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

Tellus (2011), 63A, 564–584
Printed in Singapore. All rights reserved

© 2011 The Authors
Tellus A © 2011 John Wiley & Sons A/S

TELLUS

Short-range probabilistic forecasts from the Norwegian limited-area EPS: long-term validation and a polar low study

By TRYGVE ASPELIEN[1]*, TROND IVERSEN[1,2], JOHN BJØRNAR BREMNES[1] and INGER-LISE FROGNER[1], [1]*Norwegian Meteorological Institute, PO Box 43 Blindern, NO-0313 Oslo, Norway*; [2]*University of Oslo, Dep. Geosciences, PO Box 1072 Blindern, NO-0316 Oslo, Norway*

(Manuscript received 13 August 2010; in final form 10 December 2010)

ABSTRACT

NORLAMEPS is a 42 member short-range EPS run operationally at met.no since February 2005. It combines TEPS, a 21 member version of ECMWF's EPS with perturbations targeted to Northern Europe, and a HIRLAM-based 21 member LAMEPS, and includes two control forecasts. NORLAMEPS has been upgraded extensively since 2005, including extensions for the IPY-THORPEX project. For a range of investigated weather parameters up to 60 h lead times, NORLAMEPS provides better probabilistic forecasts than ECMWF's 51 member EPS. Combining LAMEPS with TEPS is valuable when high spatial resolution is not crucial, that is, for precipitation except in summer and wind except during autumn and early winter. The IPY-THORPEX field campaign produced additional observations for two polar lows in March 2008. The impact of those observations is studied with the 21 member LAMEPS. For the first polar low a significant positive impact is found, and for long lead times in particular. For the more complex and the operationally poorer forecasted second polar low, the impact of extra observations was positive only in the first stage of the development. Later, for the more intense part of this polar low, slightly better results were actually achieved without the extra observations.

1. Introduction

Accurate predictions of severe weather are of particular importance for weather prediction centres. Severe weather is associated with damages to the environment and human lives and property. By nature these events are rare, otherwise damages would be unlikely. The fact that severe weather events are rare make them difficult to predict from a single deterministic numerical forecast. Furthermore, global numerical weather prediction (NWP) models are frequently too coarse to resolve severe weather events, which also tend to be confined to limited geographical areas.

It took until 1980s and onwards into the 1990s before NWP models started to represent dynamical and physical processes for quasi-geostrophic atmospheric disturbances with some realism. Lorenz (1982) demonstrated that error growth estimated with the first operational NWP model at European Centre for Medium-Range Weather Forecasts (ECMWF), was considerably smaller than the actual error growth measured against the verifying analyses. The difference in growth rate was particularly underestimated during the first 1–2 d of the forecasts. Twenty years of model development later, Simmons and Hollingsworth (2002) demonstrated increased realism in model-calculated error growth, but still an underestimate for the first 1–2 d.

Considerable development work has been invested in selecting initial state perturbations that grow sufficiently fast in the models. ECMWF developed their global EPS based on singular vectors (Buizza et al., 1993; Molteni et al., 1996) whilst National Centers for Environmental Prediction (NCEP, USA) used the breeding technique (Toth and Kalnay, 1993, 1997). Neither of these methods estimated initial state uncertainty explicitly, but as data-assimilation techniques developed and satellite information was properly included, also such methods were proposed. Bowler (2006) discussed different initial state perturbations along with random perturbations using a simplified model. The study indicated that the ensemble Kalman filter technique (EnKF) performs best, but given the huge costs for full scale NWP, the simplified ensemble transform Kalman filter (ETKF, Bishop et al., 2001) is often considered. It was introduced in the short-range limited area ensemble prediction system run at the UK MetOffice (Bowler et al., 2008). However, in full scale verification, ETKF was found to be slightly inferior to

*Corresponding author.
e-mail: Trygve.Aspelien@met.no
DOI: 10.1111/j.1600-0870.2010.00502.x

downscaling the ensemble members from their global model (Bowler and Mylne, 2009).

The contribution of model approximations to error growth was included in ECMWF EPS by the stochastic physics scheme (Buizza et al., 1999). Uncertain formulations of physical processes in the models have also been accounted for by using different model versions, or even completely different models, in the generation of global ensembles (e.g. Hagedorn et al., 2005) and limited area ensembles (Du et al., 2003; Garcia-Moya et al., 2007).

To calculate large ensembles with a global model with high resolution is very expensive. Therefore, ensemble systems based on limited area models have gradually been developed at a few centres (Du et al., 1997; Hamill and Colucci, 1997; Stensrud et al., 1999; Molteni et al., 2001; Marsigli et al., 2001, 2005; Frogner and Iversen, 2002; Walser et al., 2004; Frogner et al., 2006; Garcia-Moya et al., 2007; Bowler et al., 2008). Mullen and Buizza (2002) have shown significant improvements in ECMWF EPS performance of precipitation with finer model resolution. Different techniques are employed for initial and boundary data perturbations, but the methods have not reached the same mature stage as for global ensemble predictions.

The system discussed in this study is referred to as NORLAMEPS. The first version was described in Frogner et al. (2006). The perturbations of initial and boundary data are obtained from a dedicated version targeted EPS (TEPS) of the ECMWF EPS. The singular vectors are targeted to maximize the total energy at time +48 h to Northern Europe and adjacent sea areas (Buizza, 1994). In this way, a sufficient spread over the domain of interest can be obtained with fewer ensemble members (Frogner and Iversen, 2001). TEPS is run with the same resolution as the operational EPS (April 2010: T639L62), and the control forecast starts from a truncated version of the 4D-Var analysis for the deterministic model. In addition to the control forecast, 20 alternative forecasts from TEPS are used as input to HIRLAM.

Further ensemble spread is obtained by adding the initial TEPS perturbations to the 3D-Var analysis made by the HIRLAM model in a limited domain with 12 km grid resolution and 60 levels, whilst the full TEPS-fields are imposed at the boundaries. Hence, instead of a pure downscaling, NORLAMEPS combines the two sets of ensemble members to 42 members, two of which are control forecasts from alternative analyses. Since HIRLAM creates its own analysis, the limited-area analysis can use a different selection of observational data in a model consistent way. In particular, data newer than the observations used for the TEPS control can be used.

Since Frogner et al. (2006) (see also Jensen et al., 2006) presented the first version of NORLAMEPS, several major upgrades have been made to the system. Much of this work has been made in connection with the Norwegian IPY-THORPEX project. This paper summarizes the upgrades and their effects on the forecast performance.

In NORLAMEPS there is a consistent treatment of initial and lateral boundary perturbations. Targeted singular vectors ensure growth of spread in the target domain, and actual initial state uncertainty is to some extent included by using two analyses. Admittedly this uncertainty, as well as model and surface boundary uncertainty, is only quite arbitrarily accounted for in the present system. Nevertheless, even though progressive fine-scale features may be hard to predict also with increased resolution, the analysis of Boer (2003) still indicate enhanced skill for quasi-stationary, fine-scale patterns associated with strong local forcing such as orography.

Polar lows are small meso- to synoptic-scale weather systems which frequently cause adverse weather when cold Arctic air masses flow across the ice-edge over the relatively much warmer open sea surface in the Nordic and Barents Seas (Rasmussen and Turner, 2003). For centuries local fishermen and their relatives have learned that in winter, hazardous weather can occur on a very short notice in these sea-areas, with gale and hurricane force winds and intense precipitation (Rabbe, 1975; Økland, 1977). In spite of the developed understanding of trigger and development mechanisms for polar lows in recent decades (e.g. Bratseth, 1985; Emanuel and Rotunno, 1989; Montgomery and Farrell, 1992; Yanase and Niino, 2005), forecasters frequently fail to predict their occurrence, position, and structure. Some cases are forecasted almost perfectly deterministically, while others can be complete failures. This is probably a consequence of the sparse observation coverage in the region combined with their quick development. The situation is serious for the fishing industry and the increased level of shipping and other off-shore activity in the area, and calls for an enhanced effort to reduce the number of forecast failures.

In the on-going Norwegian International Polar Year project IPY-THORPEX (see http://www.ipy.org/projects), a field campaign was carried out in February - March 2008 (Kristjánsson et al., 2010; Linders and Sætra, 2010; Kristiansen et al., 2011). The NORLAMEPS was extended and improved to study the impact of the extra observations during the IPY-Thorpex campaign. The experiments were performed with an ensemble forecast system rather than a pure deterministic model, in order to detect the impacts of the extra campaign observations relative to the 'noise' caused by initial state and model uncertainties. This paper presents results from applying NORLAMEPS to forecast the adverse weather associated with two cases of polar lows in the Norwegian Sea in March 2008. We tentatively investigate the impact of the extra observations on the forecast quality, with the precaution that an evaluation of probabilistic forecasts cannot be adequate for a few cases. We also appreciate that the incremental gain in quality is not a linear function of the information added by an extra observation.

This paper is structured with a more detailed description of met.no's NORLAMEPS system in Section 2. In Section 3, a long-time validation of NORLAMEPS, and in Section 4 two case studies of two polar lows from the IPY-THORPEX campaign are presented. Section 5 concludes the paper.

2. The NORLAMEPS upgrade history

NORLAMEPS (Frogner et al., 2006) consists of two components; TEPS and LAMEPS. TEPS, being a dedicated version of ECMWF's global EPS, uses singular vectors targeted at final time (+48 h) to Northern Europe. The initial state perturbations are constructed from singular vectors (SVs) with resolution T42L62 and simplified physics (Buizza and Palmer, 1995; Molteni et al., 1996), which optimize the total energy norm in the target area after 48 h. The operational scheme for stochastic physics is included in TEPS (Buizza et al., 1999). The target domain for the evolved (48 h) singular vectors is the Northern Europe and the adjacent Nordic Oceans, bounded by (82N, 15W) at the north-west corner and (50N, 40E) at the southeast corner. Compared to the Northern Hemispheric SVs used in the operational ECMWF EPS, this enables TEPS to obtain similar spread over the first 2 d in the domain of interest with fewer ensemble members (Frogner and Iversen, 2001). TEPS thus produces 20 perturbed ensemble members plus the unperturbed control forecast.

The 21 member limited-area EPS (LAMEPS) is produced by using the TEPS ensemble members both as data for the lateral boundary conditions and to perturb the initial conditions for the limited-area model. The LAMEPS initial perturbations are the TEPS perturbations valid at forecast length +6 h. The LAMEPS control run is made from the HIRLAM 3D-Var analysis (Gustafsson et al., 2001; Lindskog et al., 2001) valid at 06 and 18 UTC and the forecast length is 60 h. The lateral boundaries are provided from TEPS every 3 h for the 20 ensemble members and for the HIRLAM control run (Frogner and Iversen, 2002; Frogner et al., 2006). To further account for model uncertainties, alternating LAMEPS members either use the Kain-Fritsch/Rasch-Kristjansson (Kain, 2004; Calvo, 2007; Rasch and Kristjánsson, 1998; Ivarsson, 2007) or the STRACO (HIRLAM-Unden 2002) cloud and precipitation schemes. For further documentation of HIRLAM versions, see hirlam.org.

The model versions used have varied over time as shown in Table 1. The present (April 2010) latest version of TEPS uses cy36r1 with resolution T639L62. This version is too new to be included in the present paper, and the latest version included here is cy35r3 for TEPS with resolution T399L62, and HIRLAM 7.1.4 with two alternating parametrization schemes for clouds and precipitation between the ensemble members, and with horizontal resolution 12 km and 60 vertical levels below 10 hPa. The integration domain covers Northern Europe and adjacent sea areas extended to include the Barents Sea (HIRLAM domain in Fig. 1). Before September 2009, TEPS was constructed from initial data at 00 UTC only, but since then the 72 h T399L62 forecasts are initialized both at 00 and 12 UTC. Partly as a part of the project IPY-THORPEX, and partly as a part of the regular upgrades at ECMWF, NORLAMEPS has undergone several upgrades (Table 1). TEPS follows the update cycle at ECMWF for the EPS setup.

The total number of ensemble members in NORLAMEPS is 42 when TEPS and LAMEPS are combined. Two of the members are control runs for the TEPS and LAMEPS, respectively. This combination will to some extent account for forecasts errors caused by model imperfections, as well as for actual uncertainty in the analysed initial states which is of increasing importance when forecasts are shorter. Although one of the ensembles has coarser resolution, the size of the ensemble is thereby increased without further cost. However, since TEPS-fields are used as boundary fields for LAMEPS, the two sets of forecasts are not fully independent, and the model differences are probably underestimated (Frogner et al., 2006).

