

# Combining 2-m temperature nowcasting and short range ensemble forecasting

A. Kann, T. Haiden, and C. Wittmann

Central Institute for Meteorology and Geodynamics, Vienna, Austria

Received: 6 July 2011 – Revised: 15 November 2011 – Accepted: 23 November 2011 – Published: 2 December 2011

**Abstract.** During recent years, numerical ensemble prediction systems have become an important tool for estimating the uncertainties of dynamical and physical processes as represented in numerical weather models. The latest generation of limited area ensemble prediction systems (LAM-EPSs) allows for probabilistic forecasts at high resolution in both space and time. However, these systems still suffer from systematic deficiencies. Especially for nowcasting (0–6 h) applications the ensemble spread is smaller than the actual forecast error. This paper tries to generate probabilistic short range 2-m temperature forecasts by combining a state-of-the-art nowcasting method and a limited area ensemble system, and compares the results with statistical methods. The Integrated Nowcasting Through Comprehensive Analysis (INCA) system, which has been in operation at the Central Institute for Meteorology and Geodynamics (ZAMG) since 2006 (Haiden et al., 2011), provides short range deterministic forecasts at high temporal (15 min–60 min) and spatial (1 km) resolution. An INCA Ensemble (INCA-EPS) of 2-m temperature forecasts is constructed by applying a dynamical approach, a statistical approach, and a combined dynamic-statistical method. The dynamical method takes uncertainty information (i.e. ensemble variance) from the operational limited area ensemble system ALADIN-LAEF (Aire Limitée Adaptation Dynamique Développement International Limited Area Ensemble Forecasting) which is running operationally at ZAMG (Wang et al., 2011). The purely statistical method assumes a well-calibrated spread-skill relation and applies ensemble spread according to the skill of the INCA forecast of the most recent past. The combined dynamic-statistical approach adapts the ensemble variance gained from ALADIN-LAEF with non-homogeneous Gaussian regression (NGR) which yields a statistical correction

of the first and second moment (mean bias and dispersion) for Gaussian distributed continuous variables. Validation results indicate that all three methods produce sharp and reliable probabilistic 2-m temperature forecasts. However, the statistical and combined dynamic-statistical methods slightly outperform the pure dynamical approach, mainly due to the under-dispersive behavior of ALADIN-LAEF outside the nowcasting range. The training length does not have a pronounced impact on forecast skill, but a spread re-scaling improves the forecast skill substantially. Refinements of the statistical methods yield a slight further improvement.

## 1 Introduction

In numerical weather prediction (NWP), the use of ensemble prediction systems (EPSs) has become the standard method of accounting for uncertainties in initial conditions and model formulations. In observation-based forecasting, or nowcasting, emphasis has more recently shifted from purely deterministic extrapolation towards inclusion of probabilistic methods. Dance et al. (2010), for example, describe a system which predicts thunderstorm strike probability using a bivariate Gaussian model of speed and direction errors of the cell tracking method employed. They show that the skill of the system in predicting threat areas exceeds that of the corresponding deterministic advection forecast.

Another approach to estimating the uncertainty of nowcasting products is the use of time-lagged ensembles. Using a set of hourly initialized Rapid Update Cycle forecasts, Lu et al. (2007) are able to improve NWP very short-range (1–3 h) forecasts of meteorological fields compared to the deterministic forecasts. The improvements in this case appear to result mainly from the correction of model errors due to initial spin-up. The potential benefit of including forecasters' knowledge in the generation of perturbations in an EPS has been investigated by Homar et al. (2006). Their study



Correspondence to: A. Kann  
(alexander.kann@zamg.ac.at)

shows that the experimental ensemble improved short-range numerical forecasts of severe weather and heavy precipitation. However, because of the lack of global dispersion, forecasts of non-severe weather were generally less skillful than those of the reference system.

The Integrated Nowcasting Through Comprehensive Analysis (INCA) system used in the present study is by design non-probabilistic. It provides, among other products, short range deterministic 2-m temperature analyses and forecasts on high resolution in time (60 min) and space (1 km × 1 km) with special emphasis on the nowcasting range (0–6 h). In analysis mode it combines NWP forecast fields from the operational limited area model ALADIN-AUSTRIA (Aire Limitée Adaptation Dynamique Développement InterNational) (Wang et al., 2006) with high-resolution topographic data and integrates observations from about 400 stations within the operational INCA domain. In the nowcasting range, the temporal change of the NWP model is superimposed on the latest analysis, modified by a correction which depends on the NWP cloudiness error. Outside the nowcasting range it merges into the (topographically downscaled) NWP model output through a fixed weighting function (Haiden et al., 2011). The system shows high skill in the nowcasting range (mean absolute error less than 1 Kelvin in the first 3 h) and improves on the ALADIN-AUSTRIA forecast due to bias correction and topographic downscaling outside the nowcasting range, but the NWP trend it uses is affected by uncertainties mainly due to errors in the initial condition (Buizza et al., 2005; Hamill et al., 2000) and model formulations and physical parameterizations (Buizza et al., 1999). The aim of this study is to develop proper methods to quantify these uncertainties and to provide sharp and reliable site-specific, probabilistic short range forecasts of 2-m temperature.

