A Probabilistic Approach to Forecast the Uncertainty with Ensemble Spread

BERT VAN SCHAEYBROECK AND STÉPHANE VANNITSEM

Royal Meteorological Institute, Brussels, Belgium

(Manuscript received 2 October 2014, in final form 2 July 2015)

ABSTRACT

The ensemble spread is often used as a measure of forecast quality or uncertainty. However, it is not clear whether the spread is a good measure of uncertainty and how the spread–error relationship can be properly assessed. Even for perfectly reliable forecasts the error for a given spread varies considerably in amplitude and the spread–error relationship is therefore strongly heteroscedastic. This implies that the forecast of the uncertainty based only on the knowledge of spread should itself be probabilistic.

Simple probabilistic models for the prediction of the error as a function of the spread are introduced and evaluated for different spread–error metrics. These forecasts can be verified using probabilistic scores and a methodology is proposed to determine what the impact is of estimating uncertainty based on the spread only. A new method is also proposed to verify whether the flow-dependent spread is a realistic indicator of uncertainty. This method cancels the heteroscedasticity by a logarithmic transformation of both spread and error, after which a linear regression can be applied. An ensemble system can be identified as perfectly reliable with respect to its spread.

The approach is tested on the ECMWF Ensemble Prediction System over Europe. The use of spread only does not lead to skill degradation, and replacing the raw ensemble by a Gaussian distribution consistently improves scores. The influences of non-Gaussian ensemble statistics, small ensemble sizes, limited predictability, and different spread–error metrics are investigated and the relevance of binning is discussed. The upper-level spread–error relationship is consistent with a perfectly reliable system for intermediate lead times.

1. Introduction

The use of ensemble predictions allows for the estimation of possible outcomes of specific weather events. From the ensemble, a ''best'' deterministic forecast can be extracted as the ensemble mean (Ehrendorfer 1997), while the forecast uncertainty is often reduced to one measure, called ensemble *spread*, summarizing the information on the uncertainties associated with the impact of the presence of the initial conditions and model errors (Nicolis et al. 2009). Although much information is lost by reducing the ensemble forecast to two numbers, such an approach does offer benefits since it condenses the most essential and most reliable knowledge. Moreover, conveying additional information could be counterproductive to most end users as the interpretation can be difficult. Whereas the added value of ensemble forecasting for minimizing the error of the ensemble mean is beyond doubt, the analysis of uncertainty information remains a nontrivial task (Toth et al. 2003; Wilks 2011) and forms the subject of intense ongoing research (Hopson 2014; Christensen et al. 2014).

The verification of the forecast uncertainty, based on the spread only, is intricate for two reasons: for each probabilistic forecast only one observation is realized and the spread is not meant to provide an exact prediction for the error. The statistical analysis of binned forecasts of similar spread can be used to overcome this last issue (Wang and Bishop 2003; Kolczynski et al. 2011; Leutbecher 2009; Grimit and Mass 2007). Within the context of forecast uncertainty verification much attention has been, however, devoted to verification scores that assess the *deterministic* rather than the probabilistic aspects of the forecast. For instance, the strengths and weaknesses of the spread–error correlation for different spread–error metrics were explored in detail by Grimit and Mass (2007) and Hopson (2014). Christensen Moroz and Palmer (2014), on the other hand, go beyond this approach by proposing a new proper skill score by extending the mean-square error of the (squared) error

Corresponding author address: Bert Van Schaeybroeck, Royal Meteorological Institute, Ringlaan 3, B-1180 Brussels, Belgium.
E-mail: bertvs@meteo.be

DOI: 10.1175/MWR-D-14-00312.1

forecast into a proper score that takes into account the ensemble skewness.

In the present work, we first argue that, rather than using the spread as a deterministic forecast of the uncertainty, it is preferable to construct a *probabilistic* forecast for the uncertainty. Spread–error verification can then be based on conventional probabilistic scores. Second, using such verification, one can assess whether the uncertainty estimate based on spread only degrades or even improves upon the full-ensemble uncertainty estimate. For instance, the impact on statistical scores of replacing an ensemble with a Gaussian distribution with corresponding mean and variance can be checked. Third, rather than inferring relations of the form $\hat{\mathcal{E}} = \alpha + \beta\mathcal{S}$ between the error $\hat{\mathcal{E}}$ and spread $\mathcal{S}$, inference of the type $\hat{\mathcal{E}} = \alpha\mathcal{S}^{\beta}$ is advocated since it is statistically well posed and allows for the identification of an ensemble system that is perfectly reliable conditional on spread as the only uncertainty measure.

With that perspective in mind the following questions are addressed: (i) Given a series of ensemble spreads and associated ensemble errors, how can one best model the uncertainty forecast based on the spread such that it reproduces well the uncertainty information contained in the full-ensemble uncertainty forecast (section 2)? (ii) What are the appropriate scores for the verification of the spread-based uncertainty forecast (section 3a)? (iii) Can we assess to what extent uncertainty estimates that are based on spread only degrade or even improve the full-ensemble estimates (section 3b)? (iv) Is the ensemble spread useful as a predictor for the uncertainty (section 4)?

The approach is then tested within the context of the Ensemble Prediction System (EPS) of the European Centre for Medium-Range Weather Forecasts (ECMWF), in order to clarify whether information is lost when using the ensemble spread instead of the full ensemble for uncertainty estimation (section 5a), and at what lead times the spread is a good predictor for the EPS uncertainty (section 5b). Finally, the influence of limited ensemble size, non-Gaussian ensemble statistics, limited predictability, the spread–error metric, and the spread-based model are also investigated in sections 5a and 5b, and our methodology is related to the binning approach in section 5c. The results are further discussed in section 6 from the perspective of previous works and our conclusions are drawn in section 7.

2. The forecast of the uncertainty and its modeling

a. General approach

In this section a minimal probabilistic framework for the spread–error relationship is proposed. Assume an ensemble forecast with F^e the forecast of member e, and $\overline{F}$ the ensemble mean. The most commonly used error metric $\mathcal{E}_0^O$ is the error of the ensemble mean:

$$\mathcal{E}_0^O = \overline{F} - O, \tag{1}$$

with O being the observation. Once the definition of the error metric $\mathcal{E}^O$ is given, the error forecast or uncertainty forecast $\mathcal{E}^F$ can be readily constructed. The *full-ensemble error forecast or uncertainty forecast* $\{\mathcal{E}^{F^e}\}_e$ is obtained by replacing, in the expression of the error, the observation with the ensemble-member prediction F^e. For the ensemble-mean error, therefore,

$$\mathcal{E}_0^{F^e} = \overline{F} - F^e. \tag{2}$$

For a perfectly reliable forecast, $\mathcal{E}_0^O$ is drawn from the same distribution as $\mathcal{E}_0^{F^e}$. The purpose is now to find statistical models that represent well the full-ensemble uncertainty forecast only based on the ensemble spread (and not the members F^e). Because of the reduction in information, it is impossible to verify perfect reliability based on spread as the only uncertainty estimate; however, the aim here is to identify behavior that is consistent with perfect reliability. Therefore, we define *perfect spread consistency* as perfect reliability conditional on spread as the only uncertainty measure.

A conventional spread metric is the ensemble standard deviation $\mathcal{S}_0$:

$$\mathcal{S}_0 = \langle(\overline{F} - F^e)^2\rangle_e^{1/2}. \tag{3}$$

Here, the brackets $\langle\cdot\rangle_e$ denote the average over the ensemble members. For a perfectly reliable ensemble the scatterplot of $\mathcal{E}_0^O$ against $\mathcal{S}_0$ is shown in Fig. 1a for the Lorenz'96 model with 36 grid points and a forcing parameter equal to 10 (Lorenz 1996). Clearly, the average of the uncertainty forecast [Eq. (2)] is zero while its standard deviation is given by the spread [Eq. (3)].

For the purpose of constructing the spread-based model, consider general spread–error metrics, $\mathcal{S}$ and $\mathcal{E}$, and the following formal decomposition of the uncertainty forecast, conditional on the spread, into an average value and a stochastic term:

$$\mathcal{E}^F \,|\, \mathcal{S} = \mu_{\mathcal{E}^F|\mathcal{S}} + \epsilon \,|\, \mathcal{S}. \tag{4}$$

Here, $\mu_{\mathcal{E}^F|\mathcal{S}} = \langle\mathcal{E}^{F^e}\rangle_e$ is the average uncertainty for an ensemble of a given spread, obtained by averaging over the ensemble of uncertainty forecasts. Likewise, one defines $\sigma_{\mathcal{E}^F|\mathcal{S}}$ as the standard deviation of the random term $\epsilon \,|\, \mathcal{S}$. It must be stressed that $\mu_{\mathcal{E}^F|\mathcal{S}}$ and $\sigma_{\mathcal{E}^F|\mathcal{S}}$ do not depend on the observation; they depend on the *error metric* but not on the error (with respect to the observation) itself. The first proposed spread-based model for

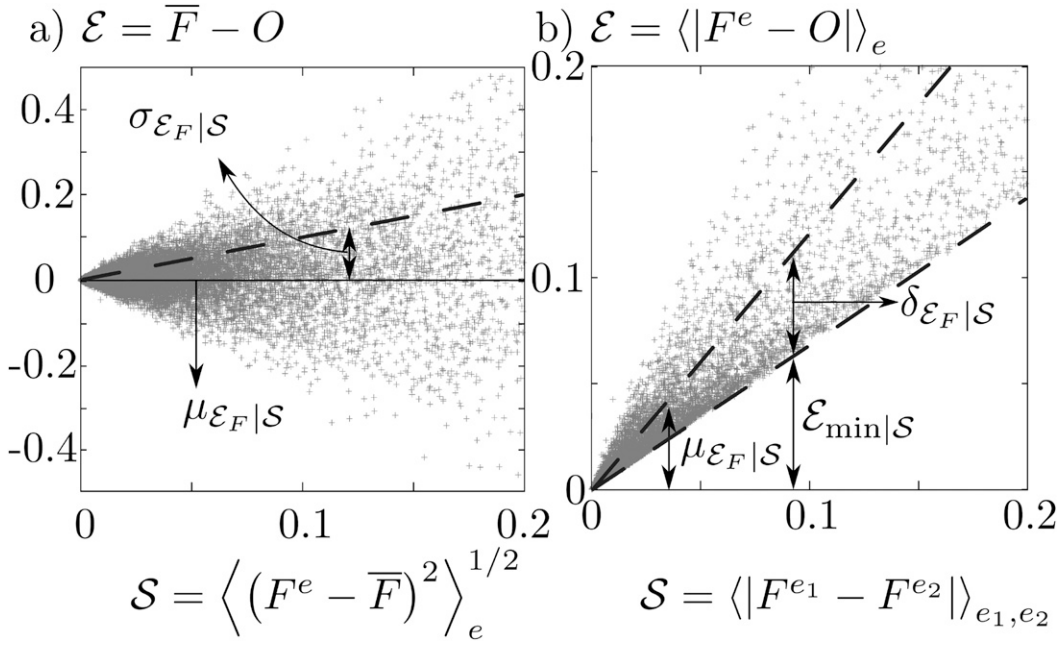

FIG. 1. Uncertainty forecast for the error $\mathcal{E}^O$ against spread $\mathcal{S}$ for a perfectly reliable ensemble system generated using the spatially extended 36-variable system and forcing parameter as specified in Lorenz (1996). (a),(b) Results for different spread–error metrics. At lead time zero all variables are slightly perturbed separately with Gaussian-distributed random noise with standard deviation 10^{-3} in order to generate 50 ensemble members. A set of 300×36 spread–error pairs at lead time 2.5 time units are shown.

the *uncertainty forecast* assumes Gaussian-distributed errors:

$$\text{Model I: } \mathcal{E}^F \,|\, \mathcal{S} \sim \mathcal{N}(\mu_{\mathcal{E}^F|\mathcal{S}}, \sigma^2_{\mathcal{E}^F|\mathcal{S}}). \tag{5}$$