3. Long-time validation of NORLAMEPS

NORLAMEPS has been operational at met.no and run once per day (18 UTC based on boundary-data from ECMWF at 12 UTC) since February 2005. As detailed in Section 2, the system has undergone a range of minor and major upgrades, some of which were bug corrections. In short, the changes that should have considerable impacts on the performance of NORLAMEPS, are:

(i) an upgrade of the ECMWF Integrated Forecasting System (IFS) in November 2007 (cy32r3), amongst other things with reduced vertical diffusion in the free atmosphere which probably contributed to the increased level of diagnosed dynamic activity in the model, with positive impacts on EPS;

(ii) an upgrade to new version of HIRLAM and an increased vertical and horizontal resolution in February 2008;

(iii) a correction of an error introduced during the February 2008 upgrade in the HIRLAM surface fields in May 2008;

(iv) an increased resolution of the sea-surface temperature (SST) data used in HIRLAM in October 2008.

Furthermore, a preliminary version of a second run per day (06 UTC) was started in February 2008. This was deliberately an inferior product since TEPS still was run only once per day (12 UTC). On 8 September 2009 these two product became a priori comparable since TEPS is run also at 00 UTC from this date.

In this section a selection of main probabilistic verification results are shown for NORLAMEPS for the period November 2007 through December 2009 for the 18 UTC production. Note that forecast timings and lead-times are those valid for LAMEPS and NORLAMEPS unless otherwise stated. These start on either 06 UTC or 18 UTC, and since the operational ECMWF EPS and TEPS are calculated at 00 UTC or 12 UTC, forecasts of lead time N hours in the text, will in reality be (N+6) hours for EPS and TEPS. This time-lag is introduced because of the timing of the operational TEPS-production at ECMWF renders

Table 1. Upgrades for met.no's NORLAMEPS since becoming operational in February 2005

Upgrades for TEPS (http://www.ecmwf.int/products/data/technical/model_id/index.html)

Date	IFS cycle	Additional upgrade for met.no
05 Apr. 2005	29r1	
28 Jun. 2005	29r2	
01 Feb. 2006	30r1	
12 Sep. 2006	31r1	
05 Jun. 2007	32r2	
06 Nov. 2007	32r3	The TEPS target area was extended and moved to the north (82N, 15W, 50N, 40E) to provide better initial perturbations for adverse Arctic weather. The initial perturbation amplitude was reduced by 30% for ECMWF EPS. For TEPS the reduction was set to 50% as TEPS previously had an additional overspread in the ensemble.
11 Mar. 2008	32r3	
03 Jun. 2008	33r1	
30 Sep. 2008	35r1/33r2	
10 Mar. 2009	35r2	
08 Sep. 2009	35r3	TEPS being run twice a day for met.no (00 UTC and 12 UTC)

Upgrades for LAMEPS since operational (see hirlam.org for documentation of HIRLAM versions)

Date	Upgrade (previous values in brackets)
20 Mar. 2006	Model version to HIRLAM 6.4.2 (6.2). Model domain extended to include the Barent Sea.
13 Feb. 2008	Model version to HIRLAM 7.1.3 (6.4.2)
13 Feb. 2008	Increased horizontal and vertical resolution: 12km = 0.108 degrees (0.2 degrees); and 60 levels (40).
13 Feb. 2008	Doubled frequency of boundary data from TEPS: 3 hours (6 hours)
13 Feb. 2008	LAMEPS runs twice per day: 06 UTC (new) and 18 UTC (18 UTC). The 06 UTC uses 18 hour old TEPS-perturbations from 00 UTC.
27 Mar. 2008	Model version to HIRLAM 7.1.4 (7.1.3). Alternating cloud parameterizations between ensemble members: Rasch-Kristjansson/Kain-Fritch and STRACO.
23 May 2008	An error in the treatment of surface fields was corrected.
01 Oct. 2008	Increased horizontal resolution of SST from UK OSTIA project. (NCEP)
09 Sep. 2009	LAMEPS for 06 UTC run with 6 hour TEPS-perturbations from 12 UTC (18 hour from 00 UTC).

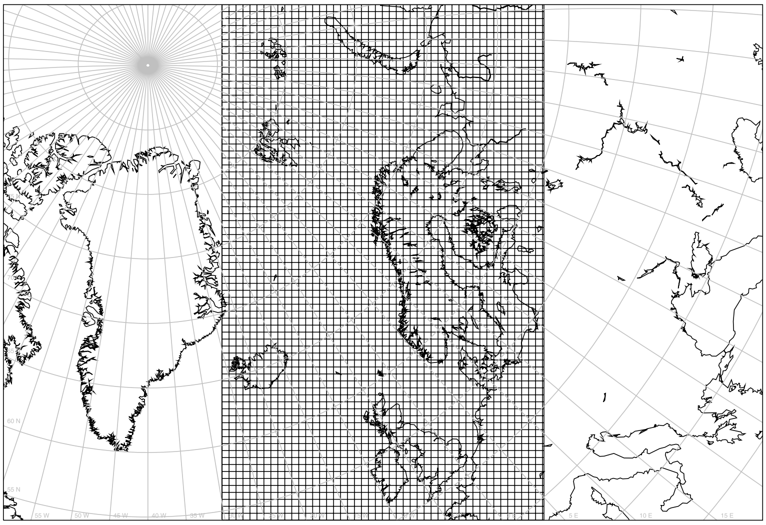

Fig. 1. Integration domain for the HIRLAM model used to produce LAMEPS. Each square show 5 × 5 grid cells.

the 00 UTC or 12 UTC analysis from HIRLAM to be old when TEPS-results become ready for use. We therefore allow the latest observation data in the HIRLAM analysis when the TEPS data are ready. The chosen EPS and TEPS are therefore the newest results to compare with LAMEPS and NORLAMEPS at any time, even though their lead-time is 6 h longer (for 06 UTC before 8 September 2009, the TEPS lead-time is even 18 h longer).

Verification statistics are calculated for mainly land-based observation sites. Observational uncertainty is not explicitly accounted for, and this may penalize the estimated model quality (Sætra et al., 2004; Bowler, 2006), in particular that the ensemble spread is underdiagnosed. Common monthly statistics are calculated for all the observation sites in the integration domain (see Fig. 1). In addition we show separate diagnostics for sites in Norway, and we also present selected verification results for sites inside the model domain which are also north of 65°N. The latter is done because the paper explicitly addresses NWP in the Arctic in connection with the International Polar Year (IPY).

Aggregated statistics for sites across different climate regimes may mask regional differences in the skill. This is partly counteracted by using a range of skill parameters, although geographically resolved statistics should be used. However, such an ultimate evaluation is hampered by the lack of sufficiently long data series from observations and the models. The aggregated statistics provides an overall picture of the system's performance as a whole, given that the sites are biased towards land areas and the thus identified probabilistic quality may be exaggerated compared to those that would be diagnosed regionally.

Little verification is shown for the forecasts started at 06 UTC, realizing that these clearly were inferior until 8 September 2009 when TEPS was started to be run twice per day. There are only about 4 months of results with a system which is a priori comparable to that run at 18 UTC.

3.1. Continuous Rank Probability Skill Score relative to EPS

Continuous Rank Probability Score (CRPS) provides an overall measure of skill for probabilistic forecasts which can be traced over time (e.g. Hersbach, 2000; Ferro et al., 2008). It can be viewed as a ranked probability score over an infinite number of classes of zero width, or it can be interpreted as the integral of the Brier Score over all possible threshold values for the considered parameter. The score is negatively oriented, hence its skill score relative to a reference forecast is CRPSS = 1-[CRPS/CRPS$_{\mathrm{ref}}$]. As we want to compare the quality of the 42-member NORLAMEPS with the 51-member ECMWF operational EPS valid at the same time, we simply use CRPS$_{\mathrm{ref}}$ = CRPS$_{\mathrm{EPS}}$.

Figure 2 shows the CRPSS for 18 UTC+30h predicted screen temperature (2 m height), which is shown here to demonstrate effects of upgrades (iii) and (iv), but also (ii). The graphs show the CPRSS development of 21 member TEPS, 21 member LAMEPS, and 22 and 42 member NORLAMEPS relative to the 51 member EPS over the period from November 2007 to December 2009. Positive values indicate better probabilistic predictions than EPS. In terms of CRPS, the improvement of TEPS with 21 members over EPS with 51 members is modest. Throughout the period, NORLAMEPS has the best skill score, and this is also true for the other lead times (not shown), but the improvement relative to EPS decreases with forecast lead time.

The improvement we expected to see in LAMEPS after increasing the resolution in February 2008, was probably counteracted by the bug introduced in the surface fields. NORLAMEPS showed a much smaller quality reduction because of the combination with TEPS. After the bug was removed in May, LAMEPS showed better results. Further considerably better results were seen after the upgrade implemented for SST in October 2008, and since then the impact of combination with TEPS has been small for CPRSS, and even slightly negative for the sites north of 65°N. Thus, for the aggregated CRPSS measure for screen temperature, NORLAMEPS clearly benefits from the higher resolution in LAMEPS, and in particular over the complex terrain in Northern Scandinavia. The tiny improvement of TEPS over EPS is partly due to the much smaller ensemble size for TEPS; to use 20 of the EPS-members would be the alternative to using TEPS as boundary data for LAMEPS.

For operational weather forecasting the modelling of extreme precipitation is a difficult but important challenge. Figure 3 show CRPSS for 24 h accumulated precipitation over the range +12 h to +36 h lead time for forecasts started at 18 UTC. NORLAMEPS generally has higher CRPSS throughout the whole period, and even for periods when the skill of LAMEPS or TEPS each are worse than EPS. One minor exception is seen in the Arctic subarea north of 65°N. for the period with coarse LAMEPS resolution, and before the surface bug removal in May 2008. Clearly, the impact of resolution is generally smaller for precipitation than for 2 m temperature, and in particular in the Arctic subarea where TEPS and LAMEPS frequently have comparable CRPSS (after the bug-fix in May 2008). During such periods, the combination of TEPS and LAMEPS proves particularly beneficial for NORLAMEPS.

One exception to this general trend, is the annual variation in CRPSS for LAMEPS with a pronounced maximum in summer and minimum in late autumn. The same is seen for longer forecast lead times (not shown). To understand the significance of the autumn minimum thoroughly requires a dedicated investigation, but since TEPS and LAMEPS both have low and even worse scores than EPS, the targeted SVs over 48 h may fail to catch the fast-moving extra-tropical systems in that season, In summer, however, precipitation has smaller spatial scales in general, and the forecasts benefits from the higher spatial resolution in LAMEPS, which is consistent with the results for 2 m temperature.

While extreme precipitation events may be particularly harmful over land areas, events with strong wind speed may also be hazardous over sea and along coastlines. The CRPSS for 10 m

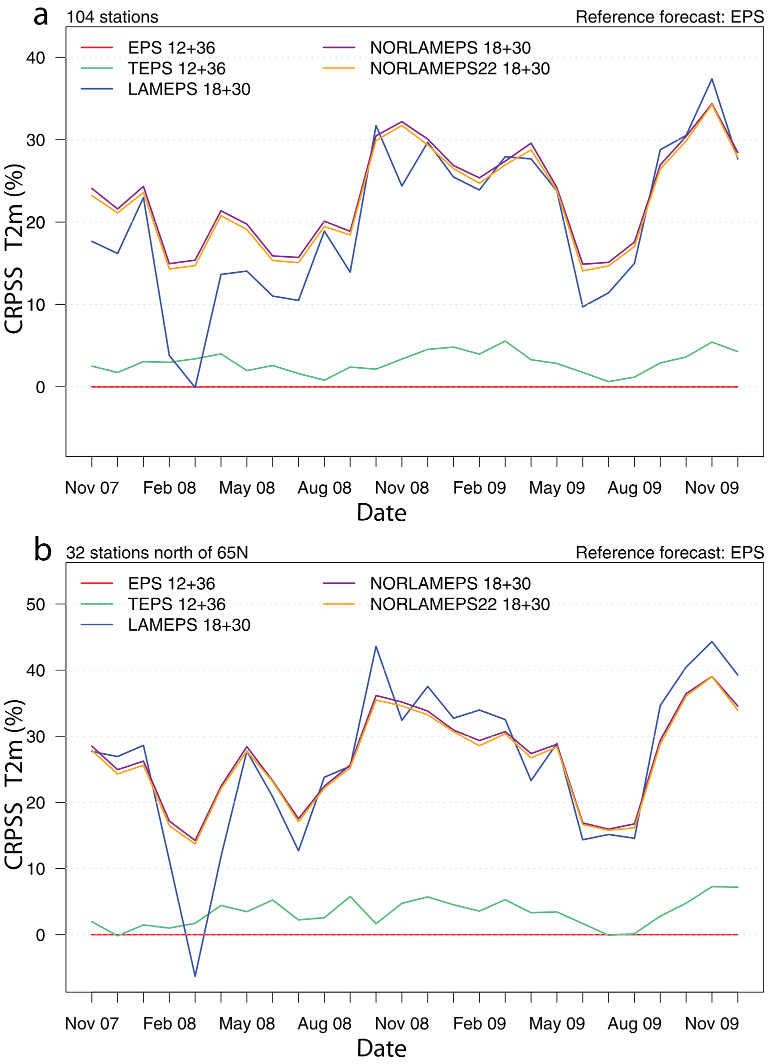

Fig. 2. Continuous rank probability skill scores (CRPSS) with operational (at actual time) ECMWF EPS as reference. (a) Temperatures at 2 meter height at 104 verification sites and 30 h LAMEPS lead time starting from 18 UTC. EPS (red), TEPS(green), LAMEPS (blue), NORLAMEPS (purple), and NORLAMEPS22 (orange). Add 6 h to the lead-times for EPS and TEPS. (b) As for (a), but at 32 sites north of 65 degrees N.

wind speed for 18 UTC+30 h forecasts, is shown in Fig. 4. Also in this case, NORLAMEPS scores better than EPS throughout the period. There is a slight annual variation in the quality of NORLAMEPS with a minimum in the autumn. In agreement with the results for precipitation, the fact that both TEPS and LAMEPS is worse in autumn may be a sign that the targeted SVs over 48 h fail to properly catch the swift and fast-moving developments in that season. The short validation time before spring 2008 does not allow to firmly conclude if the improvement is due to model upgrade or fluctuations in the LAMEPS quality. However, there is a generally higher level of CRPSS after Summer 2008 than before, and the scores were in particular smaller before the increased resolution in February 2008. This is also the case for other forecast lead times (not shown).