## 2 The INCA system

The INCA temperature analysis and forecast methodology has been described elsewhere (Kann et al., 2009; Haiden et al., 2011), so only a short summary is given here. Hourly analyses on a 1 km × 1 km grid are generated from a combination of numerical weather prediction (NWP) model output, surface station data, and high-resolution topographic data. Spatial interpolations of observation corrections are based on distance-weighting in physical and potential temperature space. NWP fields are taken from the Austrian version of the ALADIN model (Wang et al., 2006), which has a horizontal resolution of 9.6 km, and is run four times a day out to a forecast range of 72 h. Observations are provided by ~350 real-time surface weather and hydrological stations, with an average distance between the stations of about 15 km.

The three-dimensional analysis of temperature in INCA starts with an NWP short-range forecast as first guess, which is corrected based on observation-forecast differences. The

corrections are spatially interpolated using inverse-distance-squared weighting (IDW) in the horizontal and IDW in potential temperature in the vertical (Haiden et al., 2011). A “surface-layer index” ensures that corrections derived at a certain type of location (e.g. valley floor) have low or zero weight at other types of location (e.g. slope).

The nowcast of temperature is obtained by adding the trend given by the NWP model to the INCA analysis. A coefficient between 0 and 1 reduces the trend in case of negative cloudiness errors in the NWP forecast. Outside the nowcasting range the trend extrapolation is smoothly blended into the NWP forecast via a weighting function. The time-scale of the weighting function depends on static stability to account for observed variations in the persistence of temperature forecast errors under different synoptic conditions (Haiden et al., 2011).

## 3 The Limited Area Ensemble System ALADIN-LAEF

The ALADIN-LAEF (Aire Limitée Adaptation Dynamique Développement InterNational Limited Area Ensemble Forecasting) system is being developed at the Central Institute for Meteorology and Geodynamics (ZAMG), which is part of the ALADIN consortium and the Regional Cooperation for Limited-Area modeling in Central Europe (RC LACE). The very first version of ALADIN-LAEF was put into operation in March 2007, and in 2009 a major upgrade of the system took place. A detailed description of the ALADIN-LAEF system can be found in Wang et al. (2011). In the following, just the major characteristics are briefly described.

The ALADIN-LAEF system uses the spectral limited area model ALADIN (Wang et al., 2006) to produce a 17 member forecast ensemble: 16 perturbed members and one control member. The geographical domain covers Europe and large parts of the North Atlantic. The ensemble system is run twice per day (00:00 and 12:00 UTC) producing forecasts up to 60 h. It has 18 km horizontal resolution and 37 vertical levels.

In order to create the initial conditions (IC) for the 16 perturbed members of ALADIN-LAEF the following components are combined:

- Large-scale atmospheric perturbations coming from the first 16 ECMWF EPS members are created using the singular vector approach (Leutbecher and Palmer, 2008).
- Small-scale atmospheric perturbations are applied on ALADIN-LAEF resolution with a breeding technique (Wang et al., 2011).
- Small-scale initial surface perturbations are generated on the ALADIN-LAEF resolution by using the so called Non-Cycling Surface Breeding (NCSB) method (Wang et al., 2010).

The large-scale and small-scale atmospheric perturbations are combined using a spectral blending technique (Brozkova et al., 2001). In addition to the uncertainty in the initial condition, a multi-physics approach is used to account for model errors. The lateral boundary conditions are provided by the first 16 members of the ECMWF EPS.

#### 4 Methods and implementation

The basic concept is to use the deterministic INCA 2-m temperature forecast as ensemble mean and to determine the ensemble variance by different statistical and/or dynamical approaches. Three different methods for constructing an 18 member ensemble are tested over a 1-month period (1–30 November 2010): a purely statistical method, a dynamical one, and a combined dynamic-statistical approach.

##### 4.1 Statistical method “INCA-EPS<sub>stat</sub>”

This method uses the spread-skill relation and follows the idea of Cui et al. (2005), who use the root mean squared error of the ensemble mean for calibrating the ensemble spread, in a modified way. The operational INCA run is used as ensemble mean and its RMSE over the past 30 training days is used as ensemble spread. The generation of 18 ensemble members is achieved by taking the quantile values from the Gaussian cumulative distribution function (CDF) centred about the mean in such way that the ensemble spread of the new, re-scaled ensemble is limited by a fractional amount ( $f_{re-scale} = 3/4$ ) of the root mean squared error of the training data (Kann et al., 2009). Formally the distribution can be expressed by the Gaussian CDF

$$F(x) = \frac{1}{\text{RMSE}_{\text{INCA}} \sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{1}{2} \left(\frac{t - F_{\text{INCA}}}{\text{RMSE}_{\text{INCA}}}\right)^2\right) dt, \quad (1)$$

