Comparison and Combination of Regional and Global Ensemble Prediction Systems for Probabilistic Predictions of Hub-Height Wind Speed

CONSTANTIN JUNK, STEPHAN SPÄTH, AND LUEDER VON BREMEN
ForWind—Center for Wind Energy Research, University of Oldenburg, Oldenburg, Germany

LUCA DELLE MONACHE
National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 5 February 2015, in final form 22 April 2015)

ABSTRACT
The objective of this paper is to compare probabilistic 100-m wind speed forecasts, which are relevant for wind energy applications, from different regional and global ensemble prediction systems (EPSs) at six measurement towers in central Europe and to evaluate the benefits of combining single-model ensembles into multimodel ensembles. The global 51-member EPS from the European Centre for Medium-Range Weather Forecasts (ECMWF EPS) is compared against the Consortium for Small-Scale Modelling’s (COSMO) limited-area 16-member EPS (COSMO-LEPS) and a regional, high-resolution 20-member EPS centered over Germany (COSMO-DE EPS). The ensemble forecasts are calibrated with univariate (wind speed) ensemble model output statistics (EMOS) and bivariate (wind vector) recursive and adaptive calibration (AUV). The multimodel ensembles are constructed by pooling together raw or best-calibrated ensemble forecasts. An additional postprocessing of these multimodel ensembles with both EMOS and AUV is also tested. The best-performing calibration methodology for ECMWF EPS is AUV, while EMOS performs better than AUV for the calibration of COSMO-DE EPS. COSMO-LEPS has similar skill when calibrated with both EMOS and AUV. The AUV ECMWF EPS outperforms the EMOS COSMO-LEPS and COSMO-DE EPS for deterministic and probabilistic wind speed forecast skill. For most thresholds, ECMWF EPS has a comparable reliability and sharpness but higher discrimination ability. Multimodel ensembles, which are constructed by pooling together the best-calibrated EPSs, improve the skill relative to the AUV ECMWF EPS. An analysis of the error correlation among the EPSs indicates that multimodel ensemble skill can be considerably higher when the error correlation is low.

1. Introduction
The goal of ensemble forecasting is the quantification of flow-dependent forecast uncertainties, which arise from initial-condition errors and inadequacies of the numerical weather prediction model (Leutbecher and Palmer 2008). An ensemble prediction system (EPS) provides a dynamically based estimate of these uncertainties. During the past few decades, different global and regional ensemble prediction systems have been developed with different spatial and temporal resolutions and a range of strategies to simulate initial-condition and model uncertainties.

Centers such as the National Centers for Environmental Prediction (Toth and Kalnay 1993; Wei et al. 2008) or the European Centre for Medium-Range Weather Forecasts (ECMWF; Palmer 1993; Buizza and Palmer 1995; Molteni et al. 1996; Buizza et al. 1999; Leutbecher and Palmer 2008) run global ensemble prediction systems. Several verification studies indicated that ECMWF EPS has the best overall performance relative to other global ensemble prediction systems (Park et al. 2008; Hagedorn et al. 2012).

In addition to global EPSs, in recent years many regional EPSs have been developed (e.g., Du and Tracton 2001; Grimit and Mass 2002; Eckel and Mass 2005; Marsigli et al. 2008).
Regional EPSs are designed to provide probabilistic predictions for a limited geographical area, often at a higher spatial and temporal resolution than global EPSs. Regional EPSs that are operationally running in Europe include the Consortium for Small-Scale Modelling’s (COSMO) limited-area EPS (COSMO-LEPS; Marsigli et al. 2005; Montani et al. 2011); the COSMO short-range EPS (COSMO-SREPS; García-Moya et al. 2011); and a high-resolution, convection-permitting EPS centered over Germany (COSMO-DE EPS; Gebhardt et al. 2011; Peralta et al. 2012).

There exist only a few studies that attempt to compare regional and global EPSs to quantify the possible added value of regional EPSs on global EPSs (Frogner et al. 2006; Bowler et al. 2008; Montani et al. 2011; Wang et al. 2012; Alessandrini et al. 2013; Röpnack et al. 2013). However, most of these studies do not undertake a comparison of EPSs for wind forecasting at heights relevant to wind energy applications but, rather, for meteorological variables such as precipitation, mean sea level pressure, 10-m wind, temperature, and geopotential height. Only Alessandrini et al. (2013) compare COSMO-LEPS and ECMWF EPS at one complex-terrain wind farm in terms of wind and wind power, but they use 10-m wind ensemble forecasts from the EPSs. Thus, the first objective of our study is to compare regional and global EPSs for probabilistic forecasts of 100-m wind, by

- comparing the regional COSMO-LEPS and COSMO-DE EPS with the global ECMWF EPS;
- postprocessing the ensemble forecasts with state-of-the-science calibration methods to remove systematic errors and increase reliability of the EPSs; and
- verifying the EPS performance against 100-m wind speed measurements from six towers in central Europe, which are situated in regions with differing terrain complexity.

Another strategy to reduce forecast errors from ensemble forecasts and increase their reliability is the application of multimodel strategies. A multimodel ensemble is defined here as a combination of different EPSs, in contrast to a “poor man’s ensemble” that combines deterministic forecasts from different models (e.g., Hagedorn et al. 2005). We define a single-model EPS as an EPS whose members are based on a single dynamical-core version of a numerical weather prediction model. With this definition, ECMWF EPS, COSMO-LEPS, and COSMO-DE EPS are single-model ensembles even though initial and lateral boundary conditions for the generation of COSMO-DE EPS members make use of information from different global models.

Several authors (e.g., Park et al. 2008; Johnson and Swinbank 2009; Fraley et al. 2010; Hagedorn et al. 2012; Flowerdew 2012) have investigated whether the calibration of a single- or a multimodel ensemble is preferable for global short- to medium-term EPSs. These studies show that under certain conditions a global multimodel ensemble can outperform the best-calibrated single-model ensemble. Weigel et al. (2008) and Weigel and Bowler (2009) apply a Gaussian stochastic toy model and conclude that the combination of reliable ensemble forecasts can still improve prediction skill compared to the best single model. Applying both calibration and the multimodel ensemble approach together was investigated by Johnson and Swinbank (2009) and Fraley et al. (2010).

