AGU PUBLICATIONS

Geophysical Research Letters

RESEARCH LETTER

10.1002/2014GL062472

Key Points:

- Evolution of raw ensemble forecast skill
- Future benefits from statistical postprocessing
- Global distribution of forecast skill development

Supporting Information:

- Readme
- Figure S1
- Figure S2

Correspondence to:
S. Hemri,
stephan.hemri@h-its.org

Citation:
Hemri, S., M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden (2014), Trends in the predictive performance of raw ensemble weather forecasts, *Geophys. Res. Lett.*, *41*, 9197–9205, doi:10.1002/2014GL062472.

Received 7 NOV 2014
Accepted 5 DEC 2014
Accepted article online 10 DEC 2014
Published online 29 DEC 2014
Corrected 30 JUL 2015

This article was corrected on 30 JUL 2015. See the end of the full text for details.

Trends in the predictive performance of raw ensemble weather forecasts

S. Hemri[1,2], M. Scheuerer[3], F. Pappenberger[2,4,5], K. Bogner[2,6], and T. Haiden[2]

[1] Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany, [2] European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK, [3] Physical Sciences Division, NOAA/ESRL, Boulder, Colorado, USA, [4] College of Hydrology and Water Resources, Hohai University, Nanjing, China, [5] School of Geographical Sciences, University of Bristol, Bristol, UK, [6] Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), Birmensdorf, Switzerland

Abstract This study applies statistical postprocessing to ensemble forecasts of near-surface temperature, 24 h precipitation totals, and near-surface wind speed from the global model of the European Centre for Medium-Range Weather Forecasts (ECMWF). The main objective is to evaluate the evolution of the difference in skill between the raw ensemble and the postprocessed forecasts. Reliability and sharpness, and hence skill, of the former is expected to improve over time. Thus, the gain by postprocessing is expected to decrease. Based on ECMWF forecasts from January 2002 to March 2014 and corresponding observations from globally distributed stations, we generate postprocessed forecasts by ensemble model output statistics (EMOS) for each station and variable. Given the higher average skill of the postprocessed forecasts, we analyze the evolution of the difference in skill between raw ensemble and EMOS. This skill gap remains almost constant over time indicating that postprocessing will keep adding skill in the foreseeable future.

1. Introduction

Over the last two decades the paradigm in weather forecasting has shifted from being deterministic to probabilistic [see, e.g., *Palmer*, 2000; *Hamill et al.*, 2000]. Accordingly, numerical weather prediction (NWP) models have been run increasingly as ensemble forecasting systems. The goal of such ensemble forecasts is to approximate the forecast probability distribution by a finite sample of scenarios [*Leith*, 1974]. Global ensemble forecast systems, like the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble, are prone to probabilistic biases and are therefore not reliable. They particularly tend to be underdispersive for surface weather parameters [*Bougeault et al.*, 2010; *Park et al.*, 2008]. In order to correct for forecast underdispersion and bias in NWP ensembles different statistical postprocessing methods have been developed, of which ensemble model output statistics (EMOS) [*Gneiting et al.*, 2005] is among the most widely applied. EMOS yields a parametric forecast distribution by linking its parameters to ensemble statistics. Due to its simplicity and low computational cost, we focus on EMOS for this study.

The ECMWF ensemble is under continuous development, and hence, its forecast skill improves over time [*Buizza et al.*, 1998, 2007; *Richardson et al.*, 2013; *Haiden et al.*, 2014]. Parts of these improvements may be due to a reduction of probabilistic biases. From this we deduce the following hypothesis: As the raw forecasts continuously improve, it is hypothesized that the gap in skill between raw ensemble and postprocessed forecasts narrows, because systematic errors typically captured by postprocessing are reduced by those improvements. In other words, probabilistic biases, which can be reduced by statistical postprocessing methods, decrease over time. Assuming that the raw ensemble forecasts continue to improve in the future, the gap in skill may eventually be closed when the raw ensemble forecasts become reliable and unbiased.

In this work we analyze the evolution of the global performance of the operational ECMWF raw ensemble and the corresponding postprocessed EMOS forecasts for 2 m temperature (T2M), 24 h precipitation (PPT24), and 10 m wind speed (V10). We verify the forecasts against globally distributed surface synoptic observations (SYNOP) data over a period of about 10 years. We first evaluate the monthly average skill in terms of continuous ranked probability score (CRPS) [*Matheson and Winkler*, 1976] for both the raw and the EMOS forecasts. In order to assess the extent to which the results depend on the choice of the postprocessing method, Bayesian model averaging (BMA) [*Raftery et al.*, 2005; *Fraley et al.*, 2010] is additionally applied to the T2M raw ensemble forecasts. Note that the CRPS is a negatively oriented skill measure (i.e., the lower the value the higher the skill). As the CRPS assesses both reliability and sharpness and is a proper score [*Gneiting and Raftery*, 2007], we rely on it for model fitting and verification throughout this study. Note that skill and reliability are linked in that given constant sharpness an improvement in reliability leads to an improvement in skill and vice versa. We finally analyze the evolution of the gap in CRPS between raw ensemble and postprocessed forecasts.