where  $F_{\text{INCA}}$  denotes the (deterministic) INCA forecast and  $\text{RMSE}_{\text{INCA}}$  the root mean square error of the INCA forecast, respectively, evaluated separately for each lead time and initialization. Finally, the 18 re-calibrated ensemble members are generated for  $i = 1, 18$  by

$$\text{EPS}(i) = \mu + Q[p(i)](\sigma^2), \quad (2)$$

where  $\mu$  denotes the ensemble mean,  $\sigma^2$  the variance, and  $Q$  is the quantile function (or inverse cumulative distribution function; CDF) of the standard normal distribution, evaluated at the probability  $p$ , given by

$$p(i) = \frac{1-z}{2} + (i-1) \times \left(\frac{z}{n-1}\right), \quad (3)$$

where  $n$  is the ensemble size of 18 members. The value of  $z$ , representing the re-scaled area around ensemble mean, is obtained iteratively, satisfying the constraint  $\sigma_{re-scaled} \leq f_{re-scale} \times \text{RMSE}_{\text{INCA}}$  (Kann et al., 2009), where  $\sigma_{re-scaled}$

denote the standard deviation of the re-scaled ensemble. In other words,  $z$  determines the amount of reduction of the standard deviation of the Gaussian distribution. Practically,  $z$  is modified until the standard deviation of the new ensemble is smaller than a fractional amount of the root mean square error of the ensemble mean. For example, if  $z = 0.5$ ,  $p(i = 1) = 0.25$  and  $p(i = 18) = 0.75$  etc. In general, well-tuned probabilistic forecasts show similar values of RMSE and ensemble spread. Thus, poor ensemble forecast quality in a well-tuned system is reflected by a large ensemble variance. Although statistically consistent, the practical usability is reduced if the spectrum of the ensemble member forecasts covers an unrealistic range from a synoptical point of view. Restricting the standard deviation should help to reduce this deficiency without loss of forecast skill. The value  $f_{re-scale} = 3/4$  turned out to be a reasonable choice of several experiments where the CRPS reaches a minimum. The impact of spread re-scaling is further discussed in Sect. 6.2.

##### 4.2 Dynamical method “INCA-EPS<sub>dyn</sub>”

The deterministic INCA run is used as control run and the individual differences between the 17 ensemble members and the ensemble mean of the last available ALADIN-LAEF (Wang et al., 2011) model initialization is added to the INCA run. Formally, the creation of the  $i$ -th ensemble member for a specific forecast projection  $t$  can be expressed by

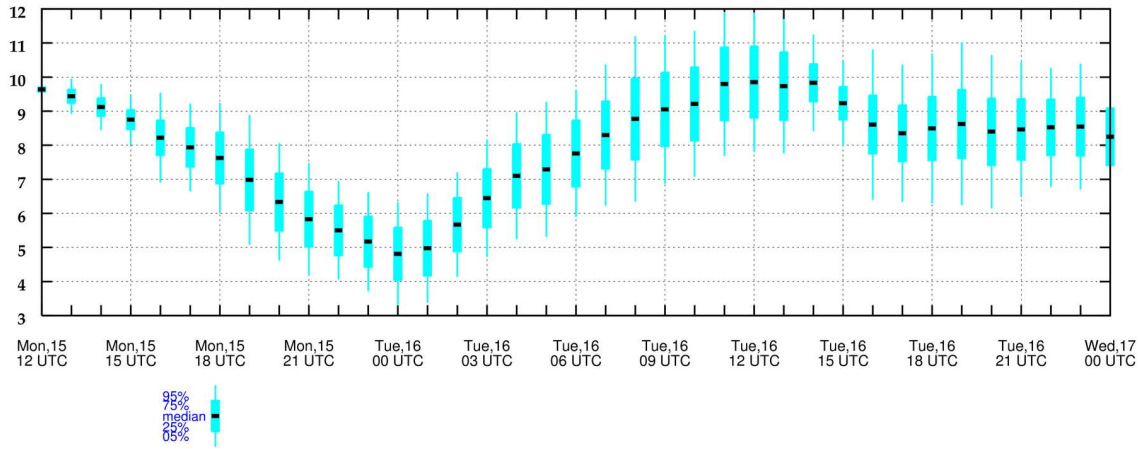
$$\text{EPS}(i) = F_{\text{INCA}}(t) - (\text{LAEF}_{\text{mean}}(t) - \text{LAEF}(i, t)). \quad (4)$$

In other words, the ALADIN-LAEF spread serves as spread of the INCA-EPS. Thus, only the uncertainties are taken from the limited area ensemble system.