Previous studies examined the benefit of multimodel ensembles for variables such as temperature, precipitation, geopotential height, and wind. However, an investigation of such systems, along with the combination of global and regional EPSs, for hub-height wind speed forecasting is missing and this work is an attempt at filling this gap. Thus, the second objective of this study is to further investigate the benefits of the multimodel ensemble approach, by

- combining ensemble forecasts from regional and global EPSs and comparing multimodel ensembles to calibrated single-model ensembles and
- testing different strategies for the construction of multimodel ensembles, and evaluating conditions under which a multimodel ensemble improves over the best single-model ensemble.

The tower measurements used for verification and the ensemble prediction systems are presented in section 2. The calibration methods and the different approaches to combine the EPSs into a multimodel ensemble are described in section 3. The performance of the uncalibrated and calibrated single- and multimodel ensembles is analyzed in section 4. The main findings of this study are discussed in section 5 and conclusions are provided in section 6. The verification methods are described in the appendix.

2. Data
   a. Tower measurements

   The measurements used for the verification of the ensemble forecasts are from the onshore measurement tower at Cabauw, the Netherlands; the onshore measurement towers at Falkenberg, Karlsruhe, and Hamburg, Germany; and the offshore research platforms Fino2 and Fino3, located in the Baltic and North Seas, respectively (Fig. 1). The quality control of the data, which are available at a temporal resolution of 10 min, is...
done following Jiménez et al. (2010). The forecasts are evaluated at 100-m height, which is representative of the hub height of modern wind turbines. The wind measurements at Karlsruhe and *Fino3* are directly available at 100-m height. At Falkenberg, measurements are taken at 98-m height. The measurements are linearly interpolated to 100 m from 80- and 140-m levels at the Cabauw tower and from 50- and 110-m levels at the Hamburg tower. At *Fino2*, wind speed is measured at 102 m and wind direction is measured at 91 m.

The onshore meteorological towers are located in regions with differing site characteristics. At the Karlsruhe tower, the measurements are strongly influenced by the surrounding mountains, the alignment of the Rhine valley, and the vicinity of a forest. Urban and industrial areas strongly influence the measurements at the Hamburg tower. For this reason, we refer to the Hamburg and Karlsruhe sites as high-roughness sites. The area around the Falkenberg tower is characterized by mixed farmland and forest vegetation. The Cabauw tower is surrounded by flat terrain.

As pointed out by Candille and Talagrand (2008), the effect of measurement uncertainty might have a significant impact on absolute verification scores particularly when the forecast errors are small. In this study, however, uncertainties of the measurements are not taken into account.

**b. Ensemble prediction systems**

1) ECMWF EPS

The global ECMWF EPS has been run operationally since 1992 and has been developed gradually (Palmer...
1993; Buizza and Palmer 1995; Molteni et al. 1996; Buizza et al. 1999; Leutbecher and Palmer 2008). Ensemble forecasts from ECMWF EPS consist of 50 perturbed predictions and one unperturbed control forecast for forecast horizons up to 15 days. The current horizontal resolution of ECMWF EPS is T639, which corresponds to a spectral truncation at wavenumber 639 (approximately 30-km horizontal resolution). On 19 November 2013, the vertical resolution of ECMWF EPS was increased from 62 to 91 vertical levels. We use 3-hourly wind vector forecasts at 100-m height from the ECMWF EPS run initialized at 1200 UTC.

Initial-condition uncertainty in ECMWF EPS is spanned by a linear combination of singular vectors (Buizza and Palmer 1995) and ensemble data assimilation–based perturbations (Buizza et al. 2008). Model uncertainty in the free atmosphere is simulated using the stochastically perturbed parameterized tendency scheme (Buizza et al. 1999) and the backscatter scheme (Berner et al. 2009). Note that the distinction between initial-condition uncertainty and model uncertainty is done for simplicity here. Since initial conditions are constructed using model-based data assimilation procedures, initial-condition uncertainty is not clearly separable from model uncertainty (Leutbecher and Palmer 2008).

2) COSMO-LEPS

COSMO-LEPS is a regional EPS of COSMO, which has been run operationally at ECMWF since 2002 (Marsigli et al. 2005). The COSMO-LEPS forecasts used in this study are composed of 16 integrations of the nonhydrostatic COSMO model at a horizontal resolution of 0.0625° × 0.0625° (approximately 7 km) with 40 vertical levels (Montani et al. 2011). The model domain covers central and southern Europe and provides short- and medium-range ensemble forecasts up to 132-h forecast horizons. We use 3-hourly wind forecasts of the 1200 UTC run and linearly interpolate from model levels 37 and 38 up to 100-m height.

The regional COSMO-LEPS can be viewed as a dynamical downscaling of ECMWF EPS. Sixteen representative members from ECMWF EPS are selected to provide initial and lateral boundary conditions to the COSMO runs (Montani et al. 2011). The ensemble reduction algorithm groups 102 members of two subsequent ECMWF EPS runs into 16 clusters and selects one representative member from each cluster to achieve an optimal diversity of initial and boundary conditions from ECMWF EPS. To account for model uncertainties, selected parameters of the parameterization schemes (turbulence, microphysics, convection, and soil model) are perturbed (Marsigli et al. 2014).

3) COSMO-DE EPS

The regional COSMO-DE EPS is a short-range EPS, which has been run operationally at the German Meteorological Service since May 2012 (Gebhardt et al. 2011; Peralta et al. 2012; Bouallégue et al. 2013). It is based on the nonhydrostatic COSMO-DE model (Baldauf et al. 2011), which is a convection-permitting model with a horizontal resolution of 0.025° × 0.025° (approximately 2.8 km), 50 vertical levels, and a computational domain centered over Germany. We use wind vector forecasts of the 1200 UTC run up to a lead time of 27 h and linearly interpolate from model levels 47 and 48 to 100-m height. Although the forecasts are available at an hourly resolution, we only evaluate 3-hourly intraday (3–24 h) forecasts for the comparison with ECMWF EPS and COSMO-LEPS.