After presenting the data set in section 2, we summarize the methods for postprocessing and for the assessment of the global skill evolution in section 3. In section 4 the results are shown. This is followed by a discussion in section 5 along with some concluding remarks. Any analyses have been performed using the statistical software R [*R Development Core Team*, 2013].

2. Data

We have selected a large number of SYNOP stations for verification to perform a study which covers the entire globe as ECMWF forecasts are issued on the global domain. SYNOP stations with suspicious or too many missing data are removed from the data set following the approach used by *Pinson and Hagedorn* [2012] with some modifications. The main criterion for removal of a station from the data set for a particular variable is the percentage of data points that are equal to the previous 10 data points. If this exceeds 20% a station is considered to be unreliable. In case of PPT24 and V10 this is applied only for nonzero values. Additionally, T2M stations with values outside the range [-70°C, 60°C], PPT24 stations with values outside [0 mm, 1826 mm] and V10 stations with values outside [0 m/s, 113.2 m/s] are removed. Those ranges extend from the lowest to the highest measurements recorded on Earth. With these removal criteria, 4160 out of 4586, 2917 out of 2956, and 4387 out of 4509 stations are considered to be of reasonable quality for T2M, PPT24, and V10, respectively.

In this study we focus on observations for 12 UTC and ECMWF ensemble forecasts initialized at 12 UTC with lead times of 3, 6, and 10 days. This selection of forecast ranges covers the transition from higher predictability at lead time 3 days to considerably lower predictability at 10 days. The raw ensemble consists of the ECMWF high-resolution (HRES), the corresponding 50-member ensemble (ENS) and the control (CTRL) runs. During the time period considered (1 January 2002 to 20 March 2014) the forecast model, which is the same for ENS, HRES, and CTRL, has undergone several upgrades. Additionally, the ENS has been reconfigured several times over that period. The ECMWF ensemble system is described in detail in *Molteni et al.* [1996] and *Buizza et al.* [2007]. Since for the postprocessed forecasts some data has to be put aside for training (see section 3), the verification periods for the following analyses are somewhat shorter and extend from January 2004 to March 2014 for T2M and V10, and from January 2007 to March 2014 for PPT24.

3. Methods

3.1. Postprocessing Using EMOS

Post-processing using EMOS converts a raw ensemble of discrete forecasts into a probability distribution. Let y be the variable to be forecast (here: T2M, PPT24, or V10) and let $\boldsymbol{f} = (f_1, f_2, \dots, f_K)^T$ be the vector of the K member raw ensemble forecasts (here: HRES, ENS, and CTRL). Then the EMOS predictive density can be written as

$$y|\boldsymbol{f} \sim g(m, \sigma), \tag{1}$$

where $g(\cdot)$ is a parametric density function with location and scale parameters m and σ, respectively, which depend on the raw ensemble.

3.1.1. Temperature

For T2M forecasts $g(\cdot)$ is a normal density distribution with mean m and variance σ^2. Here we use a variant of the original EMOS approach similar to the one proposed by *Scheuerer and Büermann* [2014] where the departures of observed temperatures from their climatological means are related to those of the forecasts. Specifically, let $T = \{t_1, \dots, t_n\}$ be a training period of n days preceding the forecast initialization and denote by f_{tk} the forecast of the kth ensemble member and by y_t the observation on day $t \in T$. As a first step, we fit a regression model

$$y_{t_j} = c_0 + c_1 \sin\left(\frac{2\pi j}{365}\right) + c_2 \cos\left(\frac{2\pi j}{365}\right) + \varepsilon_{t_j}, \quad j = 1, \dots, n \tag{2}$$

which captures the seasonal variation of T2M. The residual terms ε_{t_j} are likely correlated over time, but for simplicity an ordinary least squares fit is performed. We denote by $\tilde{y}_t$ the fitted value of this periodic regression model on day t and interpret it as the climatological mean temperature on this day. This model can easily be extrapolated to future days $t_{d+1}, t_{d+2}, \ldots$ The above regression includes both a sine term and a cosine term which is equivalent to a cosine model with variable phase, and amplitude. Since $j = 1, \ldots, n$ is just a numbering of the days in T, different training periods have different phase parameters and hence, c_1 and c_2 evolve over the calendar year. We fit the same type of model also to the ensemble mean, control, and high-resolution run and obtain climatological means $\tilde{f}_{\overline{\mathrm{ENS}},t}$, $\tilde{f}_{\mathrm{CTRL},t}$, and $\tilde{f}_{\mathrm{HRES},t}$. The mean of the forecast distribution is then

$$m = \tilde{y} + a_1(f_{\mathrm{HRES}} - \tilde{f}_{\mathrm{HRES}}) + a_2(f_{\mathrm{CTRL}} - \tilde{f}_{\mathrm{CTRL}}) + a_3(f_{\overline{\mathrm{ENS}}} - \tilde{f}_{\overline{\mathrm{ENS}}}). \tag{3}$$

The variance of the forecast distribution is linked to the raw ensemble by

$$\sigma^2 = b_0 + b_1 s^2, \tag{4}$$

where $s^2 = \frac{1}{K}\sum_{k=1}^{K}(f_k - \frac{1}{K}\sum_{k=1}^{K} f_k)^2$. The parameters $\theta_{\mathrm{T2M}} = (a_1, a_2, a_3, b_0, b_1)^T$ are constrained to be nonnegative, and hence, $a_k / \sum_{k=1}^{K} a_k$ can be understood as the weight of model k.