As for 2 m temperature, there are small benefits for NORLAMEPS to combine TEPS with LAMEPS for 10 m wind speed. Combination with TEPS even reduces the score relative to EPS over extended periods in the Arctic subregion. This shows that the spatial resolution in LAMEPS is crucial for CRPSS. However, as for precipitation when the LAMEPS- and TEPS-scores drop in late autumn and early winter, the combination between TEPS and LAMEPS enhances CRPSS considerably. This is even seen when TEPS is worse than EPS.

As mentioned above, before 8 September 2009, NORLAMEPS used 18 h old TEPS perturbations for the 06 UTC LAMEPS runs. Nevertheless, CRPSS for 10 m wind speed, precipitation and 2 m temperature shows comparable results for NORLAMEPS for forecasts started 06 UTC as for those started 18 UTC (not shown). However, for mean sea level pressure (mslp), NORLAMEPS, and in particular the TEPS part, scored significantly worse than EPS for forecasts started at 06 UTC. This is seen in Fig. 5 which shows CRPSS for forecasts 06 UTC+6 h before and after September 2009. The CRPSS in the lower panel is valid after September 2009 when TEPS started to run twice per day, and shows similar results as for forecasts started at 18 UTC. Since the pressure-field is particularly

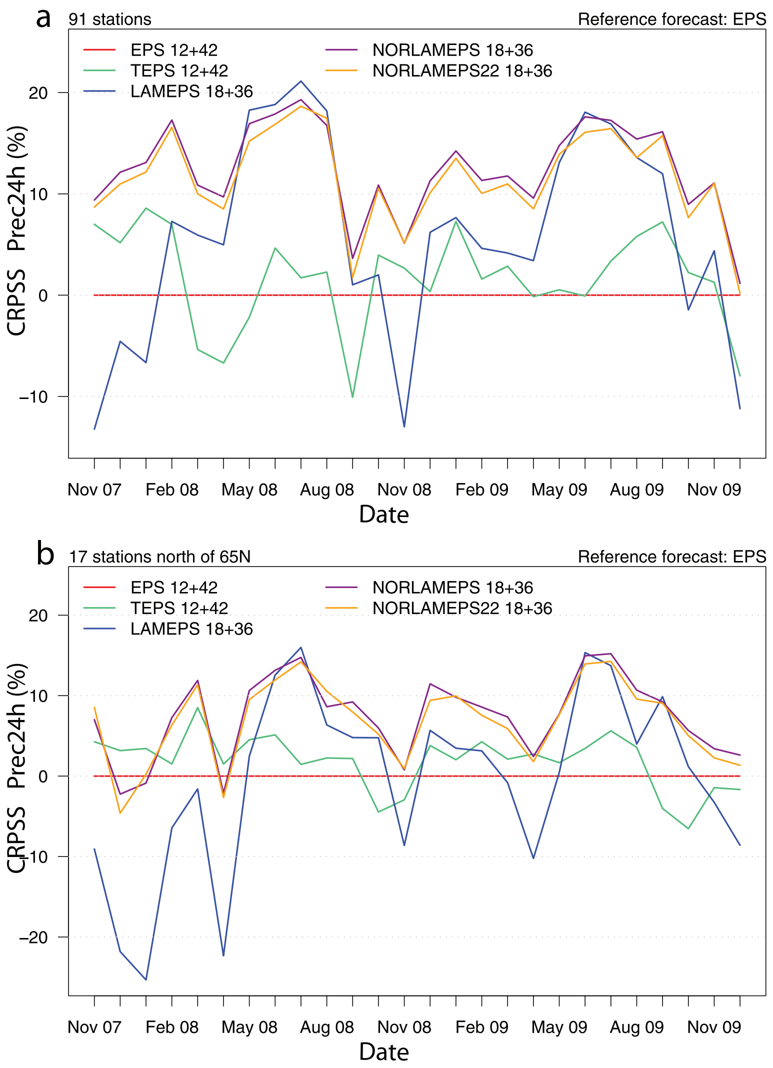

Fig. 3. As for Fig. 2, but here (a) is for daily accumulated precipitation at 91 verification sites and 12–36 h LAMEPS lead time starting from 18 UTC. EPS (red), TEPS(green), LAMEPS (blue), NORLAMEPS (purple), and NORLAMEPS22 (orange). Add 6 h to the lead-times for EPS and TEPS. (b) As for (a) but at 17 sites north of 65°N.

associated with large and synoptic scale dynamics, the benefit of increased resolution is much smaller for mean sea level pressure than for 2 m temperature, wind and precipitation.

3.2. Impacts of model diversity and ensemble size on CRPSS

A relevant question in this study is to what extent a multisystem synthesis of ensemble members may improve the probabilistic qualities relative to the better of each single system, when the latter has the same ensemble size. Intuitively, one could expect this, if (1) the total ensemble spread does not exceed the root mean square error of the synthesized ensemble mean (i.e. no inflation beyond ensemble calibration) and (2) there is a negative temporal correlation between the errors of the different systems' ensemble means for the same forecast lead time. Weigel et al. (2008) and Weigel and Bowler (2009) discussed these conditions for multimodel ensembles based on simplified 'toy models'. They argued that in general, multimodel combinations can only be systematically better than all the single-model ensembles if the latter are underdispersed. In other words, the intuitive condition (1) may be crucial. However, Weigel and Bowler (2009) also found that for normally distributed variables, such as in the short range, multimodel combinations may also improve over single-model, well calibrated ensembles.

So far, we have not been able to investigate conditions (1) and (2) explicitly for the bi-system synthesis NORLAMEPS. Nevertheless, we have calculated CRPSS also for a 22-member version, NORLAMEPS22. In this way we may to some extent quantify and better understand how the quality of NORLAMEPS emerges from combining TEPS and LAMEPS. Each of these systems have 21 members, of which one is a control run. The two control runs are produced from two different analyses for the initial state. Since NORLAMEPS has 42 members and include both the control runs, its probabilistic quality relative to the pure LAMEPS or pure TEPS may partly be due to the double

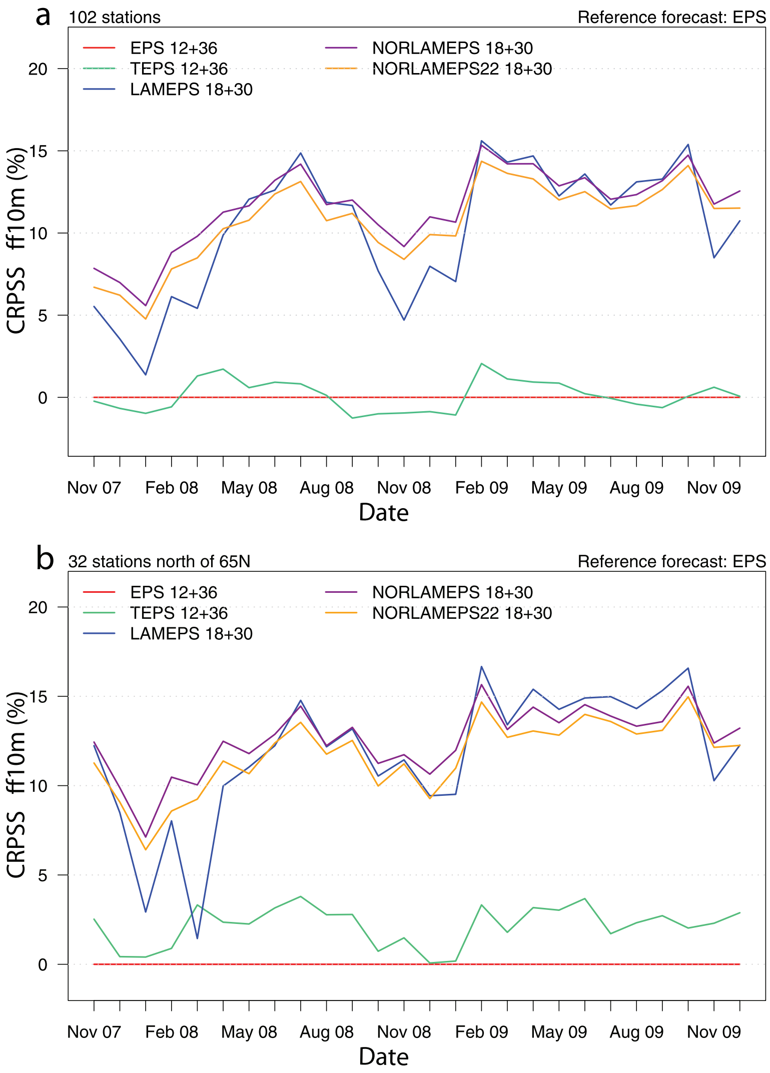

Fig. 4. As for Fig. 2, but here (a) is for wind speed at 10 m height at 102 verification sites and 30 h LAMEPS lead time starting from 18 UTC. EPS (red), TEPS(green), LAMEPS (blue), NORLAMEPS (purple) and NORLAMEPS22 (orange). Add 6 h to the lead-times for EPS and TEPS. (b) As for (a) but at 32 sites north of 65°N.

size of the ensembles, and partly due to the diversity between the two systems. Thus, NORLAMEPS22 both has the same system diversity as NORLAMEPS and (approximately) the same ensemble size as each of he systems TEPS and LAMEPS. The difference between NORLAMEPS and NORLAMEPS22 basically diagnoses the effect of ensemble size on the CRPSS, while the difference between NORLAMEPS22 and either LAMEPS or TEPS quantifies the effects of model diversity.

By comparing the purple and orange curves in the diagrams for CRPSS in Figs 3–5, we see that there are only small reductions in the CRPSS by using only half the ensembles size as long as the multisystem diversity is the same. Comparing the orange and green curves, generally documents a considerable benefit from adding LAMEPS information to TEPS. There are only very few exceptions for precipitation during periods with very low LAMEPS scores. The difference between the orange and blue curves, however, clearly shows mixed signals concerning the benefit of adding TEPS to LAMEPS. Clearly, positive impacts of TEPS is seen when LAMEPS is of lower quality than normal, such as during autumn and early winter for wind and precipitation.

In some of those cases, adding a TEPS which is worse than EPS improves the LAMEPS even without increasing the ensemble size. This contra-intuitive result is probably associated with negative correlations between each system's average behaviour so that the combined system spans a wider portion of the actual prediction uncertainty than each single system, but without inflating the system beyond calibration. We need a further study to verify if this hypothesis is correct.