The generation of COSMO-DE EPS members comprises the variation of initial and lateral boundary conditions, model physics, and surface moisture. The lateral boundary conditions are provided by four different global models (Gebhardt et al. 2011). To include model uncertainties, different physics configurations of the COSMO-DE model are used (Gebhardt et al. 2011). Details on the initial-condition perturbation scheme can be found in Peralta et al. (2012).

4) HORIZONTAL INTERPOLATION

To obtain local wind ensemble forecasts from ECMWF EPS, COSMO-LEPS, and COSMO-DE EPS, the nearest grid point to the geographical coordinate of each measurement tower is used. Since the horizontal resolutions of the regional COSMO-LEPS and COSMO-DE EPS are considerably higher relative to the global ECMWF EPS, we tested a smoothing of COSMO-DE EPS forecasts by averaging over the nearest 4, 16, and 64 grid points. For most sites, the wind forecast errors turned out to be insensitive to the averaging procedure and neither a clear decrease nor a clear increase in forecast errors could be observed (not shown).

3. Methods

a. Calibration methods

To reduce systematic errors and to obtain reliable ensemble forecasts, a variety of univariate and bivariate postprocessing methods have recently been developed for the calibration of wind ensemble forecasts (Thorarinsdottir and Gneiting 2010; Sloughter et al. 2010; Pinson 2012; Schuhen et al. 2012). Junk et al. (2014) compared state-of-the-science calibration methods for 100-m ensemble forecasts of ECMWF EPS and concluded that the bivariate recursive and adaptive
calibration (AUV; Pinson 2012) outperformed other univariate (for wind speed) and bivariate (for wind vector) calibration methods. However, at a few sites univariate ensemble model output statistics (EMOS; Thorarinsdottir and Gneiting 2010) yielded a similar performance. For this reason, both bivariate AUV and univariate EMOS are selected to calibrate ECMWF EPS, COSMO-LEPS, and COSMO-DE EPS.

1) EMOS

EMOS or nonhomogeneous regression is a parametric regression model that was first introduced by Gneiting et al. (2005). Thorarinsdottir and Gneiting (2010) proposed EMOS based on the truncated normal distribution to take the nonnegativity of wind speed into account. Other variants of EMOS for wind speed are based on the general-extreme-value distribution, the lognormal distribution, or a mixture of distributions (Lerch and Thorarinsdottir 2013; Baran and Lerch 2015). In this study, the approach of Thorarinsdottir and Gneiting (2010) is considered by fitting a truncated normal distribution

$$\mathcal{N}^0(\mu, \sigma^2) = \mathcal{N}^0(a + b\bar{x}, c + dS^2)$$ (1)

around the ensemble members $x_1, \ldots, x_M \in \mathbb{R}$, where $M$ is the number of ensemble members. The quantity $\sigma^2$ is the squared scale parameter, which is a linear function of the ensemble variance $S^2$ with the coefficients $c$ and $d$; $\mu$ is the location parameter modeled as a linear function of the ensemble mean $\bar{x}$ with regression coefficients $a$ and $b$. The EMOS coefficients $a$, $b$, $c$, and $d$ are fit with a training function based on the continuous ranked probability score (CRPS). The minimization of the objective function is done using the Nelder–Mead algorithm as implemented in the GNU R language. The coefficients are estimated for each lead time separately with a sliding training window approach; that is, a forecast is calibrated taking into account only the previous 60 days (60 samples) of forecast–measurement pairs. We tested the sensitivity to the length of the training window and found that 60 days yields the overall best results (not shown). To generate multimodel ensembles with the equal- and implicit-weighting approach (see section 3b), a synthetic ensemble of realizations is recovered by sampling equidistant quantiles from the cumulative distribution function at levels $m/(M + 1)$ for $m = 1, \ldots, M$ following Gneiting et al. (2005).

2) AUV

AUV is a method where the wind components $\mathbf{v} = (u, v)^T$ are recursively estimated via bivariate bias correction (translation model) and variance correction along $\alpha$ and $\nu$ (dilation model) within a bivariate-normal framework $\mathcal{N}_2(\mathbf{u}, \Sigma)$, with the means $\mathbf{u} = (\mu_u, \mu_v)^T$ and the variance–covariance matrix $\Sigma$ (Pinson 2012). While EMOS minimizes an objective function based on the CRPS, bivariate AUV minimizes an objective function within a recursive maximum likelihood framework. This is equivalent to minimizing the logarithmic scoring rule known as ignorance for bivariate probabilistic forecasts (Pinson 2012). The recursiveness of the calibration approach has the advantage that an update of the model coefficients requires only the last set of ensemble forecasts and measurements, which is computationally more efficient than the sliding training window approach. Furthermore, AUV uses exponential forgetting of past forecast–measurement pairs, which results in a smooth adaptation of model coefficients to changes in wind patterns. The speed of adaptivity is controlled by $\lambda \in (0, 1)$, which is chosen to be 0.980 at all sites following Junk et al. (2014). We refer to Pinson (2012) for more details on the AUV method.

b. Combination methods

One objective of this paper is to further explore the benefits of the multimodel ensemble approach by combining ensemble forecasts from regional and global ensemble prediction systems and by comparing multimodel ensembles to the best-calibrated single-model ensemble. In the following, we describe the combination methods used for constructing the multimodel ensembles.

1) IMPLICIT WEIGHTING

A straightforward way of constructing a multimodel ensemble is to simply pool together the ensemble members of the participating single-model ensembles. This can be thought of as an implicit weighting since EPSs with a higher number of members implicitly receive a higher weight (Park et al. 2008; Hagedorn et al. 2012). We later refer to the combination of calibrated single-model ensembles with the implicit-weighting approach as impl-cal. We also tested the combination of raw single-model ensembles with implicit weighting (impl-raw). Since the forecast skill of the latter is significantly lower compared to impl-cal, those results are not shown in section 4.

