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

Tellus (2011), 63A, 513–530
Printed in Singapore. All rights reserved

TELLUS

Evaluation of 'GLAMEPS'—a proposed multimodel EPS for short range forecasting

By TROND IVERSEN[1]*, ALEX DECKMYN[2], CARLOS SANTOS[3], KAI SATTLER[4], JOHN BJØRNAR BREMNES[1], HENRIK FEDDERSEN[4] and INGER-LISE FROGNER[1]
[1]*Norwegian Meteorological Institute (met.no), Oslo, Norway;* [2]*Royal Meteorological Institute (KMI), Brussels, Belgium;* [3]*Spanish Meteorological Agency (AEMET), Madrid, Spain;* [4]*Danish Meteorological Institute (DMI), Copenhagen, Denmark*

(Manuscript received 3 May 2010; in final form 16 December 2010)

ABSTRACT

Grand Limited Area Model Ensemble Prediction System (GLAMEPS) is prepared for pan-European, short-range probabilistic numerical weather prediction of fine synoptic-scale, quasi-hydrostatic atmospheric flows. Four equally sized ensembles are combined: EuroTEPS, a version of the global ECMWF EPS with European target; AladEPS, a downscaling of EuroTEPS using the ALADIN model; HirEPS_K and HirEPS_S, two ensembles using the HIRLAM model nested into EuroTEPS including 3DVar data-assimilation for two control forecasts. A 52-member GLAMEPS thus samples forecast uncertainty by three analysed initial states combined with 12 singular vector-based perturbations, four different models and the stochastic physics tendencies in EuroTEPS. Over a 7-week test period in winter 2008, GLAMEPS produced better results than ECMWF's EPS with 51 ensemble members. Apart from spatial resolution, the improvement is due to the multimodel combination and to a smaller extent the dedicated EuroTEPS. Ensemble resolution and reliability are both improved. Combining uncalibrated ensembles is seen to produce a better combined ensemble than the best single-model ensemble of the same size, except when one of the single-model ensembles is considerably better than the others. Bayesian Model Averaging improves reliability, but needs further elaboration to account for geographical variations. These conclusions need to be confirmed by long-period evaluations.

1. Introduction

This paper presents an evaluation of first tests of a system for short-range, European-wide probabilistic numerical weather prediction (NWP) called Grand Limited Area Modeling Ensemble Prediction System (GLAMEPS). GLAMEPS is intended for operational production as a part of the cooperation between two European consortia for short-range NWP: HIgh Resolution Limited Area Modeling (HIRLAM) and Aire Limitée Adaptation dynamique Developpement INternational (ALADIN).

Probabilistic weather prediction has gradually become realized operationally since Lorenz (1963) and later publications firmly established the intrinsic limitations of the weather prediction problem (Palmer, 2000; Lewis, 2005). Such predictions are realized by using ensembles of alternative predictions to sample the time development of the probability density function (pdf) of atmospheric states. The early proposal by Leith (1974) to use random perturbations proved inadequate, because the full dimension of the problem was drastically larger than the number of ensemble members that could be afforded when produced by numerical models that were state-of-the-art. Random perturbations did not grow fast enough and overestimated the real predictability. This shortcoming could also be due to the low physical realism of the available numerical weather prediction (NWP) models, which therefore underestimated the dynamic activity in the atmosphere. Lorenz (1982) compared forecasts of different lead times that were valid at the same time to demonstrate that error growth, which was 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.

Considerable development work has been invested in selecting initial state perturbations that grow sufficiently fast in the models when compared to prediction errors. Perturbations depending on the actual atmospheric state proved to be crucial.

*Corresponding author.
Present address: ECMWF, Shinfield Park, RG2 9AX, United Kingdom.
e-mail: trond.iversen@met.no
DOI: 10.1111/j.1600-0870.2010.00507.x

Hence ECMWF developed their ensemble prediction system (EPS) based on singular vectors that maximize the total energy in the extra-tropical Northern Hemisphere 2 d into the future (Buizza et al., 1993, 2000; Buizza, 1994; Molteni et al., 1996). More or less in parallel, National Centers for Environmental Prediction (NCEP, USA) developed the breeding technique as a non-linear generalization of the calculation methods for Lyapunov vectors (Toth and Kalnay, 1993, 1997). Neither of these methods estimated the actual initial state uncertainty explicitly, but found spatial structures of variable perturbations that either had grown fast recently (breeding) or would grow fast during the next couple of days (singular vectors). As advanced variational data-assimilation techniques developed and satellite information was properly included by the technique of observational operators, more attention was paid to different sources of initial-state errors. In particular, ensemble techniques were recognized as a possible tool to estimate the time-dependent model error in combination with Kalman filtering. This idea of ensemble Kalman filtering, EnKF, was first proposed for oceanic data-assimilation (Evensen, 1994; Evensen and van Leeuwen, 1996), but is also being pursued for potential atmospheric applications (Fisher and Courtier, 1995; Houtekamer and Mitchell, 1998; Fischer and Andersson, 2001).

A potential benefit of EnKF is a combined estimate of the initial state and its actual uncertainty. Singular vectors, on the other hand, are constructed to maximize the total energy of the perturbations after 2 d, but may be physically unrealistic. Barkmeijer et al. (1999) therefore constructed initial state singular vectors that projected onto leading eigenvectors of the error covariance matrix in the variational data-assimilation. These structures were more similar to the structures developed by breeding (Toth and Kalnay, 1997). Since compared the benefits for the medium range forecasts at ECMWF were small relative to the computational efforts the method was abandoned.

Since EnKF in its full implementation is computationally expensive for atmospheric applications, simplifications have been sought. The ensemble transform Kalman filter (ETKF) was developed by Bishop et al. (2001) for targeting campaign observations. The method can be viewed as a generalization of breeding (Wang and Bishop, 2003), where the perturbations in the breeding cycle are rotated towards leading eigenvectors of the error covariance matrix when rescaled. The method operates with a low number of degrees of freedom which is compensated by using a perturbation inflation factor.

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 considerable realism. Magnusson et al. (2009) demonstrated that the recent versions of the ECMWF Integrated Forecast System (IFS) model had become dynamically active enough to sustain fast growth of arbitrary but dynamically realistic perturbations. This triggers increased attention towards estimating real initial state uncertainties rather than constructing potentially unrealistic perturbations that grow fast. In such a regime, running ensembles of data-assimilation (EDA) cycles in parallel are considered ideal for generating alternative initial states (e.g. Hamill et al., 2000), although it is expensive. In this approach one may use the same model version and perturb the observations, one may use different models or model versions without perturbing the observations, or one may do both. Applying EDA in combination with singular vectors is a new development at ECMWF (Buizza et al., 2008).

In short range weather prediction actual analysis errors may indeed dominate over fast-growing structures. A short comparison of various techniques for initial state perturbations with random perturbations using a simplified model is discussed by Bowler (2006). The study indicated that the EnKF technique performs best. Nevertheless, the computationally much cheaper ETKF is often considered, and it was introduced in the limited area ensemble prediction system (LAM-EPS) run for the short range at UK MetOffice (Bowler et al., 2008). Nevertheless, in full scale verification ETKF was found to be slightly inferior to straightforward downscaling of the ensembles produced with their global model (Bowler and Mylne, 2009).

Even though (Simmons and Hollingsworth, 2002) demonstrated the increased level of realism in model-calculated error growth 20 yr after the study of Lorenz (1982), the contribution of model approximations to error growth needs to be taken more explicitly into account (Orrell et al., 2001). ECMWF therefore introduced the stochastic physics scheme (Buizza et al., 1999), which has later been further considerably refined. Inaccuracies in the formulations of physical processes in NWP models have also been accounted for with considerable success by using different model versions or models in the generation of ensembles (Du et al., 2003; Doblas-Reyes et al., 2005; Hagedorn et al., 2005; Garcia-Moya et al., 2007). Weigel et al. (2008) and Weigel and Bowler (2009) discussed under which conditions multimodel ensembles can outperform single-model ensembles. Based on simplified 'toy models' they argue that in multimodel combinations in general only can be expected to be better than all the single-model ensembles if the latter are underdispersed. However, for normally distributed variables, for example, in the short range, multimodel combinations may also improve over single-model, well-calibrated ensembles. The discussions in Hagedorn et al. (2005) emphasized that reliability and consistency were strong contributing factors in addition to error compensation for the success of multimodel ensembles over single-model, and that these factors could not be replaced by relatively trivial statistical calibration. Preliminary unpublished results (R. Hagedorn, pers. comm.) indicate that calibration using reforecasts (Hamill et al., 2006) may come out better than multimodel combinations of single-model ensembles with uneven quality, but not when they are comparable.

The aim of GLAMEPS is to construct a well-calibrated, pan-European ensemble for short-range NWP by accounting for both initial state and model inaccuracies. Model uncertainties are

presently taken into account by using a small number of different model versions and different models. Initial state uncertainties are taken into account in two ways. Ensemble perturbations are imported from a global system, EuroTEPS, based on singular vectors that maximize total energy in a target domain after 24 h in addition to the regular Northern Hemispheric singular vectors (Leutbecher, 2007; Frogner and Iversen, 2011). EuroTEPS also provides perturbations at the lateral boundaries during the prediction period. Additional initial state perturbations are included as three different assimilation cycles are run in parallel with different models and model versions, but without perturbing the observations.

GLAMEPS aims at predicting atmospheric features which spatial scales are intermediate between the synoptic presently covered by leading global EPS and the so-called convection-permitting scales. Whilst GLAMEPS will operate with approximate pan-European coverage (excluding eastern parts and Greenland), several LAM-EPS systems presently operate on these scales for parts of Europe (Marsigli et al., 2005; Frogner et al., 2006; Garcia-Moya et al., 2007; Hágel and Horányi, 2007; Bowler et al., 2008; Kann et al., 2009; Aspelien et al., 2011) and in other parts of the world (Du et al., 1997; Hamill and Colucci, 1997; Stensrud et al., 1999; Seko et al., 2007); see also http://www.smr.arpa.emr.it/tiggelam/.

This paper presents results from preparatory experiments for constructing a first operational version of GLAMEPS. The test period consists of seven consecutive weeks starting from 17 January 2008, and results are only taken as indications in order to investigate if full-scale experiments in an operational setting are recommendable. For a range of standard probabilistic verification parameters at observational sites, GLAMEPS-results are compared to results from the operational 51-member EPS run at ECMWF (EPS51) during the winter period in 2008. The results are also compared to a few potentially alternative options for the ensemble system. Thereby the relative contributions to skill enhancement by GLAMEPS are investigated, including the multimodel approach, and the targeted singular vectors in EuroTEPS. The experiments also include a first attempt at using Bayesian model averaging, BMA, (Raftery et al., 2005) to calibrate and combine the single-model ensembles.

In Section 2, we present a likely set-up for an operational GLAMEPS along with a description of the models and tools. Results of experiments are discussed in Section 3, including comparisons with EPS51 and investigations of the benefits of the multimodel approach and the use of EuroTEPS. Section 4 finally presents conclusions and possible developments in the near future.