##### 4.3 Dynamic-statistical method “INCA-EPS<sub>dynstat</sub>”

The ensemble created by combining the INCA 2-m temperature forecast and the ALADIN-LAEF ensemble spread given in Sect. 4.2 is additionally calibrated by applying the non-homogeneous Gaussian regression (NGR) technique (Gneiting et al., 2005; Hagedorn et al., 2008; Kann et al., 2009). This statistical calibration technique addresses both the first and second moments by modelling the mean and the ensemble variance according to the ensemble skill within a certain training period. The Continuous Ranked Probability Score (CRPS) serves as measure of skill, and optimal regression coefficients of mean and spread are estimated by minimizing the CRPS. The CRPS of the Gaussian distribution can be expressed by the coefficients  $a$ ,  $b$ ,  $c$  and  $d$  (Gneiting et al., 2005):

$$\text{CRPS}_{\text{train}} = \frac{1}{k} \sum_{i=1}^k (c + d\sigma_i^2)^{\frac{1}{2}} \left\{ Z_i [2\Phi(Z_i) - 1] + 2\varphi(Z_i) - \frac{1}{\sqrt{\pi}} \right\} \quad (5)$$



**Fig. 1.** Example of an EPSgram for 2-m temperature, forecast initialized at 15 November 2010, 00:00 UTC, for station Vienna Hohe-Warte.

with

$$Z_i = \frac{Y_i - (a + b\mu_i)}{(c + d\sigma_i^2)^{\frac{1}{2}}}. \quad (6)$$

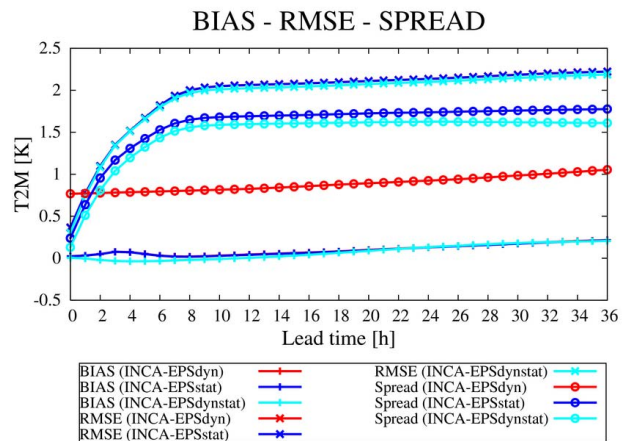
$\Phi$  and  $\phi$  denote the cumulative distribution function (CDF) and the probability distribution function (PDF) of the standard normal distribution, respectively.  $k$  is the training length in days,  $Y_i$  is the observation at day  $i$ .

The final 18-member ensemble is generated as described in Sect. 4.1 by taking the quantile values from the Gaussian CDF, centred about the mean under the constraints of the spread re-scaling factor.

#### 4.4 Implementation

The three sets of ensembles described above are generated retrospectively for one month from 1 November 2010 to 30 November 2010. Hourly INCA initializations are taken into account, each of them up to +36 h ahead. The 2-m temperature forecasts of INCA and of ALADIN-LAEF are interpolated bi-linearly to the locations of the automatic surface weather stations (about 250) in Austria, separately for each station and lead time up to +36 h.

Note that ALADIN-LAEF provides two probabilistic forecasts per day at 00:00 UTC and 12:00 UTC. The spread used for the experiments INCA-EPS<sub>dyn</sub> and INCA-EPS<sub>dynstat</sub> is calculated from the latest operationally available run, valid for the same target time as the considered INCA run (e.g.: the spread of the INCA ensemble initialized at 1 November 2010, 11:00 UTC + 06 h is provided by the ALADIN-LAEF initialization at 1 November 2010, 00:00 UTC + 17 h). There may be concern about the temporal consistency of combining different lead times as the skill of the NWP models usually decreases with time. However, the verification of the 00:00 and 12:00 UTC initializations only (i.e. combining model runs of the same initialization time) did not affect the results.

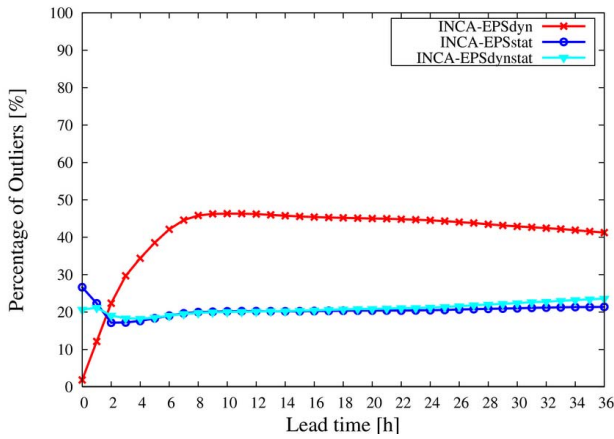


**Fig. 2.** Bias, root mean square error (RMSE) of the ensemble mean, and ensemble spread of the 3 experimental INCA ensembles (dynamical: red, statistical: dark blue, dynamic-statistical: light blue). Note that the BIAS of the experiments INCA-EPS<sub>stat</sub> and INCA-EPS<sub>dyn</sub> as well as the RMSE of INCA-EPS<sub>stat</sub> and INCA-EPS<sub>dyn</sub> are the same and therefore overlapping.

Figure 1 shows an example of a probabilistic 2-m INCA forecast for 15 November 2010, 00:00 UTC at the location of Vienna Hohe-Warte (11035). The ensemble was generated as described in Sect. 4.1 by using the root mean square error of the past 30 training days as ensemble spread and by applying the spread re-scaling.