2) EQUAL WEIGHTING

The equal-weighting approach requires that each EPS receives the same weight. This is achieved by randomly sampling a subset of members at each forecast cycle where the number of members in the subset is equal to the lowest number of members from the available EPSs. For instance, to combine COSMO-LEPS and ECMWF EPS, 16 members are randomly drawn from ECMWF EPS forecasts at each forecast cycle and then the
16-member ECMWF EPS and 16-member COSMO-LEPS are pooled together. In contrast to the implicit-weighting approach, the equal weighting reduces the dominance of ECMWF EPS. We later refer to the combination of calibrated single-model ensembles with the equal-weighting approach as equal-cal.

3) OPTIMIZED WEIGHTING

Optimizing the EPS-dependent weights with statistical postprocessing techniques is a more sophisticated approach for constructing a multimodel ensemble. The weight optimization can be done with a variant of the EMOS regression:

\[ N \sim N(0, \mu, \sigma^2) = \sum_{i=1}^{K} b_i \mathbf{x}_i + c + d_1 S_1^2 + \ldots + d_K S_K^2, \]

where \( K \) is the number of EPSs that participate in the multimodel ensemble; \( \mathbf{x}_i \) are the ensemble means; and \( S_1^2, \ldots, S_K^2 \) are the ensemble variances of the respective EPS forecasts. The quantities \( a; b_1, \ldots, b_K; c; \) and \( d_1, \ldots, d_K \) are the regression parameters.

However, results (not shown) indicate that the forecast skill is higher when applying EMOS following Eq. (1), where \( \mathbf{x} \) and \( S^2 \) are the ensemble mean and ensemble variance calculated over all ensemble members of the multimodel ensemble, rather than following Eq. (2). Thus, applying EMOS according to Eq. (1) corresponds to the generation of the implicit-weighting ensemble and the subsequent postprocessing of this ensemble with EMOS, without estimating EPS-dependent regression parameters. The motivation for this approach is that the implicit-weighting ensemble might be further improved by postprocessing. We later refer to this approach as EMOS-raw if EMOS is applied to impl-raw and as EMOS-cal if EMOS is applied to impl-cal. An additional postprocessing of the implicit-weighting ensemble with AUV was also tested (later referred to as AUV-raw and AUV-cal).

4. Results

First, we compare raw and calibrated wind speed forecasts from COSMO-DE EPS, COSMO-LEPS, and ECMWF EPS (section 4a), and then we evaluate possible benefits from combining single-model ensembles into a multimodel ensemble (section 4b). The evaluation period is from August 2013 to June 2014, excluding October 2013 when most of the COSMO-DE EPS forecasts are missing. The forecast evaluation does not consider the different schedules for the forecast dissemination adopted for each EPS considered in this study since this is not related to forecast skill. The verification methods used to evaluate the ensemble forecasts are described in the appendix. The abbreviations and corresponding descriptions of the single- and multimodel ensembles evaluated in this study are presented in Table 1.

a. Comparison of ensemble prediction systems

To compare the forecast accuracy of the single-model EPSs, the wind speed forecasts are evaluated in terms of the ensemble median MAE and the CRPS for lead times of 3–24 h (Figs. 2 and 3). At all sites, the ensemble median MAE of raw ECMWF EPS is lower compared to COSMO-LEPS and COSMO-DE EPS (Fig. 2).
higher skill of the raw ECMWF EPS is statistically significant.

A comparison of the quality of the probabilistic predictions of the raw EPSs as measured with the CRPS shows different results (Fig. 3). At Hamburg, the raw EPSs have comparable accuracy and at Karlsruhe and Fino3 raw COSMO-LEPS and ECMWF EPS perform equally well, while COSMO-DE EPS performs worse. At Fino2, raw COSMO-LEPS outperforms raw ECMWF EPS. The binned spread–skill diagrams provide one explanation for this result (Fig. 4). The raw wind speed forecasts from all EPSs are underdispersive since the curves are above the dashed diagonal. However, the raw COSMO-LEPS and COSMO-DE EPS tend to be less underdispersive particularly in the highest spread class compared to ECMWF EPS. At Fino3, raw COSMO-LEPS appears to have the best statistical consistency. In Fig. 4, the presentation is restricted to Fino3 and Hamburg since a discussion of statistical consistency at the remaining sites leads to similar conclusions.

To reduce the systematic errors of the forecasts and to increase their reliability, the ensemble forecasts are calibrated with univariate EMOS and bivariate AUV. The largest MAE decrease can be observed at the high-roughness sites Hamburg and Karlsruhe and the onshore site Falkenberg (Fig. 2), while the smallest decrease is obtained at the offshore site Fino3, where systematic errors of the raw ensemble median are small [for details, see Junk et al. (2014)]. The CRPS values are significantly decreased by calibration at all sites, with largest improvements (up to 30%) at the Hamburg site due to the substantial biases of the raw EPSs (Fig. 3).

Bivariate AUV of ECMWF EPS forecasts yields either a similar or lower MAE and CRPS than EMOS (Figs. 2 and 3). The calibration of COSMO-LEPS wind forecasts produces mixed results regarding the superiority of one calibration method over the other since AUV is preferable at Hamburg and Fino3, while EMOS is best at the remaining sites. The COSMO-DE EPS forecasts, however, have a lower CRPS at all sites when the forecasts are calibrated with EMOS. This result is discussed in more detail in section 5.

The ECMWF EPS calibration leads to good statistical consistency with only slight underdispersion for
the low-spread classes (Fig. 4). The AUV of COSMO-LEPS and COSMO-DE EPS forecasts, however, produces overdispersive ensemble forecasts for the high-spread class, while EMOS forecasts are slightly underdispersive. The overdispersion of AUV forecasts is most pronounced for COSMO-DE EPS, which confirms that EMOS performs better than AUV on COSMO-DE EPS.