2. Constructing GLAMEPS

The basic idea behind GLAMEPS is to account for all major sources of forecast inaccuracy over the next 2 d by using a multimodel approach which includes several data-assimilation cycles, and while EuroTEPS is intended to secure that atmospheric instability on the relevant synoptic scales over Europe are accounted for. The main challenge for GLAMEPS and other short-range NWP is to produce significantly better forecasts in the short range than the best available global forecasts. We therefore use the global model EPS from ECMWF as a benchmark. The next benchmark would be other similar short-range EPS.

It is a considerable practical challenge that systems for short-range forecasting have to wait for input data from a global system designed for the medium range for the open lateral boundary conditions. Only after the coarser resolution data is available, the finer resolution predictions can be prepared for production. Hence, the short-range forecasts need to be produced under excessively strict time constraints, and efficient exploitation of the computer resources is necessary in order to produce valuable additions to the global forecasts. Even though these aspects of exploring short-range predictions are of practical rather than scientific nature, they are important in the applied science of weather forecasting, in a similar way as physical parametrizations has been necessary to develop due to limited computer resources.

2.1. Technical set-up

Figure 1 shows a schematic flow diagram for data and tasks in an anticipated operational GLAMEPS. The production is launched and monitored at ECMWF computers by a script system (http://www.ecmwf.int/products/data/software/sms.html). The entire GLAMEPS is thus produced at the high-performing computer facility at ECMWF.

The anticipated real-time production chain is presented below the thick horizontal line in Fig. 1. More details about the single-model ensemble components and how they are combined are given in the next sections. EuroTEPS provides global ensemble members based on the operational EPS51 supplemented with higher-resolution singular vectors targeted to Europe (Frogner and Iversen, 2011). Three-hourly data sets at vertical coordinate levels used in by the ECMWF-IFS (IFS is the integrated forecast system) are transferred to HIRLAM and ALADIN. A program code (GL) enables flexible transfer of atmospheric data between EuroTEPS and ALADIN, while ground surface model data imported to ALADIN from the French global model ARPEGE-IFS (Courtier et al., 1991; Geleyn et al., 1995). The strategy is to produce as many as possible of the HirEPS and AladEPS (and even EuroTEPS) ensemble members by running parallel jobs, thus saving production time.

AladEPS, HirEPS_S, HirEPS_K and EuroTEPS produce ensembles of field data in a common presentation grid with the same rotated latitude–longitude coordinates as used when integrating HIRLAM. Optionally, BMA can be used to calibrate and combine all ensemble members, in which case a common probability density function for each combined and calibrated variable is constructed. At present stage the software developed by the

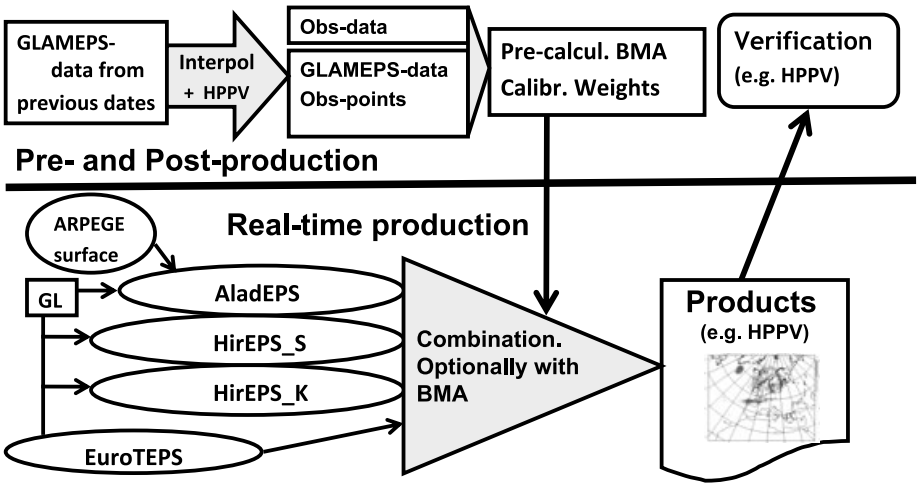

Fig. 1. Flow of data and tasks in an anticipated operational GLAMEPS. See Section 2.1 in the main text for explanations.

Spanish met-service, Hppv, is used to calibrate and combine with BMA, as well as producing selected probabilistic forecast graphics. When BMA is not applied, the multimodel ensemble is simply pooling all ensemble members.

Pre-production preparation shown above the horizontal line in Fig. 1 includes the calculation of BMA calibration weights for combining the probability density functions from the single-model ensembles. This part makes use of observational data over a pre-defined time-window. In this paper, we have applied BMA for wind speed at 10 m height with calibration statistics determined over the previous 3 d. The short period is chosen in order to maintain dependency of the actual weather situation, but since the statistics are evaluated by pooling data from the entire domain, heterogeneity is neglected. The system uses free available software for this (http://cran.r-project.org/web/packages/ensembleBMA/index.html).

Post-production evaluation, also shown above the horizontal line in Fig. 1, is the verification run on a sample of previous forecasts and relevant observations over typically a month or longer. The verification of probabilistic GLAMEPS products is run outside the real-time operation, and typically after a minimum set of cases of ensemble forecasts are ready for comparison with observations. In an operational setting, full 1-month periods or longer would be a typical running verification period. A range of probabilistic verification products are made, including rank histograms, spread-skill diagrams, reliability and sharpness diagrams, Brier skill score (BSS), discretely ranked probability skill score (DRPSS), relative operating characteristics curves (ROC) and curves for expected relative value as a function of cost-loss ratio. Skill scores are evaluated by relating to forecasts that could be produced by the statistics over the sampling period ('sample climatology'). Hppv is presently scheduled for both producing graphical output and for verification.

For the experiments developed for this paper no production was made in real time. They were performed by launching and monitoring the elements in Fig. 1 separately in sequence. This version of GLAMEPS is referred to as GLAMEPS_v0. Since March 2010 an automatic operational test production (GLAMEPS_v1) has been running twice per day at ECMWF, but with EuroTEPS replaced by selected ensemble members from EPS51. The replacement is preliminary until EuroTEPS is upgraded with ensemble data assimilation (EDA) and higher resolution as already in EPS51 (Buizza et al., 2008).

2.2. *EuroTEPS*

EuroTEPS (Frogner and Iversen, 2011) is a version of the ECMWF IFS EPS for which additional ensemble spread for the first 24 h are sought over Europe by using singular vectors targeted to three overlapping target areas that together cover major parts of Europe. The targets ensure pan-European ensemble spread with fewer ensemble members than EPS51, and Fig. 1 in Frogner and Iversen (2011) shows a map of the targets which cover the northern, middle and southern parts of Europe, respectively. The targeted SVs (TSVs) are calculated with resolution T159L62 and are by construction orthogonal, with respect to the Total Energy (TE) inner product, to the regular T42L62 Northern Hemispheric SVs optimized over 48 h and used operationally to produce EPS51 (Leutbecher, 2007). Ten TSVs are calculated per target area, and the 30 TSVs are afterwards mutually orthogonalized. Both initial and evolved TSVs and NH SVs are combined to construct a set of unique initial perturbations by the same type of Gaussian sampling as used for EPS51 (Leutbecher and Palmer, 2008).

For the experiments in this paper, EuroTEPS has been set up to produce a control prediction from the T799L91 deterministic operational analysis taken out with resolution T399L62. The alternative ensemble members are produced in parallel mode using ECMWF IFS (cycle 35r2) with resolution T159L62 for TSVs and T399L62 for the ensemble generation. Three-hourly

data suitable for input to the HIRLAM and ALADIN limited area models are provided. Furthermore, additional EuroTEPS data for provided separately for combination with ensemble members from the ALADIN and HIRLAM.

2.3. AladEPS and HirEPS

Figure 2 shows the present integration domains and the common domain for GLAMEPS output. The ALADIN model (ALADIN International Team, 1997) is used to downscale (without data-assimilation) the EuroTEPS atmospheric control and ensemble members after interpolation. The surface fields are taken from the ARPEGE IFS (Courtier et al., 1991; Geleyn et al., 1995), since the surface model in ALADIN, ISBA (Interaction Soil–Biosphere–Atmosphere, Noilhan and Planton, 1989), is incompatible with the scheme used with the ECMWF IFS models. The version of ALADIN used for the GLAMEPS experiments presented in this paper is cycle 31t1. ALADIN is a spectral limited area model (Haugen and Machenhauer, 1993) and is run with hydrostatic semi-Lagrangian dynamics. To counteract the periodicity of the basis functions used with the spectral technique, artificial damping is applied in a horizontal extension zone. The model employs 37 hybrid-coordinate, terrain-following levels in the vertical direction.

The parametrization of turbulent diffusion is a diagnostic 1st order closure scheme (Louis et al., 1981). Deep convection is parametrized by the diagnostic mass flux scheme of Bougeault (1985) and Bougeault and Geleyn (1989). The cloud microphysics includes five phases of the water substance, and the radiation scheme, which is the same as used in ECMWF IFS, is called once per hour. The solution of importing coarse-resolution ground surface data for ISBA is preliminary until a surface data-assimilation will be introduced together with a new surface soil scheme. First results with the latter indicate improvements compared to importing data from ARPEGE. Further description of ALADIN can be found in http://www.cnrm.meteo.fr/aladin/spip.php?article129.

The HIRLAM model (see http://Hirlam.org) is version 7.2 revision r6270; see Yang (2008). The model employs 40 levels in the vertical and is set up to run with two different choices for the cloud physics parametrization. HirEPS_S employs the stratiform and convective cloud and precipitation scheme STRACO (Sass et al., 1999; Undén et al., 2002), while HirEPS_K uses the Kain-Fritsch schemes for deep cumulus (Kain and Fritsch, 1990; Kain, 2004; Calvo, 2007) and Rasch and Kristjansson (1998) for stratiform clouds and precipitation (Ivarsson, 2007).

A prognostic scheme for turbulent kinetic energy is employed. For the ground surface a tile approach is used for 7 surface types together with the two-layer force-restore ISBA scheme (Noilhan and Planton, 1989). Each model version produces a separate control run starting at 00 and 12 UTC from analyses produced by two independent 6-hourly 3DVar assimilation cycles (Gustafsson et al., 2001; Lindskog et al., 2001). Twice the number of ensemble members from EuroTEPS is thus produced by HirEPS. Ensemble perturbations for the initial state and lateral and lower boundary data are taken from EuroTEPS. Visit http://hirlam.org/index.php?option=com_content&view=category&layout=blog&id=36&Itemid=99 for further information about HIRLAM.

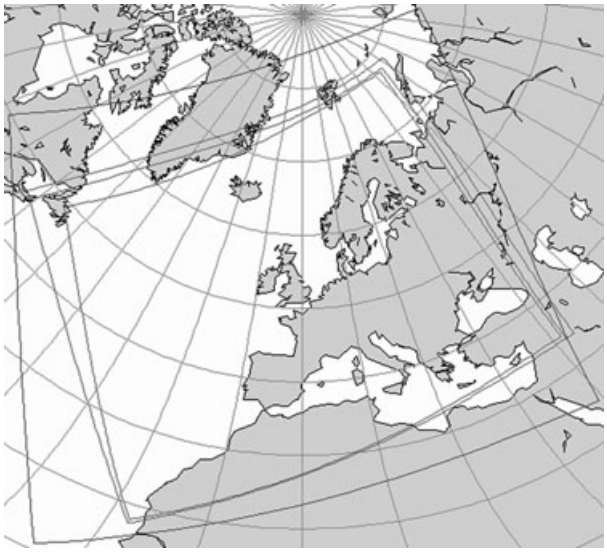

Fig. 2. Integration areas for the limited-area models used in GLAMEPS. Outer area: output domain for data from EuroTEPS in model levels; Medium area: ALADIN domain with extension zones; Inner area: HIRLAM domain and output domain for common products.

