Calibrating 2-m Temperature of Limited-Area Ensemble Forecasts Using High-Resolution Analysis

ALEXANDER KANN, CHRISTOPH WITTMANN, AND YONG WANG
Central Institute for Meteorology and Geodynamics, Vienna, Austria

XULIN MA
China Meteorological Administration, Beijing, China

(Manuscript received 12 September 2008, in final form 13 May 2009)

ABSTRACT

Although the quality of numerical ensemble prediction systems (EPS) has greatly improved during the last few years, these systems still show systematic deficiencies. Specifically, they are underdispersive and lack both reliability and sharpness. A variety of statistical postprocessing methods allows for improving direct model output. Since 2007, Aire Limitée Adaptation Dynamique Développement International Limited Area Ensemble Forecasting (ALADIN-LAEF) has been in operation at the Central Institute for Meteorology and Geodynamics (ZAMG), and its 2-m temperature model output subject to calibration. This work follows the approach of nonhomogeneous Gaussian regression (NGR) that addresses a statistical correction of the first and second moment (mean bias and dispersion) for Gaussian-distributed continuous variables. It is based on the multiple linear regression technique and provides a predictive probability density function (PDF) in terms of a normal distribution. Fitting the regression coefficients, a minimum continuous ranked probability score (CRPS) estimation has been chosen instead of the more traditional maximum likelihood technique. The use of high-resolution analysis data on a 1 km × 1 km grid as training data improves the forecast skill in terms of CRPS by about 35%, especially on the local scale. The percentage of outliers decreases significantly without loss of sharpness. Sensitivity studies confirm that about half of the total improvement can be attributed to the effect of a bias correction. The training length plays a minor role, at least for the chosen verification period. A rescaling of the predictive PDF is important in order to obtain sharp forecasts, especially in the short range. Applying the same method to the global ensemble from the European Centre for Medium-Range Weather Forecasts (ECMWF) gives improvements of similar magnitude. However, the calibrated 2-m temperature of ALADIN-LAEF still remains slightly better than the 2-m temperature from calibrated ECMWF-EPS, which leads to the conclusion that statistical downscaling of EPS cannot replace dynamical downscaling. Finally, an advanced version of NGR, the so-called NGR-TD, which uses time-weighted averaging within minimum CRPS estimation, is able to yield a further improvement of about 5% in terms of the CRPS.

1. Introduction

During the last few years great efforts have been made to increase the forecast skill of operational ensemble prediction systems. Such attempts address the construction of perturbed initial and boundary conditions, and the uncertainty in model physics and assimilation processes. Different approaches in initial condition generation [i.e., by singular vectors (Molteni et al. 1996), breeding vectors (Toth and Kalnay 1997), ensemble transform Kalman filter (Wang and Bishop 2003; Bishop et al. 2001), and ensemble transform (Bishop and Toth 1999, Wei et al. 2008)], are discussed and compared in a variety of papers (Buizza et al. 2005; Hamill et al. 2000). Methods of simulating the uncertainty in model physics (i.e., stochastic physics or multiphysics approaches) have also been developed (Buizza et al. 1999).

In the recent past, the development of limited-area ensemble prediction systems has gained momentum. Analogous to deterministic limited area modeling, they focus on high-impact weather in the short and medium
range, especially on the regional scale. The construction of skillful limited-area model ensemble prediction systems (LAM-EPS) is the subject of research and is debated intensely within the community of researchers of numerical weather prediction. Compared to its global counterpart, the construction of LAM-EPS requires special treatment addressing the higher horizontal and vertical resolution, different physical parameterizations, and associated perturbations.

A review of the state-of-the-art in current operational LAM-EPS leads to the preliminary conclusion that statistical postprocessing of the direct model output is inevitable for operational usability. Typically, ensemble systems tend to lack both reliability and sharpness and they are biased and underdispersive.

The main objective of this study is to demonstrate the potential added value of calibrated limited-area ensemble forecasts in operational weather prediction. The paper describes a calibration method that addresses both the first and second moment of Gaussian-distributed continuous variables. Using high-resolution analysis training data, the method is applied to 2-m temperature. This application allows for statistical downscaling and adaptation on a local scale. Section 2 gives a short overview about the limited-area ensemble system Aire Limitée Adaptation Dynamique Développement International Limited Area Ensemble Forecasting (ALADIN-LAEF), which has been in operation at the Central Institute for Meteorology and Geodynamics (ZAMG) since 2006. The main characteristics and properties of this system will be summarized. Section 4 provides a deeper insight into the nonhomogeneous Gaussian regression (NGR) calibration method designed to compensate for systematic deficiencies of normal-distributed variables. The calibrated ensemble forecast will be evaluated by using measurements of the Austrian automatic station network and compared to the direct ensemble model output (section 5). Section 6 discusses issues concerning

1) the impact of training length,
2) the effect of spread rescaling,
3) the contribution of a simple bias correction to full calibration,
4) calibrating LAM-EPS versus its global counterpart, and
5) approaches toward refining the minimum continuous ranked probability score (CRPS) estimation.