Note that this model applies to the *error* distribution and not the *ensemble* distribution. Both are equivalent only for the ensemble-mean error metric $\mathcal{E}_0^O$. As we will see in the following, the model represents well the combination of the ensemble-mean error $\mathcal{E}_0^O$ [Eq. (1)] and the ensemble standard deviation $\mathcal{S}_0$ [Eq. (3)]. However, it is less appropriate for other spread–error metrics. For instance, the AV-ABS metric involves nontrivial ensemble averages for both spread and error (Grimit and Mass 2007; Hopson 2014):

$$\begin{aligned} \text{AV}-\text{ABS: } \mathcal{E}^O &= \langle |F^e - O| \rangle_e, \quad \text{and} \\ \mathcal{S} &= \langle |F^{e_1} - F^{e_2}| \rangle_{e_1,e_2}. \end{aligned} \tag{6}$$

Figure 1b displays the corresponding spread–error relation for the case of a forecast system that is perfectly spread consistent, that is, perfectly reliable conditional on spread as the only uncertainty measure. Clearly, because of the conspicuous lower bound on the error, model I associated with Eqs. (4) and (5) is inappropriate. The members $\mathcal{E}^{F^e} = \langle |F^{e_1} - F^e| \rangle_{e_1}$ constitute the ensemble members of the uncertainty forecast for AV-ABS. The following decomposition takes into account a minimal error $\mathcal{E}_{\min|\mathcal{S}}$:

$$\mathcal{E}^F \,|\, \mathcal{S} = \mathcal{E}_{\min|\mathcal{S}} + \epsilon \,|\, \mathcal{S} \quad \text{with} \quad \epsilon(\mathcal{S}) \geq 0. \tag{7}$$

Two simple spread models for the distribution of the residual $\epsilon \,|\, \mathcal{S}$ are considered here: a truncated Gaussian distribution and a shifted exponential distribution. Therefore, the second and third *spread-based models for the uncertainty forecast* are

$$\begin{aligned} &\text{Model II: } \mathcal{E}^F \,|\, \mathcal{S} \\ &\quad \sim 2\mathcal{N}(\mathcal{E}_{\min|\mathcal{S}}, \delta^2_{\mathcal{E}^F|\mathcal{S}}) \quad \text{for} \quad \mathcal{E}^F \geq \mathcal{E}_{\min|\mathcal{S}} \quad \text{and} \end{aligned} \tag{8a}$$

$$\begin{aligned} &\text{Model III: } \mathcal{P}(\mathcal{E} = \mathcal{E}^F \,|\, \mathcal{S}) \\ &\quad = \frac{e^{-[(\mathcal{E}^F - \mathcal{E}_{\min|\mathcal{S}})/(\mu_{\mathcal{E}^F|\mathcal{S}} - \mathcal{E}_{\min|\mathcal{S}})]}}{\mu_{\mathcal{E}^F|\mathcal{S}} - \mathcal{E}_{\min|\mathcal{S}}} \quad \text{for} \quad \mathcal{E}^F \geq \mathcal{E}_{\min|\mathcal{S}}. \end{aligned} \tag{8b}$$

Here, $\delta_{\mathcal{E}^F|\mathcal{S}}$ is the square root of the second moment of $\mathcal{E}^F - \mathcal{E}_{\min|\mathcal{S}}$.

All quantities $\mathcal{E}_{\min|\mathcal{S}}$, $\mu_{\mathcal{E}^F|\mathcal{S}}$, $\sigma_{\mathcal{E}^F|\mathcal{S}}$, and $\delta_{\mathcal{E}^F|\mathcal{S}}$ are properties of the ensemble forecast only and independent of the observations. Figure 1 illustrates these quantities in the case of a perfectly reliable system. Clearly, for the system under consideration all quantities are linearly

TABLE 1. Different definitions for combinations of spreads and error metrics are given in the first three columns. For the given metrics, the last column suggests the preferred spread-based model, along with its parameters. For the last two spread–error metrics (EM-GEO and AV-GEO) the regression coefficients of the log-transformed regression as defined in Eq. (20) for a perfectly spread-consistent forecast are independent of any hypothesis on the nature of the ensemble distribution. Note that more results are given in Table 2 but the metrics specified here are restricted to the ones that have universal parameters (i.e., independent of assumptions on the ensemble distribution). We denote member e of the ensemble forecast as F^e, $\overline{F} = \langle F^e \rangle_e$ is the ensemble mean (EM), and the observations as O.

Metric	Error metric $\mathcal{E}^O$	Spread metric $\mathcal{S}$	Suggested spread-based model
EM-RAW	$\mathcal{E}_0^O = O - \overline{F}$	$\mathcal{S}_0 = \langle (F^e - \overline{F})^2 \rangle_e^{1/2}$	Model I [Eq. (5)]: $\mu_{\mathcal{E}^F\|\mathcal{S}} = 0$, $\sigma_{\mathcal{E}^F\|\mathcal{S}} = \mathcal{S}$
EM-SQU	$\|O - \overline{F}\|$	$\langle (F^e - \overline{F})^2 \rangle_e^{1/2}$	Model II [Eq. (8a)]: $\mathcal{E}_{\min\|\mathcal{S}} = 0$, $\delta_{\mathcal{E}^F\|\mathcal{S}} = \mathcal{S}$
EM-ABS	$\|O - \overline{F}\|$	$\langle \|F^e - \overline{F}\| \rangle_e$	Model III [Eq. (8b)]: $\mathcal{E}_{\min\|\mathcal{S}} = 0$, $\mu_{\mathcal{E}^F\|\mathcal{S}} = \mathcal{S}$
EM-GEO	$\|O - \overline{F}\|$	$\exp\langle \ln(\|F^e - \overline{F}\|) \rangle_e$	Regression coefficients for Eq. (20): $\alpha_{\text{perf}} = \beta_{\text{perf}} = 1$
AV-GEO	$\exp\langle \ln(\|O - F^e\|) \rangle_e$	$\exp\langle \ln(\|F^{e_1} - F^{e_2}\|) \rangle_{e_1 \neq e_2}$	Regression coefficients for Eq. (20): $\alpha_{\text{perf}} = \beta_{\text{perf}} = 1$

proportional to $\mathcal{S}$. Such proportionality implies strong heteroscedasticity and constitutes the most prominent reason why the uncertainty forecast should be considered probabilistically instead of deterministically. Table 1 specifies for a few spread–error metrics the preferred spread-based model along with the associated quantities that depend on the spread. Additional spread–error metrics and associated parameters are given in Table 2. Most spread–error metrics used in this work have appeared previously in Grimit and Mass (2007), Scherrer et al. (2004), or Hopson (2014). Note that many spread–error metrics are not following a (proper) scoring rule. Indeed, the most conventional error metric, $\mathcal{E}_0^O = \overline{F} - O$, is minimized for $\overline{F} \to -\infty$.

b. *Computation of model parameters for two specific error metrics*

For general spread–error metrics, in order to obtain the different parameters of the spread-based model, one must consider the uncertainty forecast $\mathcal{E}^{F^e}$ of member e based on the spread–error metrics provided in the second column of Table 2. Then, the parameters $\mathcal{E}_{\min|\mathcal{S}}$, $\mu_{\mathcal{E}^F|\mathcal{S}}$, $\sigma_{\mathcal{E}^F|\mathcal{S}}$, and $\delta_{\mathcal{E}^F|\mathcal{S}}$ are obtained using the following averages:

$$\mu_{\mathcal{E}^F|\mathcal{S}} = \langle \mathcal{E}^{F^e} \rangle_e, \tag{9a}$$

$$\sigma_{\mathcal{E}^F|\mathcal{S}} = \langle (\mathcal{E}^{F^e} - \mu_{\mathcal{E}^F|\mathcal{S}})^2 \rangle_e^{1/2}, \tag{9b}$$

$$\mathcal{E}_{\min|\mathcal{S}} = \min_e(\mathcal{E}^{F^e}), \quad \text{and} \tag{9c}$$

$$\delta_{\mathcal{E}^F|\mathcal{S}} = \langle (\mathcal{E}^{F^e} - \mathcal{E}_{\min|\mathcal{S}})^2 \rangle_e^{1/2}. \tag{9d}$$

Again, all parameters are independent of observations and therefore do not rely on the assumption of the ensemble being perfectly reliable.

Now we need to express all model parameters of Eq. (9) in terms of the spread $\mathcal{S}$. In general this must be

TABLE 2. For different combinations of spread and error metrics the parameters of the spread-based models in section 2 are specified. The approach for their calculation is given in the section 2b. In addition, the regression coefficients α_{perf} and β_{perf} of the log-transformed regression as defined in Eq. (20) and corresponding to a perfectly reliable ensemble forecast are given. Member e of the ensemble forecast is denoted as F^e, $\overline{F} = \langle F^e \rangle_e$ is the ensemble mean (ME), and the observation is O. Results that are independent of any hypothesis on the nature of the ensemble distribution are denoted with check mark (✓) while results that are considered "not applicable" are denoted as N/A. Note that $\gamma = 0.5772$ is the Euler–Mascheroni constant.

Metric	Error metric $\mathcal{E}^O$	Spread metric $\mathcal{S}$	$\mu_{\mathcal{E}^F\|\mathcal{S}}$		$\sigma_{\mathcal{E}^F\|\mathcal{S}}$		$\mathcal{E}_{\min\|\mathcal{S}}$		$\delta_{\mathcal{E}^F\|\mathcal{S}}$		α_{perf}		β_{perf}	
EM-RAW	$\mathcal{E}_0^O = O - \overline{F}$	$\mathcal{S}_0 = \langle (F^e - \overline{F})^2 \rangle_e$	0	✓	$\mathcal{S}$	✓	N/A		N/A		N/A		N/A	
EM-SQU	$\|O - \overline{F}\|$	$\langle (F^e - \overline{F})^2 \rangle_e^{1/2}$	$\sqrt{\frac{2}{\pi}}\mathcal{S}$		$\sqrt{1 - \frac{2}{\pi}}\mathcal{S}$		0	✓	$\mathcal{S}$	✓	0.5298		1	✓
AV-SQU	$\langle (O - F^e)^2 \rangle_e^{1/2}$	$\langle (F^{e_1} - F^{e_2})^2 \rangle_{e_1,e_2}^{1/2}$	$0.9577\mathcal{S}$		$0.2874\mathcal{S}$		$\frac{\mathcal{S}}{\sqrt{2}}$		$0.381\mathcal{S}$		0.9232		1	✓
EM-ABS	$\|O - \overline{F}\|$	$\langle \|F^e - \overline{F}\| \rangle_e$	$\mathcal{S}$	✓	$\sqrt{\frac{\pi}{2} - 1}\mathcal{S}$		0	✓	$\sqrt{\frac{\pi}{2}}\mathcal{S}$		0.6640		1	✓
AV-ABS	$\langle \|O - F^e\| \rangle_e$	$\langle \|F^{e_1} - F^{e_2}\| \rangle_{e_1,e_2}$	$\mathcal{S}$	✓	$0.3575\mathcal{S}$		$\frac{\mathcal{S}}{\sqrt{2}}$		$0.462\mathcal{S}$		0.9514		1	✓
EM-GEO	$\|O - \overline{F}\|$	$\exp\langle \ln(\|F^e - \overline{F}\|) \rangle_e$	$\frac{2e^{\gamma/2}}{\sqrt{\pi}}\mathcal{S}$		$\sqrt{2e^{\gamma}\left(1 - \frac{2}{\pi}\right)}\mathcal{S}$		0	✓	$\sqrt{2e^{\gamma}}\mathcal{S}$		1	✓	1	✓
AV-GEO	$\exp\langle \ln(\|O - F^e\|) \rangle_e$	$\exp\langle \ln(\|F^{e_1} - F^{e_2}\|) \rangle_{e_1 \neq e_2}$	$1.083\mathcal{S}$		$0.5186\mathcal{S}$		$\frac{\mathcal{S}}{\sqrt{2}}$		$1.201\mathcal{S}$		1	✓	1	✓

done by introducing an assumption for the distribution of the ensemble. However, for some combinations of parameters and uncertainty forecasts, such additional assumptions are not required. This is clarified here for the EM-ABS and AV-ABS uncertainty forecasts.