3.1.2. Precipitation

For PPT24 we use the EMOS approach proposed by *Scheuerer* [2014], where $g(\cdot)$ is a left-censored (at zero) generalized extreme value (GEV) distribution. While the shape parameter ξ of the GEV is kept constant ($\xi = 0.2$), the location and the scale parameters m and σ are linked to the raw ensemble via

$$m = a_0 + a_1 f_{\mathrm{HRES}} + a_2 f_{\mathrm{CTRL}} + a_3 f_{\overline{\mathrm{ENS}}} + a_4 \pi_0, \tag{5}$$

$$\sigma = b_0 + b_1 \mathrm{MD}_f, \tag{6}$$

where π_0 is the fraction of ensemble members predicting zero precipitation and $\mathrm{MD}_f := K^{-2}\sum_{k,k'=1}^{K} |f_k - f_{k'}|$ is the ensemble mean difference. Again, the parameters are denoted by $\theta_{\mathrm{PPT24}} = (a_0, \ldots, a_4, b_0, b_1)^T$. The parameters a_1, a_2, a_3, b_0, b_1 are constrained to be nonnegative, and hence, the normalized parameters a_1 to a_3 can be understood as weights.

3.1.3. Wind Speed

For V10 we use a modified version of the EMOS model based on a left-truncated (at zero) normal distribution by *Thorarinsdottir and Gneiting* [2010]. A truncated normal distribution on the square root-transformed space seems to be an appropriate choice for $g(\cdot)$, as it outperformed both the untransformed truncated normal model and a model with predictive gamma distributions in preliminary tests. We model the distribution of $\sqrt{y}$ by a truncated normal distribution with parameters:

$$m = a_0 + a_1\sqrt{f_{\mathrm{HRES}}} + a_2\sqrt{f_{\mathrm{CTRL}}} + a_3\sqrt{f_{\overline{\mathrm{ENS}}}} \tag{7}$$

$$\sigma^2 = b_0 + b_1 \mathrm{MD}_{\sqrt{f}}, \tag{8}$$

where $\mathrm{MD}_{\sqrt{f}} := K^{-2}\sum_{k,k'=1}^{K} |\sqrt{f_k} - \sqrt{f_{k'}}|$. The parameters $\theta_{V10} = (a_0, \ldots, a_3, b_0, b_1)^T$ are constrained to be nonnegative; thus, the normalized parameters a_1 to a_3 can be understood as model weights.

3.1.4. Model Fitting and Evaluation

For all three variables the parameter vector $\hat{\theta}$ is estimated by CRPS minimization over the training period T. Rationales for using the CRPS can be found in, e.g., *Hersbach* [2000], *Gneiting et al.* [2005], or *Gneiting and Raftery* [2007]. The training period for each verification day consists of the n days preceding the initialization date. Tests using a subset of European stations indicate that for T2M forecasts a training period of 720 days is appropriate, while for PPT24 and V10 training periods of 1816 and 365 days, respectively, performed best. Following *Scheuerer* [2014], we try to avoid overfitting by using the parameter estimates $\hat{\theta}_{t-1}$ as starting values for the estimation of $\hat{\theta}_t$ for verification day t and then stopping the optimization process after a few iterations. This sliding window model fitting approach generally results in good parameter estimates, but it may be affected by sudden changes in the raw ensemble models during the training period. Nevertheless, the good performance of the postprocessed forecasts as shown in section 4 indicates that this effect can be neglected for the majority of stations. The average CRPS over the training period T is calculated by CRPS $= T^{-1} \sum_T \text{crps}_t$, where the crps for a single training day is given by

$$\text{crps}(P, y) = \int_{-\infty}^{\infty} \left[P(x) - \mathbb{1}_{[x \geq y]} \right]^2 \mathrm{d}x, \tag{9}$$

where P denotes the cumulative predictive distribution function and y is the associated observation [*Hersbach*, 2000; *Gneiting et al.*, 2007]. A closed-form expression for the crps for the normal model for T2M can be found in *Gneiting et al.* [2005]. For the censored GEV model used for PPT24 a closed-form expression has been derived by *Friederichs and Thorarinsdottir* [2012] and *Scheuerer* [2014]. For the square root-transformed truncated normal model used for V10, the crps can be calculated using formulae by *Gneiting et al.* [2004]. With $q = \Phi(-\mu/\sigma)$, $p = 1 - q$, and $w = (\sqrt{y} - \mu)/\sigma$, the crps can be written as follows:

$$\begin{aligned} \text{crps}(y, \mu, \sigma) = {} & \frac{\sigma}{p^2}\left(\sigma - \frac{2\mu}{\sqrt{\pi}}\right) - 2\sigma^2 \left\{ \frac{w^2}{2} - \frac{1}{p}\left[(w^2 - 1)\Phi(w) + w\varphi(w)\right] + \frac{qw^2}{p} \right\} \\ & - 2\sigma\mu \left\{ w - \frac{2}{p}\left[w\Phi(w) + \varphi(w)\right] + \frac{2qw}{p} \right\} \\ & + \frac{q\sigma^2}{p^2}\left[-\frac{1}{q}\varphi\left(\frac{-\mu}{\sigma}\right)^2 + q\left(\frac{\mu^2}{\sigma^2} - 1\right) \right] + \frac{2\sigma\mu}{p^2\sqrt{\pi}}\Phi\left(-\frac{\sqrt{2}\mu}{\sigma}\right) - \frac{\mu^2 q^2}{p^2}, \end{aligned} \tag{10}$$

where $\Phi(\cdot)$ and $\varphi(\cdot)$ denote cumulative and probability density functions of the standard normal distribution, respectively.

