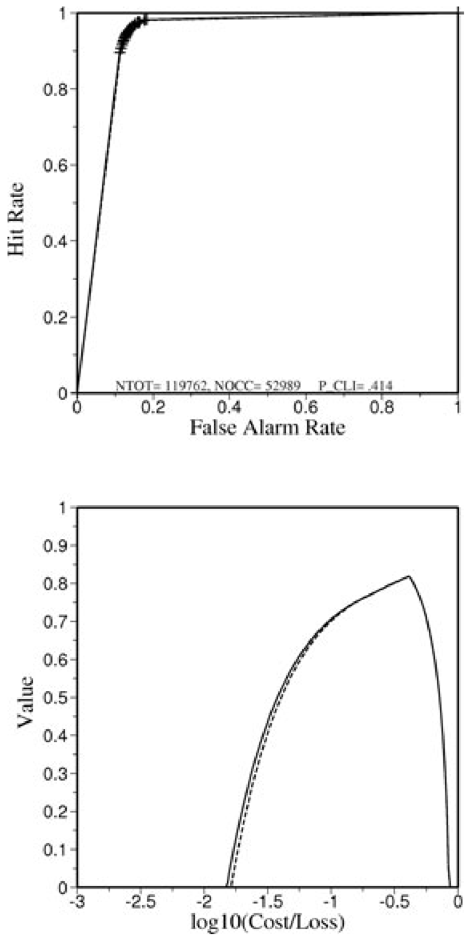

Fig. 12. ROC-area (top panel) and cost/loss analysis (lower panel) for EuroTEPS11 (solid) and EPS11 (dashed) for forecast time +24 h for MSLP for a threshold of 5 hPa compared to climatology. NTOT is the total number of times the event can possibly occur (the number of observations in the verification area multiplied with the number of cases), and NOCC is the number of times the event actually occurred in the sample. P_CLI is the climatological frequency of the event for the experiment period.

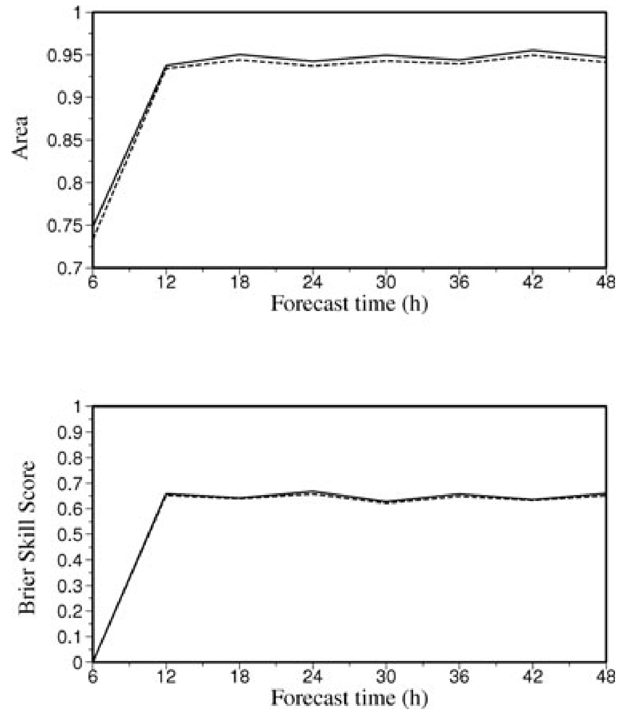

Fig. 13. ROC area (top panel) and Brier Skill Score (lower panel) for 2 m temperature for a threshold of colder than 8° compared to climatology. EuroTEPS51 (solid), EPS51 (dashed).

systems have almost equal hit rates and false alarm rates (Fig. 10, upper panel). Figure 11 indicates that EuroTEPS generally predicts more reliable probabilities than EPS. However, both systems underpredict the lower probabilities and overpredict the higher probabilities. EuroTEPS is generally better having smaller overpredicted high probabilities. EuroTEPS is also better when sets with fewer ensemble members are compared to subsets of operational EPS with the same number of ensemble members. An example is shown in Fig. 12, where ROC and cost/loss is shown for EuroTEPS11 and EPS11, for the same threshold and forecast time as in Fig. 10.