For the uncertainty forecast associated with the EM-ABS spread–error metric (with $\mathcal{E}^{F^e} = |F^e - \overline{F}|$), the spread is

$$\mathcal{S} = \langle |F^e - \overline{F}| \rangle_e . \tag{10}$$

Two quantities of Eq. (9) can then be easily expressed in terms of the spread $\mathcal{S}$,

$$\mu_{\mathcal{E}^F|\mathcal{S}} = \langle |F^e - \overline{F}| \rangle_e = \mathcal{S} \quad \text{and} \tag{11a}$$

$$\mathcal{E}_{\min|\mathcal{S}} = 0, \tag{11b}$$

without any assumptions on the ensemble distribution. Since $\mu_{\mathcal{E}^F|\mathcal{S}}$ and $\mathcal{E}_{\min|\mathcal{S}}$ are the only variables required for model III, the use of this model is proposed when using the EM-ABS uncertainty measure (see Table 1).

For the AV-ABS spread–error metric (with $\mathcal{E}^{F^e} = \langle |F^e - F^{e_1}| \rangle_{e_1}$), the spread is

$$\mathcal{S} = \langle |F^{e_2} - F^{e_1}| \rangle_{e_1,e_2}, \tag{12}$$

and one parameter is easily expressible in terms of spread:

$$\mu_{\mathcal{E}^F|\mathcal{S}} = \langle |F^{e_2} - F^{e_1}| \rangle_{e_1,e_2} = \mathcal{S}. \tag{13}$$

The computation of additional parameters is, however, required in order to use one of the models (I, II, and III) and needs an additional assumption. More specifically, the average $\langle \cdot \rangle_e$ is evaluated assuming the ensemble members are Gaussian distributed around the ensemble mean $\overline{F}$:

$$F^e \sim \mathcal{N}(\overline{F}, \sigma_F^2). \tag{14}$$

Based on this assumption, the other parameters for AV-ABS are

$$\mathcal{S} = \langle |F^{e_2} - F^{e_1}| \rangle_{e_1,e_2} = 2\sigma_F/\sqrt{\pi}, \tag{15a}$$

$$\sigma_{\mathcal{E}^F|\mathcal{S}} = \langle (\langle |F^{e_2} - F^{e_1}| \rangle_{e_1} - \mathcal{S})^2 \rangle_{e_2}^{1/2} = 0.403\sigma_F, \quad \text{and} \tag{15b}$$

$$\mathcal{E}_{\min|\mathcal{S}} = \langle |\overline{F} - F^e| \rangle_e = \sqrt{2/\pi}\sigma_F . \tag{15c}$$

Expression (15a) allows one to eliminate σ_F in the other equations such that the results in Table 2 are obtained. All parameters of Eq. (9) are found to be linear with respect to the spread. Note that, in Table 2, universal quantities that are independent of the assumption of Gaussianity are denoted with a check mark and the spread–error metrics of Table 1 are restricted to those for which models possess universal parameters.

3. Verification of the uncertainty forecast

a. *Verification of the uncertainty forecast based on spread*

As emphasized, the uncertainty forecast must be considered probabilistic rather than deterministic, as its verification. Strictly proper probabilistic scores exist and should be preferred. Well-known examples include the Brier score (Wilks 2011), the log-likelihood, the logarithmic score (LS; Roulston and Smith 2002; Benedetti 2010), the quantile score (QS; Gneiting and Raftery 2007; Bentzien and Friederichs 2014), and the continuous ranked probability score (CRPS). For a verification set, $\{\mathcal{S}_n, \mathcal{E}_n^O\}_n$, and the spread-based probabilistic uncertainty forecasts, the last two scores are suggested here for spread–error assessment:

$$\text{CRPS} = \left\langle \int [\mathcal{P}(\mathcal{E}_n^F < t \,|\, \mathcal{S}_n) - \Theta(t - \mathcal{E}_n^O)]^2 \, dt \right\rangle_n \quad \text{and} \tag{16a}$$

$$\text{QS}(q) = \langle [\Theta(\mathcal{E}_n^q - \mathcal{E}_n^O) - q](\mathcal{E}_n^O - \mathcal{E}_n^q) \rangle_n , \tag{16b}$$

with Θ the Heaviside function, $\langle \cdot \rangle_n$ the average over the verification set, and the quantity $\mathcal{E}_n^q$ defined by $\mathcal{P}(\mathcal{E}^F > \mathcal{E}_n^q \,|\, \mathcal{S}_n) = q$ for the quantile q. The QS for the 90th quantile is therefore QS(0.90) and quantifies the forecast quality of the events that have a probability of 10% to occur. The substitution of the three spread-based models [Eqs. (5), (8a), and (8b)] allows us to derive simple analytic expressions for QS and CRPS as given in the appendix.

Note that the reliability of the uncertainty forecast can be probed, for instance, by considering the reliability component of the CRPS (Hersbach 2000) or the histograms of the probability integral transform (PIT; Gneiting et al. 2007).

b. *Comparison of full-ensemble and spread-based uncertainty forecasts*

The skill of the spread-based uncertainty models can now be compared with that of the full ensemble uncertainty. Such comparison is useful in estimating the information loss caused by reducing the full-ensemble uncertainty information to one quantity, the spread, using the models proposed in section 2. One may expect that the ensemble distributions that are skewed or have high kurtosis, for instance wind speed, are not appropriately characterized by spread only. The modeling of

such higher moments, however, is crucial for meteorological purposes since it determines the tails of the ensemble distributions that may pertain to rare but high-impact events. On the other hand, the uncertainty estimate of ensembles of small size may be improved using a spread-based uncertainty forecast as the latter provides an overall smoother forecast distribution, as well as including nonzero probabilities for the more unlikely events.

To compare the full-ensemble and spread-based uncertainty forecasts, the use of skill scores is proposed. The skill scores associated with the scores of (16) are

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}_{\mathrm{full}}}{\mathrm{CRPS}_{\mathcal{S}}} \quad \text{and} \tag{17a}$$

$$\mathrm{QSS}(q) = 1 - \frac{\mathrm{QS}_{\mathrm{full}}(q)}{\mathrm{QS}_{\mathcal{S}}(q)}. \tag{17b}$$

Here, the subscript $\mathcal{S}$ denotes the spread-based uncertainty forecast while "full" denotes the full-ensemble score. Statistically significant positive (negative) values for the skill score indicate that the full ensemble improves (degrades) upon the spread-based model. Note that it is well known that QS(q) integrated over all quantile values q amount to the CRPS. The quantile score can thus be used to analyze the origin of CRPS changes.

The skill-score formalism as presented here analyzes if, *with respect to a particular score*, spread alone statistically improves upon or degrades the full ensemble for the uncertainty estimation. Of course, many aspects of such information reduction depend on the application, as well as on the user, but as long as these aspects can be quantified using scores, this formalism applies. The skill scores in Eq. (17) will be tested in section 5a for different ensemble sizes and for different meteorological variables including wind speed.

4. Spread as a predictor of uncertainty

a. Introduction

So far, an approach has been proposed to assess the overall forecast quality and its reliability based on spread and error only. However, a forecast distribution that corresponds to the climatological distribution satisfies perfect reliability and a good ensemble forecast is expected to provide additional information. Hence, a more advanced aspect of ensemble forecasting, called resolution, concerns whether the flow-dependent variations of spread are indicative of the variations of the uncertainty. Another useful ensemble system is one for which the spread is systematically twice the spread of a perfectly spread-consistent forecast system. This feature entails that the spread is a perfect predictor for the true uncertainty (their correlation is one), even though the spread–error scores as introduced in last section may not be good. Note that in such a case the scores could be improved by ensemble-spread calibration techniques along the lines of Gneiting et al. (2005) and Van Schaeybroeck and Vannitsem (2015).

A seemingly straightforward approach to testing whether the spread is a good predictor for the uncertainty is through the computation of the Pearson correlation between spread and error. However, since, even for a perfectly reliable ensemble system, the spreads and errors on a scatter diagram do not align, the associated Pearson correlation is not equal to one (Whitaker and Loughe 1998). The correlation coefficient also appears as the most important quantity for ordinary least squares (OLS) regression. Applied on spread and error, however, OLS regression is statistically unjustified since it assumes homoscedasticity. The latter means that the scatter variance around the regression line is constant. For a perfectly reliable and therefore also for a perfectly spread-consistent system, this would imply $\sigma_{\mathcal{E}^F|\mathcal{S}}$ is independent of $\mathcal{S}$, which is never the case.

b. Modeling spread as predictor of uncertainty

Veenhuis (2013), in an attempt to suppress the problem of heteroscedasticity between spread and error, applies a square root "variance-stabilizing transformation" of spread and error, followed by an OLS regression. Although such an approach results in more "Gaussian distributed" marginal statistics, it is easily proven that heteroscedasticity with a noise that is proportional to the predictor (here $\sigma_{\mathcal{E}^F|\mathcal{S}} \propto \mathcal{S}$) can only be overcome by a logarithmic transformation (Draper and Smith 1998, p. 292; Box et al. 2005, p. 321; Chatterjee and Hadi 2006, p. 168). The transformation to homoscedasticity can be heuristically understood by considering the logarithm of the forecast error decompositions Eqs. (4) and (7), using the fact that $\mu_{\mathcal{E}^F|\mathcal{S}} \propto \mathcal{S}$, $\mathcal{E}_{\min|\mathcal{S}} \propto \mathcal{S}$, and $\varepsilon(\mathcal{S}) = \varepsilon_0 \mathcal{S}$ with ε_0 a random term of order one. Both yield

$$\ln(\mathcal{E}^F) = \ln(\mathcal{S}) + \ln(C + \varepsilon_0). \tag{18}$$

Here, C is a constant, independent of the spread. The last term can be considered the new noise term, which now is *independent of the spread*. Therefore, the logarithmically transformed spread and error are homoscedastic against one another. This is also apparent in Fig. 2, which is simply the log–log equivalent of Fig. 1b. Note that the logarithmic transformation can only be applied to errors that are always positive and it is therefore applicable to all metrics of Table 2 except for EM-RAW.