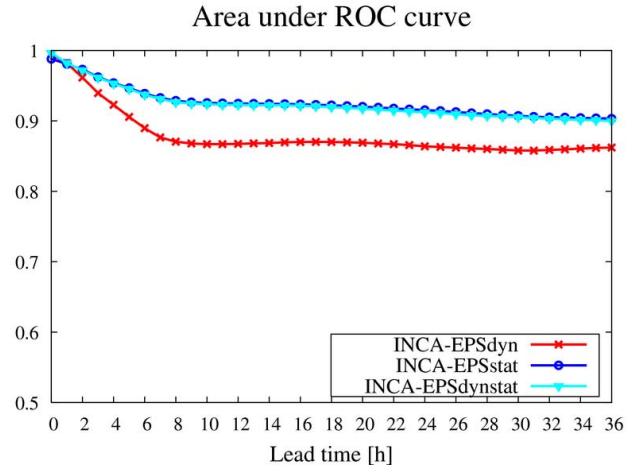
#### 5 Validation results

The evaluation was carried out for a one month period (1–30 November 2010) using 2-m temperature measurements of the network of ~250 TAWES (“Teilautomatische Wetterstationen”, Semi-automated weather stations) in Austria. The network covers most of the topographic elevation range



**Fig. 3.** Percentage of outliers as a function of lead time, for the dynamical (red), statistical (dark blue) and dynamic-statistical (light blue) method.

(100–3800 m), with the highest stations at Brunnenkogel (3440 m) and Sonnblick (3105 m). The average horizontal distance between stations is 18 km. All 24 forecast cycles (00:00, 01:00, . . . , 23:00 UTC) are combined for a specific forecast lead time. Figure 2 compares the BIAS and the RMSE of the ensemble mean and the ensemble spread of the three experiments as a function of lead time. Note that the ensemble mean error is almost identical for all experiments as they are all based on the same deterministic INCA forecasts. Thus, the BIAS of the experiments INCA-EPS<sub>stat</sub> and INCA-EPS<sub>dyn</sub> as well as the RMSE of INCA-EPS<sub>stat</sub> and INCA-EPS<sub>dyn</sub> are the same. However, as the NGR method, which is applied in experiment INCA-EPS<sub>dynstat</sub>, corrects both moments (ensemble mean and variance), marginal differences occur. The spread (standard deviation of ensemble members from ensemble mean) of the pure statistical approach and the combined dynamic-statistical approach are much more realistic with respect to the spread-skill relation. Especially in the nowcasting range, the statistical methods are able to adapt the spread consistently according to the higher skill and lower errors during the first couple of hours. In case of the dynamical approach, the percentage of outliers is about two times higher than with the statistical and the dynamic-statistical method (about 40 %–45 % and 20 %, respectively) (Fig. 3). Taken into account an expected statistical proportion of outliers of  $2/(n + 1)$ , which corresponds to approx. 10 % for an 18-member ensemble, the under-dispersive behaviour is improved by about 70 % by statistical adaptations. Another tool for assessing the forecast quality of an ensemble system is the ROC curve, and the area below it (Zhu et al., 2002), with the False Alarm Rate (FAR) on the x-axis and the Hit Rate (HR) on the y-axis. Compared to the dynamical approach, the dynamic-statistical and the pure statistical methods increase the ROC area (Fig. 4).



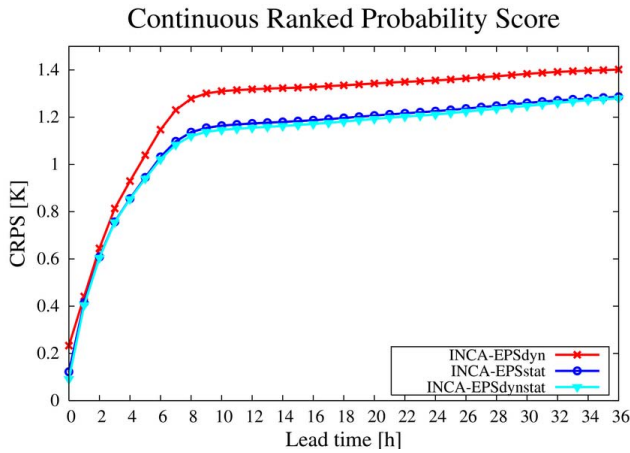
**Fig. 4.** Area under ROC curves for 2-m temperature anomaly  $> 0^\circ\text{C}$  as a function of lead time, from the dynamical (red), statistical (dark blue) and dynamic-statistical (light blue) method.