2. Ensemble forecast data

The ensemble prediction system used in this study is ALADIN-LAEF (Wang et al. 2008, manuscript submitted to Quart. J. Roy. Meteor. Soc.), which has run in preoperational mode at ZAMG since March 2007. The horizontal resolution is about 18 km, with 37 levels in the vertical; integrations are performed twice per day at 0000 and 1200 UTC up to +54 h, covering Europe and large parts of the North Atlantic as shown in Fig. 1. The choice of the integration domain is mainly based on the fact that the evolution of weather systems in the North Atlantic plays a major role in the forecast for central Europe. At the same time the domain is big enough to minimize the impact of the lateral boundary conditions on the central European region.

In ALADIN-LAEF, initial and lateral boundary conditions are provided by the first 16 out of 50 members of the ensemble system of the European Centre for Medium-Range Weather Forecasts (ECMWF), the ECMWF control member, and the high-resolution member; the forecasts are dynamically downscaled using ALADIN-AUSTRIA (Wang et al. 2006). This model is the Austrian version of the high-resolution limited-area model ALADIN, which has been developed by 14 European national weather services within an international cooperation.

Initial perturbations of the ECMWF-EPS are generated by the singular vector approach (Molteni et al. 1996); stochastic physics are used for uncertainties in model physics (Buizza et al. 1999). Further details concerning the ALADIN-LAEF system can be found in Kann et al. (2007); a short summary of the model specifications is given in Table 1.

3. Analysis data

The 2-m temperature analysis used for calibration is part of the output of the nowcasting system INCA that is being developed at ZAMG. The meteorological fields analyzed with INCA are temperature, specific humidity, wind, precipitation, cloudiness, and global radiation. The system is designed to create 2D and 3D analysis fields as closely as possible to observational data. In the following the main characteristics of the INCA temperature analysis are presented; a comprehensive description can be found in Haiden et al. (2009).

The three-dimensional analysis of temperature in INCA is performed on a domain centered over Austria, covering an area of 600 km × 350 km at a resolution of 1 km (601 × 351 grid points) on a Lambert conical projection. In the vertical, a z system is used with z describing the height above the “valley floor surface” (shown in Fig. 2), which is derived from high-resolution
topographical data by determining for every grid point the minimum elevation within a given radius. The valley floor surface coincides with the topography over flat terrain and gives an appropriate reference in mountainous areas. There are currently 21 levels used in the system with an equidistant $\Delta z$ of 200 m, so the temperature analysis covers the lowest 4000 m above the local valley floor surface.

The INCA temperature analysis starts with the temperature forecast of the ALADIN-AUSTRIA model as a first-guess field. ALADIN-AUSTRIA is the operational limited area model at ZAMG, run on a Lambert conformal projection with a horizontal resolution of 9.6 km and 60 levels in the vertical (Wang et al. 2006). The operational INCA temperature analysis is computed every hour. ALADIN-AUSTRIA 0000 and 1200 UTC runs provide the first-guess fields, depending on the availability of the model runs in the operational environment (e.g., the ALADIN-AUSTRIA forecast fields at 1200 UTC initialization +12-h lead time are used for the INCA analysis at 0000 UTC). The first guess is constructed through a trilinear interpolation of the

The INCA temperature analysis starts with the temperature forecast of the ALADIN-AUSTRIA model as a first-guess field. ALADIN-AUSTRIA is the operational limited area model at ZAMG, run on a Lambert conformal projection with a horizontal resolution of 9.6 km and 60 levels in the vertical (Wang et al. 2006). The operational INCA temperature analysis is computed every hour. ALADIN-AUSTRIA 0000 and 1200 UTC runs provide the first-guess fields, depending on the availability of the model runs in the operational environment (e.g., the ALADIN-AUSTRIA forecast fields at 1200 UTC initialization +12-h lead time are used for the INCA analysis at 0000 UTC). The first guess is constructed through a trilinear interpolation of the