3.2. Postprocessing Using BMA for T2M

As T2M predictions can be described well by a normal distribution, BMA parameters can be estimated easily using the `R` package `ensembleBMA` [*Fraley et al.*, 2014]. Hence, for T2M BMA can be used as an alternative to EMOS even on the global set of stations. BMA combines the raw ensemble forecasts to a mixture distribution of the form

$$y|\boldsymbol{f} \sim \sum_{k=1}^{K} w_k g(y \mid f_k), \tag{11}$$

where $w_1, \dots, w_K$ are model weights and $g(y \mid f_k)$ is a parametric distribution given that model k is best. In order to account for exchangeable ensemble members and to include a bias correction the BMA model is parameterized by

$$y|\boldsymbol{f} \sim \sum_{i=1}^{I} \sum_{j=1}^{J_i} w_i g(y \mid a_{i0} + a_{i1} f_{i,j}, \sigma_i), \tag{12}$$

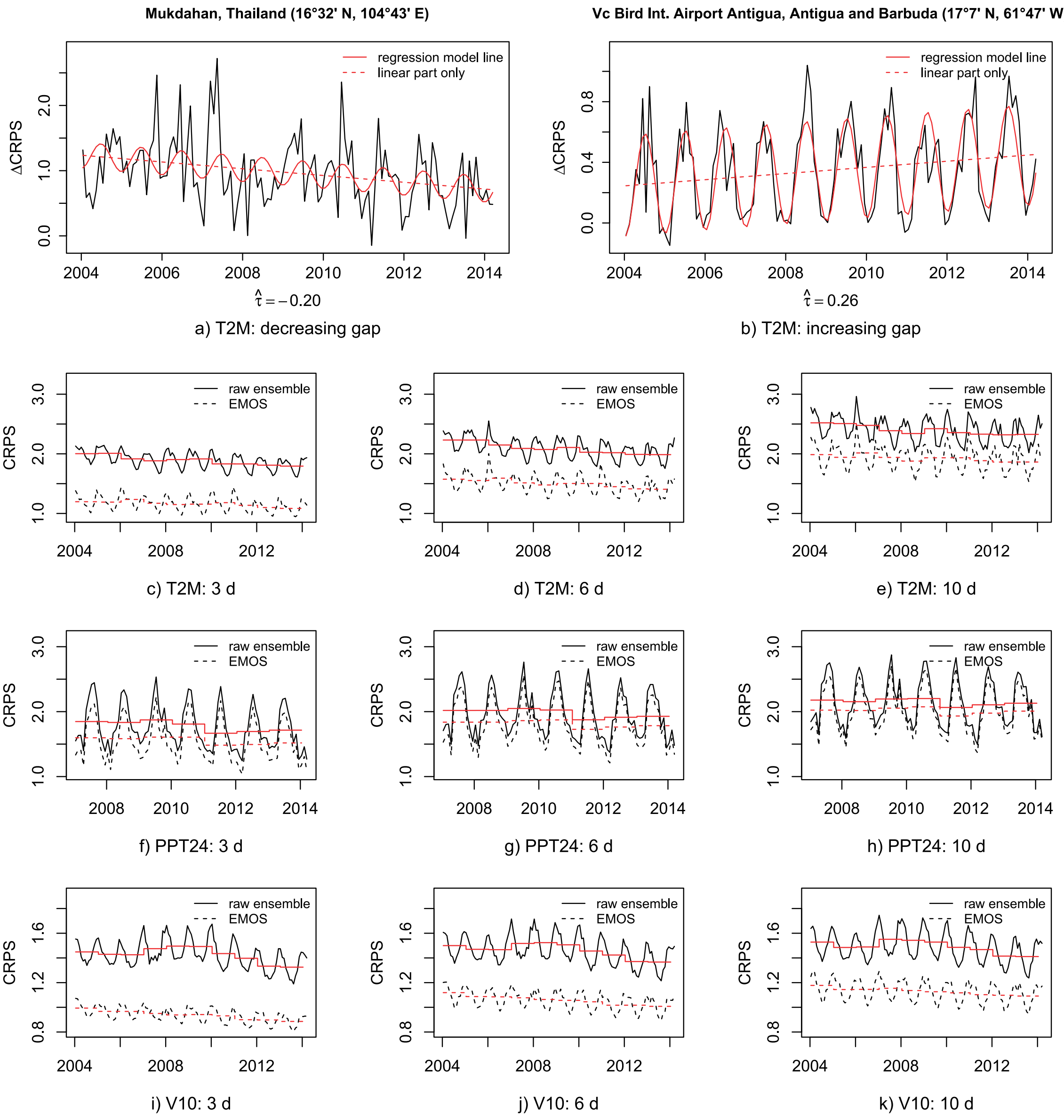
where I is the number of subgroups of the ensemble within which all members are exchangeable and J_i is the number of members in this group [*Fraley et al.*, 2010]. In case of T2M $g(\cdot)$ is a normal kernel distribution. The parameters $\hat{a}_{i0}, \hat{a}_{i1}$, $i =$ HRES, CTRL, ENS are estimated by linear regression and $\hat{w}_i, \hat{\sigma}_i$ by the Expectation-Maximization algorithm [*Dempster et al.*, 1977; *McLachlan and Krishnan*, 1997]. The BMA models for this study are fitted using a training period of 365 days prior to the verification day. The estimates for day $t - 1$ are used as starting values for the estimation of the parameter values for day t.