One should be careful not to conclude that there is no need for 42 member ensembles based on this single CRPSS study. There may be regional differences hidden in the aggregated CRPSS statistic, and, possibly even more important, the impact for extreme (and thus rare) events is not well characterized by CRPSS which integrates over all event thresholds. Brier skill score would give a better diagnostic for such particular events,

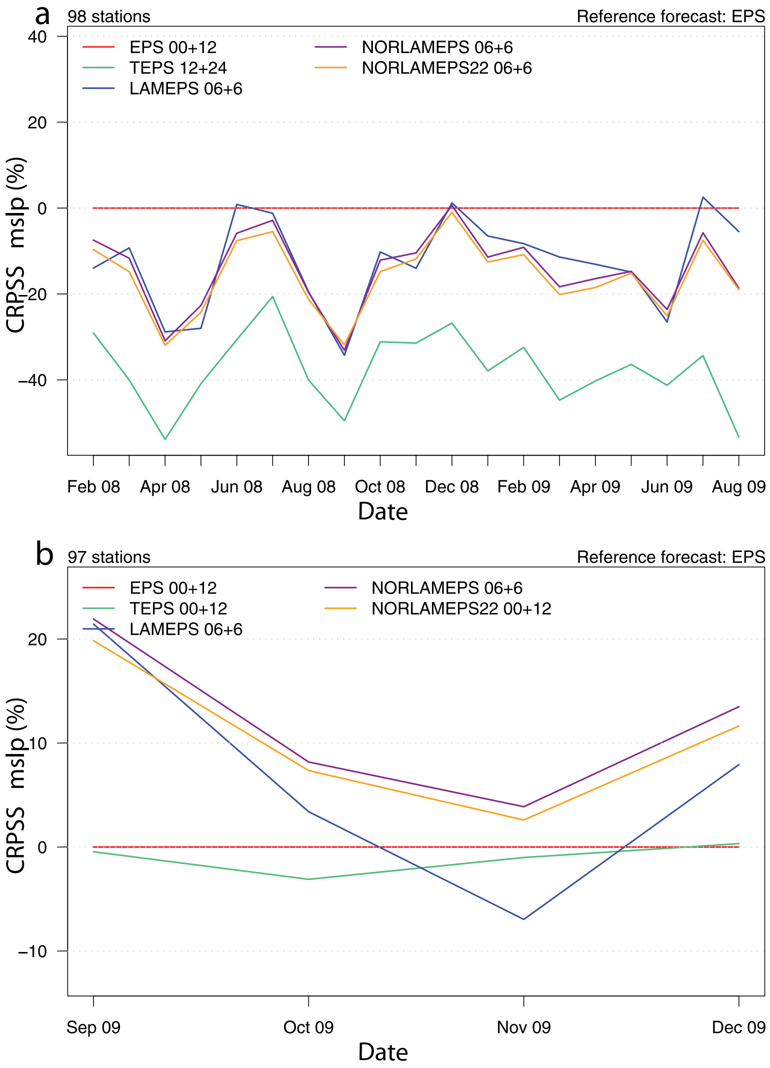

Fig. 5. As for Fig. 2, but here (a) is for mean sea level pressure at 98 verification sites and 6 h LAMEPS lead time starting from 06 UTC between February 2008 to August 2009 when TEPS was run at 00 UTC only. EPS (red), TEPS (green), LAMEPS (blue), NORLAMEPS (purple), and NORLAMEPS22 (orange). Add 6 h to the lead-times for EPS and 18 h for TEPS. (b) As for (a) but at 97 sites and from September 2009 to December 2009 when TEPS also was run at 12 UTC. In this case: add 6 h to the lead-times for both EPS and TEPS.

but due to the infrequent occurrence of such events, statistical significance is hard to obtain.

3.3. *Spread-skill relation, Reliability and ROC*

TEPS and EPS use initial state perturbations generated from total energy singular vectors. Such perturbations emphasize the fast growth of disturbances rather than the actual analysis inaccuracies. In the early phase of the forecasts, for example, up to 12–24 h, actual inaccuracies frequently dominate over the fast-growing modes, and the ensemble mean forecast error can be expected to be larger than inferred from the ensemble spread (standard deviation) if the spread is calibrated for longer lead-times. Ideally the dashed curves for ensemble spread and the solid curved for Root Mean Square Error (RMSE) of the ensemble mean in Fig. 6 (*Note:* the *x*-axis shows lead-times for EPS) should overlap for all forecast lengths, sites, and variables. For mslp this is almost the case when averaged over the sites in the integration domain, except for the EPS.

In accordance with the arguments above, there is under-spread for TEPS, LAMEPS, and NORLAMEPS for the early lead times, but soon later there is a slight overspread. The spread for EPS is on the average much too low, but this is because the amplitudes of the SVs used in EPS are chosen to fit in the medium-range of the forecast. Analysing how the spread fits with the skill of the ensemble mean in different sites and different times after the first phase of the forecasts (not shown), there is a good agreement for small and medium-sized ensemble spread while in the relatively few cases with large spread, the spread should have been larger. The only exception is EPS, which has a better fit for the few cases with large spread but generally too low spread otherwise.

The amplitudes of the targeted SVs used in TEPS are adjusted to fit the forecasts errors close to the surface in the short range, As a consequence, for the geopotential height of the 500 hPa surface the spread is too small, and in particular during the

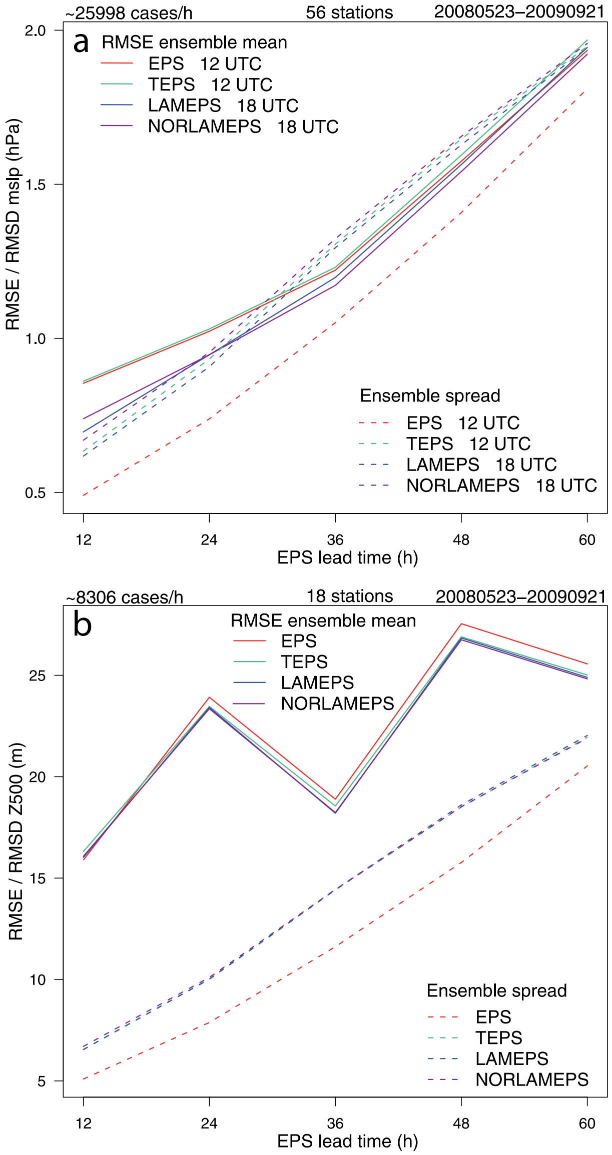

Fig. 6. (a) Ensemble standard deviation (dashed) and root mean square error (RMSE) of ensemble mean forecasts (solid) for mean sea level pressure at 56 European sites from 23 May 2008 to 21 September 2009 as a function of EPS lead-time (subtract 6 h for LAMEPS and NORLAMEPS). EPS (red), TEPS(green), LAMEPS (blue) and NORLAMEPS (purple). (b) As for (a) but for geopotential height of the 500 hPa-surface at 18 European sites.

early range of the forecast. The saw tooth shape in Fig. 6b is due to few radiosonde sites and a variable availability of data between 00 (+12 h, +36 h, +60 h, fewer data) and 12 UTC (+24 h, +48 h).

The rank histogram (Talagrand diagram) is an alternative and more adequate way to diagnose the agreement between spread and skill, since it shows the frequency of observation within the (n+1) intervals defined by the n ensemble members. Intervals number 1 and n+1 cover the semi-infinite intervals outside the range defined by the ensemble members, and 2/(n+1) of the observations should ideally have values outside this range. For a well calibrated EPS and for a large enough sample of cases, there should be an equal chance for the observed variable to be close to any of the ensemble members when observation uncertainty is accounted for (Sætra et al., 2004; Bowler, 2006). Such diagrams are shown in Fig. 7 for 10 m wind speed. The shapes indicate that all the ensembles are under-confident (i.e. too large spread) with a slight negative bias for all lead times for this variable. The same diagrams for mslp (not shown) are almost flat in close agreement with Fig. 6. It is well established that HIRLAM forecast which are not subject to post-processing, tend to underestimate 10 m wind speed over continental surfaces due to the parametrization of subgrid scale topography by enhancing the roughness parameter.

Too large ensemble spread (compared to skill of ensemble mean) can lead to overestimated probabilities of outliers. The slight negative bias may increase the problems further for low wind speed thresholds. For the events of wind speeds exceeding 15 m s^{-1}, Fig. 8 shows indeed that for +42 h forecasts EPS and TEPS exaggerate the probabilities, and the over-prediction is more pronounced for the highest probabilities. The problem still exists but is considerably smaller for the Norwegian sites, and LAMEPS and NORLAMEPS are furthermore considerably more reliable than EPS and TEPS for this event. Similar results are also obtained for other forecast lead times. Unfortunately, for the event threshold 20 m s^{-1}, there are too few cases to give stable statistics.

For 24 h-accumulated precipitation, we have only considered Norwegian sites, and we show results for precipitation accumulated from +36 h to 60 h in Fig. 8. Similar results are found for 24 h shorter lead-times (not shown). The curves for the event >0.1 mm are actually chosen to identify the opposite event (<0.1 mm) as the occurrence of no precipitation. A weakness in many NWP models is their tendency to spread out small precipitation amounts over grid squares, implying that the frequency of occurrence of precipitation is over-estimated while the effective precipitation intensity is under-estimated. A consequence of this is seen in Fig. 8 for threshold 0.1 mm, where the predicted probabilities are considerably exaggerated for all ensembles. Similar features are also found for the 1 mm threshold, and even to some extent for 5 mm threshold (not shown). For higher thresholds (10 and 20 mm) the curves are close to the diagonal, see Fig. 8, and TEPS and EPS are slightly inferior to NORLAMEPS.

Table 2 gives a list of Brier skill scores (BSS) relative to sample climatology as a reference, and area under the Relative Operating Characteristic (ROC) curves (e.g. Mason, 1982; Joliffe and Stephenson, 2003) for the events and forecast lead times shown in Fig. 8 and for the four different ensemble systems. The numbers generally confirms the impression that NORLAMEPS is the better of the systems, even though exceptions are seen. All systems show skill above the reference, except for

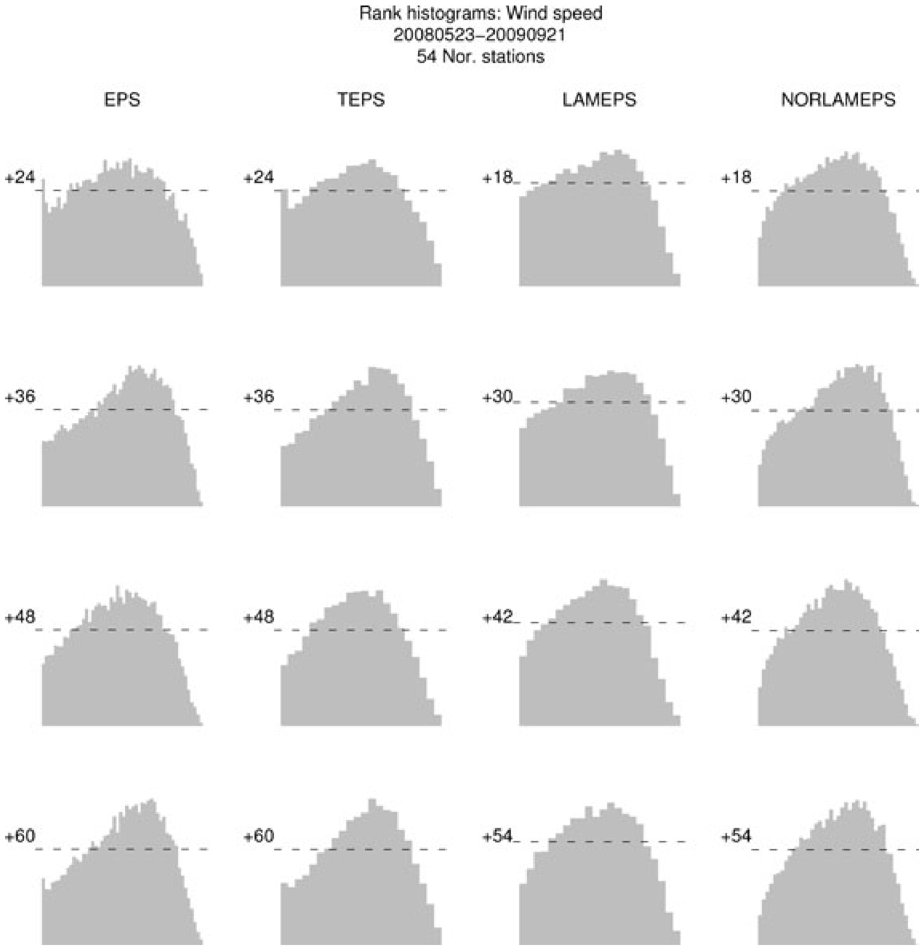

Fig. 7. Rank histograms (Talagrand diagrams) for 10 m wind speed for the four ensemble prediction systems and for 4 forecast lead times, valid at the same time. Data are taken from 54 Norwegian observation sites. The horizontal dashed lines indicate perfectly calibrated ensembles.

LAMEPS and the 0.1 mm threshold where BSS is 0. Note, however, that by combining with TEPS, the NORLAMEPS score increases to a larger value than for TEPS alone. BSS is also very small for that threshold for all systems. Judged from these numbers (and the diagrams in Figs 8–10) LAMEPS and NORLAMEPS represent more convincing improvements over EPS and TEPS for wind than for precipitation.