To assess specific probabilistic attributes of the EPS forecasts in more detail, Fig. 5 shows reliability diagrams and Fig. 6 presents relative operating characteristics (ROC) diagrams with the median of the measured wind speed chosen as the event threshold. The presentation of the reliability and ROC diagrams is again restricted to Fino3 and Hamburg for the same reasons as for the binned spread–skill diagrams. The raw ECMWF EPS is overconfident for probability classes smaller than 0.5 at Fino3 and for probability classes larger than 0.5 at Hamburg. The calibration of ECMWF EPS forecasts strongly improves the reliability. The overconfidence of the raw ensemble is also noticeable for COSMO-LEPS and COSMO-DE EPS at Hamburg. At Fino3, however, raw COSMO-LEPS and COSMO-DE EPS forecasts have a slightly better reliability than raw ECMWF EPS forecasts. A comparison of the best-calibrated ensemble forecasts indicates a similar reliability of all ensemble prediction systems. The ROC diagrams and the ROC skill score (ROCSS), however, show that AUV ECMWF EPS has a slightly better discrimination ability for the median threshold compared to the best-calibrated COSMO-LEPS and COSMO-DE EPS (Fig. 6). We also analyzed the reliability and ROC diagrams for other event thresholds. The above results are confirmed for thresholds lower than the 90th percentile of measured wind speed. For the extreme thresholds such as the 90th and 95th percentiles, however, AUV ECMWF EPS appears to lose its advantage in terms of discrimination compared to the best-calibrated COSMO-LEPS and COSMO-DE EPS (not shown).

Figure 7 shows a comparison of raw and calibrated COSMO-LEPS and ECMWF EPS for forecast horizons up to 120 h. The comparison of the raw EPSs indicates that COSMO-LEPS can compete with the ECMWF EPS for short-range forecast horizons (except at the Falkenberg site), which was also shown in Fig. 3. For forecast days 2–5, however, the raw ECMWF EPS has a higher forecast skill than raw COSMO-LEPS (Fig. 7). After the calibration, AUV ECMWF EPS outperforms the best-calibrated COSMO-LEPS even in the short range.
When ensemble forecasts are compared with each other, it is important to note that the ensemble size affects the forecast skill. The question arises, which portion of the skill of an ensemble with more members is attributed to the increased ensemble size. Figure 8 shows the percentage of the CRPS decrease relative to the full 51-member AUV ECMWF EPS when the number of members is decreased by randomly selecting members from the full ensemble at each forecast cycle. For short-range forecast horizons of 3–24 h, a 20-member ECMWF EPS has a CRPS approximately 0.5%–1.5% lower compared to the CRPS of a 51-member ECMWF EPS. Since the AUV COSMO-LEPS has CRPS values approximately 10% (Hamburg), 5% (Fino3), and 4% (Karlsruhe) higher relative to the AUV ECMWF EPS (Fig. 3), the larger ensemble size of the AUV ECMWF EPS explains only a minor portion of its higher skill.

For forecast horizons larger than 24 h, the CRPS is more sensitive to the reduction of the ensemble size (Fig. 8), which was also shown by Marsigli et al. (2014). A 20-member ECMWF EPS, for instance, has a CRP skill score (CRPSS) of 2%–3% lower than the 51-member ECMWF EPS for 27–120-h forecasts. The effect of ensemble size on verification scores is also discussed by Ferro et al. (2008). According to Eq. (26) in Ferro et al. (2008), an adjustment factor of about 3% CRPS reduction has to be applied to a 20-member ensemble compared to a 51-member ensemble, which is in line with the discussion of the analysis reported in Fig. 8 [see also the discussion in Hagedorn et al. (2012)].

**b. Evaluation of multimodel ensembles**

In this section, the possibility of increasing the forecast accuracy relative to the best-calibrated single-model ensemble (AUV ECMWF EPS) is explored by constructing a multimodel ensemble based on the regional and global EPSs.

Figure 9 presents wind speed MAE skill score (MAES) and CRPSS values of the multimodel ensembles relative to the 51-member AUV ECMWF EPS for forecast horizons of 3–24 h. The 87-member implicit-weighting multimodel ensemble, which is constructed by pooling together AUV ECMWF EPS, EMOS COSMO-LEPS, and EMOS COSMO-DE EPS, yields overall the highest MAES and CRPSS values. The impl-cal multimodel ensemble improves the MAE by 5% at Karlsruhe
and by 4% at Fino2. Improvements are slightly lower at the remaining sites. Its CRPS values are also improved at all sites. The equal-weighting approach applied to the three calibrated global and regional EPSs (equal-cal) performs worse than the implicit-weighting approach. This implies that assigning a higher weight to the most skillful ensemble forecast (here AUV ECMWF EPS) is preferable for the construction of the multimodel ensemble. Note again, that the higher number of ensemble members of impl-cal also contributes to the higher CRPSS compared to the 48-member equal-cal ensemble. Equation (26) in Ferro et al. (2008) gives an adjustment factor of approximately 0.9%, a CRPS increase that has to be applied to an 87-member ensemble compared to a 48-member ensemble from a theoretical point of view. This means for instance that the ~6% CRPSS of the impl-cal ensemble relative to AUV ECMWF EPS at the Karlsruhe site (Fig. 9, bottom) is reduced to ~5.2%.

An additional calibration of impl-raw and impl-cal with EMOS following Eq. (1) or AUV yields an overall lower skill than is found with the implicit-weighting approach (Fig. 9). At some sites, however, the AUV-raw is slightly more skillful than the implicit-weighting multimodel ensemble (e.g., Hamburg and Fino3). The EMOS-raw has a higher level of skill than the implicit-weighting multimodel ensemble at the Karlsruhe site. Furthermore, it is clear that AUV-raw performs significantly better than EMOS-raw at Hamburg, which might be attributed to the bidimensional bias-correction model employed in AUV. Since the large bias of impl-raw at Hamburg is strongly dependent on the wind direction (not shown), a bidimensional bias correction might be more beneficial than the univariate (wind speed) bias correction employed in EMOS [see also discussion in Junk et al. (2014)]. As has been mentioned in section 3b, we also tested the optimization of EPS-dependent weights with EMOS following Eq. (2). However, this approach yields lower forecast skill than EMOS following Eq. (1) (not shown).