In Fig. 13 the ROC area (top panel) and BSS (lower panel) for forecast times from +6 to +48 h are shown for T2m for a threshold of 8° below the climatological value. Whereas BSS is almost the same for the two systems, the ROC area is largest for EuroTEPS for all forecast times, although the difference is not large. The cost/loss analysis shows an improvement of EuroTEPS51 over EPS51 (see Fig. 14, lower panel). The reliability is also better for EuroTEPS (see Fig. 15). For this parameter the overforecasting is evident already from forecast probabilities of about 30%, but EuroTEPS has slightly less overforecasting. For ff10m the conclusions depend very much on the chosen score parameter. The ROC-area and BSS as a function of forecast time are shown in Fig. 16 for the event of more than 5 m s^{-1} than the climatological mean. The improvement of EuroTEPS51 over EPS51 is evident for BSS, whilst EPS51 is better if the ROC area is used. In Fig. 17 it is furthermore seen that the prediction value of EPS51 vs. cost/loss ratio is slightly higher than the value of EuroTEPS51 (bottom left figure). On the other hand, EuroTEPS predicts more reliable probabilities than EPS (upper right figure). Also this parameter has a pronounced overforecasting from about forecast probability of 20%. The rank histograms (bottom right figure) do not show any difference between the two systems.

5. Discussion and conclusions

When designing an EPS, a good spread-skill relationship is important. We have seen that the relationship between ensemble spread and the error of the ensemble mean forecast differs considerably between the chosen variable. Thus, the confidence

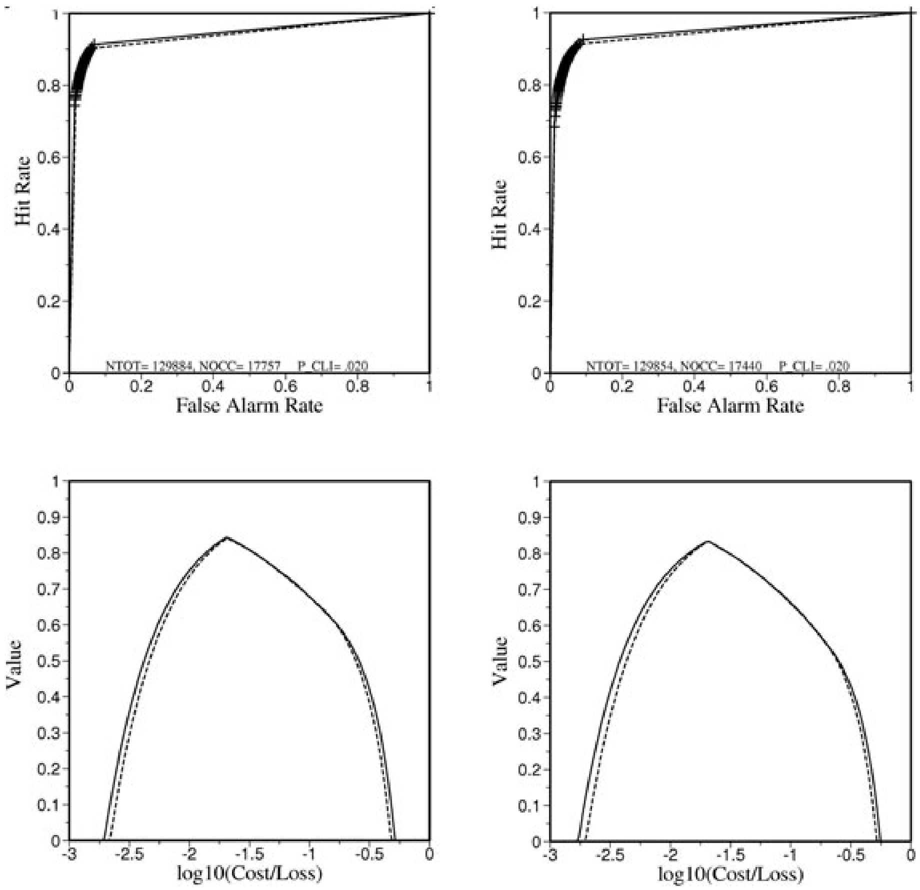

Fig. 14. ROC curves (top panel) and cost/loss analysis (lower panel) for EuroTEPS51 (solid) and EPS51 (dashed) for 2 m temperature. The threshold is colder than 8° compared to climatology. Forecast times +24 h (left) and +48 h (right). NTOT is the total number of times the event can possibly occur (the number of observations in the verification area multiplied with the number of cases), and NOCC is the number of times the event actually occurred in the sample. P_CLI is the climatological frequency of the event for the experiment period.