2.4. The GLAMEPS ensemble: combination and calibration

EuroTEPS is designed to produce one control forecast and an even number of alternative ensemble members. In this paper we have used 10, 12 and 50 for the different experiments discussed in the next sections. When combined into GLAMEPS, either by pooling together the ensemble members or by using BMA, also the two HirEPS versions and AladEPS are run for one control plus the same even number of ensemble members as produced by EuroTEPS. The AladEPS control forecast is a downscaled version of the EuroTEPS control, while the two HirEPS control forecasts starts from analyses produced with two different 3DVar assimilation cycles.

Initial state uncertainty is thus accounted for in GLAMEPS partly by the three different analyses and partly by the singular vectors used in EuroTEPS. Model uncertainty is accounted for partly by running four different models and partly by the stochastic physics used in EuroTEPS, which is the same as used in the operational EPS at ECMWF (Buizza et al., 1999).

Uncertainties entering the integration domains of the limited area models laterally are accounted for by EuroTEPS. Uncertainties originating from the ground surface are mainly taken into account by the different surface parametrizations in the models. In addition, AladEPS import surface input data from ARPEGE while HirEPS use EuroTEPS data. The uncertainty in surface data is probably the weakest part of the construction of GLAMEPS at present.

The ensemble members from the different models are combined into a unified probabilistic prediction. The combination can be done using BMA, which also enforce calibration with respect to reliability and the spread-skill relation (Raftery et al., 2005). Here we have made a few tests with BMA for wind speed at 10 m height with coefficients determined over the most recent 3 d prior to the present. Each single-model ensemble is bias-corrected before combination. In principle separate BMA weights should be estimated for each observation site or for sites with comparable climate statistics. Much longer calibration periods are then necessary for statistical significance, at the expense of flow dependency. In any case, applying BMA to extreme (and rare) weather events will be difficult due to problems with statistical confidence. If BMA is not applied, which so far is the normal situation, the single-model ensembles are combined by pooling together the single-model ensemble members without any corrections.

3. Configuration experiments

This paper presents and discusses a few experiments for different options of GLAMEPS in order to guide the set-up for full operational production. Table 1 gives a short overview of the experiments which are named EXP_0.1, EXP_0.2, EXP_0.3, EXP_0.4, where the zero refers to version 0 of GLAMEPS. In addition results from the global ensemble system with 51 members (EPS51), which was operational at ECMWF during the 7-week period, are used as reference benchmark. The experiments with GLAMEPS_v0 are set-up using the model versions and model domains described in Section 2. The forecast length is 42 h and the grid-resolution 0.115° (∼12.8 km) for HirEPS, and 12.9 km (∼0.1161°) for AladEPS. For EuroTEPS the resolution is T399L62 (∼55 km). All experiments are performed by running the links in the chain (Fig. 1) manually in sequence. We have made the experiments over 7 weeks twice per day during winter 2008 [17 January–5 March (00UTC and 12UTC)], except for EXP_0.3 which is run over the first four of the weeks.

In short, the comparison between EXP_0.1 and EXP_0.2 is done to measure the difference between, respectively, 44 and 52 ensemble members in GLAMEPS. The difference between EXP_0.2 and EXP_0.3 (calculated over 4 weeks only) diagnoses the value added by using targeted singular vectors in EuroTEPS compared to using ensemble members from EPS51. Finally, the multimodel approach in GLAMEPS is discussed based on differences between the multimodel GLAMEPS in EXP_0.2 and a range of single-model ensembles of same size in EXP_0.4. The test periods include several potentially high-impact weather events in Europe, but they are too few to yield statistically significant verification.

3.1. Probabilistic verification and reference predictions

Below we have chosen to mainly show verification results for 42 h predictions, except for some parameters for which time development is shown over the forecast range, and for precipitation where amounts accumulated from +18 to +24 h are

Table 1. Overview of the four calibration experiments. EXP_0.4 produces four different ensembles of the same size, three of which are with a limited area model

	Total no. of ensemble members	Global input	LAM-EPS	Test period	Purpose
EXP_0.1	Multimodel 44 members	EuroTEPS_11	HirEPS-K_11 HirEPS-S_11 AladEPS_11	17 January–05 March	Sensitivity to ensemble size
EXP_0.2	Multimodel 52 members	EuroTEPS_13	HirEPS-K_13 HirEPS-S_13 AladEPS_13	17 January–05 March	Control
EXP_0.3	Multimodel 52 members	EPS51, members 1–12	HirEPS-K_13 HirEPS-S_13 AladEPS_13	17 January–13 February	Sensitivity to EuroTEPS
EXP_0.4	4 alternatives, 51 members	EuroTEPS_51	HirEPS-K_51 HirEPS-S_51 AladEPS_51	17 January–05 March	Multi- vs. Single- model eps

shown, since the number of observations is considerably smaller at longer lead times. European-wide synoptic surface observations and radiosondes are used in the verification.

Results are shown for a range of probabilistic prediction scores for selected variables and events. The amount of observation data is considered sufficient for statistically confident conclusions to be drawn for frequently occurring weather events, even though generalizations to other seasons are not recommended. In addition to basing our conclusions on a single winter period, we also pool data from different subregions across various climate regimes together. Ideally all the statistics presented should be made for each observation site or for subregions in Europe with quasi-homogeneous climate statistics (Hamill, 2001). Examples also include events of relatively high wind speeds and precipitation amounts that still occur regularly over 7 weeks in winter in coastal and mountainous areas in Europe. These examples should only be taken as indications of the forecast quality of potentially high impact weather.

In any case, fully robust conclusions of the systems' properties relative to each other need longer experiment periods to account for seasonal and interannual variability in European flow patterns. This is in particular important for extreme weather events. Nevertheless, the 7 weeks in winter 2008 experienced considerable flow variations, starting with a very high index for the North-Atlantic Oscillation (Osborn, 2006), which reduced during February and switched to a large negative value in March (http://www.cru.uea.ac.uk/~timo/datapages/naoi.htm). The Scandinavian pattern, which is closely linked to patterns of precipitation variations in north-west Europe (Barnston and Livezey, 1987), changed from negative to positive over the same period (http://www.cpc.noaa.gov/data/teledoc/scand.timeseries.gif).

We include score parameters that measure consistency between ensemble spread and prediction skill, reliability of predicted probabilities, predicted ensemble resolution and the potential value of the predictions for users that may take decisions influenced by the weather forecasts. Also the brier skill score (BSS), the discrete ranked probability skill score (DPRSS) and relative operating characteristics (ROC) curves are evaluated. Reference predictions used in skill scores and the climatic occurrence of events shown in reliability diagrams are all from the 7-week sample statistics. The reference is the same for all forecasts that are verified.

3.2. *GLAMEPS_v0 compared to operational ECMWF EPS*

First we concentrate our discussions results from EXP_0.1 (44 members) and EXP_0.2 (52 members) compared to EPS51. Figure 3 shows rank histograms (Talagrand diagrams) for the 7

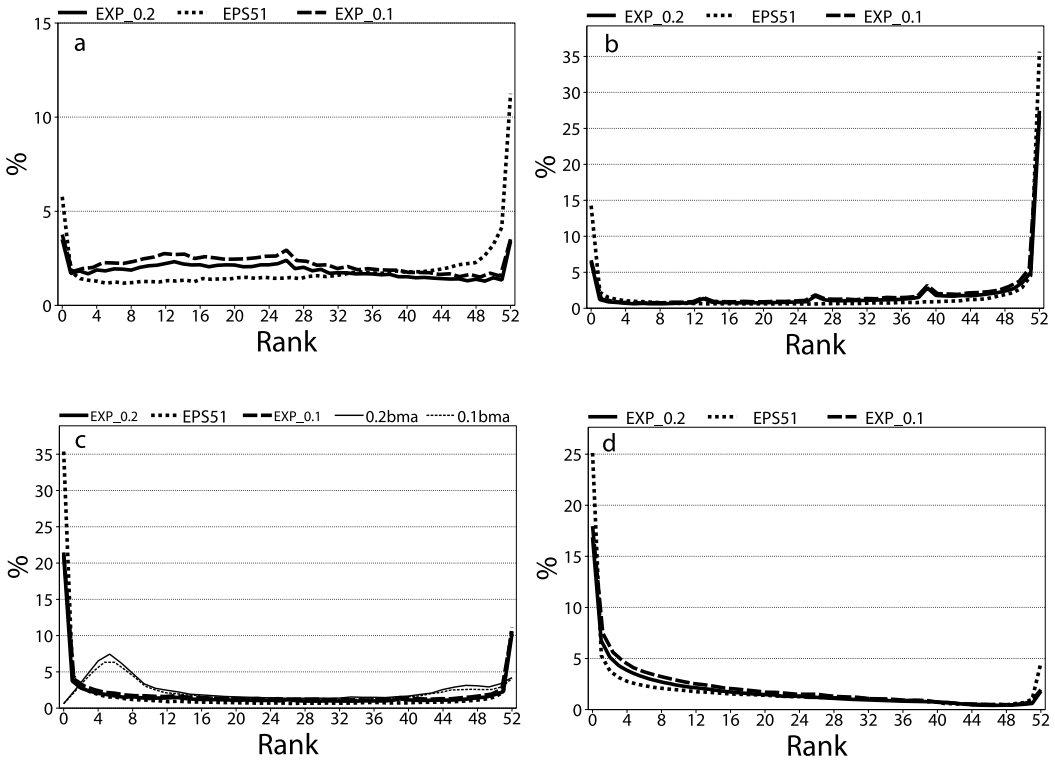

Fig. 3. Rank histograms for +42 h predicted mean sea level pressure (a), 2 m temperature (b), 10 m wind speed (c) and +18 to +24 h predicted 6-hourly precipitation (d). Thick lines: continuous: 52-member GLAMEPS (EXP_0.2); dotted: 51 member operational EPS; dashed: 44-member GLAMEPS (EXP_0.1). Thin lines in (c) are BMA-calibrated wind speed for EXP_0.2 (continuous) and EXP_0.1 (dotted).

winter weeks. The figures confirm a tendency of underspread in all the ensembles, even though some underspread should be expected when observation inaccuracy is not explicitly accounted for (Hamill, 2001; Saetra et al., 2004; Bowler, 2008). The underspread is reduced in the GLAMEPS experiments compared to EPS51, but the difference between EXP_0.1 and EXP_0.2 is negligible. For precipitation (Fig. 3d) there is a slight bias towards higher ranks, which is larger for GLAMEPS than for EPS. This is probably associated with too few occurrences of no precipitation in the models, but it is also likely that high precipitation amounts are overestimated.