The Continuous Ranked Probability Score (CRPS) measures the overall distance of the ensemble to the observations (Hersbach, 2000) and is the generalized form of the discrete ranked probability score integrated over all possible thresholds. The CRPS has the appealing property that it generalizes the mean quadratic error, to which it reduces in case of a deterministic forecast. In the nowcasting range up to approx. 4–6 h, all methods are of similar, high skill. Outside this range, the statistical and dynamic-statistical approaches yield about 10 % lower CRPS values than the dynamical method (Fig. 5). Analogously, the Continuous Ranked Probability Skill Score (CRPSS), using the deterministic ALADIN-AUSTRIA model forecast as reference, is slightly increased for both statistical and dynamic-statistical methods (not shown). Furthermore, reliability diagrams are created, to provide information about the ability of the probability forecasts to reflect the observed relative frequency (Stanski et al., 1989). The verifying events chosen are 2-m temperature anomaly exceeding 0 K, using the ERA-40 dataset as climatological reference. In comparison to the dynamical ensemble approach, both the statistical and dynamic-statistical probability forecasts are more reliable for both high and low observed relative frequencies (Fig. 6). Especially the results of the dynamical method show that high observed frequencies are slightly overforecast and low frequencies are underforecast, which is reduced by the statistical applications.

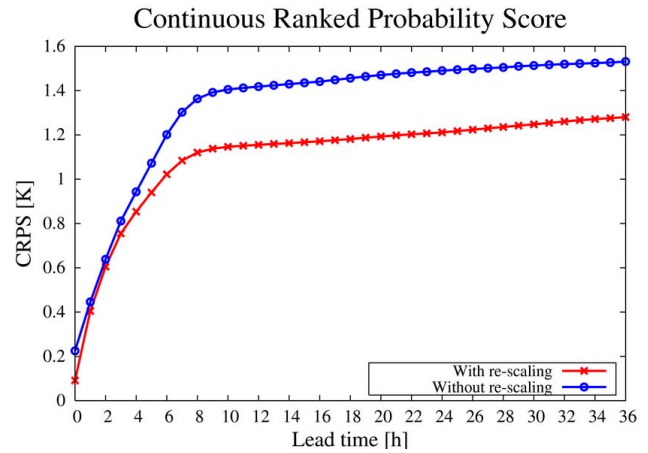
## 6 Additional considerations and sensitivity studies

### 6.1 The impact of training length

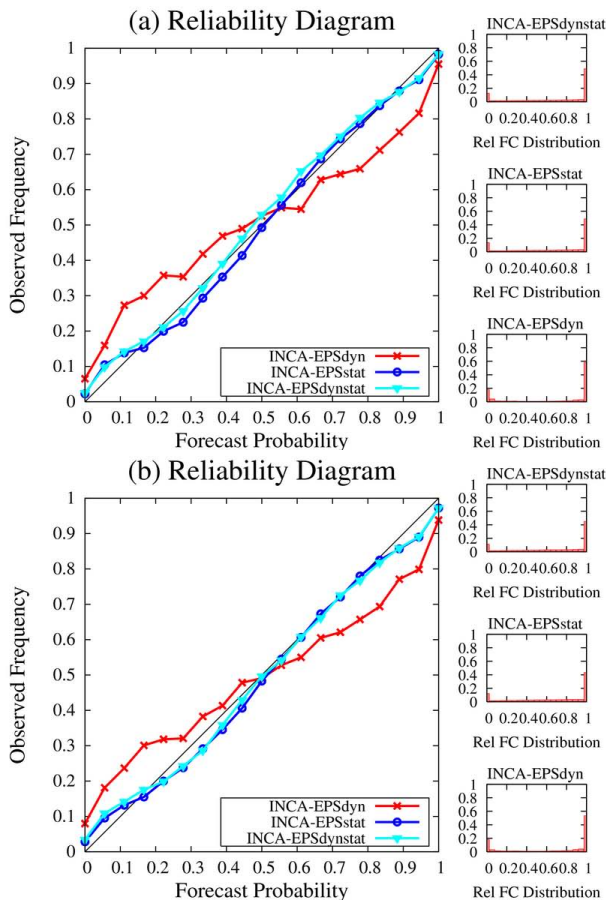
The statistical approach “INCA-EPS<sub>stat</sub>” and the dynamic-statistical method “INCA-EPS<sub>dynstat</sub>” are tested with an



**Fig. 5.** CRPS as a function of lead time, from the dynamical (red), statistical (dark blue) and dynamic-statistical (light blue) method.



**Fig. 7.** CRPS of the dynamic-statistical ensemble as a function of lead time, with (red,  $f_{re-scale} = 3/4$ ) and without (dark blue) spread re-scaling.



**Fig. 6.** Reliability diagram for +6h (a) and +24h (b) from the dynamical (red), statistical (dark blue) and dynamic-statistical (light blue) method.

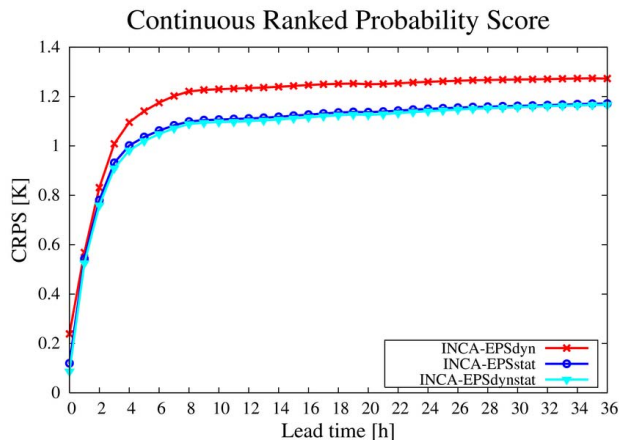
extended training length of 50 days, but in both cases without significant impact (not shown). This result agrees with previous findings which state the marginal impact of the training length for calibration of probabilistic 2-m temperature (Hamill et al., 2008), especially in the short range.