<table>
<thead>
<tr>
<th>Table 1. Model specifications and settings of ALADIN-LAEF.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Numerical model</strong></td>
</tr>
<tr>
<td>Forecast period</td>
</tr>
<tr>
<td>Horizontal mesh</td>
</tr>
<tr>
<td>Vertical levels</td>
</tr>
<tr>
<td>Initial condition</td>
</tr>
<tr>
<td>Lateral boundary</td>
</tr>
<tr>
<td>Initial perturbation</td>
</tr>
<tr>
<td>Advection</td>
</tr>
<tr>
<td>Time integration</td>
</tr>
<tr>
<td>Moist physics</td>
</tr>
<tr>
<td>Convection</td>
</tr>
<tr>
<td>Turbulence</td>
</tr>
<tr>
<td>Radiation</td>
</tr>
<tr>
<td>Ground temperature</td>
</tr>
</tbody>
</table>
model temperature field onto the three-dimensional INCA grid. The horizontal resolution (9.6 km) of the ALADIN-AUSTRIA model gives a rather poor representation of valleys and inner-alpine basins. To construct a realistic valley atmosphere on the high-resolution (1 km) grid, the boundary layer atmosphere of the model is shifted downward to the INCA valley surface floor. This is done by using a modified most unstable temperature gradient, which is defined here as the maximum temperature lapse rate found within the lowest layer of depth \( H_{\text{EXT}} \) (=1000 m) in the ALADIN-AUSTRIA model atmosphere over successively higher height intervals of thickness \( H_{\text{EXT}}/2 \). After the most unstable gradient is found, a minor modification of the gradient is performed through introduction of a correction factor (Haiden et al. 2009). The most unstable gradient is used to avoid unrealistic extrapolations in cases of strong surface or elevated inversions.

In the next step, the first guess is corrected by using the differences between forecast and observation at surface station locations. Station observations are located in the atmospheric surface layer, thus it is important to take into account the daytime temperature surplus and the deficit during night near the surface [\( DT_{\text{ALA}} \) in Eq. (1)]. This is done by separating the ALADIN-AUSTRIA 2-m temperature forecast \( T_{\text{ALA}} \) into a 3D model-level part and a 2D surface layer contribution:

\[
T_{\text{ALA}} = T_{\text{LALA}} + DT_{\text{ALA}}. \tag{1}
\]

Here \( T_{\text{LALA}} \) is the ALADIN-AUSTRIA 2-m temperature direct model output (now interpolated to the high-resolution INCA 1-km grid), \( T_{\text{LALA}} \) is the temperature at the lowest model level, and \( DT_{\text{ALA}} \) is the difference between the two temperatures. The corrections \( \Delta T \) of the first-guess field are also partitioned into a 3D part \( \Delta TL \) and a surface layer contribution \( \Delta DT \):

\[
\Delta T = \Delta TL + \Delta DT. \tag{2}
\]

In contrast to the model forecast, this partition has to be estimated relying on the assumption that the forecast error is restricted to the surface layer as long as this is physically plausible. The 3D part of the correction at the \( k \)th station is computed from

\[
\Delta TL_k = \max[0, T_{k}^{\text{OBS}} - \max(T_{k}^{\text{LALA}} + I_{\text{SFC},k} DT_{\text{LALA}}^{\text{EXT}}, T_{k}^{\text{LALA}} + I_{\text{SFC},k} DT_{\text{SCALE}}^{\text{EXT}})]
\]

if \( T_{k}^{\text{OBS}} \geq T_{k}^{\text{LALA}}, \tag{3}\)

and from

\[
\Delta TL_k = \min[0, T_{k}^{\text{OBS}} - \min(T_{k}^{\text{LALA}} + I_{\text{SFC},k} DT_{k}^{\text{ALA}}^{\text{EXT}}, T_{k}^{\text{LALA}} - I_{\text{SFC},k} DT_{k}^{\text{SCALE}}^{\text{EXT}})]
\]

if \( T_{k}^{\text{OBS}} < T_{k}^{\text{LALA}}, \tag{4}\)

The quantity \( DT_{\text{SCALE}}^{\text{EXT}} \), which is always greater than 0, is the estimation for the magnitude of the near-surface temperature surplus or deficit depending on insolation and wind speed observations. The factor \( I_{\text{SFC},k} \) is the so-called surface layer index, which takes on a continuum of values from 1 in flat terrain or valleys to 0 at mountain slopes, ridges, and peaks (Haiden et al. 2009). The 3D corrections \( \Delta TL_k \) are then spatially interpolated onto the INCA grid using geometrical distance weighting in the horizontal and a distance weighting in potential temperature space in the vertical. Using the three-dimensional squared “distance” \( r_{ijmk}^2 \) between INCA grid point \((i, j, m)\) and the \(k\)th station, the temperature difference field at grid point \((i, j, m)\) is obtained from the
weighted average (with \( n = 6 \) representing the number of nearest stations to the actual grid point):

\[
\Delta TL(i, j, m) = \frac{\sum_{k=1}^{n} \Delta TL_k}{\sum_{k=1}^{n} r_{ijmk}^2}.
\]  

After that the temperature difference field is added to the ALADIN-AUSTRIA first-guess field

\[
\text{TL}^{\text{INCA}}(i, j, m) = \text{TL}^{\text{ALA}}(i, j, m) + \Delta TL(i, j, m).
\]  

The remaining differences \( \Delta DT_k \), which can be interpreted as surface layer contributions, are obtained from

\[
\Delta DT_k = T_k^{\text{OBS}} - \text{TL}_k^{\text{INCA}},
\]

and interpolated horizontally by using simple inverse distance weighting. Finally, the resulting INCA 2-m temperature analysis \( T^{\text{INCA}} \) is obtained by adding the surface layer correction to the corrected model-level temperature at the given topography height using

\[
T^{\text{INCA}}(i, j) = \text{TL}^{\text{INCA}}[i, j, m(z)] + I_{\text{SFC}} \Delta DT(i, j).
\]  

The INCA system is under continuous development. Verification shows that the mean absolute error of the temperature analysis as determined by cross validation over all Austrian stations is typically less than 1 K in well-mixed situations and between 1 and 1.5 K in stable situations.