The BMA applied to the 10 m wind speed is seen to inflate the ensemble spread so that observations, which were ranked outside the range of the ensemble members, become ranked amongst the lower, respectively higher, ensemble members. Even though the overall ensemble reliability will increase due to this (see Fig. 4a), the result is that too many observations are ranked in the extreme parts. The reliability plot for 10 m wind speed exceeding 10 m s^{-1} in Fig. 4a show considerably better results for GLAMEPS than EPS51 even though both are overconfident, that is, the probabilities of the events are overpredicted. The results for EPS51 are close to the line of no skill. The figure clearly demonstrates that BMA considerably improves the overall reliability, except for probabilities lower than ∼50% which are underpredicted. The corresponding sharpness diagram (not shown) shows a slightly reduced sharpness for this predicted event after BMA.

Figure 4 also shows examples of reliability diagrams for event thresholds for screen temperatures (2 m above ground). In both cases GLAMEPS is better than EPS51. For the frequently occurring event of T2m > -10 °C (∼98% of the sample), the improvement is seen for predicted probabilities lower than 50%, whilst for the quite rarely occurring event of T2m > 10 °C

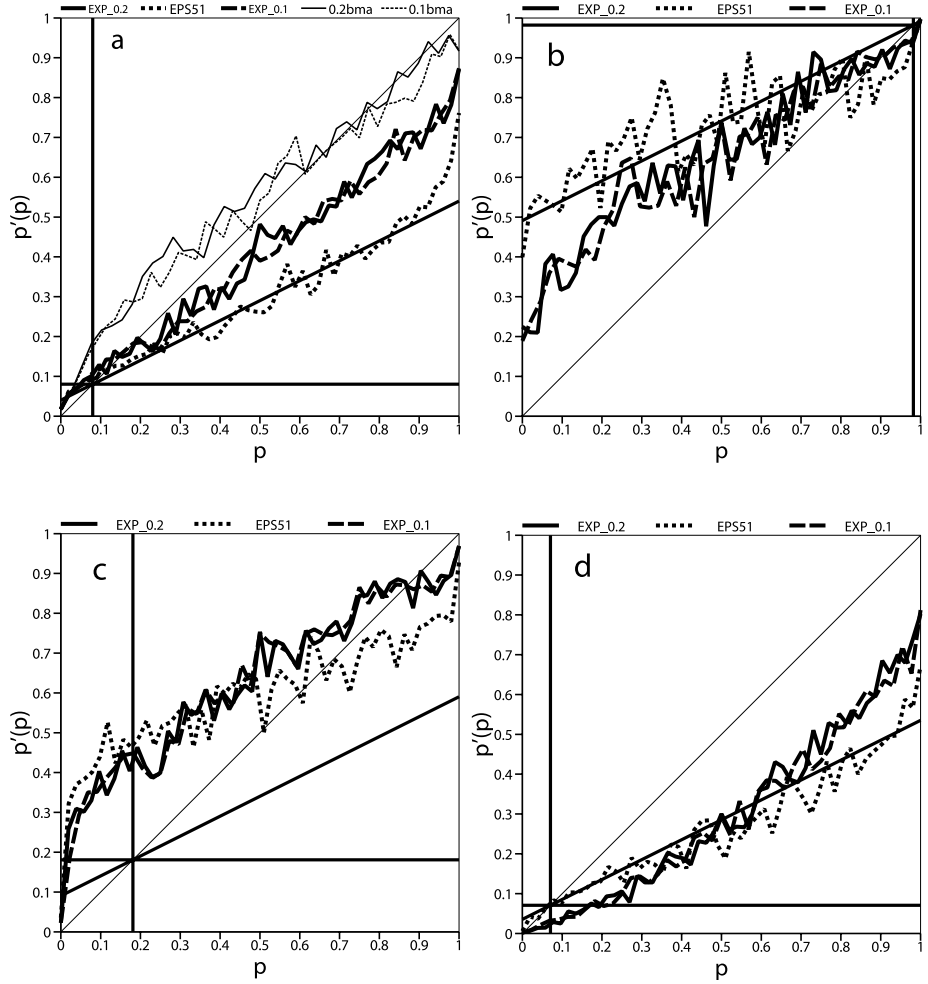

Fig. 4. Reliability diagrams, i.e. the conditional frequency p′ of observed events with predicted probabilities p. Thick lines: 52 member GLAMEPS (EXP0.2, continuous); 51 member operational EPS (dotted), 44 member GLAMEPS (EXP_0.1, dashed). Events: +42 h predicted wind speed >10 ms^{-1} (a); +42 h predicted 2 m temperature > −10 °C (b); >10 °C (c); and +18 to +24 h predicted 6-hourly accumulated precipitation >1 mm (d). Thin lines in (a): BMA calibrated wind speeds EXP_02 (continuous) and EXP_0.1 (dashed). The diagonal is perfect reliability; the half diagonal is zero brier skill score (no skill), horizontal and vertical lines is the observed frequency over the period (no resolution).

(∼18% of the sample), the improvement occurs for predicted probabilities higher than 70%. This indicates that GLAMEPS is better than EPS51 at reproducing considerable deviations from the sample average. The two GLAMEPS options, 44 members and 52 members, are indistinguishable in these plots.

The bias noticed for precipitation amounts in Fig. 3d can also be seen in the reliability diagrams for precipitation events (Fig. 4d). The overconfident predictions are larger for small precipitation amounts, and particularly large for GLAMEPS. For predicted probabilities over 70% GLAMEPS is clearly better, but still quite close to the line of no skill.

Figure 5 shows BSS for selected events associated with 2 m temperature, 10 m wind speed and 6 h accumulated precipitation, and Fig. 6 shows DRPSS over a range of event thresholds for 2 m temperature and 10 m wind speed. DRPSS is a measure of BSS integrated over many events and is as such a more robust measure of the overall probabilistic forecast quality of a system than BSS. On the other hand, DRPSS may mask information about less frequent events that can be of importance for users. We therefore include some BSS diagrams in addition to DRPSS.

All the figures show that GLAMEPS provides considerably improved probabilistic predictions compared to EPS51, and that the differences between the 44 member EXP_0.1 and the 52 member EXP_0.2 appear to be almost negligible (the blue curves appear to be missing in Fig. 6 because they are hidden by the red). Slight differences can be seen for precipitation (Figs 5e

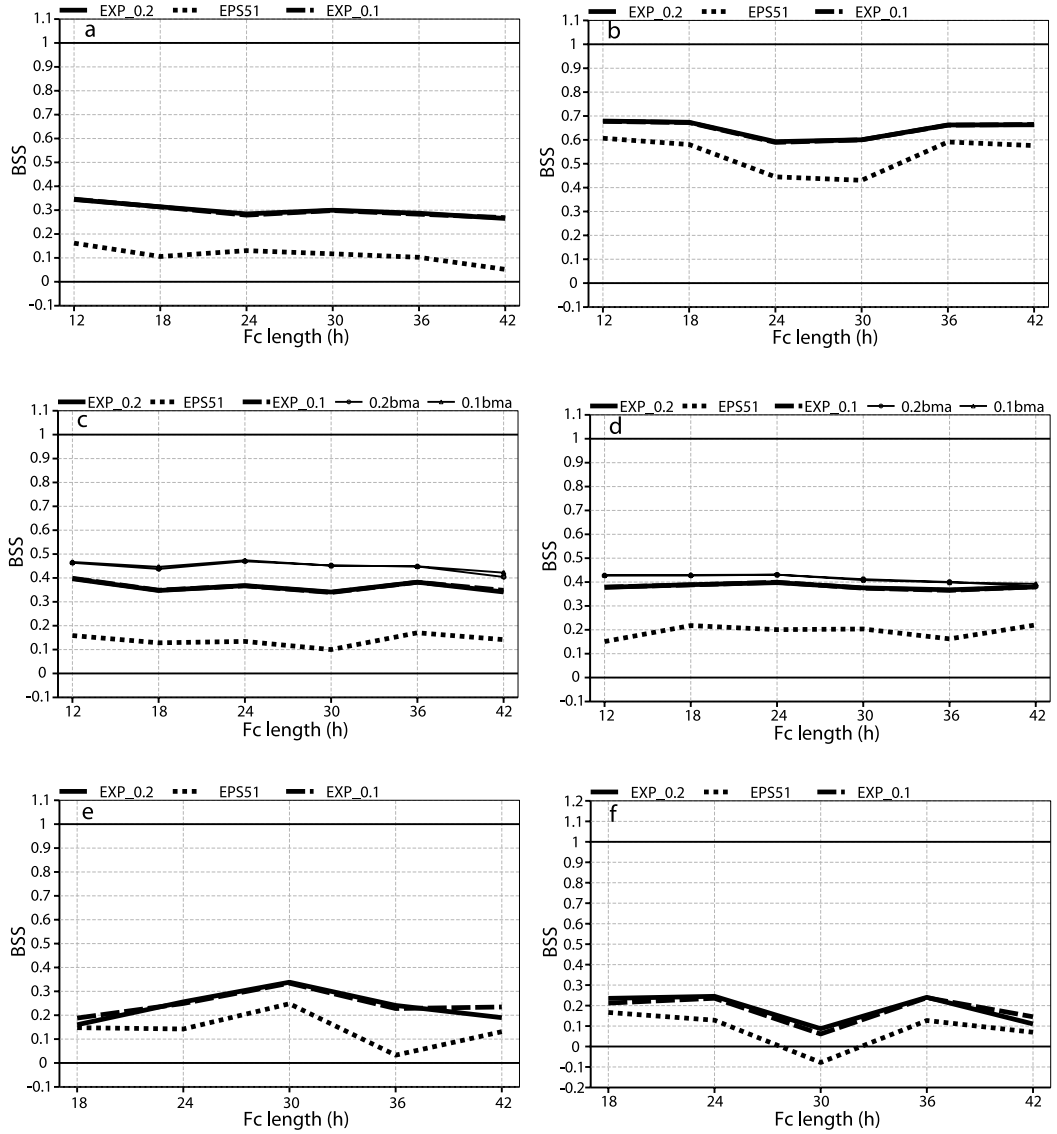

Fig. 5. Brier Skill Score (BSS) for 52 member GLAMEPS (EXP_0.2, continuous), 51 member operational EPS (dotted) and 44 member GLAMEPS (EXP_0.1. dashed) for predicted probabilities of the events: 2 m temperature > -10 °C (a), >10 °C (b), 10 m wind speed >5 ms^{-1} (c), >10 ms^{-1} (d); 6-hourly precipitation >1 mm (e), >5 mm (f). BSS for BMA-calibrated wind-speed is shown in (c) and (d) as thin lines with markers.

and f), and in favour of the higher number of ensemble members for the more extreme event. Calibration by BMA increases BSS for wind speed considerably for the most frequent event (>5 m s^{-1}), but less for the more extreme (>10 m/s). DPRSS is also increased after BMA, clearly by increasing reliability with negligible change in ensemble resolution.

The main contributor to increased DRPSS of GLAMEPS relative to EPS51 (Fig. 6) appears to be the predicted ensemble resolution, but the reliability of the predicted probabilities is also considerably larger. The same can be said for BSS of the chosen events (not shown). Predictions of the frequently occurring (98%) event T2m >-10 °C have a considerably lower BSS than predictions of the rare (18%) event T2m >10 °C. For EPS51 both ensemble resolution and reliability are considerably smaller for the frequent event, whilst for GLAMEPS the smaller BSS is mainly due to reduced ensemble resolution with a smaller reduction in reliability (not shown). This indicates that it is difficult to add essential information relative to relevant reference predictions (sample climate data) for such very frequent events without a considerably improved modelling of local effects at observation sites, such as complex topography, radiation and cloudiness and turbulence in the planetary boundary layer. In fact, the challenge is to accurately predict the complementary event T2m < -10 °C which only occurs about 2% of the time on observation sites during the 7 weeks.