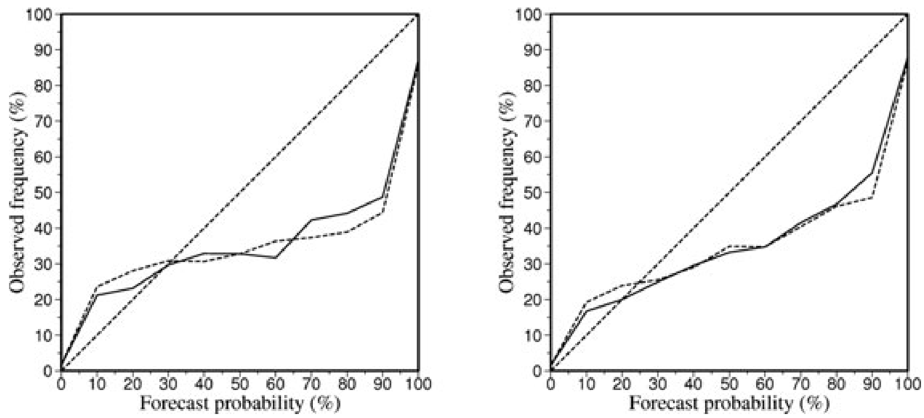

Fig. 15. Reliability diagram for 2 m temperature colder than 8 degrees compared to climatology, EuroTEPS51 (solid) and EPS51 (dashed). Forecast time +24 h (left-hand panel) and +48 h (right-hand panel).

in the probabilistic wind forecasts from the system may differ from the confidence in the temperature forecasts. Frogner et al. (2006) found that if the system produced a satisfactory spread-skill relationship for MSLP, the spread was too large for Z500 and too low for precipitation. Such apparent inconsistency between variables may vary between weather situations. Ideally, the spread-skill relationship should be well calibrated for all parameters, and the ensemble system should ideally include

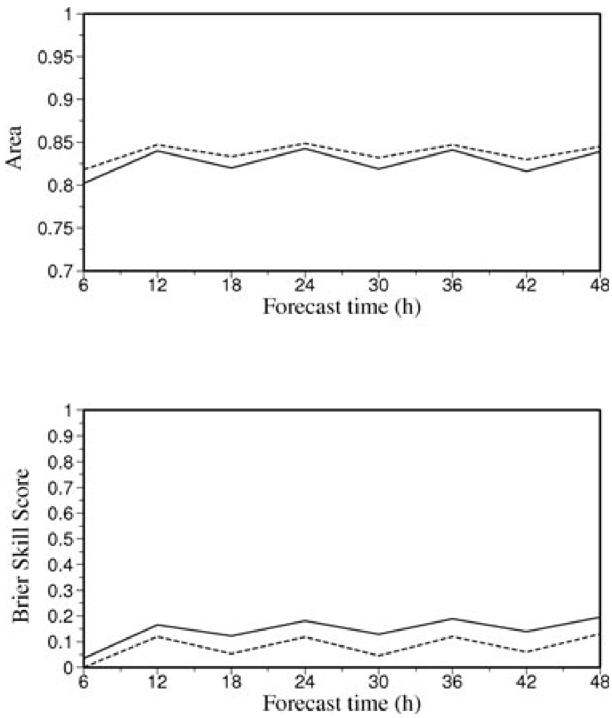

Fig. 16. Area ROC (top panel) and BSS (lower panel) for 10 m wind speed, for a threshold of 5 m s^{-1} compared to climatology. EuroTEPS51 is solid, EPS51 is dashed.

perturbations that yield the desired relationship for all parameters across all weather situations and geographical areas. This is not achieved for the ensemble weather prediction systems that are operationally available today, unless they are a posteriori calibrated statistically (e.g. Hamill et al., 2004; Raftery et al., 2005). Such calibration statistics could ideally be used to diagnose the sources of the inconsistencies in modelled ensemble spread. Such information could guide the generation of different types of perturbations of initial data, boundary data, and model parameters that improves the spread-skill relationship spatially and across model variables and associated weather parameters. Nine weeks of data is clearly too short to get significant statistics that both differentiate weather types and geographical areas. This type of diagnostics is better performed in a long-term operational setting.