The ROC curve diagnoses the combination of hit rate with false alarm rate. The area between the ROC curve and the x-axis showing the false alarm rate, is referred to as the ROC area. Skilful forecasts have ROC area larger than 0.5, which is the value for which the hit-rate of a predicted event equals the false alarm rate. Figure 9 shows the ROC area for four different thresholds of wind speed and for two lead times. Again NORLAMEPS scores better than the others for all thresholds and lead times, despite the fact that TEPS has considerably smaller ROC area than EPS for high thresholds. Similar results were found when European sites were included. ROC area curves are also shown for 24 h precipitation (Fig. 10), and the results are similar to the curves for wind speed: NORLAMEPS has the best scores, even for a few cases where both TEPS and LAMEPS were inferior to EPS. It is noteworthy that NORLAMEPS produce larger improvements for the more extreme events included in these diagrams.

4. IPY-THORPEX: Two polar low case studies

4.1. Experiments

The project IPY-THORPEX (http://www.ipy-thorpex.com/Thorpex/English/) is a Norwegian contribution to the international cluster of projects taking part in the international polar year activity, and is associated with THORPEX (The Observing system Research and Predictability EXperiment, http://www.wmo.ch/pages/prog/arep/wwrp/new/THORPEXProjectsActivities.html). An experimental field campaign was set up to study the life-cycling of polar lows and other weather systems associated with adverse weather off the ice-edge over the Nordic and Barents Seas. The aim was to contribute to a better understanding of the weather phenomena, and to investigate to what extent

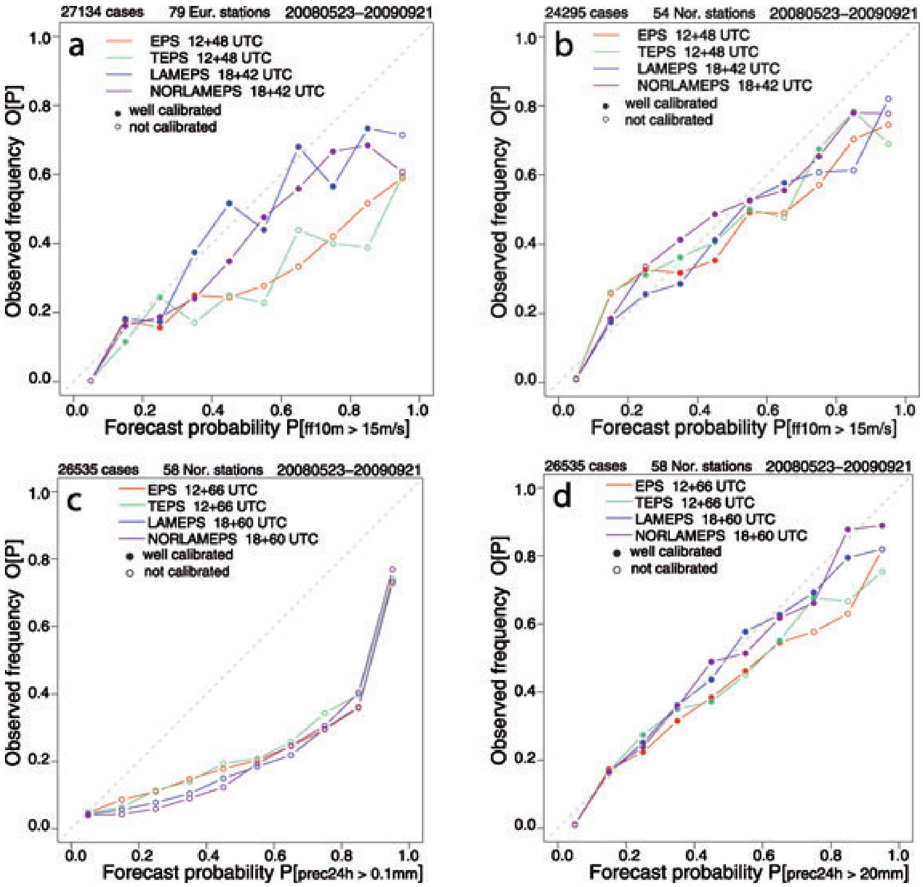

Fig. 8. Upper row: reliability diagrams for forecast probability of 2 m wind speed exceeding 15 ms^{-1} at 79 European sites (a) and 54 Norwegian sites (b). Forecast lead times are 42 h for NORLAMEPS and LAMEPS, 48 h for TEPS and EPS. Lower row: reliability diagrams for forecast probability of 24 h accumulated precipitation exceeding 0.1 mm (c) and 20 mm (d). Forecast lead times are 36–60 h for NORLAMEPS and LAMEPS, 42–66 h for TEPS and EPS. Verification period is 23 May 2008 to 21 September 2009: NORLAMEPS (purple), EPS51 (red), TEPS (green) and LAMEPS (blue).

Table 2. Brier Skill Score relative to sample climatology and area under the Relative Operating Characteristics (ROC) curve for prediction of the selected event probabilities for which reliability diagrams are shown in Fig. 8

		Wind-speed >15 ms^{-1}, 79 European sites		Wind-speed >15 ms^{-1}, 54 Norwegian sites	
		BSS (%)	ROC-area	BSS (%)	ROC-area
EPS	+48 h	14.6	0.92	24.8	0.84
TEPS	+48 h	10.5	0.88	22.3	0.81
LAMEPS	+42 h	32.1	0.90	32.6	0.87
NORLAMEPS	+42 h	27.3	0.93	30.8	0.88
		24 h-precipitation >0.1 mm, 58 Norwegian sites		24 h-precipitation >20 mm, 58 Norwegian sites	
		BSS (%)	ROC-area	BSS (%)	ROC-area
EPS	+(42-66) h	2.3	0.82	29.4	0.89
TEPS	+(42-66) h	6.1	0.83	29.2	0.87
LAMEPS	+(36-60) h	0.0	0.82	29.1	0.89
NORLAMEPS	+(36-60) h	6.9	0.85	32.8	0.92

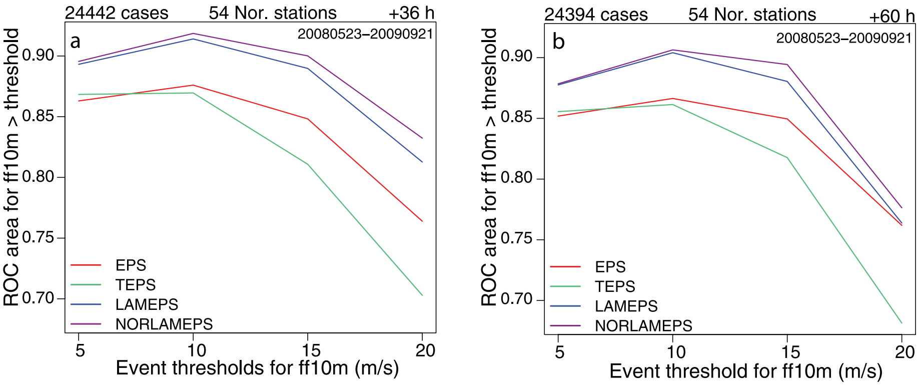

Fig. 9. Area under the Relative Operating Characteristics (ROC) curves (%) for predicted probabilities of wind speed exceeding thresholds of 5, 10, 15 and 20 ms^{-1} at 54 Norwegian observation sites. EPS (red), TEPS(green), LAMEPS (blue) and NORLAMEPS (purple). (a) 36 h lead time; (b) 60 h lead time. Add 6 h to the lead-times for EPS and TEPS.

increased availability of observational data could improve the operational weather prediction. The field campaign lasted from 25 February 2008 until 17 March 2008. The extra observations gathered through the IPY-THORPEX campaign and distributed on the GTS (Global Telecommunications System) in real time, were:

(i) Dropsondes from DLR Falcon (IPY-THORPEX research aircraft).

(ii) Extra Russian radiosondes (WMO: 20046, 22113 and 20744).

(iii) Extra radiosondes from the Norwegian coast guard ships KV Senja and KV Svalbard (identifier: LBHB and LBSV).

(iv) Extra radiosondes from Bear Island (06 UTC and 18 UTC in addition to 00 UTC and 12 UTC).

During the campaign two polar lows were closer investigated. The major challenges associated with predicting polar lows are their small horizontal scale and that they develop rapidly in an area with sparse coverage of conventional observations. Background information on polar lows can be found in for example, Rasmussen and Turner (2003), and further references can be found in Kristiansen et al. (2011). The latter paper studies a very-high resolution downscaling of the LAMEPS prediction of the first of the two polar lows during the campaign.

In this section we want to quantify the impacts of the extra campaign observations on the predictability of the two polar lows. For this purpose we use met.no's limited area contribution to the operational NORLAMEPS, the LAMEPS. As the campaign observations were distributed on GTS and thus were available for LAMEPS by default, the impact of the campaign observations can be studied by comparing with alternative LAMEPS forecasts made after removing the extra campaign data. The following experiments, CTL and IPY, are thus performed:

(i) The control experiment (CTL: the extra observations removed).

(ii) The IPY experiment (IPY: the extra campaign observations included as received on GTS).

The operational NWP systems at met.no use boundary data from operational global ECMWF forecasts. The information received through the boundaries are thus influenced from the extra campaign observations, since the data were put out on GTS in real time. The operational HIRLAM, which also provides the analysis for the LAMEPS control run, further introduce effects from the extra observations since the first guess used in the 3D-Var FGAT data assimilation is a forecast started from a six hour old ECMWF analysis. To minimize these incestuous effects, we ran alternative data assimilation cycles for HIRLAM for CTL and IPY. The lateral boundary data were still used from ECMWF, but the first guess fields in the data assimilations were HIRLAM's own 6-h forecast. The two data assimilation cycles were spun up from 25 February 2008 and the observational data usage was monitored. Also, since TEPS will already include the impacts of the extra observations, only the LAMEPS contribution to NORLAMEPS was used to analyse the impact of the extra campaign observations.

In the operational products at met.no a synthesis between OSI SAF SST (http://www.osi-saf.org/) and NCEP (before upgrade)/UK OSTIA SST (after upgrade) is used as the surface boundary over ocean in the model. The sea surface temperatures used in the LAMEPS experiments are not influenced by the bug in the surface fields which was operationally removed 23 May 2008. However, it should be noted that better results probably would be achieved if the OSTIA SST from UK Met Office had been used instead of the NCEP SST, which were operational during the campaign period. Fig. 2 indicates a better score for near surface temperature after implementing the upgraded SST product in October 2008.

The non-parametric Wilcoxon–Mann–Whitney test (Mann and Whitney, 1947) is used in this section to test the statistical significance of differences between the two experiments.

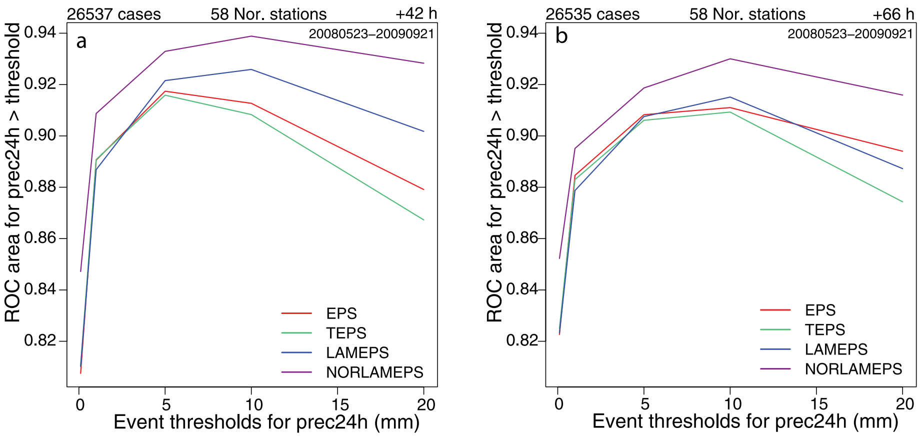

Fig. 10. Area under the Relative Operating Characteristics (ROC) curves (%) for predicted probabilities of 24 h accumulated precipitation exceeding 0.1, 1, 5, 10, 15 and 20 mm at 58 Norwegian observation sites. EPS (red), TEPS(green), LAMEPS (blue) and NORLAMEPS (purple). (a) lead-times +12 h–36 h; (b) lead times +36 h–60 h. Add 6 h to the lead-times for EPS and TEPS.

The test was revised by Yue and Wang (2002). In this section a confidence level of 95% is required.

This section focus on the prediction skill of strong winds and heavy precipitation, which are associated with high-impact weather with potential damaging effects on human life and assets. This is also the case for adverse Arctic weather. The modelled precipitation is also an indicator of how well the model is able to forecast the observed clouds from satellite images. We will also discuss parameters used as indicators for the occurrence of polar lows.