Figure 7 shows the estimated value of the ensembles for decision making for users with different ratios between protection costs and damage loss associated with predicted events for 2 m temperature, 10 m wind speed and 6 hourly accumulated precipitation amounts. The events are amongst those analysed with BSS in Fig. 5. Also shown are curves for relative operating characteristics (ROC), that is, the hit rate as a function of the false alarm rate for the same events and forecast lead times. Also in this case, the GLAMEPS experimental ensembles yield consistently better scores than the EPS51. In most cases the improvement in ROC are due to higher hit rates obtained with only minor changes in false alarm rates. The value of the GLAMEPS forecasts appears better for all users. The difference between EXP_0.1 and EXP_0.2 is very small. BMA calibration also produce a slight improvement in ROC for 10 m wind speed accompanied with a small increase in expected value for large costs for protection relative to loss when the event occurs. However, since a user with a high cost-loss ratio would have very little tolerance for false alarms, this small increase in value is negligible in practice.

An important reason for using ensembles to predict probabilities of weather events is to improve the ability to predict high impact weather. Provided that nature and society is adapted to recurrent weather events, high impact weather should be rare. One may ask if it is possible at all to properly verify probabilistic forecasts of truly rare events. For example, even though there were a few weather events with strong winds and large precipitation amounts in Europe during the 7 weeks, they were too few to obtain stable probabilistic score measures. However, the thresholds we have chosen to show in Fig. 8 as an indication, are not particularly extreme for a normal winter in Europe; and 200–300 cases (a few per thousand in frequency) were observed over the 7 weeks at the chosen sites. Even though this is too few to obtain confident estimates of the reliability over a range of forecasted probability thresholds, stable differences in the overall BSS can be seen.

The two examples in Fig. 8 are shown in order to detect effects of the larger ensemble size in EXP_0.2 compared to EXP_0.1. Figure 8 shows BSS for 10 m wind speed exceeding 20 m s^{-1} (a) and 6 hourly accumulated precipitation exceeding 10 mm (b). The BSS indicates only slight improvement compared to sample climatology. Even though conclusions need to be confirmed for much larger samples, the figure shows higher BSS for GLAMEPS than for EPS51, while the 52-member ensemble (EXP_0.2) has slightly higher values than the ensemble with 44 members (EXP_0.1). Since forecasting high-impact

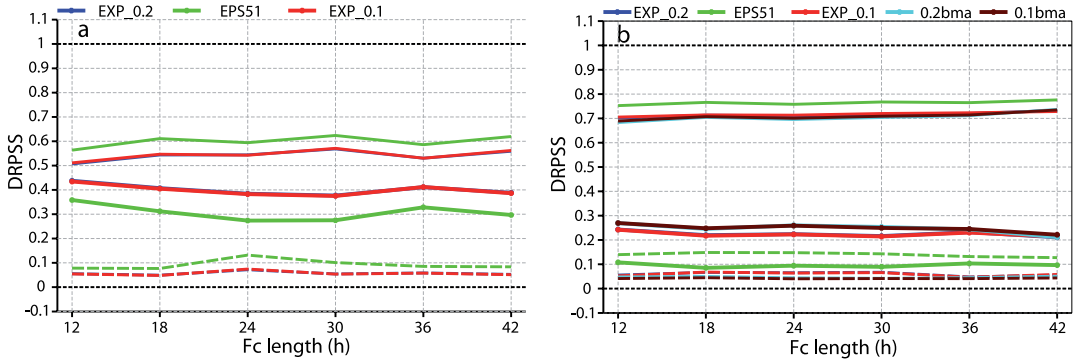

Fig. 6. Discrete rank probability skill score, DRPSS, (continuous lines with bullets) for 52 member GLAMEPS (EXP_0.2, blue), 51 member operational EPS (green) and 44 member GLAMEPS (EXP_0.1, red), of 2 m temperature (a) and 10 m wind speed (b). DPRSS for BMA-calibrated 10 m wind speed are turquoise (EXP_0.2) and brown (EXP_0.1). Events are for 2 m temperature $> -10, -5, 0, 5, 10, 15, 20, 25, 30$ °C and for 10 m wind speed $>3, 5, 10, 15, 20, 25, 30$ ms^{-1}. Contribution from resolution (continuous) and reliability (dashed), defined so that DPRSS = 1 – DPRSS(reliability) – DPRSS(resolution).

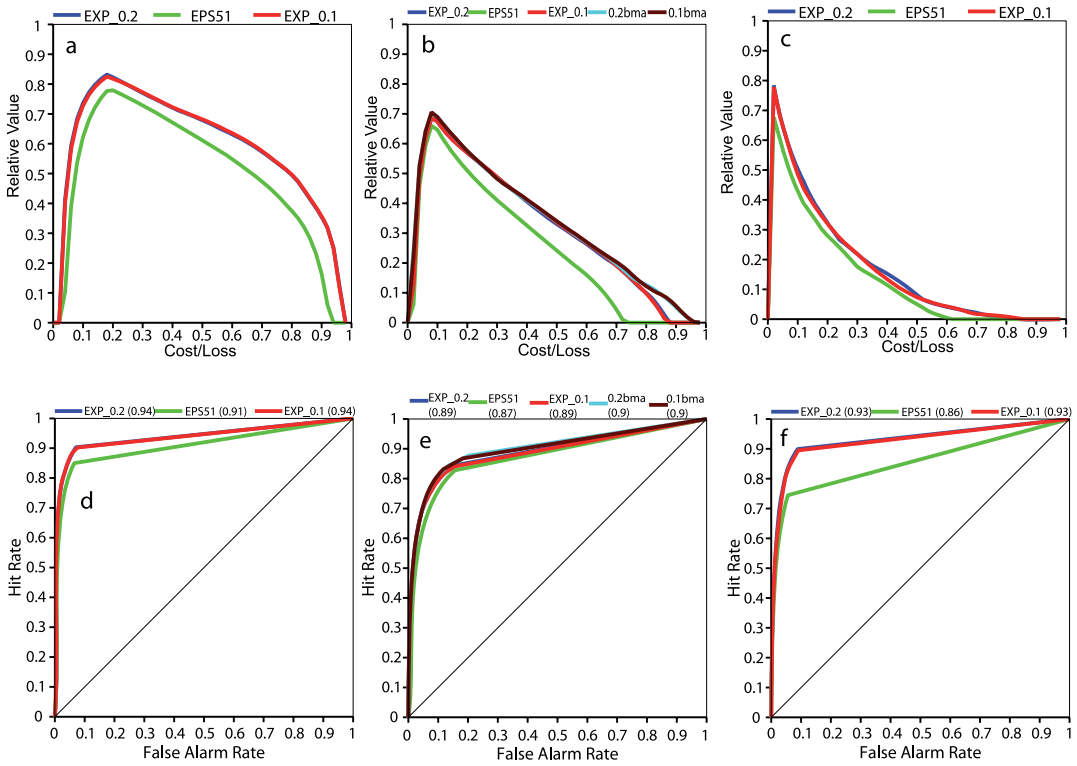

Fig. 7. Expected relative value (panels a, b, c) and relative operating characteristics (ROC) curves (panels d, e, f), for decisions based on predicted probabilities of events. Lines shown for 52 member GLAMEPS (EXP_0.2, blue), 51 member operational EPS (green) and 44 member GLAMEPS (EXP_0.1, red). BMA-calibrated forecasts of wind speed are shown for EXP_0.2 (turquoise) and EXP_0.1 (brown). The events are: +42 h predicted 2 m temperature >10 °C (a, d), 10 m wind speed >10 ms^{-1} (b, d); and +18 h to +24 h predicted 6-hourly precipitation >5 mm (c, f).

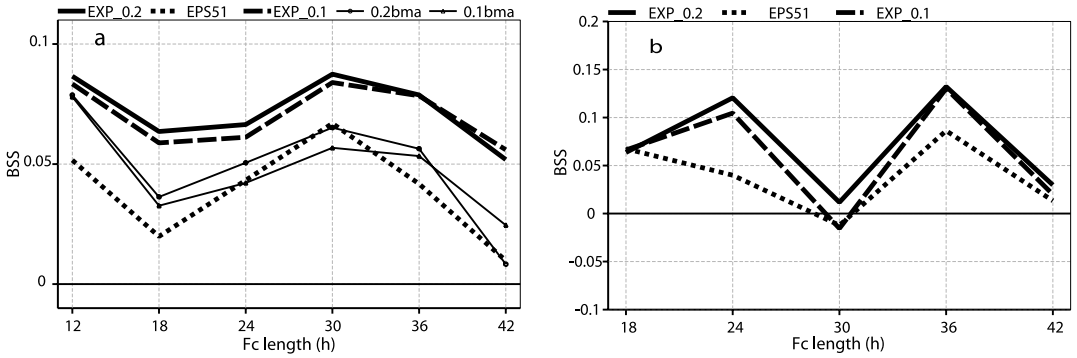

Fig. 8. Brier Skill Score (BSS) for +12 h to +42 h predicted probabilities of two rare events: 10 m wind speed >20 ms^{-1} (a) and 6 h precipitation >10 mm (b). Thick lines show EXP_0.2 (continuous), EPS51 (dotted), EXP_0.1 (dashed); thin lines for wind speed (left) indicate effects of BMA calibration for EXP_0.2 and EXP_0.1.

weather is an important purpose for GLAMEPS, the 52 member GLAMEPS is chosen for the first version even though the improvement is very small.

We already saw from Fig. 5c and d that improvements by BMA calibration were smaller for the more extreme event in Fig. 5d. It is therefore interesting to see that for the even more extreme case in Fig. 8a, BMA calibration destroys the skill of GLAMEPS beyond the first 12 h of the forecast. This can be a consequence of calibrating the BMA coefficients based on a common statistics for the entire area, thus neglecting the fact that statistical properties for upper quantiles are particularly heterogeneous. The BMA will then be dominated by the average pan-European climate and mask the statistics for strong winds. An alternative option is to estimate BMA coefficients over a longer period of time which may enable segregated statistics for subregions, but then the temporal flow dependence will be lost. Since extreme weather is the main purpose of GLAMEPS, we have decided to abandon BMA calibration in the present form.

This is also why we have so far neither run BMA calibration for other variables, nor for EPS51 as a fair comparison with BMA-calibrated GLAMEPS.

3.3. *Multimodel versus single-model ensembles*

As an example for illustration, Fig. 9 shows forecasted ensemble plumes for a coastal site in the northwestern Netherlands for a case of a storm with gale-force winds and heavy precipitation. In the morning of March 1 a storm developed over the North Sea with maximum northwesterly winds analysed around 25 m s^{-1}. At the Dutch site De Kooy in Fig. 9, the wind speed reached 15 m s^{-1} and the precipitation persisted with more than 7 mm per 6 h over many hours.

For precipitation, it is clear from these figures that the ensemble spread in GLAMEPS EXP_0.2 is larger than for EPS51, and that the observed values are clearly better encompassed by the ensemble in EXP_0.2 than for EPS51. While EPS51 overestimates the observed precipitation amounts, GLAMEPS include many ensemble members that do not. Many EuroTEPS members still overestimate, but this is compensated by most of the LAM ensemble members, which mostly underestimate. Hence, in the case of precipitation, compensating systematic errors appears to be important.