### 6.2 The impact of spread re-scaling

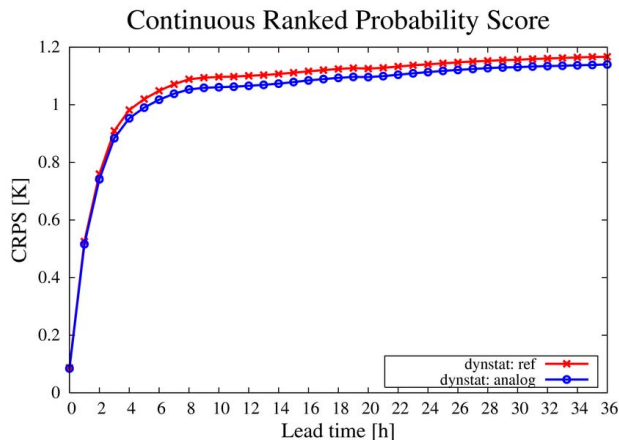
The re-scaling factor determines the PDF of the final calibrated ensemble values. Without re-scaling, i.e. by taking the quantile values of the predictive PDF at face value, the ensemble loses sharpness, and the CRPS increases compared to the experiment with spread re-scaling (using a value of  $f_{re-scale} = 3/4$  (Fig. 7). Lower values of the factor do not lead to significant improvements in terms of CRPS (not shown). However, for an optimized re-scaling factor, a seasonal- or situation-dependant factor should be calculated, based on retrospective training data.

### 6.3 Validation on another, independent sample (July 2010)

The same three approaches have been validated for another one-month period from 1 July to 31 July 2010 in order to generalize the findings and to test the statistical significance of the results gained from the period in November 2010 (Fig. 5). In terms of CRPS the results for July 2010 confirm the conclusion that all three methods provide sharp and reliable forecasts and that the statistical and dynamic-statistical methods slightly outperform the purely dynamical approach (Fig. 8).



**Fig. 8.** CRPS as a function of lead time, from the dynamical (red), statistical (dark blue) and dynamic-statistical (light blue) method. Verification period: 1 July 2010–31 July 2010.



**Fig. 9.** CRPS as a function of lead time, from the reference dynamic-statistical method (red) and dynamic-statistical method including the analogue “add-on” which uses training data selectively (dark blue). Verification period: 1 July 2010–31 July 2010.

#### 6.4 Refinements of the statistical methods

Generally, the statistical methods use historical training data to improve the skill of the raw probabilistic forecast. An optimized way of choosing proper training data or of taking into account error persistence could provide added value. A time-decaying variant of the NGR approach, i.e. giving higher weights to the latest model errors (Kann et al., 2009), was applied, but did not lead to further improvements.

An analogue-type correction that takes only into account the 30 most similar training days (from a sample of the last 60 days) for the RMSE calculation, is applied on both validation periods (July and November 2010). Note that “similar” is defined as those INCA runs with lowest sum of squared differences compared with the current run. In

case of November 2010, the analogue method does not add significant value, probably due to rather persistent weather regimes during late fall anyway (e.g. long-persisting stable situations with low stratus). During summer, the analogue add-on slightly improves the skill compared to the reference dynamic-statistical approach by about 5 % (Fig. 9).

## 7 Conclusions

The basic concept, combining INCA with statistically and dynamically derived ensemble variance, provides sharp and reliable probabilistic 2-m temperature forecasts. In accordance with the general higher skill of INCA in the nowcasting range, all methods show the highest skill during the first 12 h. The statistical and the coupled dynamic-statistical approaches give slightly better results than the pure dynamical method, i.e. statistical adaptations are able to overcome the under-dispersive behavior of the limited area ensemble system, at least from +12 h onwards. These results, obtained from a one-month validation period (November 2010), are confirmed by an independent verification sample from another season (July 2010).

Additional refinement of the statistical approaches using an analogue technique leads to slight further improvements (about 5 % in terms of CRPS) of the combined statistical-dynamical method, especially for July 2010.

Although the mean skills of the statistical and the combined dynamic-statistical method are very close, statistical methods alone are not ideally suited for handling uncertainties originating from dynamical processes. For example, temporal uncertainties concerning the passage of a front and the associated temperature de- or increase can only be addressed by an NWP model. Thus, “dynamical” uncertainties, adapted by statistical methods to reduce systematic model errors, are suitable and necessary ingredients for optimized, probabilistic 2-m temperature short range forecasts.