The raw EuroTEPS is thus so far not capable of giving the desired spread-skill relationship for all parameters. We have chosen initial state perturbation amplitudes for the sets of SVs in order to obtain a good spread-skill relationship for MSLP. Further development of the system can then focus on developing and including new types of perturbations that may improve the relation for the weather parameters. We chose MSLP rather than, for example, Z500 in order to emphasize developments in the lower troposphere, for example, similar to the MOGREPS NAE ensemble (Bowler et al., 2008). Hence, the spread for Z500 is too large in the current setup of EuroTEPS, despite being relatively well tuned for MSLP. Judging from the probabilistic scores of EuroTEPS compared to EPS, the choice of emphasizing the spread for MSLP seems reasonable. The scores for EuroTEPS are predominantly better than or almost equal to those for EPS, except for ff10m for which results are inconclusive.

Early experience with the Norwegian TEPS system showed an improvement of the simpler targeted system TEPS over EPS (Frogner and Iversen, 2001, 2002; Frogner et al., 2006). The only difference between TEPS and EPS was that the SVs in TEPS were targeted at +48 h to northern Europe instead of the extra tropical northern hemisphere and that the amplitude of the initial perturbations was different. This could indicate that the improvement (although small) of EuroTEPS over EPS in this study could be a consequence of the targeting and increased spread relative to EPS. We have not performed experiments for separating the impacts of the targeting, the increased spread, the higher resolution of the SVs, and the inclusion of several sets of SVs. For instance we could, as an alternative to EuroTEPS, run separate experiments with EPS by simply increasing the initial amplitudes in the regular SVs for EPS. However, such experiments have not been performed, due to the computational costs. When developing the system further, it is planned to quantify more accurately the relative impact of the different changes included in EuroTEPS.

As the improvements of EuroTEPS over EPS are quite small, one has to consider the cost of running the system relative to its usefulness. EPSs are computationally expensive, and the computer time could possibly be better used on other aspects. As EuroTEPS is a part of the GLAMEPS ensemble (Iversen et al., 2011), one possibility is to use the computer time on the LAMs instead, and then use the operational EPS already available as input. On the other hand, another possibility could be to emphasize more the EuroTEPS part of GLAMEPS, for example, to run EuroTEPS with higher resolution than EPS (Super-EuroTEPS) and then skip the limited area runs with HIRLAM/ALADIN and instead run very high-resolution EPS with convective permitting models directly from Super-EuroTEPS. This is so far not tested. In addition to the cost issue, there is also an issue of timing. The TSVs are dependent of the NH SVs being run first, and hence it introduces a slightly longer computation time. This is, however, not substantial (about 10 min in research mode) and is therefore not believed to be important. GLAMEPS based on the present version of EuroTEPS is compared to GLAMEPS based on the operational EPS in Iversen et al. (2011). Also in that case, using EuroTEPS is slightly better than using EPS. However, the improvement is modest and the question remains if it is worth the computational production cost.

We believe that it is still worth continuing the work on EuroTEPS despite the modest improvements so far obtained from the present system. The main reason is that further extensions and refinements are likely once the dedicated system is in place, and several developments are already underway. One important