For the forecasted wind speed the improvement in the combined ensemble is due to EuroTEPS being better than EPS51, and that extra spread is provided by some of the LAM ensemble members. Most GLAMEPS ensemble members overestimate the

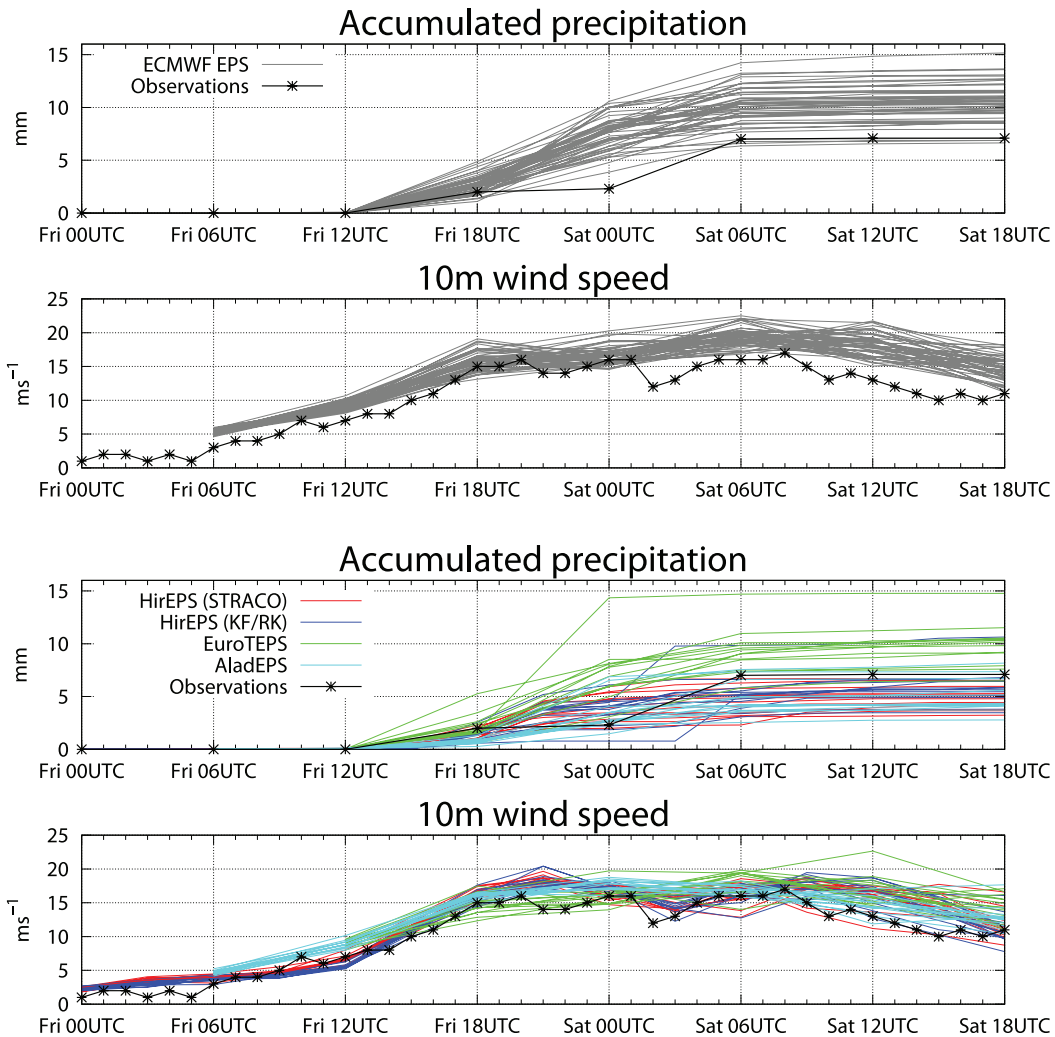

Fig. 9. EPS-meteograms for the site 06235 De Kooy in the north-western coast of The Netherlands for an extreme weather case. The dates on the *x*-axes start on 29 February 2008 00UTC and end 42 h later, with 6 h between tick marks. Grey curves in the two uppermost diagrams are from the operational EPS51 and multicoloured curves in the two lowermost are from the different model components of GLAMEPS EXP_0.2. Black curves with markers are observations. In each case, the upper curves show predicted accumulated precipitation for each ensemble member and the lower show wind speed at 10 m height.

observed wind-speed but to a smaller extent than those of EPS51. It should be borne in mind in this context that for a well performing ensemble, observations should occur with equal frequency over the variable intervals implied by the ensemble members. This property is diagnosed with rank histograms. Figure 9 can be taken to indicate that using several models in the ensemble generation may, at least occasionally, improve the predictive performance by adding spread and thus increase reliability, and by compensation of flow-dependent systematic errors.

In this section we want to further investigate if the multimodel ensemble of GLAMEPS adds predictive skill compared to single-model ensembles. In this context we consider single-model ensembles of similar size as the multimodel, and not the small single-model ensembles that are combined in GLAMEPS. Separate experiments are made for the 7-week winter period in 2008, for which each of the three different single-model LAMEPS have been rerun with 51 ensemble members. To prepare for this, EuroTEPS was run to produce 50 alternative ensemble members as input to, respectively, AladEPS, HirEPS_S and HirEPS_K in addition to each model's control forecast.

Figure 10 shows DRPSS for the three 51 member AladEPS, HirEPS_S and HirEPS_K, respectively compared to the 52 member GLAMEPS (EXP_0.2). In all cases, the multimodel GLAMEPS gives better scores than any of the single-model ensembles of the same size, and the improvement is particularly evident for the ensemble resolution component (thin continuous lines in the upper parts of the diagrams). The reliability is also generally better for GLAMEPS, but the increased ensemble resolution is particularly encouraging, since this a non-trivial benefit that cannot be achieved by statistical calibration methods.

For wind speed at 10 m height the improvement of the multimodel GLAMEPS, as measured by DPRSS, is modest compared to AladEPS. For forecast lead time +36 h and for BSS for a certain event thresholds (not shown) AladEPS is even slightly better. Figure 10 clearly shows that the two 51-member HirEPS ensembles are of lower quality than the one 51-member AladEPS. It is possible that this is due to bias errors for wind speed in HIRLAM (e.g. de Bruijn and van Meijgaard, 2005; Yang, 2007), for which low wind speeds tend to be overestimated and high wind speeds underestimated. Over rough topography the underestimates dominate. It should be borne in mind that the verification presented here is obtained as statistics for observation sites directly, and may not represent the model quality in data-sparse areas.

The point to be made here is that an uneven quality between the models may undermine the potential quality enhancement of the multimodel ensemble combinations. As discussed in connection with Figs 5 and 6, combining and calibrating the ensembles with BMA, which reduces systematic errors at observation sites, increases DRPSS for the GLAMEPS ensemble. However, as indicated from Fig. 8, calibration statistics, which is both flow dependent and accounts for heterogeneous climatology, is not trivially obtained. On the other hand, fully flow-dependent bias-free and well-calibrated single-model ensembles would have little complementary information to contribute when combined into a multimodel ensemble.

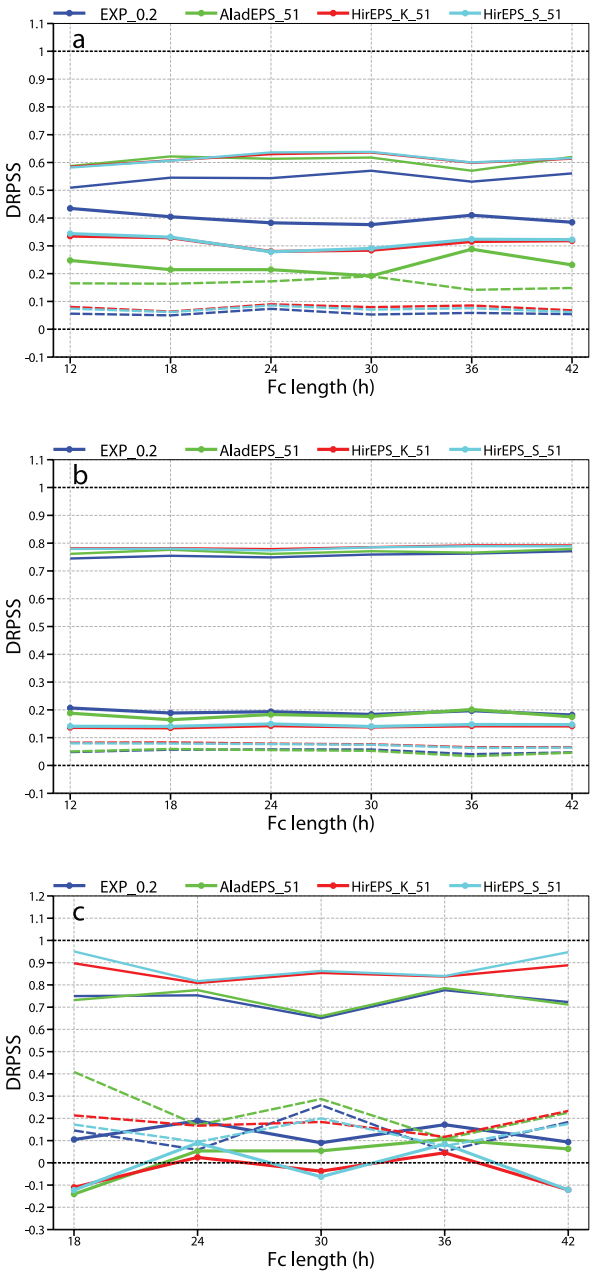

Fig. 10. Discrete rank probability skill score, DRPSS, (continuous lines with bullets) for 52 member EXP_0.2 (blue), 51 member AladEPS (green), 51 member HirEPS_K (red) and 51 member HirEPS_S (turquoise). Events include 2 m temperature $> -10, -5, 0, 5, 10, 15, 20, 25, 30$ °C (a), 10 m wind speed $>3, 5, 10, 15, 20, 25$ and 30 ms^{-1} (b) and precipitation $>0.1, 1, 2, 5, 10, 15, 20, 25$ mm over 6 h (c). Contributions from resolution and reliability are as in Fig. 6.

There are also uneven quality between the models for other variables and events. For temperature at 2 m height AladEPS has

consistently lower reliability and smaller DPRSS. One reason may be that AladEPS does not use a control from its own analysis but is a downscaling of the coarse-resolution EuroTEPS. Thus, small-scale features, which in particular influence 2 m temperature in complex topography, are less well represented in AladEPS than in HirEPS. In this case, however, there are twice as many ensemble members with better quality in GLAMEPS, and the combination is clearly better than the best single-model ensemble. For 6-hourly accumulated precipitation the difference in quality is considerably smaller, even though the single-model AladEPS has higher DPRSS for lead-times +30 to +42 h. In this case, the single-model AladEPS has considerably better ensemble resolution (sharpness) than both the HirEPS ensembles, similar to the multimodel GLAMEPS combination. However, the improved ensemble resolution is combined with a reduced reliability for AladEPS. Hence GLAMEPS has higher DRPSS for all lead-times.