An assessment of the statistical consistency of the ensemble spread indicates that the equal- and implicit-weighting multimodel ensembles have a high statistical
consistency, while the multimodel ensembles additionally calibrated with EMOS are underdispersive for low-spread classes (Fig. 10, left). While EMOS-raw and EMOS-cal appear to be similar in reliability and sharpness for the chosen median threshold (Fig. 10, middle), the ROC diagram and ROCSS values indicate a slightly higher discrimination ability of impl-cal compared to the other multimodel ensembles, though it is statistically not significant.

To quantify the contribution of each EPS to the improvements achieved with the multimodel ensembles, we analyze different EPS combinations using the implicit-weighting combination method. As shown in Fig. 11, a multimodel ensemble based on AUV ECMWF EPS and EMOS COSMO-DE EPS performs as well as the multimodel ensemble based on all three EPSs, while a multimodel ensemble based on AUV ECMWF EPS and EMOS COSMO-LEPS performs slightly worse. The AUV ECMWF EPS is the main contributor, as evident by the negative CRPSS of the CL–DE system, which is formed by COSMO-LEPS and COSMO-DE EPS. Also, these results imply that COSMO-DE EPS contributes slightly more to the multimodel ensemble skill than COSMO-LEPS. One reason might be the dependence of COSMO-LEPS and ECMWF EPS since the initial and boundary conditions of COSMO-LEPS are based on ECMWF EPS members, which causes common biases and information redundancy. The positive impact of the boundary conditions as provided by four different global models in COSMO-DE EPS could be even larger for lead times beyond hour 27 (not available for this analysis). As discussed by Hagedorn et al. (2005, 2012) and Solazzo et al. (2013), lack of independence undermines the performance of a multimodel ensemble. Additional contributing factors might include coarser horizontal and vertical resolution of COSMO-LEPS and its smaller number of members.

To test the dependence of forecast errors between the EPSs, we analyze forecast-error correlations and relate the amount of correlation to the improvement, which can be achieved with a multimodel ensemble over the best single-model ensemble. We calculate the correlation between the CRPS time series of AUV ECMWF EPS and EMOS COSMO-DE EPS or EMOS COSMO-LEPS for each month of the evaluation period for 27-h forecasts (Fig. 12). The error correlation is strongly dependent on the time of the year and ranges between 0.0 and 0.8 at Fino3 and 0.0 and 0.9 at Karlsruhe. On average, the CRPS of COSMO-LEPS has
slightly stronger correlations with ECMWF EPS than COSMO-DE EPS with ECMWF EPS. This can be explained by the fact that COSMO-LEPS initial and boundary conditions are based on ECMWF EPS members, while COSMO-DE EPS uses information from different global models.

To relate the error correlation to improvements achieved with the multimodel ensemble, Fig. 12 additionally shows the CRPSS of the implicit-weighting multimodel ensemble relative to the AUV ECMWF EPS. At Karlsruhe and Fino3, the CRPSS and the error correlation are anticorrelated. The CRPSS reaches almost 12% for a relatively low error correlation at the Karlsruhe site in December 2013, while a higher correlation causes the CRPSS to drop to close to 0%. Results at other sites and forecast horizons lead to similar conclusions (not shown).

5. Discussion

We have shown that the bivariate AUV method yields a higher increase in ECMWF EPS forecast performance than univariate EMOS [also see Junk et al. (2014)], while calibration of COSMO-LEPS yields comparable results with EMOS and AUV. The COSMO-DE EPS forecasts, however, appear to be more accurate and statistically consistent after calibration with univariate EMOS than with bivariate AUV. This result might be explained by the tendency of COSMO-DE EPS forecasts to cluster with the respective global model (not shown),
which are used to generate its initial and boundary conditions (Peralta et al. 2012), leading to multimodal ensemble forecast distributions. The EMOS method fits truncated normal distributions around the wind speed ensemble forecast with the parameters of the distribution being estimated from the empirical ensemble mean and ensemble variance over the training period. Thus, EMOS corrects the uncalibrated ensemble toward a truncated normal distribution. In contrast, the AUV method requires the \((u, v)\)-ensemble forecasts to be distributed bivariate normal. Applying AUV to COSMO-DE EPS forecasts therefore means that the translation and dilation parameters are estimated within a bivariate-normal framework despite the nonnormally distributed ensemble forecasts. This might explain the poorer performance of AUV relative to EMOS when applied to COSMO-DE EPS wind ensemble forecasts.

The EPS comparison indicates that calibrated wind speed forecasts from the global ECMWF EPS are more skillful than calibrated forecasts from the regional COSMO-LEPS and COSMO-DE EPS. Global and regional EPSs were also compared by Bowler et al. (2008), Montani et al. (2011), Wang et al. (2012), and Alessandrini et al. (2013). Bowler et al. (2008) compared the short-range, Met Office Global and Regional EPS (MOGREPS) with the ECMWF EPS. In contrast to our study, they found that the regional MOGREPS compares favorably to ECMWF EPS for weather parameters such as temperature, precipitation, and wind speed. However, Bowler et al. (2008) note that ECMWF EPS forecasts are transferred to the Met Office on a coarse \(1.5\times1.5^\circ\) grid, which may decrease the scores of that system. Montani et al. (2011) compared the raw ECMWF EPS with the raw COSMO-LEPS for upper-air variables such as 700-hPa geopotential height and 850-hPa temperature and concluded—similar to our study—that COSMO-LEPS has overall larger forecast errors than ECMWF EPS. However, ECMWF analyses were chosen as verification dataset, which might favor the scores provided by ECMWF EPS.