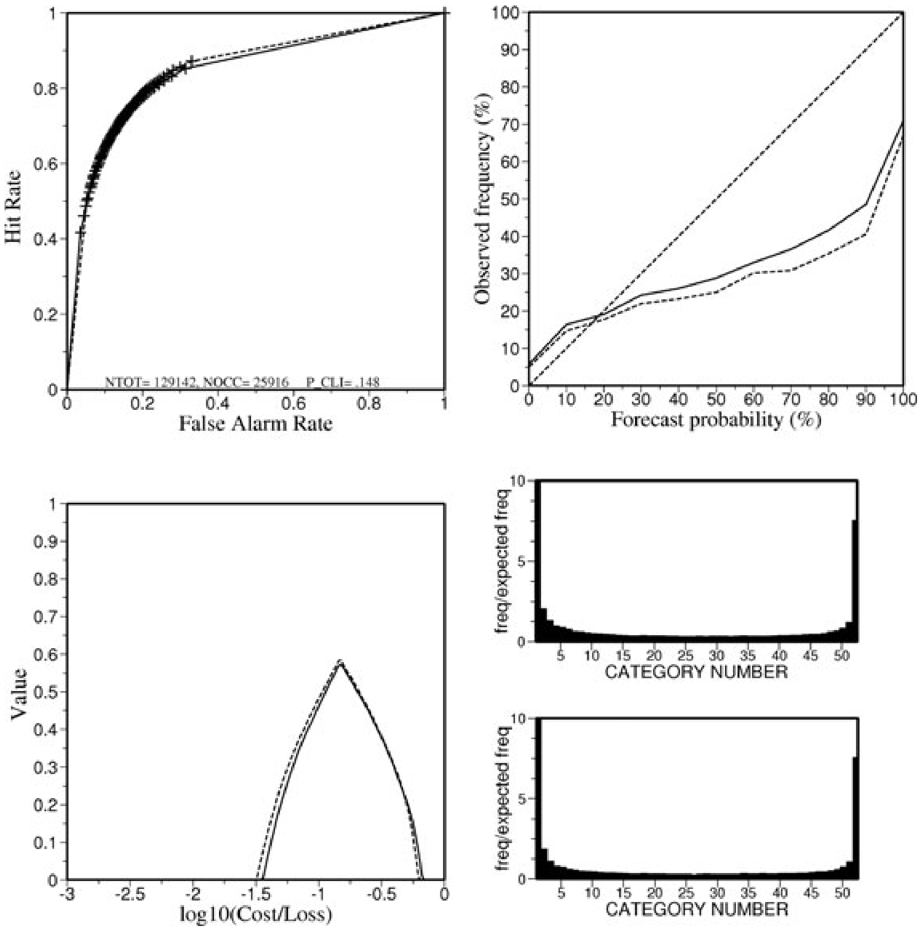

Fig. 17. Figures for 10 m wind speed, for a threshold of 5 m s^{-1} compared to climatology, forecast time is +24 h. EuroTEPS51 is solid, EPS51 is dashed. Upper left: ROC area, lower left: value vs. cost/loss ratio, upper right: reliability diagram and lower right: rank histograms for EuroTEPS51 (top one) and EPS51 (bottom one). NTOT is the total number of times the event can possibly occur (the number of observations in the verification area multiplied with the number of cases), and NOCC is the number of times the event actually occurred in the sample. P_CLI is the climatological frequency of the event for the experiment period.

change is to include Ensemble Data Assimilation (EDA) perturbations (Buizza et al., 2008) into the system, in line with the upgrades of the operational EPS. Experiments have shown that including these perturbations has impacts on the short range, including an increase in the spread for the first hours of the forecast. This indicates an improved spread-skill relation in the early phase due to the actual analysis uncertainty included by EDA. As seen in Fig. 3 it is in the early phase of the forecast range that EuroTEPS selectively needs increased spread. Also the inclusion of moist and diabatic SVs (Mafouf, 1999, Walser et al., 2006) are being tested out at the time of writing this paper. This is expected to be potentially important for the spread of weather parameters associated with smaller scales in the short range. Finally, it is also considered to increase the resolution of the EuroTEPS ensemble itself, and not only for the generation of SVs as is presently done (i.e. Super-EuroTEPS).

We have tested EuroTEPS for four different ensemble sizes, and the selected scores do not appear to crucially depend on the ensemble size. However, the detection of rare events is a very important issue for EuroTEPS and GLAMEPS. Even though the general probabilistic scores do not vary much with ensemble sizes, the size of the ensemble can still be important in order to detect rare events. This is not adequately tested in the selected scores, simply because such events are too few to yield significance for 9 weeks.