The homoscedasticity resulting from the logarithmic transformation justifies the use of OLS regression. A

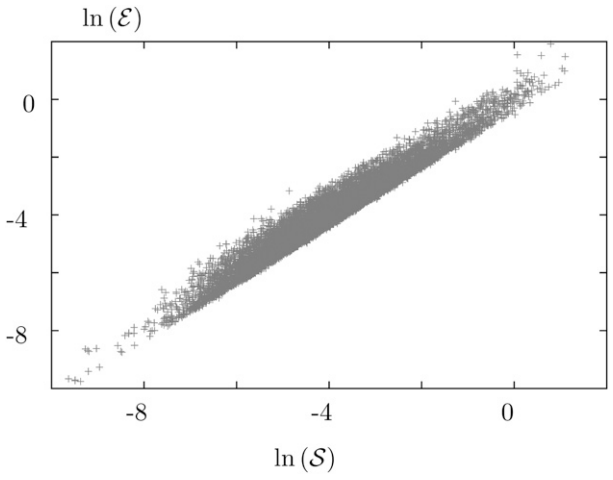

FIG. 2. As in Fig. 1b, but after logarithmic transformation of the AV-ABS spread and error [see Eq. (6)]. The linear relationship is clearly visible.

good estimator for the error, $\hat{\mathcal{E}}^O$, must then satisfy the following relation:

$$\hat{\mathcal{E}}^O = \alpha \mathcal{S}^\beta \quad \text{such that} \quad \ln(\hat{\mathcal{E}}^O) = \beta \ln(\mathcal{S}) + \ln(\alpha). \tag{19}$$

OLS analysis leads to

$$\beta = \frac{\langle \ln(\mathcal{E}_n^O) \ln(\mathcal{S}_n) \rangle_n - \langle \ln(\mathcal{E}_n^O) \rangle_n \langle \ln(\mathcal{S}_n) \rangle_n}{\langle [\ln(\mathcal{S}_n)]^2 \rangle_n - \langle \ln(\mathcal{S}_n) \rangle_n^2} \quad \text{and} \tag{20a}$$

$$\alpha = \langle \ln(\mathcal{E}_n^O) \rangle_n - \beta \langle \ln(\mathcal{S}_n) \rangle_n. \tag{20b}$$

Note that both a climatological forecast and a forecast that is dressed using past error statistics suffer from a lack of variability of the spread. This yields a vanishing denominator in Eq. (20a) and therefore a diverging $|\beta|$, and Eq. (20) can not be used. Generally, α and β should be compared to α_{perf} and β_{perf} associated with a perfectly spread-consistent forecast with nonzero spread variability. For different spread–error metrics the latter are listed in Table 2. For all metrics $\beta_{\text{perf}} = 1$ while values for α_{perf} are derived using the assumption of Gaussian ensembles [Eq. (14)]. Our approach is able to go beyond such assumptions through the use of the geometric-mean spread–error metrics:[1]

$$\begin{aligned} &\text{EM} - \text{GEO}: \mathcal{E}^O = |O - \overline{F}| \quad \text{and} \quad \mathcal{S} = \exp\langle \ln(|F^e - \overline{F}|) \rangle_e, \\ &\text{AV} - \text{GEO}: \mathcal{E}^O = \exp\langle \ln(|O - F^e|) \rangle_e, \quad \text{and} \\ &\qquad \mathcal{S} = \exp\langle \ln(|F^{e_1} - F^{e_2}|) \rangle_{e_1 \neq e_2}. \end{aligned} \tag{21}$$

The perfect-model regression parameters for these are independent of any hypothesis on the nature of the ensemble distribution and satisfy $\alpha_{\text{perf}} = \beta_{\text{perf}} = 1$. Therefore, if there is sufficient day-to-day variation of the ensemble spread, the calculation of α and β can be used to quantify the predictive power of the spread as a measure of uncertainty. And, even more importantly, it allows for evaluating the quality of any ensemble system as compared to a perfectly spread-consistent ensemble as will be demonstrated in the next section. For systems that are close to perfectly spread consistent the heteroscedasticity is expected to be present, so Eq. (19) provides a statistically firm relation, to estimate a spread–error relationship.

Note that even for a perfectly spread-consistent ensemble, the random term of order one that appears in Eq. (18) may be non-Gaussian. Additional transformations may improve Gaussianity (Yeo and Johnson 2000; Draper and Smith 1998) but care should be taken that this is effective for all values of the spread.

5. Application on ECMWF EPS

The spread–error assessments are now applied to forecasts for 500-hPa geopotential height (denoted Z500), 500-hPa wind velocity (V500), and forecasts of wind velocity at 10 m (V10m) over Europe based on the ECMWF's EPS. These are compared with their analysis. Daily forecasts (at 0000 UTC) from 20 June 2012 until 19 March 2013 (240 cases), during which no model change was performed and with lead-times intervals of 12 h, are considered. Note also that a resolution change occurs after the 10-day lead time.

The choice of variables allows us to study the potential impact of statistical characteristics and predictability on the theory proposed in section 2. Indeed, upper-air variables (Z500 and V500) are known to have much better predictability features than surface variables (V10m). In addition, the statistical characteristics of Z500 are expected to be well represented using Gaussian statistics, in contrast to wind speed (V500 and V10m).

a. *Verification of spread-based models*

Figure 3 shows the CRPSSs [see Eq. (17a)] against lead time for different meteorological variables (different columns: Z500, V500, and V10m) and different

[1] For an ensemble of discrete size E, the AV-GEO and EM-GEO spreads and errors can be reduced to a product over the ensemble members. For example, $\mathcal{E}^O = \prod_e |O - F^e|^{1/E}$ for AV-GEO. The geometric mean of a series of E numbers $\{x_1, \ldots, x_E\}$ may be considered to be the limit for $q \to 0$ of the Hölder mean $(\langle |x_e|^q \rangle_e)^{1/q}$ by means of which, using $q = 1$ and $q = 2$, all averages used in Table 1 could be rewritten (Abramowitz and Stegun 1972).

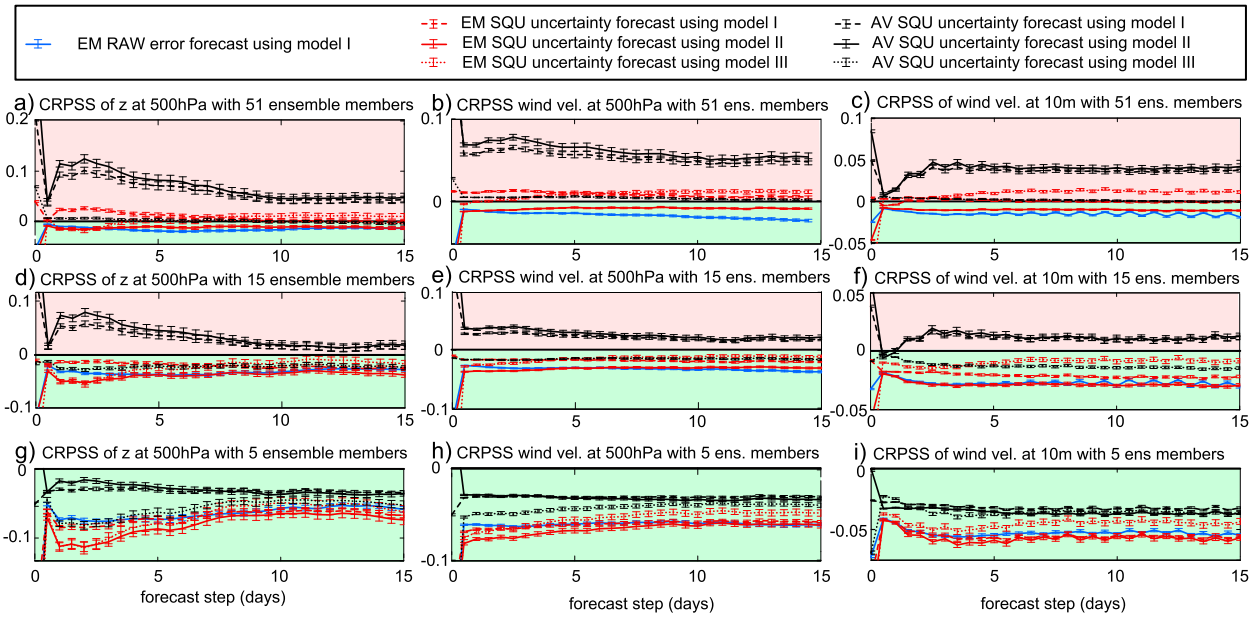

FIG. 3. CRPSS against lead time of the uncertainty forecast for different meteorological variables (columns), different ensemble sizes (rows), and different spread-based uncertainty forecasts. The ECMWF EPS is used over Europe. The green-shaded area indicates values where the spread-based uncertainty forecast improves upon the full-ensemble uncertainty forecast. The variables are (a),(d),(g) geopotential height at 500 hPa; (b),(e),(h) wind speed at 500 hPa; and (c),(f),(i) wind speed at 10 m. The considered ensemble sizes are (a)–(c) 51, (d)–(f) 10, and (g)–(i) 5 members. The uncertainty forecasts (EM-RAW, EM-SQU, and AV-SQU) are associated with different spread–error metrics as defined in Table 2. The full-ensemble forecast associated with each metric is compared with the three spread-based models specified in Eqs. (5), (8a), and (8b). The latter make use of the analytic expressions for the CRPS, as given in the appendix. An average is taken over 240 consecutive forecast days and the 95% confidence intervals are also provided.

ensemble sizes (different rows: 51, 15, and 5 ensemble members). In each panel in Fig. 3 three uncertainty forecasts (EM-RAW, EM-SQU, and AV-SQU) are also considered to be associated with different spread–error metrics, each of which is modeled using spread-based uncertainty forecasts. More specifically, EM-RAW is modeled using model I (blue line) while all models (I, II, and III) are used for EM-SQU (red lines) and AV-SQU (black lines). The full-ensemble uncertainty forecast improves upon the spread-based uncertainty forecast for positive CRPSS (shaded in red). The analytic expressions in Eq. (A6) were used to calculate the CRPS of the spread-based models.

Surprisingly, almost all conclusions that can be drawn from Fig. 3 are independent of the lead time and, most importantly, of the variable under consideration. These conclusions are as follow:

1) For all cases including the ones with a five-member ensemble, the CRPS skill improvement or degradation is at most 10%. This indicates that the uncertainty information results of the spread-based models and the full ensemble are largely comparable. Skill changes among the full-ensemble forecasts due to changes of the ensemble sizes are usually larger (not shown).
2) For EM-RAW the spread-based model I shows an improvement for all variables and ensemble sizes. Therefore replacing each ensemble by a Gaussian distribution with its corresponding mean and variance results in an improvement in the CRPS score.
3) For each uncertainty forecast at least one spread-based model is able to improve upon or be marginally worse than the full-ensemble forecast. For the EM-SQU uncertainty, forecast model II leads to the best results. The quality of this model is related to the fact that the associated model parameters are independent of any hypothesis on the nature of the ensemble distribution, as indicated in Table 1. For AV-SQU, model III is far better than models I or II and induces only a marginal skill loss upon use of 51 ensemble members. Note that although the errors of EM-SQU and AV-SQU are always positive, model I allows for negative values. This may explain why model I is never the best model for the AV-SQU and EM-SQU uncertainties. Results for EM-ABS and AV-ABS are not shown but were very similar to those of EM-SQU and AV-SQU, respectively. Therefore, it is concluded that for the ensemble-mean-based error metrics (except for EM-RAW) the best spread-based model is model II. For all

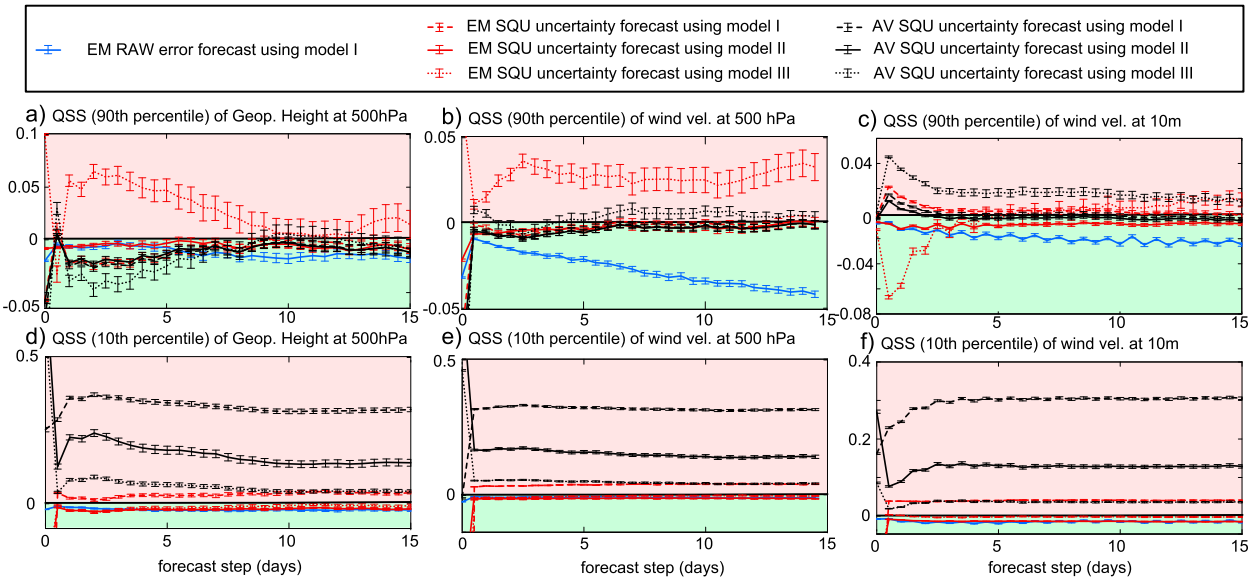

FIG. 4. QSSs for different percentiles (rows) against lead time of the uncertainty forecast for different meteorological variables (columns) and different spread-based uncertainty forecasts. The 51-member ECMWF EPS is used over Europe. The green-shaded area indicates values where the spread-based uncertainty forecast improves upon the full-ensemble uncertainty forecast. The variables are (a),(d) geopotential height at 500 hPa; (b),(e) wind speed at 500 hPa; and (c),(f) wind speed at 10 m. The QSSs of the (a)–(c) 90th and (d)–(f) 10th percentiles are shown. The uncertainty forecasts (EM-RAW, EM-SQU, AV-SQU, and AV-GEO) are associated with different spread–error metrics, as defined in Table 2. The full-ensemble forecast associated with each metric is compared with the three spread-based models specified in Eqs. (5), (8a), and (8b). The latter make use of analytic expressions for the QS, as given in the appendix. An average is taken over 240 consecutive forecast days and the 95% confidence intervals are also provided.

ensemble-average error metrics (indicated with "AV") the use of model III is advised.