3.3. Global CRPS Analysis

As stated in the introduction, the main objective of this study is to analyze whether the gap in CRPS between the raw ensemble and the post-processed forecast narrows over time. This is assessed station-wise using both a parametric and a non-parametric approach. For the former, we fit the following regression model to the monthly time series of CRPS differences ($\Delta\text{CRPS}_t = \text{CRPS}_{\text{raw},t} - \text{CRPS}_{\text{EMOS},t}$) :

$$\Delta\text{CRPS}_t = \beta_0 + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{12}\right) + \beta_3 \cos\left(\frac{2\pi t}{12}\right) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{13}$$

where ΔCRPS_t is the predictand, t is now the time in months, and σ^2 denotes the error variance. For the latter, we use Kendall's τ correlation coefficient and the associated test statistics [*Mann*, 1945] as



a) T2M: decreasing gap

b) T2M: increasing gap

c) T2M: 3 d

d) T2M: 6 d

e) T2M: 10 d

f) PPT24: 3 d

g) PPT24: 6 d

h) PPT24: 10 d

i) V10: 3 d

j) V10: 6 d

k) V10: 10 d

Figure 1. (a and b) Monthly averages of ΔCRPS between raw ensemble and EMOS forecasts with a lead time of 6 days, for example, at stations with a decreasing and an increasing gap. The red solid lines correspond to the fits of the regression model stated in equation (13); the red dashed lines to their linear parts. (c to k) The monthly (in black) and yearly (in red) global average CRPS of the raw ensemble and EMOS forecasts for T2M, PPT24, and V10.

implemented in the R package Kendall [*McLeod*, 2011]. In order to correct for seasonal effects, we calculate the τ statistics using the residuals of the following model:

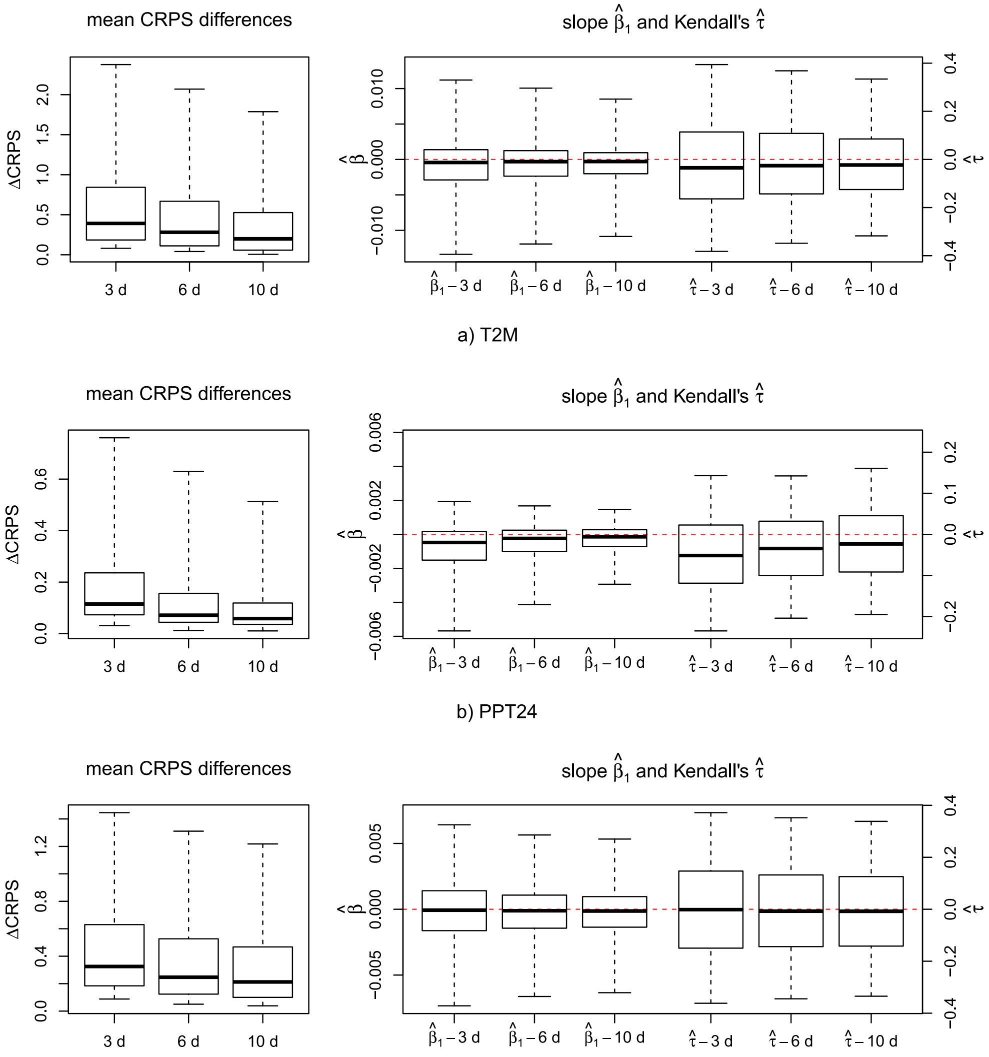
$$\Delta\text{CRPS}_t = \gamma_0 + \gamma_1 \sin\left(\frac{2\pi t}{12}\right) + \gamma_2 \cos\left(\frac{2\pi t}{12}\right) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{14}$$