In GLAMEPS it is not likely that we will use EuroTEPS51, but rather EuroTEPS11 or 13. Using EuroTEPS13 will give a GLAMEPS ensemble of 13 × 4 members = 52 members (see Iversen et al., 2011), which is comparable to EPS from ECMWF. The reason for running EuroTEPS with 51 members for the present paper was primarily to enable a clean comparison with EPS for the changes that we have made to the system. In this way it is possible to see the benefits of the changes to the system without taking into account reduction in the ensemble size. When comparing EuroTEPS with different ensemble sizes with subensembles of EPS of the same size, the conclusions are the same. Hence, when, for example, EuroTEPS51 is better than EPS51 for BSS and MSLP, this is also seen for EuroTEPS11 compared to EPS11. This is a nice feature which implies that we do not have to run a full 51 member EuroTEPS to take advantage of the changes we have included in our system. The usefulness of EuroTEPS with fewer members also has to be evaluated when used in combination in the full GLAMEPS ensemble, and then compared to the operational EPS (Iversen et al., 2011).

Recently, limited area EPSs with very high, so called convection permitting, resolution are tried out at a few meteorological centres (e.g. Hohenegger and Schär, 2007). The horizontal grid mesh width typically varies between 1.5 and 4 km. Such systems will also need perturbations at the lateral boundaries, and Super-EuroTEPS with higher resolution than the operational EPS at ECMWF can be a candidate. More realistically is perhaps a setup where EuroTEPS is used to perturb an EPS with resolution about 10 km (like GLAMEPS), which is then used to perturb the very high resolution ensemble system (e.g. Kristiansen et al., 2011).

It is seen in this paper that although EuroTEPS is more reliable than EPS, the system is still not very reliable in absolute terms for T2m and wind speed (see Figs. 15 and 17). This is quite serious, since being able to trust the forecast probabilities is a crucial part of probabilistic forecasting. This is, however, the state-of-the-art of both our system and EPS at the moment. It is possible to increase reliability by statistical calibration during post-processing (Hamill et al., 2004). However, such calibration can be computationally costly, and should not be made at the expense of reduced resolution of the ensemble.

EuroTEPS seems to perform reasonably well compared to EPS for MSLP and T2m (see e.g. Figs. 3 and 9 for MSLP and, e.g. Fig. 13 and 15 for T2m). Further work is needed to improve the scores for ff10m, which may be influenced by model biases. Precipitation is not analysed so far, and clearly this important parameter has to be looked at. The observations available for precipitation were presently not sufficient for stable results.

It is our belief that EuroTEPS, when further developed, can be a valuable part of EPSs, and in particular in defining the links between the global-scale EPS for the medium range with very-high resolution EPS for the short range.

6. Acknowledgments

This research benefits from the use of ECMWF's computing and archive facilities through the special project SPNOGEPS. In particular, we thank Martin Leutbecher from ECMWF for many discussions and valuable help in setting up the system. We also thank two reviewers for valuable comments and suggestions.

References

Anwender, D., Jones, S. C. , Leutbecher, M. and Harr, P. A. 2010. Sensitivity experiments for ensemble forecasts of the extratropical transition of typhoon Tokage 2004. *Q. J. R. Meteorol. Soc.* **136**, 183–200.

Aspelien, A., Iversen, T., Bremnes, J. B. and Frogner, I-L. 2011. Short-range probabilistic forecasts from the Norwegian limited-area EPS. Long-term validation and a polar low study. *Tellus* **63A**, this issue.

Bishop, C. H., Etherton, B. J. and Majumdar, S. J. 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. *Mon. Wea. Rev.* **129**, 420–436.

Bowler, N. E. 2006. Explicitly accounting for observation error in categorical verification of forecasts. *Mon. Wea. Rev.* **134**, 1600–1606.

Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B. and Beare S. 2008. The MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.* **134**(632), 703–722. doi:10.1002/qj.234.

Buizza, R. 1994. Localization of optimal perturbations using a projection operator. *Q. J. R. Meteorol Soc.* **120**, 1647–1681. doi:10.1002/qj.49712052010.

Buizza, R., Tribbia, J., Molteni, F. and Palmer, T. 1993. Computation of optimal unstable structures for a numerical weather prediction model. *Tellus* **45A**, 388–407.

Buizza, R., Miller, M. and Palmer, T. N. 1999. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **125**, 2887–2908. doi:10.1256/smsqj.56005.