4) All spread-based models improve upon the full-ensemble forecast for five-member ensembles as a result of the smoother representation of the ensemble distribution by the spread-based model.

Figure 4 shows the QSSs as given in Eq. (17b) for the same variables (columns: Z500, V500, and V10m), spread–error metrics, and associated models as in Fig. 3. The first and second rows in Fig. 4 are for the 10th and 90th percentiles, respectively. All 51 ensemble members are used. The main conclusions that were drawn from Fig. 3 remain valid for Fig. 4: for EM-RAW the spread-based model I shows an overall improvement for all variables and quantiles while the same is valid for model II applied on the EM-SQU uncertainty forecast. Finally, model III is the best model for the AV-SQU uncertainty forecast.

However, differences are present between the QSSs for the different quantiles and the different variables. Good skill scores are found for all models for the 90th percentile and thus for predicting events that are unlikely according to the forecast. Exceptions exhibiting minor skill decreases include model III applied on the EM-SQU forecast of Z500 and V500 and model III applied on the AV-SQU forecast of V10m. This indicates that at least for the cases considered, the reduction of uncertainty information to one quantity (spread) does not strongly affect the prediction of events that are unlikely according to the forecast. The spread-based model I applied on the EM-RAW uncertainty forecast even gradually improves the full-ensemble forecast as a function of lead time. The fact that models I and II strongly deteriorate the overall AV-SQU forecast quality (see Fig. 3) is a result of from their bad ensemble predictions of small errors (10th percentile). As mentioned before, model I allows for negative values while the EM-SQU and AV-SQU uncertainties are always positive, and it is therefore worse than models II and III, as well as the full-ensemble forecast.

Note that again the results for EM-ABS and AV-ABS are not shown but were very similar to those for EM-SQU and AV-SQU, respectively.

The 10th and 90th percentiles pertain to the ensemble distribution of the error forecast, which is expected to be asymmetric—except for EM-RAW—and are therefore skewed because of the positive character of the error. Therefore, even for a variable that is expected to be described reasonably well using Gaussian statistics (Z500), the results for the 10th and 90th percentiles

strongly differ. For EM-RAW the results at the 10th and 90th percentiles are similar for Z500 because of the symmetric character of the ensemble distributions, but different for the wind variables for which skewed distributions are still present.

To summarize, the use of spread as the only measure of uncertainty, combined with the spread-based models in section 2, does not lead to a strong uncertainty information loss. Replacing each ensemble by a Gaussian distribution with corresponding mean and variance yields a small but significant improvement in the statistical scores. A similar result, based on the Brier score, was obtained by Atger (1999) for the ECMWF EPS. Also, model II applied on the uncertainty forecasts involving ensemble-mean errors (EM metrics) systematically improves the full-ensemble uncertainty forecasts, even for 51 members. Model III is suitable for uncertainty forecasts involving ensemble-average errors (AV metrics) but overall induces weak skill degradation.

b. *Verification of spread as a predictor of uncertainty*

The spread–error relation measured with the regression parameters as specified in Eqs. (20a) and (20b) can now be evaluated for the spread of the ECMWF EPS. Each panel in Fig. 5 shows the coefficients $\ln(\alpha)$ and β for the EPS as well as for a ''persistence'' forecast over Europe. The persistence forecast is constructed by subtracting from each ensemble member at each lead time a constant value such that its ensemble mean is the one of the forecast at lead time zero. Therefore, only the ensemble mean and not the ensemble spread is affected. The different columns in Fig. 5 correspond to different variables (Z500, V500, and V10m) while the different rows are for different spread–error metrics (EM-SQU, AV-SQU, and AV-GEO). Also shown are the regression values corresponding to a perfectly reliable forecast $\ln(\alpha_{\text{perf}})$ (black dashed line) and β_{perf} (black full line) as given in Table 2. Note that the proximity with respect to these lines is a measure of spread consistency. The 95% confidence interval is determined for each point by performing 1000 block bootstraps with each block consisting of a week of consecutive forecast days.

It is found that for a fixed variable the qualitative behavior of the regression coefficients pertaining to different spread–error metrics is very similar. This indicates that the results are independent of the assumption of Gaussianity [Eq. (14)]. For instance for Z500 (Figs. 5a,d,g), α and β for all spread–error metrics are far from α_{perf} and β_{perf}, respectively, before 48-h lead time. This is to be expected, as the initial perturbations (lead time zero) of the EPS consist of perturbed analyses using the ECMWF ensemble data assimilations together with initial singular vectors that maximize a total energy norm at 48 h (Buizza et al. 2010). Therefore, the uncertainty estimates are expected to be representative of the true ones only after 2 days. Indeed, for Z500, α and β are stationary and very close to α_{perf} and β_{perf} between 2- and 7-day lead times. For longer lead times the values of α and β depart again from α_{perf} and β_{perf}. After day 11, the 95% confidence intervals for β of the EPS and the persistence forecasts overlap one another.

As can be seen in Figs. 5b, 5e, and 5h, the evolution of the coefficients characterizing consistency for wind velocity at 500 hPa qualitatively differs from that of Z500. Good spread consistency is visible already at 12 h and persists up to a lead time of 5 days, after which time a progressive degradation sets in. As opposed to the upper-air variables (Z500 and V500), the wind velocity at 10 m (see Figs. 5c,f,i) is much less predictable, which is visible through the close proximity of β for the EPS and the one of the persistence for all spread–error metrics. Remarkably, V10m becomes more and more consistent as the lead time increases.

A general feature for all variables and spread–error metrics is the fact that α is generally larger than α_{perf} and β is smaller than $\beta_{\text{perf}} = 1$. Care must be taken, however, to interpret correctly this result. Since $\hat{\mathcal{E}} = \alpha\mathcal{S}^{\beta}$ is assumed, $\alpha > \alpha_{\text{perf}}$ naturally implies that the ensemble system is underdispersive (overconfident) for very small values of $\mathcal{S}$ while $\beta < \beta_{\text{perf}}$ implies overdispersion for large $\mathcal{S}$. However, such conclusions are meaningless if very small or very large values of the spread are not realized. This occurs for instance at long lead times when small spread values are generally not encountered. Nevertheless, $\alpha > \alpha_{\text{perf}}$ together with $\beta < \beta_{\text{perf}}$ means that ensembles of small spread are more confident than those of large spread. Additional conclusions require at least partial knowledge of the distribution of spread and, as illustrated in next section, the binning approach can help us in understanding the dependence on the spread–error relationship.

Let us now focus specifically on the spread consistency of Z500. Strong deviations from the perfectly spread-consistent behavior still appear at 24 h when spread and error are related by $\hat{\mathcal{E}}^{O} \approx 1.5\mathcal{S}^{0.8}$, with underdispersion of about half of the ensembles (the ones with the smallest spread) and realistic spread consistency for the other half (not shown). Between days 2 and 7 the spread and error are related by $\hat{\mathcal{E}}^{O} \approx 1.3\mathcal{S}^{0.95}$. Therefore, the linearity is very strong and implies near-perfect spread consistency. This relation indicates that the ensembles are slightly underdispersive, whatever their spread.

c. *Impact of ensemble size*

Limited ensemble sizes are known to have a strong impact on spread–skill measures (Grimit and Mass 2007;

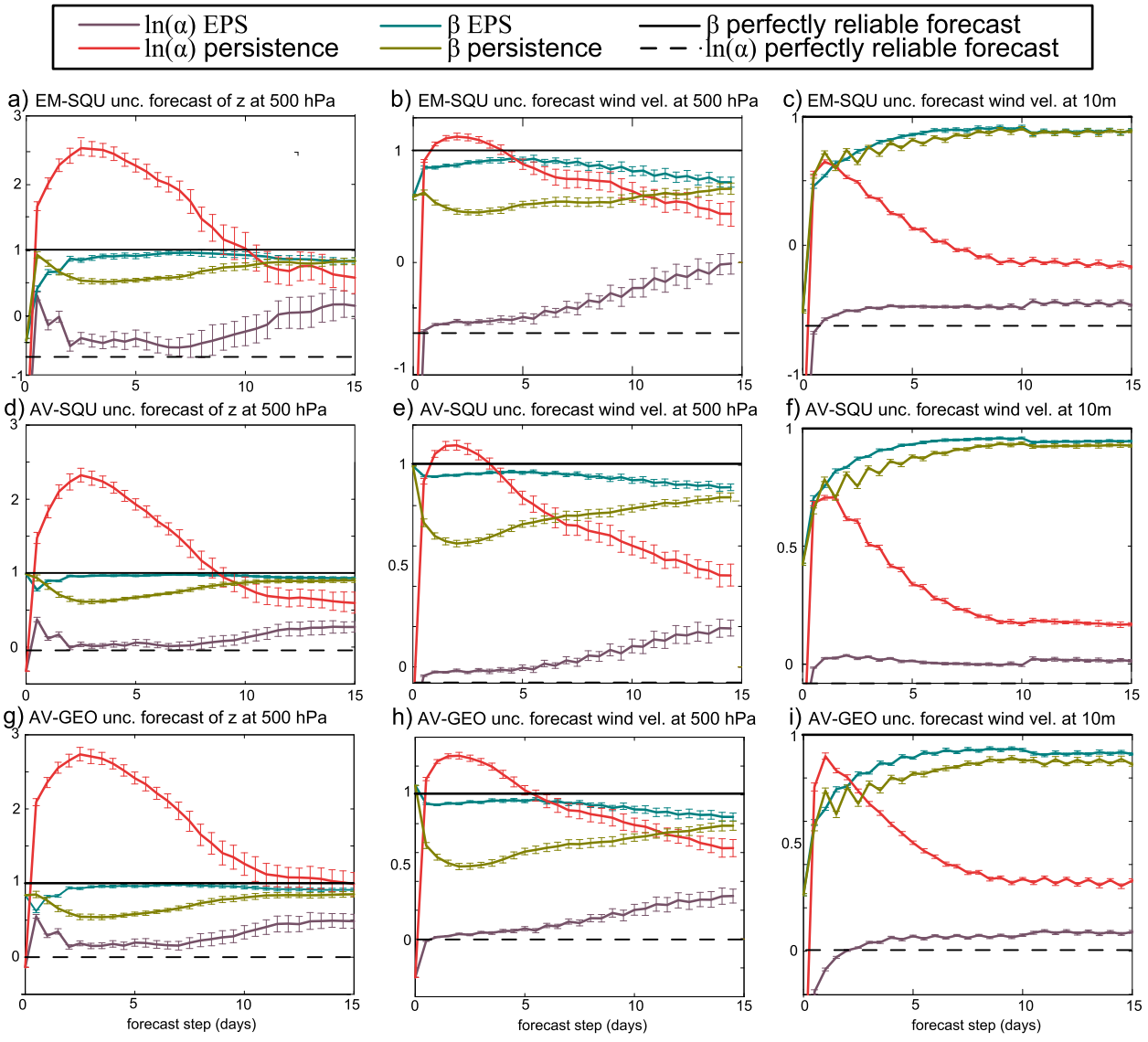

FIG. 5. Regression parameters β and $\ln(\alpha)$ as given in Eqs. (20a) and (20b) for forecasts of different meteorological variables (columns) and different spread–error measures (row). The ensemble forecasts analyzed are the EPS and a persistence forecast (see text for its definition) both over Europe and taken from the ECMWF. The variables are (a),(d),(g) geopotential height at 500 hPa; (b),(e),(h) wind speed at 500 hPa; and (c),(f),(i) wind speed at 10 m. The considered spread–error metrics are (a)–(c) EM-SQU, (d)–(f) AV-SQU, and (g)–(i) AV-GEO, as defined in Table 2. The values for β and α are associated with a perfectly reliable forecast system and are shown with black full and black dashed lines, respectively. All scores are based on a period of 240 consecutive forecast days and the 95% confidence intervals are also provided.