Interpretations of experiments like this must be made with considerable precaution due to non-linear saturation effects of observational data in the data-assimilation. The incremental impact of one extra observation in an operational data-assimilation system may appear to be much smaller than its real value contributed to the system. The first few observations will normally impact the system much more than an extra set of observation on top of hundreds and thousands pre-existing observations. A full evaluation of the added value of an extra observation system should therefore ideally be made by comparing with the added value of other (pre-existing) observations. This would, however, render such experiments too computationally expensive. In addition, here we only study two cases, and a generalization of the results is therefore difficult.

4.2. *Polar Low I (3–4 March 2008)*

The first of the two polar lows (Polar Low I) observed during the campaign started to develop during the night and morning on 3 March 2008. It was fairly well forecasted by most operational models, but it is not known whether this was because of the campaign observations. To illustrate the development of the polar low, three snapshots from 3 March 2008 at 06 UTC and 17 UTC, and 4 March 2008 at 02 UTC are shown in Fig. 11. The days before the polar low developed, a synoptic-scale low propagated along the Norwegian coast, and this contributed to an outbreak of cold polar air from the ice-covered Arctic ocean southwards over the open ocean between Greenland and Svalbard (The Fram Straight). The synoptic-scale frontal zone was parallel to the Norwegian coast, and the polar low developed in the cold air outbreak just southwest of Svalbard to the west side of the synoptic-scale cyclone. The polar low made landfall in the middle of Norway in the evening on 4 March 2008.

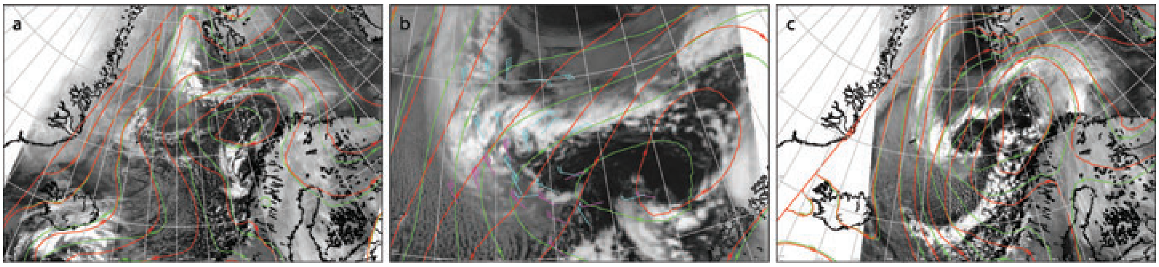

Fig. 11. Development of Polar Low I. Ensemble mean of mean sea level pressure from the forecast run started on 1 March 2008 at 18 UTC for CTL (red) and IPY (green) valid on 3 March 2008 at 06 UTC (a), 17 UTC (b) and 4 March 2008 02 UTC (c). Equidistance is 5 hPa. At 17 UTC (b), the figure is zoomed and wind vectors for the two IPY-THORPEX flights on 3 March 2008 are plotted. The first flight in cyan and the second in magenta.

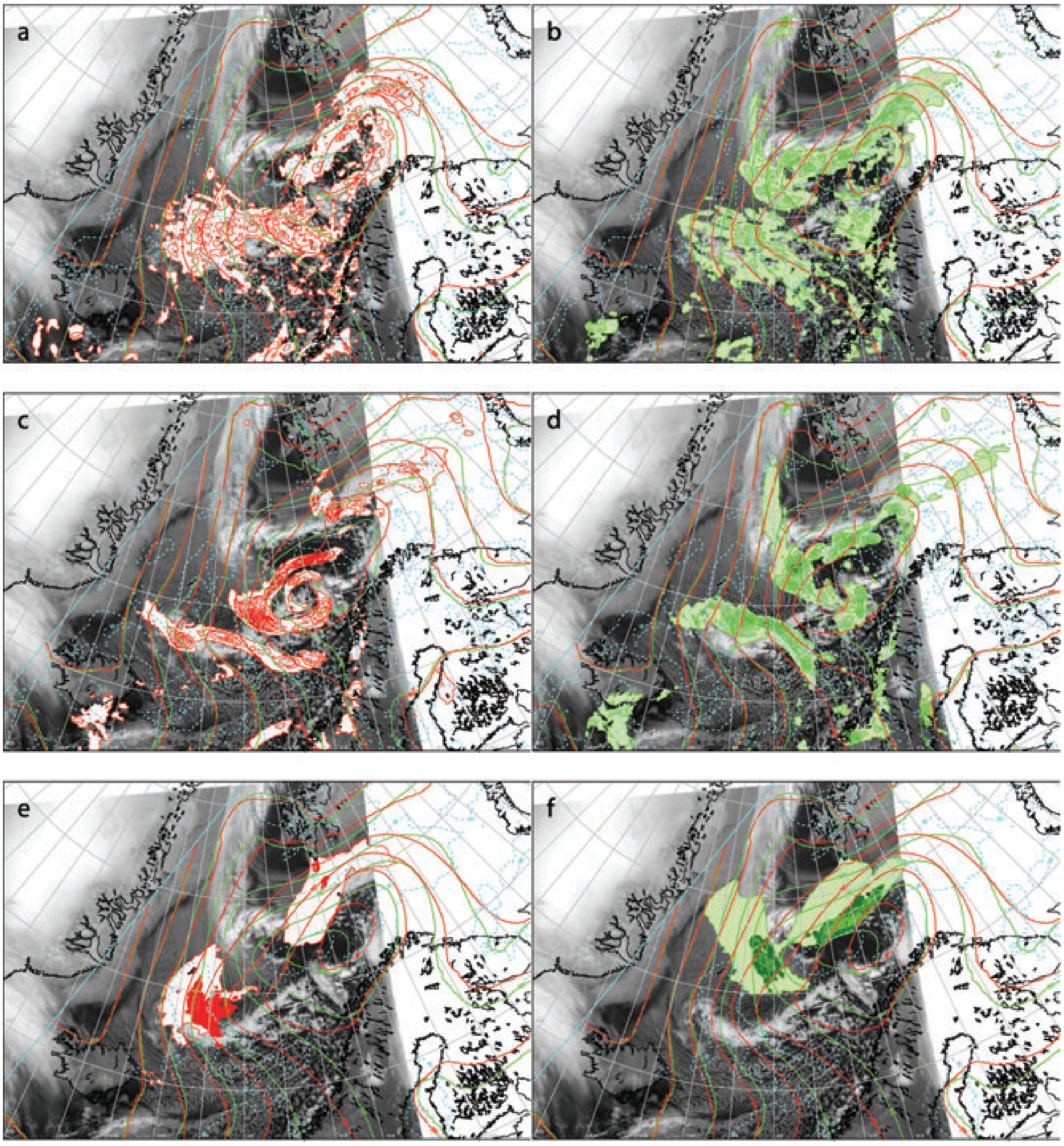

Fig. 12. In all panels the predicted ensemble mean of mean sea level pressure for CTL (red) and IPY (green) are plotted with 5 hPa contour intervals for the corresponding lead times all valid at 3 March 17 UTC. The 95% confidence level for differences (precipitation/10 m wind) between CTL and IPY are the dashed cyan lines. Top panel: light red/green are probabilities above 10% of precipitation rate exceeding 0.5 mm h^{-1}. Darker red/green are probabilities of precipitation rate exceeding 1.0 mm h^{-1}. Left-hand panel is CTL and the right IPY for forecast lead time +47 h. Middle panel: same as top, but forecast lead time is +23 h. Bottom: Light red/green are probabilities above 30% of wind speed exceeding 15 m s^{-1}. Darker red/green are probabilities of precipitation rate exceeding 20 m s^{-1}. Left-hand panel is CTL and the right IPY for forecast lead time is +47 h.

From an operational weather forecaster's point of view, both long and short lead times are important. However, an early warning can be crucial when dealing with adverse weather. In Fig. 11 the ensemble mean of mean sea level pressure for the CTL experiment is shown in red and the ensemble mean for IPY in green from the run started on 1 March 2008 at 18 UTC. Thus, the early polar low development stage in Fig. 11b has a lead time of 47 h. On 3 March 2008 two flights were conducted and dropsondes were released in the area of the developing polar low. The wind vectors in 10 m height of the dropsondes are shown in Fig. 11b. If one assumes geostrophic wind, the vectors correspond fairly well with the mass field of the IPY experiment. For CTL, however, the trough in the marine cold air outbreak is completely missing.

The impact of the extra observations from the IPY-THORPEX campaign on wind and precipitation is shown in Fig. 12. The two upper panels show the probabilities of precipitation exceeding 0.5 and 1 mm h^{-1}. The upper panel represent the longer lead times (+47 h) and the middle panel the shorter (+23 h). Areas with high probabilities of precipitation should coincide with clouds in the satellite images. Thus, a visual inspection can indicate differences in the quality between IPY and CTL. The longest lead time (Figs 12a and b) coincides with the indications we saw in Fig. 11; the probabilities in IPY match the observed cloud pattern from the satellite image, but for CTL only the clouds connected with the synoptic cyclone are predicted. Interestingly, when the lead time is shorter (Fig. 12c), the CTL experiment does predict a polar low development, however, it appear to propagate too fast and has reached too far to the southeast at 17 UTC. IPY has extra observations in its initial state and the polar low predicted in the short range (Fig. 12d) has many similarities with the one predicted at 24 h longer lead time (Fig. 12b).

It is important that the predicted probabilities of strong winds are of high quality, especially in the area where the polar low developed. This area is much used for commercial fishing, there is considerable ship transport and in addition off-shore industry is under prospect. The predicted probabilities of wind speed stronger than 15 and 20 m s^{-1} are shown in the lower panel of Fig. 12. IPY (Fig. 12f) predicts high probability of wind speeds exceeding these thresholds to the northeast and to the west of the polar low centre. The high probabilities compare well with the wind observations taken during the campaign (Fig. 11b). CTL (Fig. 12e) also predicts probabilities of strong winds, but the wind is apparently not connected with the polar low.

A frequently used indicator among duty forecasters when evaluating the possibility of polar low occurrence, is based on the difference between the sea surface temperature and the temperature at 500 hPa (Gunnar Noer, operational forecaster, personal communication). A typical minimum value of this temperature difference causing increased attention, is 43 °C. For the first polar low during this campaign there are large areas where the probability of this temperature difference to exceed 43 °C is high. This is the case for the IPY experiment as well as for CTL, but is not shown here. Thus, a large vertical temperature difference is not sufficient for a polar low development. Such conditions need to be combined with dynamical trigger mechanisms, for example associated advection of potential vorticity or a horizontal jet shear. An alternative index to the vertical temperature difference discussed here is Kolstad-Bracegirdle index defined in Kolstad (2006) and Kolstad and Bracegirdle (2008), but this is not investigated in this study.

Polar low I is not a hurricane-like polar low, but rather a shallow baroclinic disturbance developing along the edge of the cold air outbreak southwest of Svalbard. The cold air flows southwards over warm ocean water causing large amounts of latent heat energy to be released in the deep moist convection created in the area. Probabilities of 2 m temperature reaching lower temperatures than different thresholds predicted at lead time +47 h indicated a larger area of high probabilities for the threshold −2 °C for IPY than CTL. However, for other temperature thresholds the probabilities are more similar for CTL and IPY (not shown).

Several flights were conducted during IPY-THORPEX, however, for Polar Low I the most relevant flights were done at a later stage than relevant for improving the analysis used for the forecasts. Nevertheless, the results indicate that informations from earlier flights and the extra radiosondes improved the forecast significantly, especially for long lead times. For operational purposes, Sensitive Area Prediction (SAP) systems could be used to target sensitive areas for guiding extra observations. However, the targeted observations must be taken before the operational analyses are made. Also as a part of IPY-THORPEX, SAP studies were performed and more information can be found in Kristjánsson et al. (2010).

4.3. *Polar Low II (15–17 March 2008)*

The second polar low was in fact two polar lows and will in this section be referred to as Polar Low IIa and Polar Low IIb. Three snapshots for different states of the development are presented in Fig. 13. Similar to Polar Low I, a synoptic low pressure area propagated along the coast of northern Norway as a cold air outbreak propagated southwards between Greenland and Svalbard. The synoptic low is further to the east, compared to the situation for Polar Low I, when Polar Low IIa starts to develop (Fig. 13a). However, as Polar Low IIa starts to intensify and move southwards, a new vortex develops to the east of Polar

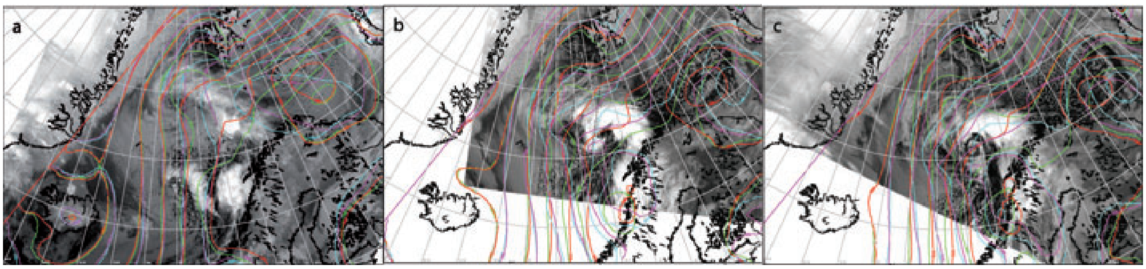

Fig. 13. Development of Polar Low II as shown by predictions valid at 16 March 2008 02 UTC (a), 12 UTC (b), and 18 UTC (c). Ensemble mean of mean sea level pressure for CTL (red) and IPY (green) for runs started at 14 March at 18 UTC. Magenta is CTL for run started at 13 March 18 UTC in (a) and runs started from 15 March 18 UTC in (c). Cyan is the same as magenta, but for IPY. All curves are plotted with 5 hPa contour interval.