Buizza, R., Leutbecher, M. and Isaksen L. 2008. Potential use of analyses in the ECMWF Ensemble Prediction System. *Q.J.R. Meteorol. Soc.* **134**(637), 2051–2066. doi: 0.1002/qj.346.

Fjørtoft, R. 1953. On the change in the spectral distribution of kinetic energy for two dimensional, non-divergent flow. *Tellus* **5**, 225–230.

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

Frogner, I.-L. and Iversen, T. 2002. High-resolution limited-area ensemble predictions base don low-resolution targeted singular vectors. *Q. J. R. Meteorol. Soc.* **128**, 1321–1341.

Frogner, I.-L., Haakenstad, H. and Iversen, T. 2006. Limited-area ensemble predictions at the Norwegian Meteorological Institute. *Q. J. R. Meteorol Soc.* **132**, 2785–2808. doi:10.1256/qj.04.178.

Gebhardt, C, Theis, S., Krahe, P. and Renner, V. 2008. Experimental ensemble forecasts of precipitation based on a convection-resolving model. *Atmos. Sci. Lett.* **9**, 67–72.

Hamill, T. M. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.* **129**(3), 550–560. doi:10.1175/1520–0493(2001)129 2.0.CO;2.

Hamill, T. M. , Whitaker, J. S. and Wei, X., 2004. Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.* **132**, 1434–1447.

Hohenegger C. and Schär, C. 2007. Atmospheric predictability at synoptic versus cloud-resolving scales. *Bull. Amer. Meteor. Soc.* **88**(11), 1783–1793. doi:10.1175/BAMS-88-11-1783.

Iversen, T., Deckmyn, A., Santos, C., Sattler, K., Bremnes, J. B. and co-authors. 2011. A grand LAM-EPS (GLAMEPS) for operational use. *Tellus* **63A**, this issue.

Kristiansen, J., Sørland, S. L. , Iversen, T., Bjørge, D. and Køltzow, M. Ø. 2011. High-resolution ensemble prediction of a polar low development. *Tellus* **63A**, this issue.

Leutbecher, M. 2007. On the representation of initial uncertainties with multiple sets of singular vectors optimized for different criteria. *Q. J. R. Meteorol. Soc.* **133**, 2045–2056. doi:10.1002/qj.174.

Lorenz, E. N. 1963. Deterministic non-periodic flow. *J. Atmos. Sci.* **20**, 130–141.

Mafouf, J-F. 1999. Influence of physical processes on the tangent linear-approximation. *Tellus* **51A**:147–166.

Marsigli, C., Boccanera, F., Montani, A. and Paccagnella, T. 2005. The COSMO-LEPS mesoscale ensemble prediction system: Validation of the methodology and verification. *Nonlin. Process. Geophys.* **12**, 527–536.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M. and co-authors. 2004. Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). *Bull. Amer. Meteorol. Soc.* **85**, 853–872.

Palmer, T. N. and Williams, P. (eds.) 2010, *Stochastic Physics and Climate Modeling.* Cambridge University Press. UK. 480.

Raftery, A. E. , Gneiting, T., Balabdaoui, F. and Polakowski, M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.* **133**, 1155–1174.

Sætra, Ø., Hersbach, H., Bidlot, J-R. and Richardson, D. S. 2004. Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Wea. Rev.* **132**, 1487–1501.

Schwartz, C., Kain, J., Coniglio, M., Weiss, S., Bright, D. and co-authors. 2010. Toward improved convection-allowing ensembles: model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Am. Meteorol. Soc.* **25**(1), 263–280.

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

Vitart, F., Buizza, R., Balmaseda M. A., Balsamo, G., Bidlot, J.-R. and co-authors. 2008. The new VarEPS-monthly forecasting system: a first step towards seamless prediction. *Q. J. R. Meteorol. Soc.* **134**(636), 1789–1799. doi:10.1002/qj.322.

Walser, A., Arpagus, M, Appenzeller, C. and Leutbecher, M. 2006. The impact of moist singular vectors and horizontal resolution on short-range limited-area ensemble forecasts for two European winter storms. *Mon. Wea. Rev.* **134**. 2877–2887.