Kolczynski et al. 2011). The spread–error relation is examined in Fig. 6 for different ensemble sizes. Each panel in Fig. 6 shows α and β as a function of lead time for the ECMWF EPS with 51 members (solid lines), 10 members (long-dashed lines), and 5 members (short-dashed lines). The influence of a change in ensemble size is independent of the spread–error metric used (Fig. 6, rows) and independent of the variable (Fig. 6, columns). More specifically a decrease of ensemble size induces a larger departure of α and β to their perfect values α_{perf} and β_{perf}. The impact on the regression coefficients of an ensemble size reduction from 51 to 10 is also very strong and induces strong deviations from the perfect-model behavior. The weakest impacts are observed for the AV-LOG uncertainty forecast of V10m. Remarkably, a change of ensemble size does not strongly affect the 95% confidence intervals and can therefore be seen as mostly a biasing effect.

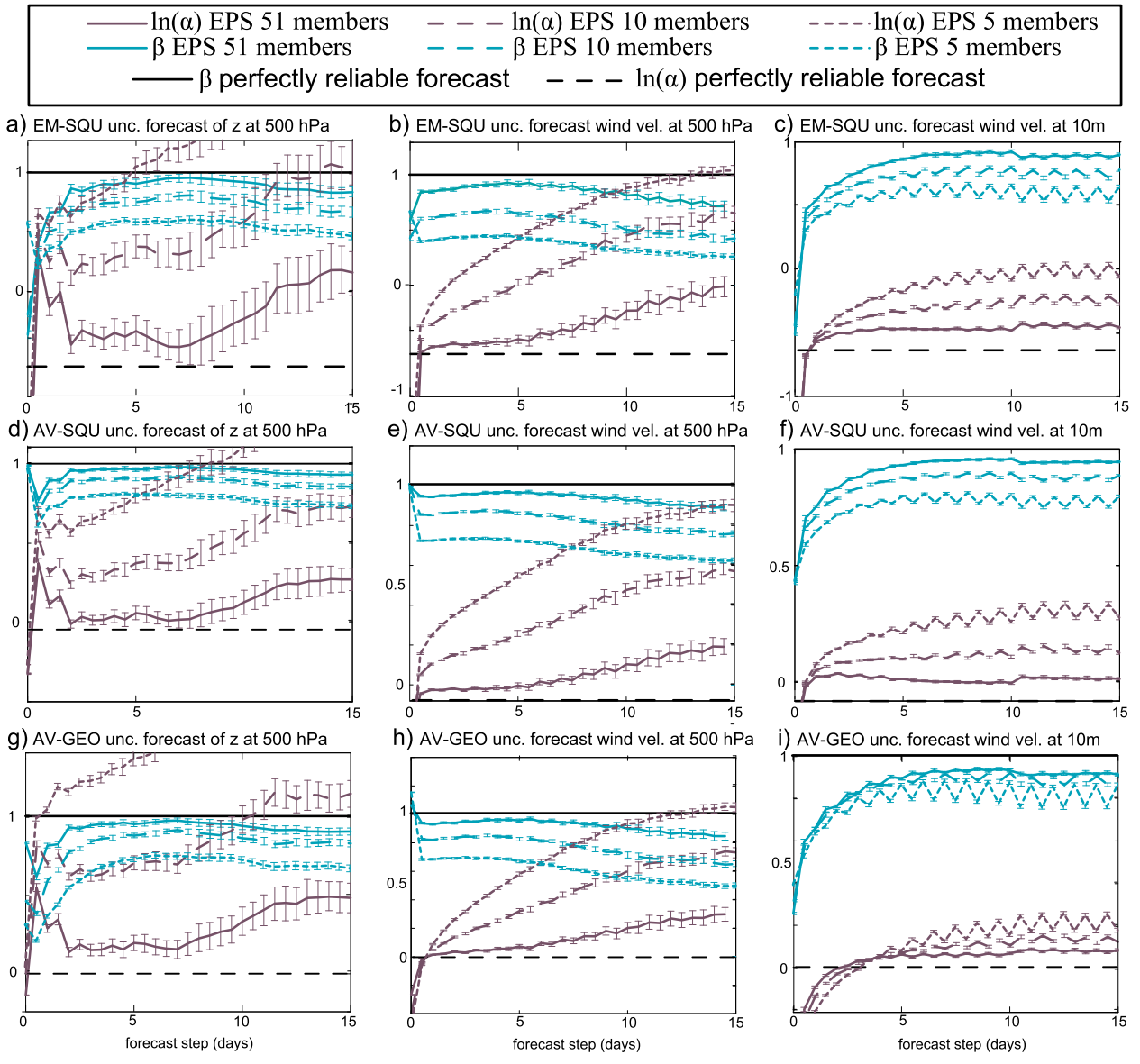

FIG. 6. As in Fig. 5, but without the persistence forecast and comparing ensembles of 5, 10, and 51 members.

d. Comparison with a binning approach

As specified before, for perfectly reliable ensemble forecasts the error for a forecast of fixed ensemble spread may vary over a wide range of values proportional to the spread. Approaches that partially avoid the problem of heteroscedasticity are the categorization of forecasts depending on the state (Ziehmann 2001; Toth et al. 2001), or binning, and averaging over errors with comparable spread (Leutbecher 2009; Wang and Bishop 2003; Kolczynski et al. 2011; Grimit and Mass 2007).

Figure 7 uses the approach of binning for 500-hPa geopotential height and we consider how such an approach can supplement and support our findings on the spread–error relationship. At a fixed lead time, the entire verification set $\{\mathcal{S}_n, \mathcal{E}_n^O\}_n$ is divided into $N = 100$ equally sized bins of comparable spread. Therefore, the errors in each bin can be considered to constitute the error distribution conditional on the spread $\mathcal{E}^O \,|\, \mathcal{S}$. To detect ensemble underdispersion or overdispersion for a fixed ensemble spread, the most common approach is to compare the average error $\langle \mathcal{E}^O \rangle_b$ against the average spread $\langle \mathcal{S} \rangle_b$ within each bin. This is shown with the gray dots in Fig. 7a for the 2-day forecast using the EM-ABS spread–error metric; that is, $\mathcal{E}^O = |O - \overline{F}|$ and $\mathcal{S} = \langle |\overline{F} - F^e| \rangle_e$. For this metric a perfectly reliable forecast satisfies

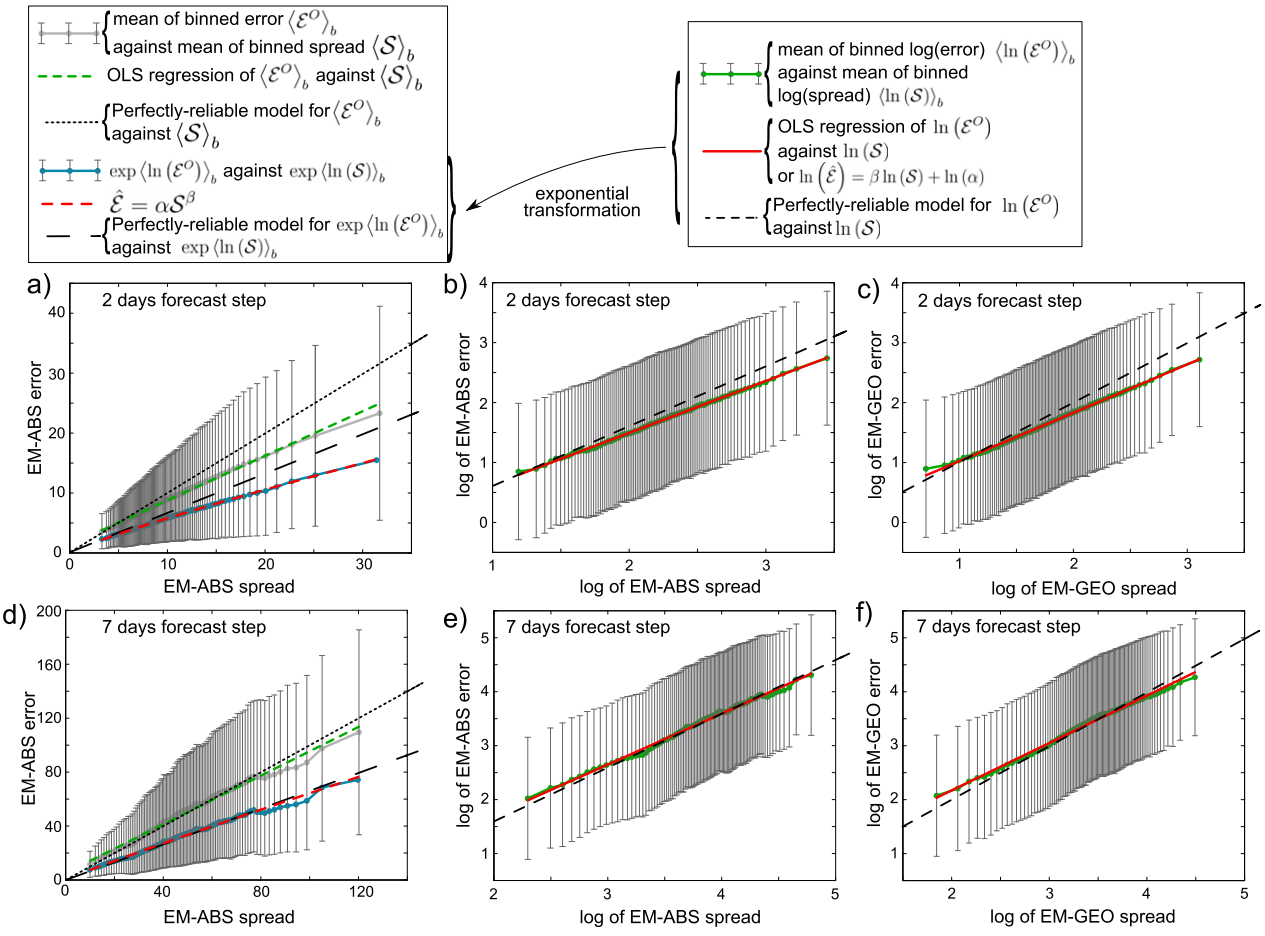

FIG. 7. (a),(d) Error against spread and (b),(c),(e),(f) logarithmically transformed error against spread for geopotential height at 500 hPa at lead times of (a)–(c) 2 and (d)–(f) 7 days. The spread–error metric used in (a),(b),(d),(e) is EM-ABS while in (c),(f) it is EM-GEO. The gray dots indicate bin averages of (100) equally sized spread and error bins with comparable spread size while the green dots (middle and right) connect averages of logarithmically transformed spread and error. Note that the full red lines are superimposed on the green dots. The intervals around the dots indicate the standard deviation of the error distribution (left) and the logarithmically transformed error distribution within the bin. The dashed green lines (left) and the solid red lines (middle and right) are OLS regressions to the gray dots and the green dots, respectively. The dashed red lines and blue dots (left row) are exponential transformations of the full red lines and the green dots of the middle row, respectively. Dotted and dashed black lines indicate the behavior corresponding with a perfectly reliable ensemble system. ECMWF ensemble forecasts over Europe of 240 consecutive forecast days are used.