Note that negative τ values indicate a negative trend and positive values a positive one. Figures 1a and 1b show the regression lines estimated by model (13) for monthly averages of ΔCRPS and the corresponding Kendall's τ test statistics for an example with decreasing and increasing gap.

4. Results

4.1. General Features of ΔCRPS

Before assessing the stationwise evolution of ΔCRPS over time, we consider first the evolution of global average CRPS values of both raw ensemble and EMOS forecasts. As shown in Figures 1c to 1k the average



c) V10

Figure 2. Box plots over all stations representing the 5, 25, 50, 75, and 95% quantiles of (left) the average CRPS differences between raw ensemble and EMOS forecasts and (right) the slope coefficients of the linear model fits and the Kendall's τ statistics of monthly ΔCRPS averages. Depicted are (a) T2M, (b) PPT24, and (c) V10; the red dashed lines on the right-hand panels indicate the zero line.

CRPS for both forecasts increases with increasing lead time regardless of the variable of interest. Note that all three variables exhibit seasonal oscillations in average CRPS. In the case of T2M and V10 postprocessing by EMOS obviously improves the average CRPS, whereas for PPT24 the improvement is much smaller relative to its seasonal oscillations in average CRPS. In any case, further analyses on the temporal evolution of ΔCRPS should correct for seasonal effects. Note that ΔCRPS depends on the performance of the postprocessing method selected. If alternative postprocessing methods perform better, ΔCRPS will be further increased by using them.

Let us now focus on a stationwise analysis. According to the box plots on the panels on the left of Figures 2a to 2c, more than 95% of the stations benefit from EMOS in terms of ΔCRPS averaged over the entire verification period regardless of lead time and variable of interest. Note also the positive skewness and the decrease in ΔCRPS with increasing lead time. The box plots on the panels on the right of Figures 2a to 2c describe the empirical distributions among the set of all stations considered of the slope coefficients $\hat{\beta}_1$ and the $\hat{\tau}$ test statistics of ΔCRPS against time for the parametric and the nonparametric model, respectively. For T2M and, in particular, PPT24 negative trends are more common than positive trends, whereas the

Table 1. Percentages of Stations Showing No, Negative, or Positive Trend in ΔCRPS[a]

			Parametric Model				Kendall's τ Statistics			
			T2M		PPT24	V10	T2M		PPT24	V10
			EMOS	BMA	EMOS	EMOS	EMOS	BMA	EMOS	EMOS
Forecast lead time	3 days	no significant trend	42%	42%	76%	41%	44%	43%	77%	42%
		negative trend	34%	34%	19%	31%	32%	32%	18%	29%
		positive trend	24%	25%	5%	28%	24%	25%	5%	29%
	6 days	no significant trend	46%	48%	82%	43%	48%	49%	82%	45%
		negative trend	31%	28%	14%	31%	29%	27%	13%	29%
		positive trend	23%	24%	4%	26%	23%	24%	5%	27%
	10 days	no significant trend	54%	58%	83%	45%	54%	58%	82%	46%
		negative trend	27%	23%	11%	31%	26%	23%	11%	28%
		positive trend	19%	18%	6%	25%	20%	19%	7%	26%

[a] Percentages of stations (totals are 4160 (T2M), 2917 (PPT24), and 4387 (V10)) showing no, negative, or positive trend in monthly ΔCRPS values against time at a significance level of 0.05.

corresponding box plots for V10 are almost symmetric around the zero line. In general, the medians of the $\hat{\beta}_1$ and the $\hat{\tau}$ values seem to converge to zero with increasing lead time.