### 4. Calibration procedure

A variety of methods exists for statistical adaptation of the direct model output of ensemble forecasts. Focusing especially on severe weather, the parameters of primary interest for statistical calibration are precipitation, 10-m wind speed, and 2-m temperature. As the observed relative frequency of precipitation is characterized by a high degree of skewness, logistic regression techniques are found to be adequate for many applications (Hamill et al. 2008). In case of wind speed, calibration experiments have been carried out that use a power transformation combined with a cutoff normal distribution approach (Gneiting et al. 2006).

The potential of different calibration methods has been examined and summarized by Wilks and Hamill (2007). Analog techniques (Hamill and Whitaker 2006), Bayesian model averaging (Raftery et al. 2005; Sloughter et al. 2007), and ensemble dressing methods (Wang and Bishop 2005) are found to improve EPS skill substantially. Nonhomogeneous Gaussian regression, a generalization of the standard linear regression, proposed in a modified way by Gneiting et al. (2005), turns out to be a promising candidate for calibrating continuous variables that are well represented by a Gaussian PDF, particularly in the short range (Wilks 2006).

Statistical evaluation of observed 2-m temperature data confirms the Gaussian distribution to be a good approximation to the observed cumulative distribution (Wilks 2006). Position and shape are characterized by the ensemble mean and variance. In the case of a nonhomogeneous Gaussian regression, both mean and variance are modeled by linear regression (thus, the regression is nonhomogeneous as the variance is varying with different values of the predictor).

If \( \mu \) denotes the ensemble mean and the variance is given by \( \sigma^2 \), the predictive, Gaussian PDF (modeled by NGR) can formally be written as a normal distribution \( N \):

\[
\text{PDF} = N(a + b\mu, c + d\sigma^2).
\]  

The regression coefficients \( a \) and \( b \) determine the behavior of the bias, the coefficients \( c \) and \( d \) reflect the spread–skill relationship. Fitting Eq. (9) means that the values of the regression coefficients \( a, b, c, \) and \( d \) have to be found. Traditionally, estimation of the coefficients is carried out by maximum likelihood estimation (MLE; Wilks 2006). Gneiting et al. (2005) propose a new approach based on minimization of the CRPS, which is a proper score (Gneiting and Raftery 2007). The CRPS is the generalized form of the discrete ranked probability score, simulating the mean over all possible thresholds. Let the variable \( x \) denote the parameter of interest, for example 2-m temperature or 10-m wind speed. If the PDF forecast by an ensemble system is given by \( P(x) \), and \( x_a \) is the value that currently occurs, then the CRPS expresses the distance between the probabilistic forecast \( P \) and the “truth” \( x_a \) (Hersbach 2000):

\[
\text{CRPS}(P, x_a) = \int_{-\infty}^{x_a} [P(x) - P_a(x)]^2 \, dx.
\]  

The cumulative distributions \( P \) and \( P_a \) can be written as

\[
P(x) = \int_{-\infty}^{x} \rho(y) \, dy \quad \text{and} \quad P_a(x) = H(x - x_a),
\]  

with the Heaviside function:

\[
H(x - x_a) = \begin{cases} 
0 & \text{for } x < x_a, \\
1 & \text{for } x \geq x_a.
\end{cases}
\]
The CRPS compares the overall distribution with observation and generalizes the mean absolute error, to which it reduces in the case of a deterministic forecast. In the present case of a normal distribution, the CRPS can be described analytically by the coefficients \(a, b, c,\) and \(d\) (Gneiting et al. 2005):

\[
\text{CRPS}_{\text{train}} = \frac{1}{k} \sum_{i=1}^{k} (c + d\sigma_i^2)^{1/2} \\
\times \left\{ Z_i [2\Phi(Z_i) - 1] + 2\phi(Z_i) - \frac{1}{\sqrt{\pi}} \right\},
\]

with

\[
Z_i = \frac{Y_i - (a + b\mu)}{(c + d\sigma)^{1/2}}.
\]

Both \(\Phi\) and \(\phi\) denote the cumulative distribution function (CDF) and the PDF of the standard normal distribution (with \(\mu = 0, \sigma^2 = 1\), respectively. Here \(k\) is the number of days in the training sample and \(Y_i\) corresponds to the observation at day \(i\).