In general, these methods are not restricted to 2-m temperature only, although the assumption of a Gaussian distribution is a limiting factor. However, applying similar methods on 10-m wind speed using a non-negative variant (with a cut-off at zero) of the non-homogeneous Gaussian regression (Thorarinsdottir and Gneiting, 2010) confirms the results obtained for 2-m temperature and leads to the result that a combined dynamical-statistical approach performs best.

The experiments have been carried out based on station locations, but the method can easily be extended to gridded fields if this is required for operational purposes.

*Acknowledgements.* The authors are thankful to the editor and the anonymous reviewers for their suggestions and comments that led to substantial improvement of the manuscript.

Edited by: O. Talagrand

Reviewed by: two anonymous referees

## References

- Brozkova, R., Klaric, D., Ivatek-Sahdan, S., Geleyn, J. F., Casse, V., Siroka, M., Radnoti, G., Janousek, M., Stadlbacher, K., and Seidl, H.: DFI Blending, an alternative tool for preparation of the initial conditions for LAM. PWRP Report Series No. 31 (CAS/JSCWGNE Report), WMO-TD, No. 1064, 1.7, 2001.
- Buizza, R., Miller, M., and Palmer, T. N.: Stochastic simulation of model uncertainties, *Q. J. Roy. Meteor. Soc.*, 125, 2887–2908, 1999.
- Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, N., and Zhu, Y.: A comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems, *Mon. Weather Rev.*, 133, 1076–1097, 2005.
- Cui, B., Toth, Z., Zhu, Y., Hou, D., and Wobus, R.: Bias correction methods – adjusting moments, *Geophys. Res. Abstr.*, Vol. 7, 05997, 2005.
- Dance, S., Ebert, E., and Scurrah, D.: Thunderstorm strike probability nowcasting, *J. Atmos. Ocean. Tech.*, 27, 79–93, 2010.
- Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, *Mon. Weather Rev.*, 133, 1098–1118, 2005.
- Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B., and Gruber, C.: The Integrated Nowcasting through Comprehensive Analysis (INCA) system and its validation over the Eastern Alpine region. *Weather Forecas.*, 26, 166–183, doi:10.1175/2010WAF2222451.1, 2011.
- Hagedorn, R., Hamill, T. M., and Whitaker, J. S.: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts, Part I: temperature, *Mon. Weather Rev.*, 136, 2608–2619, 2008.
- Hamill, T. M., Snyder, C., and Morss, R. E.: A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles, *Mon. Weather Rev.*, 128, 1835–1851, 2000.
- Hamill, T. M., Hagedorn, R., and Whitaker, J. S.: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: precipitation, *Mon. Weather Rev.*, 136, 2620–2632, 2008.
- Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather Forecast.*, 15, 559–570, 2000.
- Homar, V., Stensrud, D. J., Levit, J. J., and Bright, D. R.: Value of human-generated perturbations in short-range ensemble forecasts of severe weather, *Weather Forecast.*, 21, 347–363, 2006.
- Kann, A., Wittmann, C., Wang, Y., and Ma, X.: Calibrating 2-m Temperature of Limited Area Ensemble Forecasts Using High-Resolution Analysis, *Mon. Weather Rev.*, 137, 3373–3387, 2009.
- Leutbecher, M. and Palmer, T. N.: Ensemble forecasting, *J. Comput. Phys.*, 227, 3515–3539, 2008.
- Lu, C., Yuan, H., Schwartz, B. E., and Benjamin, S. G.: Short-range numerical weather prediction using time-lagged ensembles, *Weather Forecast.*, 22, 580–595, 2007.
- Stanski, H. R., Wilson, L. J., and Burrows, W. R.: Survey of common verification methods in meteorology. WMO/WWW Tech. Rep. 8, 114 pp., 1989.
- Thorarindottir, T. L. and Gneiting, T.: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroskedastic censored regression, *J. R. Stat. Soc. A. Sta.*, 173, 371–388, 2010.
- Wang, Y., Haiden, T., and Kann, A.: The operational Limited Area Modelling system at ZAMG: ALADIN-AUSTRIA, *Österr. Beiträge Meteorol. Geophys.*, 37, ISSN 1016-6254, 2006.
- Wang, Y., Kann, A., Bellus, M., Pailleux, J., and Wittmann, C.: A strategy for perturbing surface initial conditions in LAMEPS, *Atmos. Sci. Lett.*, 11, 108–113, doi:10.1002/asl.260, 2010.
- Wang, Y., Bellus, M., Wittmann, C., Steinheimer, M., Weidle, F., Kann, A., Ivatek-Sahdan, S., Tian, W., Ma, X., Tascu, S., and Bazile, E.: The Central European limited-area ensemble forecasting system: ALADIN-LAEF, *Q. J. Roy. Meteor. Soc.*, 137, 483–502, doi:10.1002/qj.751, 2011.
- Zhu, Y., Toth, Z., Wobus, R., Richardson, D., and Mylne, K.: The economic value of ensemble based weather forecasts, *B. Am. Meteorol. Soc.*, 83, 73–83, 2002.