4.2. Are There Any Significant Temporal Trends?

The above results indicate a tendency of a decrease in ΔCRPS over time at least for T2M and PPT24. In the following, we check the percentages of stations with decreasing, an absence of, or increasing trend in ΔCRPS over time at a significance level of 0.05. In order to be more confident about the results, this analysis is performed using both the parametric regression model and the nonparametric Kendall's τ correlation coefficient test. As already mentioned, both approaches correct for seasonal effects. Furthermore, in case of T2M the same analysis has been performed additionally using BMA instead of EMOS in order to relax the dependence on one particular postprocessing method. As shown in Table 1 the stations with no significant trend outnumber the stations with either negative or positive trend for all three variables and lead times considered. Note that the percentage of stations without any significant trend increases with increasing lead time. In line with the results shown in Figure 2, significantly negative trends are more common than positive ones for T2M and PPT24. The difference between the number of stations with negative and those with positive trend reduces with increasing lead time, but is still greater than zero for a 10 day forecast. Note that the high number of nonsignificant stations in case of PPT24 is likely to be due to the high variability of precipitation amounts, and hence variability of CRPS values, which leads to a large residual standard error in case of the parametric regression model and to a lot of pairs (a pair denotes here a value of ΔCRPS and its associated time stamp) opposite to the estimated direction in case of the τ test statistics. In case of V10 the stations with a negative trend and those with a positive trend are almost equally frequent regardless of the lead time. Figures of the global distributions of stations with no, significantly negative, and significantly positive trend in ΔCRPS are available as supporting information to this paper.

5. Discussion and Conclusions

According to the above analyses, the gap in CRPS between the raw ensemble and the EMOS forecasts remains almost constant over time. For T2M and PPT24 ΔCRPS shows a slightly decreasing tendency. The higher the lead time, the less accentuated is this tendency. For V10 such a tendency cannot be detected. The parametric regression model and the nonparametric τ test yield similar results. Hence, a linear model that is overlaid by seasonal fluctuations seems to be reasonable. Note that the skill of the raw ensemble and the EMOS forecasts may sometimes be negatively affected by upgrades to the atmospheric model. Model upgrades may deteriorate raw ensemble skill at some individual stations. For instance, a resolution increase may introduce new issues with statistical downscaling of the forecasts to some specific observation sites. But more importantly, the skill of the postprocessed forecasts can be lowered dramatically if a model update happens between the training and the verification period. These issues may result in positive trends in ΔCRPS. Ideally, postprocessing would be based on a cascade of reforecasts. That is, for each atmospheric model version, training of the postprocessing model would be done using a corresponding time series of reforecasts made with that same model version. Furthermore, the observations may be affected by measurement errors. If these errors change over time, they may also influence the estimates of the trends in ΔCRPS. As the problems introduced by statistical downscaling may be mitigated by verifying against model analysis, a similar study that replaces observations by model analysis, as proposed by *Ghelli and Lalaurette* [2000] and *Pappenberger et al.* [2009], may give further insights.

Additionally, verification scores are affected by ensemble size [e.g., *Richardson*, 2001]. Let us assume the hypothetical case of a perfect forecast distribution that equals the distribution of the stochastic process of interest. A raw ensemble forecast sampled from this forecast distribution would then be reliable by definition. Nevertheless, the raw ensemble would only be a stepwise approximation to the underlying forecast distribution. This would lead to an underperformance of the raw ensemble compared to the underlying forecast distribution in terms of CRPS, because CRPS is a proper skill score [*Gneiting and Raftery*, 2007]. This has to be kept in mind when comparing raw ensemble CRPS values with those values obtained from continuous forecast distributions. But note that this does not mean that the continuous forecast distributions obtained by postprocessing equal the underlying distribution mentioned above. *Ferro et al.* [2008] discuss the effect of ensemble size on CRPS. Hence, further analyses on the gap in skill between raw ensemble and postprocessed forecasts may benefit from taking this effect into account.

From the above, we conclude that the probabilistic skill of both the raw ensembles and the EMOS forecasts improves over time. The fact that the gap in skill has remained almost constant, especially for V10, suggests that improvements to the atmospheric model have an effect quite different from what calibration by statistical postprocessing is doing. That is, they are increasing potential skill. Thus, this study indicates that (a) further model development is important even if one is just interested in point forecasts and (b) statistical postprocessing is important because it will keep adding skill in the foreseeable future.

Acknowledgments
The forecast data used in this study can be made available subject to a handling charge. SYNOP observations are not available through ECMWF. We are grateful to F. Rabier, D.S. Richardson, D. Lavers, and other colleagues at ECMWF for helpful discussions and inputs. We thank David Möller of the University of Heidelberg for his contribution. Furthermore, we like to thank T. Gneiting of the Heidelberg Institute for Theoretical Studies and of the Institute for Stochastics at Karlsruhe Institute of Technology for valuable comments, and the colleagues at the Heidelberg Institute for Theoretical Studies. Additionally, we are grateful to the two reviewers for their helpful comments. We are indebted to the ECMWF for granting access to their products.

The Editor thanks Pierre Pinson and an anonymous reviewer for their assistance in evaluating this paper.

References

Bougeault, P., et al. (2010), The THORPEX Interactive Grand Global Ensemble, *Bull. Am. Meteorol. Soc.*, *91*, 1059–1072.

Buizza, R., T. Petroliagis, T. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi (1998), Impact of model resolution and ensemble size on the performance of an ensemble prediction system, *Q. J. R. Meteorol. Soc.*, *124*, 1935–1960.