Estimation of the regression coefficients is performed with amoeba.f90, which is based on the simplex method by Nelder and Mead (1965) (from Numerical Recipes in Fortran90). Figure 3 shows an example of the coefficients in the case of 0000 UTC 16 December 2007. In lowland areas the coefficients generally show rather smooth and homogeneous structure. In Alpine areas the spatial variability of the coefficients reflects the topographic influence and varying heights to a large extent. The coefficients that are found iteratively from the training data are finally applied on the current ensemble mean and variance, providing a new, calibrated probability forecast in terms of an adapted Gaussian PDF.

To allow for a fair comparison between direct and calibrated ensemble forecasts, ensemble members of the same size as the original EPS are generated from the calibrated PDF. For an ensemble size of \(n\) members, one option is to generate the \(i\)th ensemble member from the \([i/(n + 1)]\)th quantile of the standard normal distribution (Hagedorn et al. 2008). Generally, well-tuned ensemble systems show similar values of root-mean-squared error and ensemble spread. Consequently, poor ensemble forecast quality in a well-tuned system would lead to large ensemble spread. Although statistically consistent, the practical usability is somehow reduced if the spectrum of the ensemble member forecasts covers an unrealistic range (in a synoptic sense). To overcome this deficiency, the generation of the ensemble members is realized by taking the quantile values from the CDF centralized about the mean in such a way that the ensemble spread of the new, rescaled ensemble is limited by a fractional amount of the root-mean-squared error of the training data. Thus, the creation of the \(i\)th of \(n\) ensemble members, \(\text{ENSMem}(i)\) can be expressed as
\[ ENSmem(i) = Q(p(i))(c + \sigma^2) + (a + b\mu), \quad (15) \]

where \( Q \) denotes the quantile function of the standard normal distribution, evaluated at the probability \( p \) that is given by

\[ p(i) = \frac{1 - z}{2} + (i - 1) \left( \frac{z}{\mu - 1} \right). \quad (16) \]

The quantity \( z \), representing the rescaled percentile area around ensemble mean, is obtained iteratively satisfying the constraint \( \sigma_{\text{re-calibrated}} \leq f_{\text{re-scale}} \text{RMSE}_{\text{train}} \). Practically, \( z \) determines the amount of reduction of the standard deviation of the Gaussian distribution. A reasonable choice (in terms of the CRPS) of \( f_{\text{re-scale}} \) turned out to be about \( 2/3 \). For operational purposes, the optimal value of \( f_{\text{re-scale}} \) should be determined on a monthly or at least on a seasonal basis in order to take into account the seasonal variability of the ensemble system quality. Experiments on settings of the rescaling factor are further discussed in section 6b. Applying this procedure to the operational ALADIN-LAEF 2-m temperature, these calibrated (or “dressed”) ensemble members are the basis of calibrated ensemble forecast products.

The calibration procedure described above was tested for one month during 1–31 December 2007. The twice-daily model runs (0000 and 1200 UTC) were taken into account separately. The 2-m temperature fields of ALADIN-LAEF were interpolated bilinearly onto the INCA domain and grid and statistically downscaled applying a sliding training period of 30 days. This procedure was carried out separately for each cycle, grid point, and lead time. For example, applying the calibration to 0000 UTC 1 December 2007, the training period consisted of the previous 30 days for a +6-h projection. The calibration process for the +54-h projection made use of those last 30 days, for which the corresponding observations had been available yet (in this specific case the 0000 UTC model runs from 30 October 2007 to 28 November 2007). The experiments were done retrospectively, so no real-time products have been generated for the time being.

The calibration could be done with observations only, but as the INCA analysis reproduces the measurement at the station location (within the horizontal resolution), there would be no difference in verification results (on station basis). Moreover, the high-resolution INCA analysis provides the best available gridded training dataset. Figure 4 demonstrates the individual steps of the process by means of the case 0000 UTC 16 December 2007. The top-left panel shows ensemble mean of ECMWF-EPS 2-m temperature forecast (+36 h), ALADIN-LAEF (top right), which is obtained by dynamical downscaling of ECMWF-EPS members, is able to simulate more regional structures on the scale of the model’s horizontal resolution. The ensemble mean obtained by the calibrated

**Fig. 4.** Ensemble mean: production chains of ALADIN-LAEF at 0000 UTC 16 Dec 2007, +36-h lead time. Direct model output (DMO) from (a) ECMWF-EPS, (b) ALADIN-LAEF, (c) calibrated ALADIN-LAEF, and (d) verifying INCA analysis.
ensemble members (bottom left) contains information on a local scale due to the downscaling procedure on the high-resolution grid. Comparisons with the verifying INCA analysis (bottom right) confirm high correlations, especially in Alpine valleys and basins (i.e., Inn Valley, basin of Klagenfurt, etc.).