A tentative conclusion is that a multimodel GLAMEPS is generally better than a single-model ensemble, even if there are cases when uneven qualities between the models render the improvement by combination small. The multimodel improvement is particularly seen in ensemble resolution. Nevertheless, an advanced calibration (Hamill et al., 2006) of a high-quality single-model ensemble may be a more cost-efficient solution even though the ensemble resolution improvement will be lost. Such calibration should then be performed using reforecasts over several years and seasons which requires huge computer resources and man-power. For the present situation, multimodel GLAMEPS is better suited when evaluated over a range of variables and events. But maintaining and running several models of state-of-the-art quality also takes huge resources.

3.4. EuroTEPS versus regular EPS

Running EuroTEPS takes considerable computer resources and human resources for maintenance. The cost will become particularly large when upgraded with higher spatial resolution. These resources may provide larger benefits for GLAMEPS if used differently. If the same number of ensemble members is taken from the operational EPS (EPS51), the computer resources used to produce EuroTEPS can in stead be used on AladEPS and HirEPS. For the model versions used for GLAMEPS EXP_0.2, EuroTEPS takes considerably less computer resources than the limited area model EPS production. Hence, since reducing the horizontal grid-mesh width with a factor x implies an increased computer requirement of at least a factor x^3, only a very small resolution increase can be obtained in this case. In stead, the released resources could be used for longer forecasts or an increased number of ensemble members.

Here, we simply investigate the consequence of replacing the 13 member EuroTEPS used in GLAMEPS EXP_0.2 with the control and the first 12 (= 6 pairs) alternative ensemble members from EPS51. Choosing the first 12 or any other 6 pairs of ensemble members from EPS51 is not expected to yield systematic differences, since each pair is constructed from Gaussian sampling of development coefficients for the singular vectors. The comparison is only made over the first 4 weeks of the test period, that is, from 17 January to 13 February 2008.

Figure 11 shows the DPRSS for the same events and variables as shown in Fig. 10. There is a consistent but small improvement of using EuroTEPS, and the improvement stems predominantly from a slightly better ensemble resolution. The BSS shown for selected event thresholds in Fig. 12 confirms the results. Whether this improvement is worth the computational cost is not fully evaluated yet, because EuroTEPS and EPS51 are under considerable upgrade both with respect to spatial resolution and methods (EDA). Further experiments will therefore be made with the new system, for which the computational gain of abandoning EuroTEPS will be much larger. At the same time, running a dedicated EuroTEPS also enables more possibilities for GLAMEPS which is yet to be exploited. This includes using diabatic targeted singular vectors and forcing singular vectors (Barkmeijer et al., 2003). The benefits of such extensions should be included in further tests.

4. Conclusions and prospects

Based on the experiments over a 7-week period in winter 2008 the GLAMEPS ensembles of 44 ensemble members (EXP_0.1) and 52 ensemble members (EXP_0.2) are both improvements relative to the operational ECMWF 51 member EPS. In spite of the relatively limited data basis, the improvement seems considerable and consistent. The results provide substantial confidence in the potentials of GLAMEPS for pan-European, short-range probabilistic weather forecasting, and there are solid grounds for pursuing the development of a first operational GLAMEPS.

At the same time, even though sufficient evidence is provided for developing the potentials of the system operationally, the results should not be interpreted too general. In particular, results may be different for other seasons and for other winters, even though there were considerable flow regime changes in Europe during the winter in 2008. Finally, a common evaluation is made for all available observations across Europe during the 7 weeks. The evaluation may mask regional differences and temporal variations. Hence, the evaluation needs to be followed up and confirmed over much longer periods, and this can in practice be done in an operational setting. The preliminary positive results encourage the follow-up with such a long-term evaluation.

Three main aspects of GLAMEPS are believed to contribute to the improvement seen in the preliminary evaluation: the factor four (approximately) increase in horizontal resolution, the use of several models and assimilation cycles, and the impact of employing targeted perturbations in EuroTEPS. We have investigated the two latter aspects explicitly, and find that whilst the positive contribution of EuroTEPS is consistent but modest, the multimodel aspect is crucial. One exception is seen for

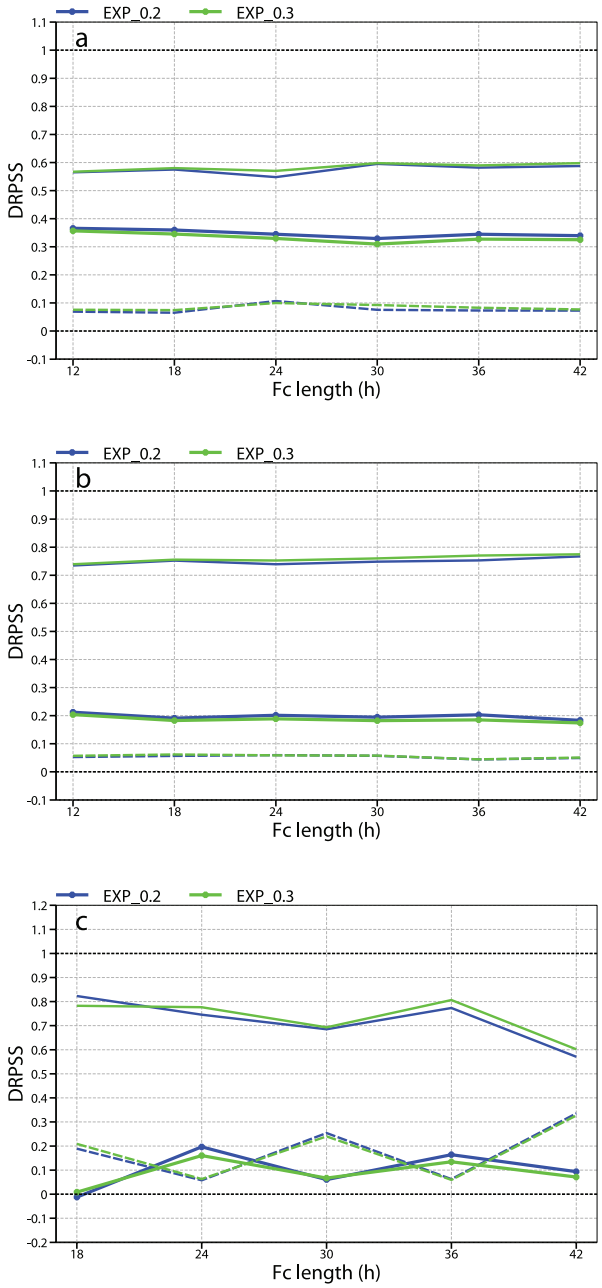

Fig. 11. Discrete rank probability skill score, DRPSS, (continuous lines with bullets) for 52 member GLAMEPS with 13 member EuroTEPS (EXP_0.2, blue) and with 13 selected ensemble members from the operational EPS (EXP_0.3, green). Event thresholds are as in Fig. 10 for 2 m temperature (a), 10 m wind speed (b) and 6 h accumulated precipitation (c). Contributions from resolution and reliability are as in Fig. 6.

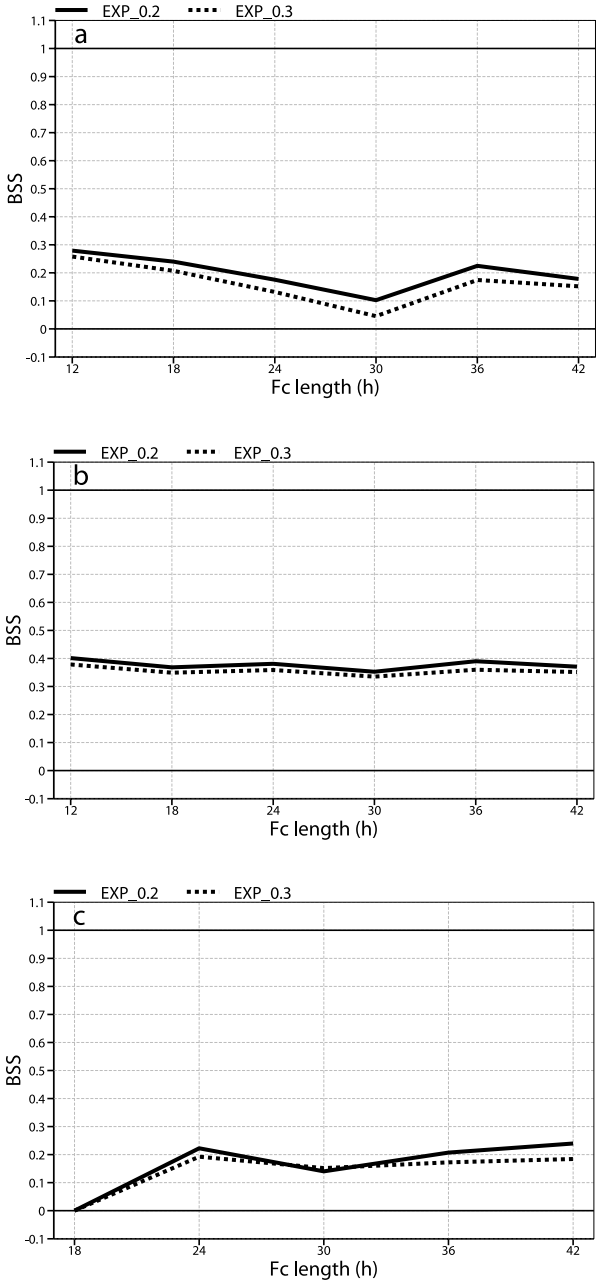

Fig. 12. Brier Skill Score (BSS) for GLAMEPS with EuroTEPS (EXP_0.2, continuous) and with members from the operational EPS (EXP_0.3, dotted) for events: 2 m temperature $> -10\,^{\circ}$C (a); 10 m wind speed >5 ms^{-1} (b); 6 h accumulated precipitation >5 mm (c).

wind speed forecasts when one of the single-model ensembles of the same size is considerably better than the others. In this case the multimodel ensemble is only slightly better than the best single-model ensemble, and this indicates that proper advanced calibration of the single-model forecasts may enhance the multimodel performance. This was also preliminary confirmed when applying BMA on these forecasts, thus improving reliability with no change of ensemble resolution.

The positive impact of the multimodel approach on the forecast quality confirms the experience of Hagedorn et al. (2005)

and Doblas-Reyes et al. (2005) who combined several single-model ensembles into a multimodel ensemble for seasonal prediction. In this case statistical calibration was seen to enhance the quality of single-model ensembles more than multimodel ensembles, since the latter was more reliable already before calibration. Still the multimodel ensemble performed better than the calibrated single-model ensembles, and increased reliability by calibration was predominantly achieved by inflated variability and thus reduced ensemble resolution.

We have not applied calibration to the single-model ensembles in our case, and we cannot claim that the statements of Hagedorn et al. (2005) and Doblas-Reyes et al. (2005) are valid in our case. Nevertheless, we have seen that single-model ensembles of the same size as the multimodel ensemble perform consistently worse. The reliability of the multimodel ensemble is better, but more importantly, the ensemble resolution is improved relative to the single-model ensembles. This means that, although the combination of models accounts for model uncertainty in a rather arbitrary way, their combination into a multimodel ensemble contributes qualities that cannot be obtained by statistical calibration. Calibration can increase reliability but often at the expense of resolution. We have also seen, however, that for cases when the single-model ensembles are of considerably different quality, the multimodel combination of same ensemble size may not be able to perform better than the best member.