Buizza, R., J.-R. Bidlot, N. Wedi, M. Fuentes, M. Hamrud, G. Holt, and F. Vitart (2007), The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System), *Q. J. R. Meteorol. Soc.*, *133*, 681–695.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc.*, *39*(B), 1–39.

Ferro, C. A. T., D. S. Richardson, and A. P. Weigel (2008), On the effect of ensemble size on the discrete and continuous ranked probability scores, *Meteorol. Appl.*, *15*, 19–24, doi:10.1002/met.45.

Fraley, C., A. E. Raftery, and T. Gneiting (2010), Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging, *Mon. Weather Rev.*, *138*(1), 190–202, doi:10.1175/2009MWR3046.1.

Fraley, C., A. E. Raftery, J. M. Sloughter, and T. Gneiting (2014), *ensembleBMA: Probabilistic Forecasting Using Ensembles and Bayesian Model Averaging*, R package version 5.0.6, Univ. of Washington, Seattle. [Available at http://CRAN.R-project.org/package=ensembleBMA, last checked: 3.12.2014.]

Friederichs, P., and T. L. Thorarinsdottir (2012), Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction, *Environmetrics*, *23*(7), 579–594.

Ghelli, A., and F. Lalaurette (2000), Verifying precipitation forecasts using upscaled observations, *ECMWF Newsl.*, *87*, 9–17.

Gneiting, T., and A. E. Raftery (2007), Strictly proper scoring rules, prediction, and estimation, *J. Am. Stat. Assoc.*, *102*, 359–378.

Gneiting, T., K. Larson, and K. Westrick (2004), Development of next-generation wind energy forecast and optimization technologies, *Tech. Rep.*, Univ. of Washington, Department of Statistics, Seattle, Wash.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman (2005), Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, *Mon. Weather Rev.*, *133*(5), 1098–1118, doi:10.1175/MWR2904.1.

Gneiting, T., F. Balabdoui, and A. E. Raftery (2007), Probabilistic forecasts, calibration and sharpness, *J. R. Stat. Soc. Ser. B*, *69*, 243–268.

Haiden, T., et al. (2014), ECMWF forecast performance during the June 2013 flood in Central Europe, *ECMWF Tech. Memo.*, *723*, 34 pp., Reading, U. K.

Hamill, T. M., C. Snyder, and R. E. Morss (2000), A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles, *Mon. Weather Rev.*, *128*(6), 1835–1851.

Hersbach, H. (2000), Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather Forecasting*, *15*(5), 559–570.

Leith, C. E. (1974), Theoretical skill of Monte-Carlo forecasts, *Mon. Weather Rev.*, *102*, 409–418.

Mann, H. B. (1945), Nonparametric tests against trend, *Econometrica*, *13*, 245–259.

Matheson, J. E., and R. L. Winkler (1976), Scoring rules for continuous probability distributions, *Manage. Sci.*, *22*, 1087–1096.

McLachlan, G. J., and T. Krishnan (1997), *The EM Algorithm and Extensions*, Wiley, New York.

McLeod, A. (2011), *Kendall: Kendall Rank Correlation and Mann-Kendall Trend Test*, R package version 2.2, Univ. of Western Ontario, London, Calif. [Available at http://CRAN.R-project.org/package=Kendall, last checked: 31.10.2014.]

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

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

Pappenberger, F., A. Ghelli, R. Buizza, and K. Bódis (2009), The skill of probabilistic precipitation forecasts under observational uncertainties within the generalized likelihood uncertainty estimation framework for hydrological applications, *J. Hydrometeorol.*, *33*, 807–819.

Park, Y.-Y., R. Buizza, and M. Leutbecher (2008), TIGGE: preliminary results on comparing and combining ensembles, *Q. J. R. Meteorol. Soc.*, *134*, 2029–2050.

Pinson, P., and R. Hagedorn (2012), Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations, *Meteorol. Appl.*, *19*(4), 484–500.

R Development Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. [Available at http://www.R-project.org/, last checked: 29.07.2014.]

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, *133*(2), 1155–1174.

Richardson, D. S. (2001), Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size, *Q. J. R. Meteorol. Soc.*, *127*, 2473–2489.

Richardson, D. S., J.-R. Bidlot, L. Ferranti, T. Haiden, T. Hewson, M. Janousek, F. Praters, and F. Vitart (2013), Evaluation of ECMWF forecasts, including 2012-2013 upgrades, *ECMWF Tech. Memo.*, *710*, 55 pp., European Centre for Medium-Range Weather Forecasts, Reading, Berkshire.

Scheuerer, M. (2014), Probabilistic quantitative precipitation forecasting using ensemble model output statistics, *Q. J. R. Meteorol. Soc.*, *140*(680), 1086–1096.

Scheuerer, M., and L. Büermann (2014), Spatially adaptive post-processing of ensemble forecasts for temperature, *J. R. Stat. Soc. Ser. C*, *63*(3), 405–422.

Thorarinsdottir, T. L., and T. Gneiting (2010), Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression, *J. R. Stat. Soc. Ser. A*, *173*, 371–388.

Erratum

In the originally published version of this article, equation 10 contained a typographical error. The "w" omega character was missing from the first line of the crps expression. The equation has since been corrected, and this version may be considered the authoritative version of record. The authors thank Maxime Taillardat for pointing out the error.