$\langle \mathcal{E}^O \rangle_b = \langle \mathcal{S} \rangle_b = \mu_{\mathcal{E}^F|\mathcal{S}}$, as indicated by the black dotted line. Although the ensemble is well calibrated for small errors, it is still underdispersive for large ensemble spread. From the gray dots in Fig. 7d it is seen that, after day 7, the ensemble calibration has improved but a small underdispersion remains for the largest ensemble spread.

In Figs. 7a and 7d the intervals around the gray dots indicate the standard deviation of the error distribution within each bin. Even though the ensemble is not perfectly spread consistent, the standard deviation is still proportional to the spread and as argued before such heteroscedasticity can be overcome by a logarithmic transformation. Therefore, similar to the binning performed in Figs. 7a and 7d, the green dots in Figs. 7b and 7e show $\langle \ln(\mathcal{E}^O) \rangle_b$ and $\langle \ln(\mathcal{S}) \rangle_b$ for lead times of 2 and 7 days, respectively. The standard deviations of the logarithmically transformed error within each bin are now constant.

Diagrams of binned error against binned spread are very informative and yield a good picture of the spread–error relationship at a fixed lead time. However, this relationship must be summarized in order to be able to see its evolution as a function of lead time. A straightforward manner to do this is to apply ordinary least squares (OLS) regression of $\langle \mathcal{E}^O \rangle_b$ against $\langle \mathcal{S} \rangle_b$ and use the associated coefficients to detect overdispersion or underdispersion. In Figs. 7a and 7d, the OLS regression of the gray points is shown with a green dashed line.

Such an analysis, however, is not statistically well posed as OLS assumes homoscedasticity; that is, the error variance of the points along the regression line must be constant. However, since the error on the mean is proportional to the standard deviation of the error distribution within each bin, this condition is violated. To overcome this heterogeneity of the error variance, one option is to apply weighted linear regression; however, in case of a linear increase in the error standard deviation, such a regression might be very restrictive (Draper and Smith 1998; see note on p. 225), as the error weights are inversely proportional to the spread, which strongly emphasizes small spread. An alternative approach is the variance-stabilizing logarithmic transformation followed by a linear regression as explained in section 4b. Whereas the weighted linear regression approach requires use of the binning procedure in order to estimate the weights, linear regression can be straightforwardly applied to all logarithmically transformed data. In that sense the binning approach combined with a regression involves some loss of information, which is bypassed in the analysis of section 4.

The full red lines in Figs. 7b and 7e are obtained using OLS regression of all logarithmically transformed error and spread (not of the binned averages) and thereby establish a relationship of the form $\ln(\hat{\mathcal{E}}^O) = \beta \ln(\mathcal{S}) + \ln(\alpha)$. For the EM-ABS spread–error metric, the regression coefficients α_{perf} and β_{perf} corresponding to a perfectly reliable ensemble system are obtained using Gaussian assumptions for the ensemble distributions in Table 2 and the corresponding lines are shown by black dashed lines in Figs. 7b and 7e. Comparing these lines with the green dots leads us to exactly the same conclusions that were drawn by the comparison of the gray dots with the black dotted lines in Figs. 7a and 7d: the ensemble is underdispersive at lead times of 2 days for large spreads but overall is well calibrated at 7 days.

An exponential transformation is performed on the green dots, the full red line, and the black dashed line in Figs. 7b and 7e, and the results are shown in Figs. 7a and 7d as the blue dots, the dashed red line, and the black dashed line, respectively. Note that the dotted line relates the spread and error based on the arithmetic mean $\langle \mathcal{E}^O \rangle$ of the error, whereas the long-dashed line uses the geometric mean $\exp\langle \ln(\mathcal{E}^O) \rangle$ of the error, both of which are conditional on the spread, corresponding to a perfectly reliable ensemble system.

Conclusions drawn from binned data in Figs. 7b and 7e remain subject to doubt as a result of the fact that the line corresponding to perfect reliability associated with the EM-ABS metric makes Gaussian assumptions for the ensemble. A more rigorous approach is therefore to consider the EM-GEO or AV-GEO schemes for which the regression coefficients associated with a perfectly reliable system are $\alpha_{\text{perf}} = \beta_{\text{perf}} = 1$, independent of any hypothesis on the nature of the ensemble distribution.

Figures 7c and 7f show the same results as Figs. 7b and 7e but for the EM-GEO spread–error metric. The strong heteroscedasticity is again removed by the logarithmic transformation. On the other hand, the dashed lines in Figs. 7c and f corresponding to perfect reliability are universal and independent of ensemble assumptions. Again, the same conclusions are drawn concerning underdispersion and overdispersion at the different lead times.

Based on Fig. 7, it is concluded that a binning of spread and error is useful for studying the spread–error relationship for fixed lead times. This approach has been extended in order to allow for a quantification of the spread–error relation for different lead times. Such quantification requires certain simple statistical relationships, for which the assertions can be checked using binning. In particular, the inference of the link between binned spread and error is feasible together with a comparison of the results at different lead times. However, such an inference must necessarily take into account the strong heteroscedasticity. One option is to apply a weighted linear regression to the binned data while OLS regression could also be applied to the unbinned but logarithmically transformed data. The latter should be preferred, in particular in cases of limited data availability.

Perfectly reliable systems satisfy certain relationships between averages of binned quantities and feature strong heteroscedasticity with the standard deviations of the binned errors being linear with respect to the spread. For cases of near-perfect spread consistency, as discussed in Fig. 7, the linear relationships are modified; however, the heteroscedasticity remains similar to a large extent. Therefore, the statistical inference on logarithmically transformed data remains justified.

6. Discussion

The fact that, for a perfectly reliable ensemble forecast the observation is statistically indistinguishable from the ensemble members should be reflected in the relation between error and spread. A common way to verify the spread–error relation is by direct comparison of the *average* ensemble variance $\langle \mathcal{S}^2_{0,n} \rangle_n$ against the mean squared error of the ensemble mean $\langle (\mathcal{E}^2_{0,n})^2 \rangle_n$. For a perfectly reliable system, both must coincide (Leutbecher and Palmer 2008). Although such verification is meaningful (in the case of unbiased forecasts) and important, it has its limitations. More specifically, a prominent added value of ensemble systems as

compared to deterministic systems is in their ability to provide *flow-dependent* uncertainties. The coincidence of the averages of spread and error does not imply a good flow-dependent or point-by-point correspondence.

One could improve upon the simple comparison between $\langle \mathcal{S}_{0,n}^2 \rangle_n$ and $\langle (\mathcal{E}_{0,n}^O)^2 \rangle_n$ by considering the squared difference: $\langle [\mathcal{S}_{0,n}^2 - (\mathcal{E}_{0,n}^O)^2]^2 \rangle_n$. The cross-product $-2\langle \mathcal{S}_{0,n}^2 (\mathcal{E}_{0,n}^O)^2 \rangle_n$ partly accounts for the flow dependence. Moreover, this score is proper (Gneiting and Raftery 2007) with respect to variations in the spread but, as Christensen Moroz and Palmer (2014) uncovered, it is not with respect to variations of the ensemble mean. They propose a proper alternative score:

$$\mathrm{ES} = \langle [\mathcal{S}_{0,n}^2 - (\mathcal{E}_{0,n}^O)^2 - \mathcal{E}_{0,n}^O \mathcal{S}_{0,n} \gamma_n]^2 \rangle_n, \qquad (22)$$

where γ_n is the skewness of ensemble n.

For unskewed ensembles, the ES score can be considered to be a mean squared error for the squared error $(\mathcal{E}_0^O)^2$ and is in that case a deterministic measure, which contrasts with the probabilistic spread–error metrics proposed in this work. Their approach, however, extends to skewed ensembles, in which case additional probabilistic information is taken into account. The ES score is strongly affected by ensembles with large errors and spreads and is therefore highly nonrobust. For instance, unskewed ensembles that have an error that is 2 times larger than the spread contribute $9\mathcal{S}_{0,n}^4$ to the ES score. Therefore, ensembles of large spread are penalized more by the ES score than are those with small spread. The viewpoint taken in the present work is therefore that, rather than comparing error and spread in absolute terms, error should be compared relative to the spread, or in other words scores should depend on the ratio $\mathcal{E}^O/\mathcal{S}$. For the EM-RAW uncertainty forecast this ratio corresponds to the reduced centered random variable (RCRV) as used in Candille et al. (2007). Although not a proper score, the first and second moments of the RCRV can be used to characterize spread–error consistency. Indeed, since $\mathcal{E}_{\min|\mathcal{S}}$, $\mu_{\mathcal{E}^F|\mathcal{S}}$, $\sigma_{\mathcal{E}^F|\mathcal{S}}$, and $\delta_{\mathcal{E}^F|\mathcal{S}}$ are all linearly proportional to $\mathcal{S}$, it can be shown that the spread-based models of Eqs. (5), (8b), and (8b) depend on $\mathcal{E}^O/\mathcal{S}$ rather than $\mathcal{E}^O$ and the scores of Eq. (16) depend only linearly on $\mathcal{S}$.

The strengths and weaknesses of the spread–error correlation as a basis of spread–error assessment are explained in Whitaker and Loughe (1998), Grimit and Mass (2007), and Hopson (2014). These works also highlight the prerequisite of sufficient day-to-day variability of ensemble spread for good spread–error assessment. Several works also noted that the spread–error correlation is mostly determined by only a few cases featuring anomalous spread (Barker 1991; Houtekamer 1993; Whitaker and Loughe 1998; Grimit and Mass 2007). This issue can be understood in terms of heteroscedasticity and is analogous to the previous argument that errors should preferably be assessed relative to the ensemble spread. Usually the spread–error Pearson correlation is compared to the perfectly reliable model, which is then considered as the highest obtainable correlation—an interpretation that has recently been criticized by Kumar et al. (2014). Moreover, correlation values comparable to these maximal values seem difficult to obtain (Barker 1991; Buizza 1997; Houtekamer 1993; Grimit and Mass 2007; Van Schaeybroeck and Vannitsem 2011; Eckel et al. 2012). Despite the well-known weaknesses of the Pearson correlation, it is still often used.

Postprocessing techniques based on spread calibration improve reliability (Gneiting et al. 2005; Raftery et al. 2005; Wilks 2009; Roulin and Vannitsem 2012) and are therefore expected to improve the spread–skill scores as proposed in this work. Moreover, along the lines of Veenhuis (2013) and Van Schaeybroeck and Vannitsem (2015), new spread-calibration techniques could be designed that directly improve the new spread–error relations. In section 5a it was concluded that for the ECMWF EPS, no substantial information is lost when using the spread as the measure of uncertainty. It is, however, unclear at this stage if this conclusion would hold after spread calibration. Note also that many spread-calibration techniques only correct the ensemble mean and ensemble spread (Gneiting et al. 2005; Wilks 2009) and therefore no additional information is retained. This is different for ''member by member'' correction schemes for which each ensemble member is corrected separately and ensemble kurtosis and skewness are naturally preserved (Johnson and Bowler 2009; Van Schaeybroeck and Vannitsem 2015). Therefore, tests are in order to see whether member-by-member postprocessing of the full-ensemble uncertainty forecast of the ECMWF EPS improves as compared with the spread-based uncertainty forecast.

When model errors are present, systematic biases of both the ensemble mean and spread are observed. In particular, a systematic underdispersion at intermediate lead times is experienced [see Van Schaeybroeck and Vannitsem (2015)]. Ensemble calibration will transform the distribution, in general, by modifying the ensemble mean and/or spread. Moreover, these modifications are strongly dependent on the lead time and will obviously affect the verification of spread. This will be the subject of future investigation.