Usually, probability forecasts are derived from ensemble forecasts. Generating such products from direct and calibrated ensemble forecasts confirms that statistical downscaling improves the forecast quality on the local scale (Fig. 5). In the next section, a comprehensive evaluation using a variety of both deterministic and probabilistic verification scores will objectively examine this issue.

5. Verification: Direct versus calibrated forecast

For assessing the forecast quality of an ensemble system, the following verification scores are used:

1) for deterministic forecasts, the bias and root-mean-square error of ensemble mean are used;
2) for probabilistic forecasts, rank histograms (“Talagrams”), percentage of outliers, relative operating characteristics (ROC), reliability diagrams, CRPS, continuous ranked probability skill score (CRPSS) are used.

A detailed description of verification scores is given by Wilks (2006).

The evaluation was carried out retrospectively for 1 month (1–31 December 2007) using 2-m temperature measurements of approximately 170 automatic stations from the Austrian network. The verification scores for both forecast cycles (0000 and 1200 UTC) are combined for a specific forecast projection. The spatial distribution of stations is shown in Fig. 6. The statistical downscaling procedure on a spatial resolution of 1 km implicitly contains a bias correction that is (partly) due to differences between model topography and station elevation. Thus, there is no need for an additional height correction as the bias (including the contribution caused by differences in elevation) is reduced automatically through the regression.

Alternatively, verification could be done based on gridded analysis fields, but an analysis error would be introduced into the verification, which is small, but not negligible. Verifying on a station basis excludes this problem.

For 1), Fig. 7 shows bias and root-mean-squared error (RMSE) of the ensemble mean of the direct output and the calibrated ALADIN-LAEF system. The bias is removed to a great extent; the RMSE is reduced from about 3 to 2.4 K. The standard deviation of the ensemble members (from ensemble mean) is increased from 0.1–0.7 to 1.6 K. Thus, the calibrated ensemble seems to reflect the model uncertainty much better.
than the direct output of the ensemble (spread–skill relation).

For 2), the rank histograms or Talagrand diagrams (Hamill 2001; Talagrand et al. 1997) alone do not give full information about the quality of ensembles, but they are providing assistance for evaluating the ability of the model to reflect the observed frequency distribution. Rank histograms for various forecast lead times indicate the lack of ensemble spread causing U-shaped rank histograms (not shown). This underdispersive behavior is confirmed by the percentage of outliers (as a function of lead time) in Fig. 8, which indicates that calibration is able to reduce the number of outliers from more than 70% to about 25%. Additionally, reliability diagrams provide information about the ability of the probability forecasts to reflect the observed relative frequency (Stanski et al. 1989). Compared to the direct ensemble output, the calibrated probability forecasts are much more reliable for both high and low observed relative frequencies (not shown). Although high observed frequencies remain slightly overforecast and low frequencies are underforecast, the calibrated ensemble output allows for probabilistic predictions that are very close to the probability of occurrence.

Another tool assessing the quality of an ensemble system is the ROC curve and the area under the ROC curve (Zhu et al. 2002), composed of false alarm rate (FAR) on the x axis and hit rate (HR) on the y axis. A perfect forecast would be obtained by FAR = 0 and HR = 1. Relative to the direct ensemble output, the calibration significantly improves ROC scores and the values of areas under the ROC curves (Fig. 9).
The CRPS, previously discussed in section 4, represents the overall distance of the ensemble to observation. The fact that the CRPS reduces to the mean absolute error in the case of deterministic forecasts allows direct and straightforward interpretation. The mean CRPS decreases from approximately 2.3 K of the direct model output (DMO) to 1.5 K by calibration (thus the relative improvement is almost 35%, Fig. 10). In the case of the CRPSS with the deterministic limited area model as a reference (operational ALADIN-AUSTRIA), the calibration achieves an improvement from about 0.3 to 0.6, so the skill score is almost doubled (Fig. 11).

To prove the potential of the calibration method on a different and independent sample, an additional verification period was chosen. A study of NGR on ensemble forecasts during the convective season offers a broader basis for assessment of the quality of the approach. Again, a sliding training period of 30 days was used and applied to both cycles (0000 and 1200 UTC) independently. For the verification period July 2008, the mean CRPS is decreased from approximately 2.4 K (DMO) to almost 1.0 K (Fig. 12), and the CRPSS is increased from about 0.35 to 0.7 (Fig. 13). Thus, the improvement of forecast quality is about the same magnitude as obtained for the winter month of December 2007.