Alessandrini et al. (2013) found that COSMO-LEPS forecasts (calibrated with a variance inflation approach) outperform calibrated ECMWF EPS forecasts for wind and wind power forecasting. Although this result appears to be in contrast with what is found in this study, this difference can be explained by the fact that Alessandrini et al. (2013) compared the ensemble forecasts at a single wind farm in a complex-terrain area. The higher resolution of COSMO-LEPS might have a larger positive impact on the scores at this complex-terrain site than at locations with less complex terrain, as used in this study. Furthermore, they use wind ensemble forecasts at the 10-m model level to forecast wind power and hub-height wind speed, which reduces the comparability to our study as we extracted upper model level forecasts at 100-m height.

Wang et al. (2012) compared raw forecasts from the central limited-area ensemble forecasting (LAEF) system (ALADIN-LAEF; Wang et al. 2011) with the raw ECMWF EPS for surface and upper-air weather variables. They conclude that ALADIN-LAEF has comparable or slightly lower skill when compared to ECMWF EPS for upper-air variables, but outperforms ECMWF EPS for precipitation, mean sea level pressure, and 10-m wind. To compare the results to our study, one has to note that the raw ECMWF EPS forecasts of 10-m wind speed, which are analyzed in Wang et al. (2011), are strongly biased (i.e., ECMWF EPS overestimates observed wind speed). A removal of systematic errors by ensemble calibration might lead to a different result since we have shown that the CRPS of the uncalibrated ECMWF EPS can be similar relative to the uncalibrated COSMO-LEPS at a few sites, whereas the calibrated ECMWF EPS outperforms the calibrated COSMO-LEPS.

As the EPS intercomparison in this study is limited to a few sites, future studies may compare the EPSs at a larger number of measurement sites to generalize the results. Furthermore, high-resolution, convection-permitting systems such as COSMO-DE EPS might show their strengths compared to coarser-resolution EPSs particularly for meteorological variables, whose predictability is strongly connected to an accurate representation of the terrain (e.g., precipitation). For instance, the predictability of vertical temperature profiles by COSMO-DE EPS was analyzed by Röpnick et al. (2013). They showed that temperature profiles predicted by COSMO-DE
EPS are more consistent with measured profiles than those of the short-range COSMO-SREPS. Thus, additional studies are needed to analyze the strengths and weaknesses of high-resolution systems, and to understand for which meteorological variables a high-resolution EPS adds value to coarser-resolution EPSs. Furthermore, the value of high-resolution systems may be better exploited by postprocessing techniques that try to take advantage of the ability of those systems to solve finer-scale (spatial and temporal) structures.

We showed that a multimodel ensemble constructed by pooling together the calibrated single-model ensemble forecasts with the implicit-weighting approach outperforms the best single-model ensemble (AUV ECMWF EPS). Only a portion of the improvement is attributed to the larger number of ensemble members in the implicit-weighting ensemble compared to AUV ECMWF EPS. This confirms the conclusions of Weigel et al. (2008) and Weigel and Bowler (2009), where the success of multimodel ensembles has been evaluated with a toy model, that the combination of reliable ensemble forecasts can still improve prediction skill. Similar results were obtained by Johnson and Swinbank (2009), where multimodel ensembles were compared to bias-corrected (but not variance adjusted) single-model ensembles for 2-m temperature, mean sea level pressure, and 500-hPa geopotential height. They found that the multimodel ensembles, which are constructed by a bias correction of each individual ensemble, an estimation of model-dependent weights, and variance adjustment, yield largest improvements for 2-m temperature as a result of a higher degree of forecast-error similarity for this weather variable. Furthermore, Johnson and Swinbank could slightly improve the ensemble skill by adjusting the EPS-dependent weights compared to the simple pooling of bias-corrected ensemble members. This is in contrast to our study, where a slightly lower skill is achieved by an optimization of EPS-dependent ENSEMBLES.
weights based on their historical performance with the EMOS method. Fraley et al. (2010) applied Bayesian model averaging (BMA) to optimize EPS-dependent weights and showed that BMA-calibrated multimodel ensembles outperform calibrated single-model ensembles for surface temperature and precipitation forecasts. Thus, a thorough comparison of different methods for constructing multimodel ensembles would be an interesting topic of future research to further investigate possible benefits of the multimodel ensemble approach.

6. Conclusions

In this study, we compared raw and calibrated 100-m wind speed forecasts from regional and global ensemble prediction systems (EPSs) at six measurement towers in central Europe and investigated the benefits of combining global and regional ensemble prediction systems by testing different approaches for the construction of multimodel ensembles. The investigated ensemble prediction systems are the global 51-member EPS from the European Centre for Medium-Range Weather Forecasts (ECMWF), the regional 16-member Consortium for Small-Scale Modelling (COSMO) limited-area EPS (COSMO-LEPS), and the high-resolution 20-member EPS centered over Germany (COSMO-DE EPS).

Calibrated forecasts of ECMWF EPS outperform calibrated COSMO-LEPS and COSMO-DE EPS in terms of the wind speed mean absolute error and continuous ranked probability score (CRPS). A portion of the higher skill of ECMWF EPS can be attributed to the larger number of ensemble members compared to COSMO-LEPS and COSMO-DE EPS. While probabilistic attributes of the calibrated forecasts such as sharpness and reliability are of similar performance, calibrated ECMWF EPS forecasts have a higher discrimination ability, as indicated by the relative operating characteristics for thresholds lower than the 90th percentile of the measured wind speed. Univariate ensemble model output statistics (EMOS) and a bivariate recursive and adaptive calibration (AUV) method are used to postprocess the ensemble forecasts. The forecast skill of the calibrated ECMWF EPS is higher after postprocessing with AUV than with EMOS, while the calibration of COSMO-LEPS yields similar skill with both AUV and EMOS. However, COSMO-DE EPS
benefits most from the calibration with univariate EMOS both in terms of forecast skill and the statistical consistency of the ensemble spread.