Low IIa. This low intensifies more rapidly and in Fig. 13b Polar Low IIa is barely visible and the new Polar Low IIb dominates and continue to grow in strength as can be seen 6 h later in Fig. 13c. Later it propagates southwards and weakens without making landfall.

Polar Low II has a longer and more complicated development than Polar Low I. In Fig. 13 the ensemble mean of mean sea level pressure is therefore shown for runs started 13 March 2008 at 18 UTC, 14 March 2008 at 18 UTC and 15 March 2008 at 18 UTC. The forecast with the longest lead time (+56 h, cyan (IPY) and magenta (CTL) in left-hand panel) has a stronger trough in IPY than CTL in the area in which Polar Low IIa develops. The runs started at 14 March 2008 capture Polar Low IIa in IPY (+32 h lead time, green (IPY) and (CTL) red in left-hand panel). This is also illustrated in the upper panel of Fig. 14 which show probabilities of precipitation exceeding 0.5 and 1.0 mm h^{-1}. Even though CTL and IPY have many similarities, IPY is better able to model the developing cyclone movement southwest of Svalbard. At this early stage of the development of Polar Low IIa, the wind speeds are low. However,

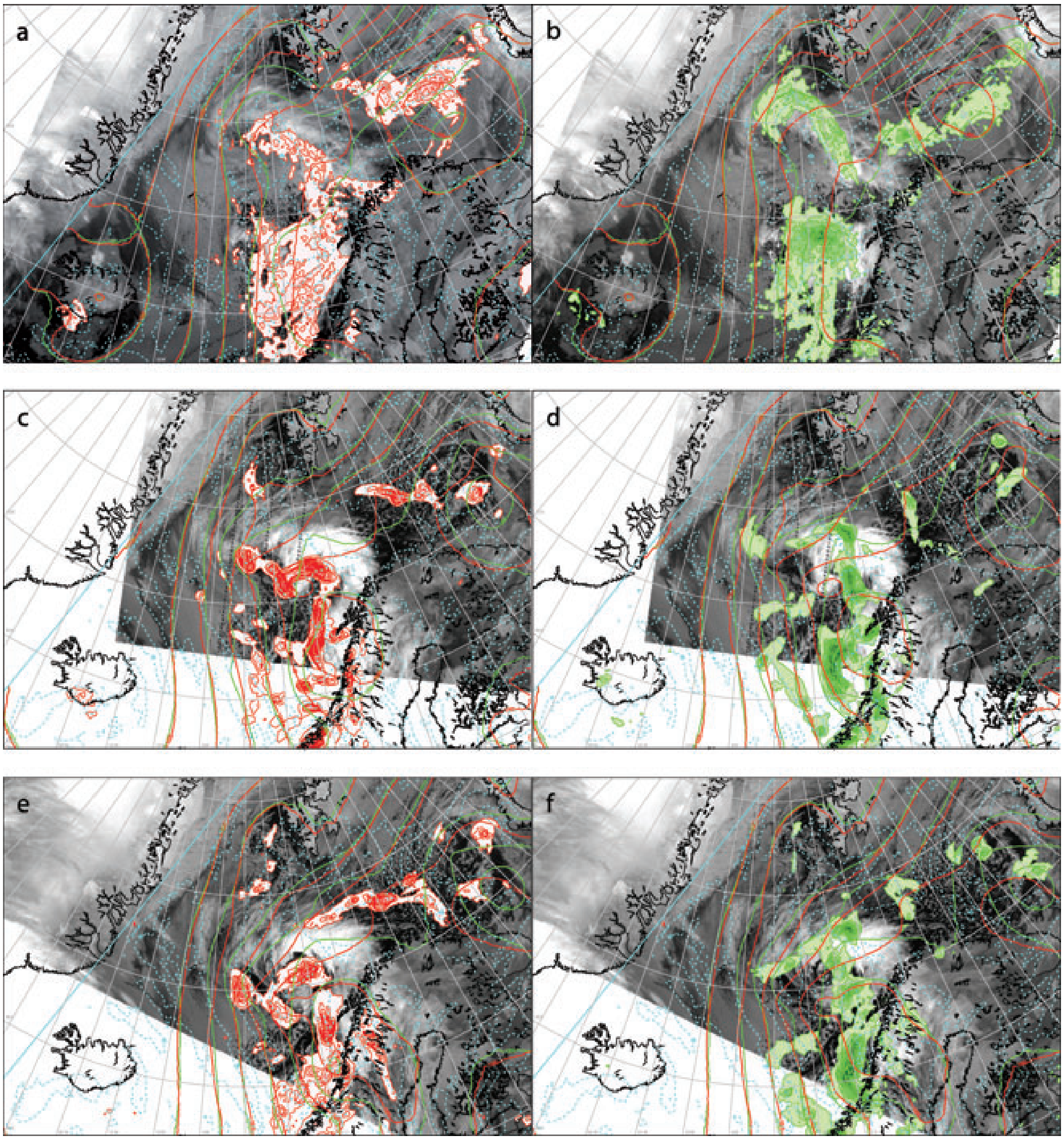

Fig. 14. In all panels the 95% confidence level for differences (precipitation) between IPY and CTL is the dashed cyan line, while red lines are ensemble mean of mean sea level pressure for CTL and green lines are for IPY. Equidistance 5 hPa. Top panels: predictions 14 March 2008 18 UTC + 32 h valid at 16 March 2008 02 UTC; Light red/green shadings are probabilities above 10% of precipitation rate exceeding 0.5 mm h^{-1} for CTL/IPY. Darker red/green are probabilities of precipitation rate exceeding 1.0 mm h^{-1}. Left-hand panel is CTL and the right IPY. Middle panels: same as top, but for predictions 15 March 2008 18 UTC +18 h valid at 16 March 2008 12 UTC. Bottom panels: Same as middle, but with lead time +24 h valid at 16 March 2008 18 UTC.

contrary to CTL, IPY does have a small area of small probabilities of wind speeds exceeding 15 m s^{-1} for Polar Low IIa (not shown).

Polar Low IIb is visible on satellite images from just around noon on 16 March 2008. For predictions 14 March 2008 18 UTC +42 h, Polar Low IIa or Polar Low IIb are not easily recognized on the maps of probabilities for precipitation, neither for the IPY nor CTL experiment (figures not shown). Even though the lead time is short, the results from the runs started 24 h later are more interesting. The two lower panels of Fig. 14 show the probabilities of the precipitation rate exceeding 0.5 and 1 mm h^{-1} for 16 March 2008 at 12 UTC (+18 h) and 16 March 2008 at 18 UTC (+24 h) for the run started at 15 March 2008 at 18 UTC. The position of Polar Low IIb in Fig. 14c for CTL corresponds fairly well with cloud pattern of the satellite image. IPY also has indications of a polar low, although too far North. Six hours later both CTL and IPY have small mismatches in the location of the polar low.

For Polar Low IIb, the probabilities of precipitation from CTL have a more cyclonic shape, and this is also seen for the probabilities of high wind speeds shown in Fig. 15. In general, the modelled wind speeds are weaker for Polar Low II than Polar Low I, but the only probabilities of wind speeds exceeding 20 m s^{-1} are found for CTL in Fig. 15a and Fig. 15c. IPY predicts only high wind speeds in areas connected with the dissipating Polar Low IIa for +18 h lead time (Fig. 15b). Six hours later, the predicted wind speeds in IPY are better located to the north of Polar Low IIb. CTL have two maxima in the probability of wind speeds exceeding 20 m s^{-1} in Fig. 15c. Unfortunately, the weather conditions at the campaign base was influenced by the close proximity of Polar Low IIb at Andøya, and prevented any campaign flights that day. A quantitative verification is therefore difficult to achieve.

The vertical temperature difference, discussed for Polar Low I as an indicator for increased possibility of polar low occurrence, has also for the period of Polar Low II large probabilities of exceeding thresholds of 43 °C. As for Polar Low I, the areas with high probabilities are large and it is difficult to determine critical areas connected to where Polar Low IIa and Polar Low IIb develop (not shown). The differences between CTL and IPY are also small. This indicator cannot be used for predicting individual polar lows, but can be used to map areas where polar lows may be sustained if triggered.

Campaign flights were prevented due to bad weather on 16 March 2008. However, on 15 March 2008 a flight was performed to Svalbard for other purposes, and sondes were dropped to the east and north of the developing polar lows before they started to develop. The IPY forecasts started on 15 March 18 UTC have these observations included, and one can only speculate why these extra observations did not improve IPY-forecasts for Polar Low IIb. Furthermore, to understand the way Polar Low II developed as well as the lack of quality of the forecasts, the

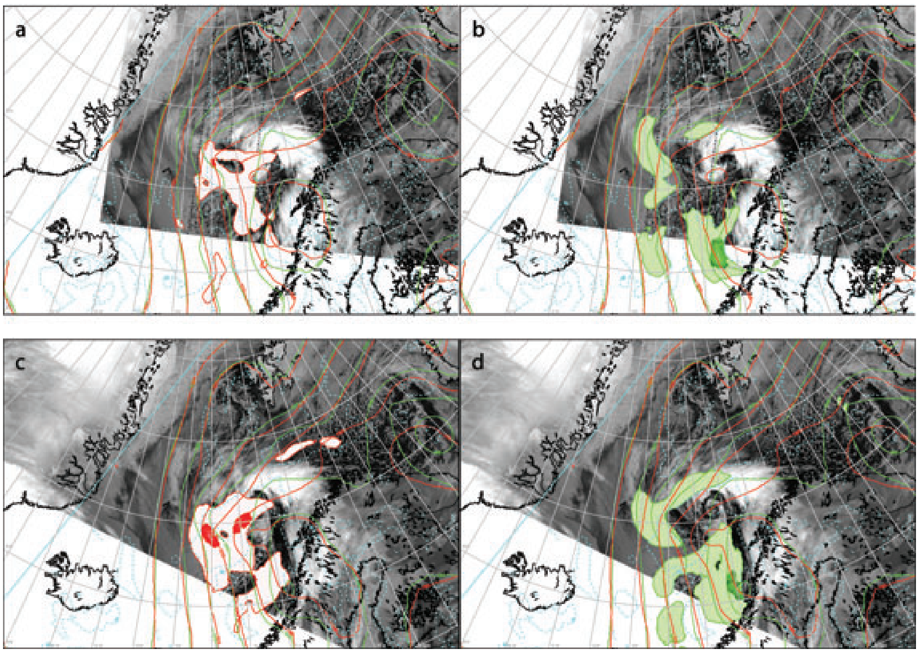

Fig. 15. The panels show predictions started at 15 March 2008 18 UTC of the ensemble mean of mean sea level pressure for CTL (red) and IPY (green) with 5 hPa contour interval. The dotted cyan line is the 95% confidence level for differences (10 m wind) between CTL and IPY. Light red/green shadings are probabilities above 30% of wind speed exceeding 15 m s^{-1}. Darker red/green are probabilities above 10% of wind speeds exceeding 20 m s^{-1}. Left-hand panels are CTL (a, c) and the right IPY (b,d). Top panels, forecast lead time is +18 h valid at 16 March 2008 12 UTC; bottom panels: forecast lead time is +24 h valid at 16 March 2008 18 UTC.

mechanisms behind the rapidly growing Polar Low IIb must be better understood. Unfortunately, this is not feasible with the available data and the fact that the model simulations are quite poor.

5. Conclusions

NORLAMEPS is an operational short-range EPS, which has been operational at met.no since February 2005. The system combines TEPS and LAMEPS, producing a 42 member EPS out of which two members are control forecasts from different analyses. The combination of TEPS and LAMEPS is supposed to partly account for forecasts errors caused by model imperfections, and it increases the size of the ensemble without further cost. This study has demonstrated that NORLAMEPS in general terms produces considerably better probabilistic forecasts than the 51 member ECMWF EPS for the investigated weather parameters and for lead times up to 60 h.

The combination of the different systems into a common ensemble is the main reason for the increased quality for cases when other aspects than increased spatial resolution in LAMEPS is of importance. Such cases are seen to include precipitation for other seasons than the summer, and 10 m wind speed in late autumn and early winter. In some cases, TEPS and NORLAMEPS are both inferior to EPS, but still combine to a better NORLAMEPS than EPS according to some probabilistic verification scores.