6. Sensitivity studies

a. The impact of training length

The choice of the training length determines the degree of improvement of calibration. This is especially true for rare events that are hardly covered by short
periods of training data (Hamill et al. 2008). In case of 2-m temperature, sensitivity studies reveal that the training length has only marginal impact on calibration success. Different sliding periods of 30, 50, and 100 days were used and their results compared, but the differences between 30 and 50 days are negligible (Fig. 14). A training period of 100 days results in even worse scores, which is probably due to the inclusion of forecast error from a time window that is not representative for the evaluation month of December. So it seems that for limited-area ensemble forecasting and in case of a low synoptic variability during the chosen verification period shorter training periods are sufficient or even slightly better. Most likely the training length plays a more important role during other seasons, when temporal variability dominates (e.g., convective season) and for rare events.

b. The effect of spread rescaling

As mentioned in section 5, the spread rescaling of the new, calibrated ensemble determines the final shape of the predictive Gaussian PDF. Figure 15 compares the calibration procedures using $f_{\text{re-scale}} = \frac{3}{4}$, $f_{\text{re-scale}} = \frac{2}{3}$, and $f_{\text{re-scale}} = \frac{1}{2}$ and $f_{\text{re-scale}} = 1$. Although similar, the first two settings turn out to be more appropriate than using half of the RMSE as a limit of spread. Higher values of $f_{\text{re-scale}}$ do not result in a significant improvement of forecast skill in terms of the CRPS.

c. The contribution of simple bias correction to full calibration

NWP models are frequently affected by systematic errors leading to biases that vary in space and time. These deficiencies may be caused by deviations of the model topography from real topography and by inaccurate representation of model physics and surface processes. The simplest way of eliminating systematic errors is achieved by removing first-order bias. Hamill et al. (2008) found the bias correction contributing to a large extent to full calibration. In our implementation, the mean bias was assessed from the same sliding training period and applied on each ensemble member. With respect to the mean CRPSS (again with the deterministic ALADIN-AUSTRIA model as a reference) about 50% of the improvement can be attributed to the bias correction (Fig. 16). The CRPSS increases from 0.3 (DMO) to roughly 0.45 (bias-corrected ensemble forecast), and furthermore to 0.6 by applying a full calibration.

d. Calibrated ALADIN-LAEF versus calibrated ECMWF-EPS

It has been successfully demonstrated that calibration significantly improves the quality of probability
forecasts of ALADIN-LAEF, especially on a local scale when using a high-resolution analysis. In section 2, the setup of the operational ALADIN-LAEF ensemble system has been discussed. Dynamical downscaling of the global ECMWF ensemble members forms the core process of this system. The question arises, if calibration is able to provide the same forecast quality applied on ECMWF-EPS. In other words: Do we actually need the process of dynamical downscaling before statistical adaptation or are we able to achieve the same results by calibrating the global ensemble itself? Trying to answer this question, the same calibration procedure (with the same training period and rescaling factor) was applied to the 18 ECMWF ensemble members that are used operationally for dynamical downscaling with ALADIN. Again, 2-m temperature ensemble forecast fields were interpolated onto the INCA grid and calibrated by using INCA 2-m temperature analysis fields. The calibration process was carried out on the same forecast period (December 2007) with the same training length. Finally, a comparative verification was done by using station observations again (Fig. 6). Deterministic scores deliver slightly better performance of the calibrated ALADIN-LAEF than calibrated ECMWF-EPS. The bias averages almost to zero in case of ALADIN-LAEF, but remains slightly negative for ECMWF ensemble mean forecasts. Both spread of the ECMWF ensemble members and RMSE of ensemble mean are somewhat higher than its limited-area counterpart (Fig. 17). In an operational system, the calibration of the ECMWF-EPS would be preferably extended to the full number of members (one high-resolution, one control, and 50 ensemble members) and not limited to an arbitrary number of 18. Thus, the practical added value of combining dynamical and statistical downscaling compared to a statistical downscaling procedure has to be evaluated by calibrating the full set of ECMWF-EPS members. In an additional experiment, the same forecast period with the same training length was used, but applied on the full set of 52 ensemble members. In terms of CRPS, the calibrated probability forecasts of ALADIN-LAEF perform slightly better than that of ECMWF-EPS, regardless of the number of calibrated ECMWF-EPS members (18 or 52 members); the relative difference is up to 8% (Fig. 18). It has to be checked carefully, whether calibrated ALADIN-LAEF forecasts slightly outperform calibrated ECMWF-EPS forecasts only in the selected verification period or whether this result persists during other seasons.
e. Refinements on the approach of minimum CRPS estimation (NGR-TD)

The NGR method, as proposed by Gneiting et al. (2005), suggests optimizing the mean CRPS over a specified time window. Although the larger the sample the more robust the score becomes, the temporal variability of the model error is unknown because of averaging. Taking into account a quasi-persistency of model error, a time-decaying averaging method was applied on CRPS estimation (called NGR-TD) instead of simple averaging over the entire period. Motivated by Cui et al. (2006) for a first moment bias correction, the CRPS was adapted by a weighted mean of a prior estimate and an update. The following steps were carried out (for a sliding training period of 50 days).