Additionally, we combined raw and calibrated single-model ensembles from the global and regional EPSs to multimodel ensembles. By simply pooling together (implicit-weighting approach) the best-calibrated ECMWF EPS, COSMO-LEPS, and COSMO-DE EPS forecasts, the multimodel ensemble outperforms the best-calibrated single-model ensemble (AUV ECMWF EPS). An analysis of the CRPS error correlation between the ensemble prediction systems indicated that the improvements achieved with the multimodel ensemble approach are higher when the error correlations are low (i.e., when the model errors are independent of each other). We showed that improvements in the CRPS relative to the AUV ECMWF EPS reach up to 12% in case of low error correlations. Future studies should be extended to a wider range of measurement sites to confirm the generality of the results obtained for 100-m wind speed.

Acknowledgments. The work presented in this study is funded by the national research project Baltic I (FKZ 0325215A, Federal Ministry for Environment, Nature Conservation and Nuclear Safety) and the Ministry for Education, Science, and Culture of Lower Saxony. The authors thank the Karlsruhe Institute of Technology (KIT), the Royal Netherlands Meteorological Institute (KNMI), the Lindenberg Meteorological Observatory–Richard Aßmann Observatory (German Meteorological Service), and the Meteorological Institute (MI) of the University of Hamburg for providing the wind measurements of the onshore meteorological measurement masts at Karlsruhe, Cabauw, Falkenberg, and Hamburg. The Project Management Jülich (PTJ) and the Federal Maritime and Hydrographic Agency (BSH) are acknowledged for providing measurements of the offshore research platforms Fino2 and Fino3. Numerical weather prediction data are provided by ECMWF and the German Meteorological Service. The authors are grateful to Stefano Alessandrini for general discussions regarding the comparison of ensemble prediction systems. Furthermore, we thank three reviewers and the editor for their valuable comments and suggestions.

APPENDIX

Verification Methods

To assess the deterministic forecast error of the ensemble median \( \hat{x} \in \mathbb{R} \), we calculate the mean absolute error,

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{x}_i - y_i|,
\]

over \( N \) forecast–measurement pairs in the verification period, where \( y \in \mathbb{R} \) is the measurement (i.e., verifying wind speed value). The continuous ranked probability score (CRPS) is a proper scoring rule for the evaluation of ensemble forecasts (Hersbach 2000). For an EPS consisting of \( M \) members, it can be evaluated as
where $F_{\text{ens}}$ is the ensemble forecast with ensemble members $x_1, \ldots, x_M \in \mathbb{R}$ (Gneiting and Raftery 2007). Since the EMOS technique yields predictive distributions instead of discrete ensemble forecasts, the closed-form expression of the CRPS for a truncated normal distribution is used for EMOS forecasts (Thorarinsdottir and Gneiting 2010). Over $N$ forecast–measurement pairs, the CRPS values are given by

$$\text{CRPS}(F_{\text{ens}}, y) = \frac{1}{M} \sum_{m=1}^{M} |x_m - y|$$

$$- \frac{1}{2M^2} \sum_{n=1}^{M} \sum_{m=1}^{M} |x_n - x_m|,$$  \hspace{1cm} (A2)

Since we are often interested in improvements that can be achieved relative to a reference forecast, we also use the skill scores of the CRPS and the MAE that are defined as

$$\text{CRPSS} = 1 - \frac{\text{CRPS}}{\text{CRPS}_{\text{ref}}},$$

where $\text{CRPS}_{\text{ref}}$ and $\text{MAE}_{\text{ref}}$ are computed with the reference ensemble forecast.

The statistical significance of the scoring rules is tested with the bootstrap resampling technique (Efron 1979; Bröcker and Smith 2007; Pinson et al. 2010). For instance, the wind speed forecasts for lead times of 3–24 h from each forecast run are pooled together over the evaluation period and the resulting wind speed time series is used as input for the bootstrap. The bootstrap is repeated 500 times and 50% and 95% confidence intervals as well as the mean are calculated from the resulting bootstrap.

While the CRPS is a scoring rule that is computed across the entire variable range, the reliability diagram and relative operating characteristics (ROC) diagram are graphical displays of forecast verification of a binary event for a given threshold. The reliability diagram includes the reliability curve and commonly the sharpness...
histogram. The former evaluates the reliability (also called calibration) of a binary event at an event threshold and plots the observed relative frequency of an event against the forecast probability for any given level of probability. An ensemble forecast is reliable if the reliability curve lies along the diagonal. Sharpness is an attribute of the forecast only, and is displayed in the sharpness histogram (relative frequency of use of each probability level). The consistency bars around the diagonal in the reliability diagram are calculated with a quantile function for a binomial distribution (Bröcker and Smith 2007; Pinson et al. 2010).

The ROC diagram is a discrimination-based display of forecast verification and evaluates the ability of ensemble forecasts to discriminate between the occurrence and nonoccurrence of the event. It is generated by plotting the false alarm rate against the hit rate. The discrimination can be summarized using the area under the ROC curve as a single scalar value, which is $A = 1$ for a perfect forecast and $A = 0.5$ for the sample climatology (Wilks 2011). The ROC skill score (ROCSS) is then defined as

$$\text{ROCSS} = \frac{A - 0.5}{1 - 0.5} = 2A - 1. \quad (A5)$$

To assess the statistical consistency of the ensemble spread, the spread–skill relationship is analyzed by comparing the ensemble mean root-mean-square error (RMSE) to the square root of the average ensemble variance (e.g., Wilks 2011; Fortin et al. 2014). The so-called binned spread–skill diagram is chosen in this study to compare the square root of the average ensemble variance and RMSE over small class intervals of spread (e.g., Van den Dool 1989).

REFERENCES


——, S. Tascu, F. Weidle, and K. Schmeisser, 2012: Evaluation of the added value of regional ensemble forecasts on global


OCTOBER 2015 JUNKETAL. 1253