The spatial resolution of LAMEPS is shown to be crucial for 2 m temperature, summer precipitation, and 10 m wind speed except late autumn and early winter. For the gross measure CRPSS, splitting the ensemble size in two, does not reduce the skill of NORLAMEPS considerably, as long as the combination of system diversity is kept. However, we have not investigated the effect on parameters directly related to extreme events, for which is expected that the ensemble size is more important. The results for the area under the Relative Operating Characteristics curves indeed indicate that the largest improvements by NORLAMEPS relative to ECMWF's EPS is obtained for high wind speeds and precipitation amounts.

NORLAMEPS has undergone many upgrades, and several as a part of the project IPY-THORPEX. We have shown long-term validation in this paper for wind at 10 meter height, precipitation, and 2 m height temperatures. All variables show a long term positive score when compared to EPS. Also the NORLAMEPS runs started 06 UTC, which earlier employed a 18 h time-lagged TEPS as input for LAMEPS ensemble perturbations, improved relative to EPS. The exception was for mean sea level pressure, for which improvements were not obtained for the predictions started at 06 UTC. After introducing TEPS twice per day, similar results are seen for both cycles (06 and 18 UTC).

Polar lows are small-scale cyclones associated with adverse Arctic winter weather. They may develop fast in areas which might have sparse coverage of conventional observations. Two polar lows were observed and investigated as a part of the project IPY-THORPEX, and this paper discusses the impact of extra campaign data on the operational weather forecasts of these polar lows. For this study, the limited area component of NORLAMEPS, LAMEPS, was used. Two parallel 3D-Var data-assimilation cycles were run, one without the campaign data (CTL) and one with the extra data included (IPY) to provide initial data for LAMEPS.

The first polar low investigated was significantly better forecasted with extra data, and especially for the longest lead times. The second polar low was more complex than the first, but the extra data from the campaign did improve the forecasts with long lead times for the first stage of this polar low. For shorter lead times both experiments had problems to predict the complex polar low, but the experiment without extra observations appeared to better predict the developing phase of the most intense part of the second polar low.

Polar lows develop in maritime cold air outbreak. Duty forecasters in Norway use a vertical temperature difference between SST and the temperature of 500 hPa to indicate an increased general risk of polar low development. In this study no clear difference between the two experiments in the prediction of the indicator is seen. Probably a high value is needed in the environment to sustain the growth of a polar low, but the trigger mechanism is probably more related to dynamics such as anomalies of potential vorticity or large horizontal jet-shears.

The observational network is important for making good forecasts. Regular in situ observations are sparse in the Arctic. This study shows a positive impact of extra observations in such areas, but the cases are very few. If a general increase in observational coverage cannot be afforded, an alternative approach could be to use targeted observations together with extra radiosondes on demand. The weakness of such an approach is that a good indication of the adverse weather needs to be known up to 2 d before it occurs. In such cases one may argue that extra observations only can contribute marginally to the forecast of an already predictable development. Real adverse surprises will remain in such a system.

At met.no further studies have been performed with downscaling LAMEPS with a even higher resolution, and preliminary good results have been achieved (Kristiansen et al., 2011). Such a setup is demanding on CPU time, and will probably be available only on demand in the first stage. The ongoing project GLAMEPS (Iversen et al., 2011) is partly based on similar ideas as employed in NORLAMEPS, but developed on a pan-European domain and with a considerably extended TEPS, called EuroTEPS (Frogner and Iversen, 2011). A further new feature could be to introduce ETKF (e.g. Bojarova et al., 2011) to rescale the perturbations used for NORLAMEPS (or GLAMEPS) and thus introduce initial perturbations that are closer to actual analysis errors than in NORLAMEPS.

6. Acknowledgments

The present paper was supported by the Norwegian Research Council through its International Polar Year program and the project IPY-THORPEX (grant number 175992/S30) - www.ipy-thorpex.no. The paper has made used of extra observations provided through the GTS during the campaign in March 2008, including data from the dropsendes relased by the DLR Falcon research aircraft. The authors would also like to thank Mariann Kløverød Foss for valuable help.

References

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

Boer, G. J. 2003. Predictability as a function of scale. *Atmos-Ocean* **41**, 203–215.

Bojarova, J., Gustafsson, N., Johansson, Å. and Vignes, O. N. 2011. The ETKF rescaling scheme in HIRLAM. *Tellus* **63A**, this issue.

Bowler, N. E. 2006. Comparison of error breeding, singular vectors, random perturbations and ensemble Kalman filter perturbation strategies on a simple model. *Tellus* **58A**, 538–548.

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

Bowler, N. E., Arribas, A. and Mylne, K. R. 2008. The MOGREPS ensemble prediction system. *Q.J.R. Meteorol. Soc.* **134**, 703–722.

Bowler, N. E. and Mylne, K. R. 2009. Ensemble transform Kalman filter perturbations for a regional ensemble prediction system. *Q.J.R. Meteorol. Soc.* **135**, 757–766.

Bratseth, A. M. 1985. A note on CISK in polar air masses. *Tellus* **37A**, 403–406.

Buizza, R. 1994. Localization of optimal perturbations using a projection operator. *Q.J.R. Meteorol. Soc.* **120**, 1647–1681.

Buizza, R. and Palmer, T. N. 1995. The Singular-Vector Structure of the Atmospheric Global Circulation. *J. Atmos. Sci.* **52**, 1434–1456.

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.

Calvo, J. 2007. Kain-Fritsch convection in HIRLAM. Present status and prospects. *HIRLAM newsletter* **52**, 57–64 Available at: http://hirlam.org/index.php?option=com_content&view=article&id=64&Itemid=101.

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

Du, J., DiMego, G., Tracton, M. S. and Zhou, B. 2003. NCEP short-range ensemble forecasting (SREF) system: multi-IC, multi-model and multi-physics approach. *Research Activities in Atmospheric and Oceanic Modelling* Ed: J. Cote, Report 33, CAS/JSC Working Group Numerical Experimentation (WGNE), WMO/TD-No. 1161, 5.09-5.10. Available at: http://www.emc.ncep.noaa.gov/mmb/SREF/srefWMO_2003.pdf.

Emanuel, K. A. and Rotunno, R. 1989. Polar lows as arctic hurricanes. *Tellus* **49A**, 1–17.

Ferro, C. A.T., Richardson, D. S. and Weigel, A. P. 2008. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Metorol. Appl.* **15**, 19–24.

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 based on low-resolution targeted singular vectors. *Q.J.R. Meteorol. Soc.* **128**, 1321–1341.

Frogner, I.-L. and Iversen, T. 2011. EuroTEPS - A targeted version of ECMWF EPS for the European area. *Tellus* **63A**, this issue.

Frogner, I.-L., Haakenstad, H. and Iversen, T. 2006. Limited-area ensemble predictions at the Norwegian Meteorolgical Institute. *Q.J.R. Metorol. Soc.* **132**, 2785–2808.

Garcia-Moya, J. A., Callado, A., Santos, C., Santos, D. and Simarro, J. 2007. Multi-model ensemble for short-range predictability. *3rd International Verification Methods Workshop, ECMWF, Reading, UK*.

Gustafsson, N., Berre, L., Hörnquist, S., Huang, X.-Y., Lindskog, M. and co-authors. 2001. Three-dimensional variational data assimilation for a limited area model. Part I: general formulation and the background error constraint. *Tellus* **53A**, 425–446.

Hagedorn, R., Doblas-Reyes, F. J. and Palmer, T. N. 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: basic concepts. *Tellus* **57A**, 219–233.

Hamill, T. M. and Colucci, S. J. 1997. Verification of eta-RSM short-range forecasts. *Mon. Weather Rev.* **125**, 1312–1327.

Hersbach, H. 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Am. Meteorol. Soc.* **15**, 559–570.

Ivarsson, K.-I. 2007. The Rasch Kristjansson large scale condensation. Present status and prospects. *HIRLAM Newslett.* **52**, 50–56 Available at: http://hirlam.org/index.php?option=com_content&view=article&id=64&Itemid=101.

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

Jensen, M. H., Frogner, I-L., Iversen, T. and Vignes, O. N. 2006. Limited area ensemble forecasting in Norway using targeted EPS. *ECMWF Newslett.* **107**, 23–29.

Joliffe, I. T. and Stephenson, D. B. 2003. In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, John Wiley and Sons, Chichester.

Kain, J. S. 2004. The Kain-Fritsch Convective Parameterization. An Update. *J. Appl. Meteor.* **43**, 170–181.

Kolstad, E. W. 2006. A new climatology of favourable conditions for reverse-shear polar lows. *Tellus* **58A**, 344–354.

Kolstad, E. W. and Bracegirdle, T. J. 2008. Marine cold-air outbreaks in the future: an assesment of IPCC AR4 model results for the Northern Hemisphere. *Clim. Dyn.* **30**, 871–885.

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.

Kristjánsson, J.-E., Barstad, I., Hov, Ø., Irvine, E., Iversen, T. and co-authors. 2010. The Norwegian IPY-THORPEX: improved forecasting of adverse weather in the Arctic. *Bull. Am. Meteorol. Soc.*, submitted.

Linders, T. and Sætra, Ø. 2010. Can Cape Maintain Polar Lows?. *J. Atmos. Sci.* **67**(8), 2559–2571 doi:10.1175/2010JAS3131.1.

Lindskog, M., Gustafsson, N., Navascués, B., Mogensen, K. S., Huang, X.-Y. and co-authors. 2001. Three-dimensional variational data assimilation for a limited area model. Part II: observation handling and assimilation experiments. *Tellus* **53A**, 447–468.

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

Mann, H. B. and Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60.

Marsigli, C., Montani, A., Nerozzi, F., Paccagnella, T., Tibaldi, S. and co-authors. 2001. A strategy for high-resolution ensemble prediction. II: limited-area experiments on four Alpine flood events. *Q.J.R. Meteorol. Soc.* **127**, 2095–2115.

Marsigli, C., Boccanera, F., Montani, A. and Paccagnella, T. 2005. The COSMO-LEPS mesoscale ensemble prediction system: validation of the methodology and verification. *Nonlinear Processes in Geophysics* **12**, 527–536.

Mason, I. 1982. A model for assesment of weather forecasts. *Austr. Meteorol. Mag.* **30**, 2069–2094.

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

Molteni, F., Buizza, R., Marsigli, C., Montani, A., Nerozzi, F. and co-authors. 2001. A strategy for high-resolution ensemble prediction I: definition of representative members and global-model experiments. *Q.J.R. Meteorol. Soc.* **127**, 2013–2033.

Montgomery, M. T. and Farrell, B. F. 1992. Polar Low Dynamics. *J. Atmos. Sci.* **49**, 2484–2505.

Mullen, S. L. and Buizza, R. 2002. The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF Ensemble Prediction System. *Weather and Forecasting* **17**, 173–191.

Økland, H. 1977. On the intensification of small-scale cyclones formed in very cold air masses heated by the ocean. *Institute Report Series* **26**, University of Oslo, Department of Geophysics.

Rabbe, Å. 1975. Arctic instability lows. *Meteorologiske Annaler* **6**, 303–329.

Rasch, P. J. and Kristjánsson, J. E. 1998. A comparison of the CCM3 model climate using diagnosed and predicted condensate parameterizations. *J. Climate* **11**, 1587–1614.

Rasmussen, E. A. and Turner, J. 2003. In: *Polar Lows: Mesoscale Weather Systems in the Polar Regions*, Cambridge University Press, Cambridge.

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.

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

Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S. and Rogers, E. 1999. Using ensembles for short-range forecasting. *Mon. Weather. Rev.* **127**, 433–446.

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

Toth, Z. and Kalnay, E. 1997. Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev* **125**, 3297–3319.

Undén, P., Rontu, L., Järvinen, H., Lynch, P., Calvo, J. and co-authors. 2002. HIRLAM-5 Scientific Documentation HIRLAM-5 Project. *Available from SMHI, S-601767 Norrkö*ping, *Sweden*.

Walser, A., Lüthi, D. and Schär, C. 2004. Predictability of precipitation in a cloud-resolving model. *Mon. Weather. Rev.* **132**, 560–577.

Weigel, A. P., Liniger, M. A. and Appenzeller, C. 2008. Can multimodel combination really enhance the prediction skill of probabilistic ensemble forecasts?. *Q.J.R. Meteorol. Soc.* **134**, 241–260.

Weigel, A. P. and Bowler, N. E. 2009. Comment on 'Can multimodel combination really enhance the prediction skill of probabilistic ensemble forecasts?'. *Q.J.R. Meteorol. Soc.* **135**, 535–539.

Yanase, W. and Niino, H. 2005. Effects of baroclinicity on the cloud pattern and structure of polar lows: a high-resolution numerical experiment. *Geophys. Res. Lett.* **32**, L02806, doi:10.1029/2004GL020469.

Yue, S. and Wang, C.-Y. 2002. The influence of serial correlation in the Mann-Whitney test for detecting a shift in median. *Adv. Water Res.* **25**, 325–333.