A prior time mean average of the CRPS is calculated between days 50 and 30. The prior average is multiplied by \((1 - w)\) and added to the most recent CRPS, multiplied by the weight \(w\). As a result of earlier experiments with time-decaying averaging method for a first-order bias correction, the factor \(w = 0.1\) turned out to be a proper value (not shown). The update procedure described above is repeated from days 29 to 1. Finally, the obtained time-weighted mean CRPS is used in the NGR method for calibration of temperature forecast of day 0.

The verification of the calibrated probability forecasts shows a slight further improvement compared to the reference version with mean CRPS estimation. Both the bias and root-mean-squared error are reduced (Fig. 19); the CRPS is improved up to 5% by applying NGR-TD instead of NGR (Fig. 20). To test the robustness of the refinement during another independent sample, the experiment is carried out for July 2008. Similarly, the time-decaying method is able to improve forecast skill (in terms of CRPS, not shown), although less pronounced than for December 2007 (about 1%–2%). It has to be checked carefully if this refinement of NGR is able to increase ensemble skill during all seasons and if further improvements can be achieved by seasonal variations of the weighting factor \(w\). Furthermore, the importance of the training length in combination with this refinement has to be investigated before an operational implementation.

7. Conclusions

The direct model output of ALADIN-LAEF is biased and shows deficiencies concerning both reliability and sharpness. Consequently, there is a need for calibration and postprocessing techniques that address these issues and provide to the forecasters more reliable products for operational use. Previous papers demonstrated a variety of methods that are able to improve probability forecasts by calibration techniques. A promising candidate that is appropriate for continuous, Gaussian-distributed variables is known as the nonhomogeneous Gaussian regression. Within NGR, a rather new approach by Gneiting et al. (2005) proposes a minimum CRPS estimation instead of the classical MLE. This technique was applied on the limited-area ensemble system ALADIN-LAEF, which has been in operation at ZAMG since spring 2007. High-resolution analysis data from the analysis and nowcasting system INCA were used as training data, leading to calibrated probability forecasts on a very high-resolution (1 km × 1 km) grid. The verification of the calibrated probability forecasts
was carried out by using approximately 170 automatic stations within Austria. In summary, the calibrated ensemble was improved with respect to both reliability and sharpness. Rank histograms became flatter; the percentage of outliers decreased. Despite larger spread the system did not lose sharpness. On the contrary, the continuous ranked probability score, which summarizes the overall behavior of the ensemble compared to observation, was significantly reduced by 35%. The usage of high-resolution analyses as training data had an added value especially on a local scale, where additional information was missing in the DMO, but obtained by statistical downscaling (i.e., implicit height correction). Hence probability products of calibrated ensemble forecasts provide useful information for daily operational weather prediction in a flexible way.

Sensitivity studies pointed out several results:

1. The impact of the training length (30, 50, or 100 days) is rather negligible (at least for the selected month). Similar results for short lead times have been found by Hagedorn et al. (2008). Results may be different in other seasons when the persistence of weather regimes decreases.

2. A spread rescaling factor of $f_{\text{rescale}} = \frac{2}{3}$, limiting the final predictive PDF, as a fraction of the RMSE of the training data, was found appropriate. Lower values resulted in lower reliability; higher values gave no significant further improvement.

3. About half of the improvement achieved by full calibration can be attributed to a correction of the mean bias.

4. Calibrating the ECMWF-EPS did not produce the same results as calibrating ALADIN-LAEF, although the differences were not very large. However, there is still need for dynamical downscaling as the statistical downscaling alone is not able to guarantee the same forecast quality.

5. Refining the NGR method with respect to minimum CRPS estimation still allows for further improvement. Replacing the average CRPS by a time-decaying average (defined as NGR-TD) includes persistence of model error and results in reduction of CRPS up to 5% relative to NGR.

In the near future, an operational implementation of the calibration method and real-time products for daily weather prediction are planned. Prior to this, some open issues have to be clarified: Do we need a seasonally dependent spread rescaling factor? Do we have to reconsider the influence of the training length if we apply NGR-TD instead of the standard NGR method? What is the optimal weighting factor for NGR-TD? For the two independent verification periods (December 2007 and July 2008) the calibration method provides promising results. Are the conclusions generally applicable? Further studies on different samples have to be carried out in order to guarantee the robustness of the results during all seasons.

A major goal of limited-area ensemble prediction is to provide reliable and sharp forecasts in case of high-impact weather such as heavy precipitation and high wind speeds. Both are rare events requiring long training periods for calibration. The need for a large set of hindcasts is currently under discussion. The future challenge and main focus of statistical postprocessing of ensemble prediction systems will be the development of proper methods addressing the characteristics of both the synoptic phenomena and the ensemble systems.

Acknowledgments. The authors are thankful to Thomas Haiden for providing useful comments during production of the manuscript and to Renate Hagedorn for fruitful and helpful discussions. We thank the anonymous reviewers for their suggestions and comments that led to substantial improvements of the manuscript. Part of the work was funded by OEAD (Austrian Academic Exchange Service Project CN 29/2007).

REFERENCES