The first improvements of GLAMEPS compared to the version investigated in this paper are already underway. This concerns the upgrading of EuroTEPS in line with the upgrade of the operational EPS at ECMWF, and includes increased spatial resolution and the use of ensemble data assimilation (EDA). The basic development work for the latter is already done, and a version of the upgraded EuroTEPS will probably start in experimental production mode in the autumn of 2010. At the same time (and starting before), further tests of the benefits of EuroTEPS will be made, including even higher resolution.

We will also further investigate ways of improving the application of BMA (e.g. Johnson and Swinbank, 2009) or other ways of combining and calibrating the ensembles. Even though the improvement we obtained for the BMA calibrated ensemble for wind speed is encouraging, we also saw a need for calibration statistics that depend on the actual atmospheric state in any geographical position. This is particularly evident for extreme weather since the definition of rare events is geographically highly variable. Elaborated techniques may thus be developed to obtain calibration coefficients with statistical confidence for frequent events, while for rare and potentially high-impact weather events this is hard to obtain. Another further complicating aspect for calibration (as well as evaluation), is the uneven distribution of observations. Over oceans, in rough terrain, and in the Arctic, where high-impact weather probably is more prevalent than average, observations are often sparse. Therefore, technique of using data from reforecasts over several years (Hamill et al., 2006) is probably the best way to determine the calibration coefficients, but this is computationally too costly for GLAMEPS. In stead an improved BMA can be done by partly making statistics for different geographical regions based and by partly stratifying the samples used for estimating the calibration coefficients into quantiles of the climatic frequency distributions. We intend to pursue this way of implementing calibration, including investigating the feasibility of using downscaled reanalyses.

In GLAMEPS there are also ongoing efforts to develop methods for more mesoscale initial perturbations that are more directly developed for the short range than the global EuroTEPS (or EPS51). This includes ETKF (Bojarova et al., 2011) and LAM-specific singular vectors that either maximize or suppress the convective available potential energy (CAPE) at final time (Stappers and Barkmeijer, 2011). Experiments on forcing perturbations and surface parameters, in particular soil moisture, are also being developed. These are intended for a later version of GLAMEPS if proven valuable and computationally affordable.

5. Acknowledgments

GLAMEPS depends on dedicated resources from member countries in the HIRLAM and ALADIN, both personnel and computer resources. Experiments are made at ECMWF computers through a special project (SPNOGEPS), but considerable computer billing units are dedicated from national member quota. Apart from the co-authors, many contribute to GLAMEPS on aspects not covered in this paper. We are grateful for scientific advice by Martin Leutbecher, and strong technical support by Umberto Modigliani and Dominique Lucas, ECMWF.

References

ALADIN International Team. 1997. The ALADIN project: mesoscale modelling seen as a basic tool for weather forecasting and atmospheric research. *WMO bull.* **46**, 317–324.

Aspelien, T., 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.

Barkmeijer, J., Buizza, R. and Palmer, T. N. 1999. 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **125**, 2333–2351.

Barkmeijer, J., Iversen, T. and Palmer, T. N. 2003. Forcing singular vectors and other sensitive model structures. *Q. J. R. Meteorol. Soc.* **129**, 2401–2423.

Barnston A. G. and Livezey, R. E. 1987. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea. Rev.* **115**, 1083–1126.

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.

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

Bougeault, P. 1985. A simple parameterization of the large-scale effects of cumulus convection. *Mon. Wea. Rev.* **113**, 2108–2121.

Bougeault, P. and Geleyn, J.-F. 1989. Some problems of closure assumption and scale dependency in the parameterization of moist deep convection for numerical weather prediction. *Meteorol. Atmos. Phys.* **40**, 123–135.

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. 2008. Accounting for the effect of observation errors on verification of MOGREPS. *Meteorol. Appl.* **15**, 199–205.

Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B. and Beare, S. E. 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.

de Bruijn, C. and van Meijgaard, E. 2005. Verification of HIRLAM with ECMWF physics compared with HIRLAM reference versions. HIRLAM Technical Report No. 63. Available at: hirlam.org.

Buizza, R. 1994. Localization of optimal perturbations using a projection operator. *Q. J. R. Meteor. Soc.* **120**, 1647–1682.

Buizza, R., Barkmeijer, J., Palmer, T. N. and Richardson, D. 2000. Current status and future developments of the ECMWF ensemble prediction system. *Meteorol. Appl.* **6**, 1–14.

Buizza, R., Leutbecher, M. and Isaksen L. 2008. Potential use of analyses in the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.*, **134**, 2051–2066, doi:10.1002/qj.346.

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.

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

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

Courtier, P., Freydier, C., Geleyn, J.-F., Rabier, F. and Rochas, M. 1991. The ARPEGE project at Météo-France. In: *Proceedings of the 1991 ECMWF seminar on numerical methods in atmospheric models*. Vol. II, 193–231. ECMWF, Shinfield Park, Reading, England.

Doblas-Reyes, F. J., Hagedorn, R. and Palmer, T. N. 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting II. Calibration and combination. *Tellus* **57A**, 234–252.

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 (edited by 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/reference.html.

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

Evensen, G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *J. Geophys. Res. Oceans* **99**, 10143–10162.

Evensen, G. and van Leeuwen, P. 1996. Assimilation of geosat altimeter data for the Agulhas current using the ensemble Kalman filter with a quasigeostrophic model. *Mon. Wea. Rev.* **124**, 85–96.

Fisher, M. and Courtier, P. 1995. Estimating the covariance matrices of analysis and forecast error in variational data assimilation. *ECMWF Tech. Memo.* **220**. ECMWF, Shinfield Park, Reading, RG2 9AX, UK.

Fischer, M. and Andersson, E. 2001. Developments in 4D-Var and Kalman filtering. *ECMWF Tech. Memo.* 347. ECMWF, Shinfield Park, Reading, RG2 9AX, UK.

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

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

Garcia-Moya, J. A., Callado, A., Santos, C., Santos, D. and Simarro, J. 2007. Multi-model ensemble for short-range predictability. In: *Proceedings of the 3rd International Verification Methods Workshop, ECMWF*, Shinfield Park, Reading, RG2 9AX, UK.

Geleyn, J.-F., Bazile, E., Bougeault, P., Déqué, M., Ivanovici, V. and co-authors 1995. Atmospheric parametrization schemes in Meteo-France's ARPEGE N.W.P. model. In: *Proceedings of the 1994 ECMWF Seminar on Physical Parametrizations in Numerical Models, ECMWF*, 385–400. Shinfield Park, Reading, England.

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—I. Basic concepts. *Tellus* **57A**, 219–233.

Hágel, E. and Horányi, A. 2007. The ARPEGE/ALADIN limited area ensemble prediction system: the impact of global targeted singular vectors. *Meteorologische Zeitschrift* **16**, 653–663.

Hamill, T. M. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.* **129**, 550–560.

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

Hamill, T. M., Mullen, S. L., Snyder, S., Toth, Z. and Braumhefner, D. P. 2000. Ensemble forecasting in the short to medium range. Report from a workshop. *Bull. Am. Meteor. Soc.* **81**, 2653–2664.

Hamill, T. M., Whitaker, J. S. and Mullen, S. L. 2006. Reforecasts, an important dataset for improving weather predictions. *Bull. Am. Meteorol. Soc.* **87**, 33–46.

Haugen, J. E. and Machenhauer, B. 1993. A spectral Limited Area model formulation with time-dependent boundary conditions applied to the shallow-water equations. *Mon. Wea. Rev.* **121**, 2618–2630.

Houtekamer, P. L. and Mitchell, H. L. 1998. Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.* **126**, 796–811.

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.

Johnson, C. and Swinbank, R. 2009. Medium-range multimodel ensemble combination and calibration. *Q. J. R. Meteorol. Soc.* **135**, 777–794.

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

Kain, J. S. and Fritsch, J. M. 1990. A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.*, **47**, 2784–2802.

Kann, A., Wittmann, C., Wang, Y. and Ma, X. 2009. Calibrating 2-m temperature of limited-area ensemble forecasts using high-resolution analysis. *Mon. Wea. Rev.* **137**, 3373–3387, doi:10.1175/2009MWR2793.

Leith, C. E. 1974. Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.* **102**, 409–418.

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.

Leutbecher, M. and Palmer, T. N. 2008. Ensemble forecasting. *J. Comp. Phys.* **227**, 3515–3539.

Lewis, J. M. 2005. Roots of ensemble forecasting. *Mon. Wea. Rev.* **133**, 1865–1885.

Lindskog, M., Gustafsson, N., Navascu'e Ns, 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. 1963. Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141.

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

Louis, J.-F., Tiedke, M. and Geleyn, J.-F. 1981. A short history of the operational PBL-parameterization at ECMWF. In: *Workshop Proceedings on Planetary Boundary Layer Parameterizations*, Nov. 1981, pp. 59–79.ECMWF, Shinfield Park, Reading, RG2 9AX, UK.

Magnusson, L., Nycander, J. and Källén, E. 2009. Flow-dependent versus flow-independent perturbations for ensemble prediction. *Tellus* **61A**, 194–209, doi:10.1111/j.1600–0870.2008.00385.x.

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

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. 1996. The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.* **125**, 73–119.

Noilhan, J. and S. Planton. 1989. A simple parameterization of land surface processes for meteorological models. *Mon. Wea. Rev.* **117**, 536–549.

Orrell, D., Smith, L., Barkmeijer, J. and Palmer T. N. 2001. Model error in weather forecasting. *Nonl. Process. Geophys.* **8**, 357–371.

Osborn, T. J. 2006. Recent variations in the winter North Atlantic Oscillation. *Weather* **61**, 353–355.

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

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.

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

Saetra, 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.

Sass, B. H., Nielsen, N. W., Jørgensen, J. U. and Amstrup, B. 1999. The Operational HIRLAM System at DMI. DMI *Tech Rep. no 99–21*. Danmarks Meteorologiske Institutt, Lyngby v. 100, Copenhagen, Denmark.

Seko, H., Saito, K., Kunii, M., Hara, T., Kyouda, M. and Yamaguchi, M. 2007. Japan area mesoscale ensemble experiments using JMANHM. CAS/JSC WGNE Research Activities in Atmospheric and Oceanic Modelling. 37, 5.31–5.32.

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.

Stappers, R.J.J. and Barkmeijer, J. 2011. Properties of singular vectors using convective available potential energy as final time norm. *Tellus* **63A**, this issue.

Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S., Rogers, E. 1999. Using ensembles for short-range forecasting. *Mon. Wea. 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 Norrköping, Sweden.

Wang, X. G. and Bishop, G. H. 2003. A comparison of breeding and ensemble transform Kalman filter forecast schemes. *J. Atmos. Sci.* **60**, 1140–1158.

Weigel, A. P. and Bowler, N. E. 2009. Comment on 'Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?'. *Q. J. R. Meteorol. Soc.* **135**, 535–539, doi:10.1002/qj.381

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

Yang, X. 2007. Status update on operational HIRLAM. *HIRLAM Newslett.* 52, 114–122. (Can be downloaded from http://Hirlam.org).

Yang, X. 2008. Status of the Reference System. *HIRLAM Newslett.* **55**, 202–203. (Can be downloaded from http://Hirlam.org).