7. Conclusions

As most end users do not require all the uncertainty information contained in an ensemble, it is common to

provide only a single uncertainty measure, the ensemble spread. Statistically justified models are proposed to assess the spread–error relation. The uncertainty forecast should be considered probabilistically rather than deterministically, the main reason being that both the mean and the standard deviation of the observed error are proportional to the spread. The verification of such forecasts can then be done by using proper scores like the logarithmic score or the CRPS. Because of the sole dependence of the uncertainty forecast on the spread, simple analytic expressions for these scores can be derived (see the appendix). A methodology is outlined that allows us to assess whether the uncertainty estimation based on spread only is worse or better than the full-ensemble estimation.

To verify whether spread is a good predictor for uncertainty, a regression of log-transformed spread and error is proposed. Such a transformation is justified as it overcomes the problem of heteroscedasticity. Two spread–error metrics are proposed [see Eq. (21)] for which the regression coefficients are independent of any hypothesis on the nature of the ensemble distribution. For other spread–error metrics the regression coefficients are also provided, and are derived based on the assumption of Gaussian-distributed ensembles. Note that, given spread as the only measure of uncertainty, the identification of perfect reliability is not possible and one can only identify perfect spread consistency, that is, behavior that is consistent with a perfectly reliable system. Binning was shown to provide additional insights into the approach and is recommended for the determination of ensemble overdispersion or underdispersion as a function of spread.

For the two upper-air variables (Z500 and V500) and one surface variable (V10m) of the ECWMF 51-member EPS, the uncertainty forecasts based on spread (spread-based models) do not result in substantial loss of uncertainty information, at least not with respect to the CRPS and quantile scores. More specifically for the conventional EM-RAW spread–error metric the spread-based model I [Eq. (5)] even improves upon the full-ensemble uncertainty estimate. This implies that replacing each ensemble by a Gaussian distribution with corresponding mean and variance amounts to an improvement in the statistical scores. For other error metrics that involve the ensemble mean, model II [Eq. (8a)] also systematically improves the full-ensemble forecast. Moreover, for error metrics related to nontrivial ensemble averages, model III [Eq. (8a)] is the best, even though it still slightly degrades the full-ensemble uncertainty forecast. For small ensemble sizes, all spread-based models are found to improve upon the full ensemble.

Concerning spread as a predictor of uncertainty, the best results are obtained for Z500. For lead times between 2 and 7 days the regression coefficients indicate that the ensemble is close to a perfectly spread-consistent ensemble.

The sensitivity of the new scores on the ensemble-mean state (Ziehmann 2001; Toth et al. 2001), spatial averages (Barker 1991), and the impact of systematic biases and extreme events (Whitaker and Loughe 1998) are important issues worth investigating in the future.

Acknowledgments. This work has benefited from fruitful discussions with Martin Leutbecher, Olivier Talagrand, Hannah Christensen, Geert Smet, Alex Deckmyn, and Lesley De Cruz. Michael Scheuerer, Hannah Christensen, Pieter Smets, and three anonymous referees are thanked for their constructive comments on the manuscript. This work is partially supported by the Belgian Federal Science policy office under Contract SD/CA/04A.

APPENDIX

Analytic Derivation of Probabilistic Scores

a. *Proposed models*

The three probabilistic models for the uncertainty forecast as introduced in section 2 are detailed here and analytic expressions for the CRPS and QS are derived for the three models. For each forecast the uncertainty forecast $\mathcal{E}^F$ is assumed to be dependent on the spread $\mathcal{S}$. The following elementary models are therefore introduced:

1) The first model for the predictive density of the error is well suited to model the error distribution when the error distribution is approximately Gaussian distributed. The error metric corresponds to the error of the ensemble mean and the spread is the ensemble standard deviation. The predictive density of the error or uncertainty forecast $\mathcal{E}^F$ is

$$\mathcal{P}(\mathcal{E}=\mathcal{E}^F\,|\,\mathcal{S})=\frac{e^{-z^2/2\sigma^2_{\mathcal{E}^F|\mathcal{S}}}}{\sqrt{2\pi\sigma^2_{\mathcal{E}^F|\mathcal{S}}}}\quad\text{with}\quad z=\frac{\mathcal{E}^F-\mu_{\mathcal{E}^F|\mathcal{S}}}{\sigma_{\mathcal{E}^F|\mathcal{S}}}. \tag{A1}$$

2) The second model is suitable when the error is bound from below by $\mathcal{E}_{\min|\mathcal{S}}$ (all metrics in Table 2 except EM-RAW):

$$\mathcal{P}(\mathcal{E}=\mathcal{E}^F\,|\,\mathcal{S})=\frac{2e^{-x^2/2\delta^2_{\mathcal{E}^F|\mathcal{S}}}}{\sqrt{2\pi\delta^2_{\mathcal{E}^F|\mathcal{S}}}}\quad\text{with}$$
$$x=\frac{\mathcal{E}^F-\mathcal{E}_{\min|\mathcal{S}}}{\delta_{\mathcal{E}^F|\mathcal{S}}}\quad\text{and}\quad\mathcal{E}^F\geq\mathcal{E}_{\min|\mathcal{S}}. \tag{A2}$$

3) The third model is a variant of the latter:

$$\mathcal{P}(\mathcal{E}=\mathcal{E}^F\,|\,\mathcal{S})=\frac{e^{-y}}{\mu_{\mathcal{E}^F|\mathcal{S}}-\mathcal{E}_{\min|\mathcal{S}}}\quad\text{with}$$
$$y=\frac{\mathcal{E}^F-\mathcal{E}_{\min|\mathcal{S}}}{\mu_{\mathcal{E}^F|\mathcal{S}}-\mathcal{E}_{\min|\mathcal{S}}}\quad\text{and}\quad\mathcal{E}^F\geq\mathcal{E}_{\min|\mathcal{S}}. \tag{A3}$$

The model parameters $\mu_{\mathcal{E}^F|\mathcal{S}}$, $\sigma_{\mathcal{E}^F|\mathcal{S}}$, $\delta_{\mathcal{E}^F|\mathcal{S}}$, and $\mathcal{E}_{\min|\mathcal{S}_n}$ are all linear functions of $\mathcal{S}$. The associated proportionality coefficients are given in Table 2 and are derived using the assumption that the ensemble is Gaussian distributed. Note that for some spread–error metrics the linear constant of proportionality is independent of the Gaussian assumption.

b. *Scores used for verification of the spread-based forecast models*

To assess the forecasts, probabilistic scores are used. For a verification set $\{\mathcal{S}_n,\mathcal{E}_n^O\}_n$, the scores are the quantile score (QS) and the continuous ranked probability score (CRPS), as given in Eqs. (16b) and (16a), respectively. These require the derivation of the cumulative distribution $\mathcal{P}(\mathcal{E}^F<t\,|\,\mathcal{S}_n)$. For the three models these are

Model I, Eq. (A1): $\mathcal{P}(\mathcal{E}^F<t\,|\,\mathcal{S})$
$$=\frac{1}{2}[1+\mathrm{erf}(z_t/\sqrt{2})]\quad\text{with}\quad z_t=\frac{t-\mu_{\mathcal{E}^F|\mathcal{S}}}{\sigma_{\mathcal{E}^F|\mathcal{S}}}, \tag{A4a}$$

Model II, Eq. (A2): $\mathcal{P}(\mathcal{E}^F<t\,|\,\mathcal{S})=\mathrm{erf}(x_t/\sqrt{2})$ with
$$x_t=\frac{t-\mathcal{E}_{\min|\mathcal{S}}}{\delta_{\mathcal{E}^F|\mathcal{S}}}\quad\text{and}\quad t\geq\mathcal{E}_{\min|\mathcal{S}},\quad\text{and} \tag{A4b}$$

Model III, Eq. (A3): $\mathcal{P}(\mathcal{E}^F<t\,|\,\mathcal{S})$
$$=1-e^{-y_t}\quad\text{with}\quad y_t=\frac{t-\mathcal{E}_{\min|\mathcal{S}}}{\mu_{\mathcal{E}^F|\mathcal{S}}-\mathcal{E}_{\min|\mathcal{S}}}\quad\text{and}\quad t\geq\mathcal{E}_{\min|\mathcal{S}}. \tag{A4c}$$

For these distributions, one can determine the quantile function for the quantile q as used in Eq. (16b):

$$\text{Model I: }\mathcal{E}^q=\mu_{\mathcal{E}^F|\mathcal{S}}+\sqrt{2}\sigma_{\mathcal{E}^F|\mathcal{S}}\mathrm{erf}^{-1}(2q-1), \tag{A5a}$$
$$\text{Model II: }\mathcal{E}^q=\mathcal{E}_{\min|\mathcal{S}}+\sqrt{2}\delta_{\mathcal{E}^F|\mathcal{S}}\mathrm{erf}^{-1}(q),\quad\text{and} \tag{A5b}$$
$$\text{Model III: }\mathcal{E}^q=\mathcal{E}_{\min|\mathcal{S}}-(\mu_{\mathcal{E}^F|\mathcal{S}}-\mathcal{E}_{\min|\mathcal{S}})\ln(1-q). \tag{A5c}$$

For the CRPS one must substitute the cumulative distribution functions of Eq. (A4) into Eq. (16a). Taking into account the appropriate minimal integration value for the models of Eqs. (A2) and (A3), one gets

Model I (Gneiting et al. 2005): CRPS
$$=\left\langle\sigma_{\mathcal{E}^F|\mathcal{S}_n}\left(z_n[2\Phi(z_n)-1]+2\phi(z_n)-\frac{1}{\sqrt{\pi}}\right)\right\rangle_n, \tag{A6a}$$

Model II: CRPS
$$=\left\langle\delta_{\mathcal{E}^F|\mathcal{S}_n}\left[4x_n^+\Phi(x_n^+)+4\phi(x_n^+)-3x_n^+-\frac{2}{\sqrt{\pi}}-x_n^-\right]\right\rangle_n, \tag{A6b}$$

and

Model III: CRPS
$$=\left\langle(\mu_{\mathcal{E}^F|\mathcal{S}_n}-\mathcal{E}_{\min|\mathcal{S}_n})\left(y_n^++2e^{-y_n^+}-\frac{3}{2}-y_n^-\right)\right\rangle_n. \tag{A6c}$$

Here, ϕ is the standard normal PDF and Φ is its CDF, and

$$z_n=\frac{\mathcal{E}_n^O-\mu_{\mathcal{E}^F|\mathcal{S}_n}}{\sigma_{\mathcal{E}^F|\mathcal{S}_n}}, \tag{A7a}$$
$$x_n^\pm=\left(\frac{\mathcal{E}_n^O-\mathcal{E}_{\min|\mathcal{S}_n}}{\delta_{\mathcal{E}^F|\mathcal{S}_n}}\right)\Theta[\pm(\mathcal{E}_n^O-\mathcal{E}_{\min|\mathcal{S}_n})],\quad\text{and} \tag{A7b}$$
$$y_n^\pm=\left(\frac{\mathcal{E}_n^O-\mathcal{E}_{\min|\mathcal{S}_n}}{\mu_{\mathcal{E}^F|\mathcal{S}_n}-\mathcal{E}_{\min|\mathcal{S}_n}}\right)\Theta[\pm(\mathcal{E}_n^O-\mathcal{E}_{\min|\mathcal{S}_n})]. \tag{A7c}$$

Note that for the EM-RAW spread–error metric, the CRPS of the full-ensemble uncertainty forecast is identical to the CRPS of the full-ensemble forecast F.

REFERENCES

Abramowitz, M., and I. A. Stegun, 1972: *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. 9th ed. Dover, 1047 pp.

Atger, F., 1999: The skill of ensemble prediction systems. *Mon. Wea. Rev.*, **127**, 1941–1953, doi:10.1175/1520-0493(1999)127<1941:TSOEPS>2.0.CO;2.

Barker, T. W., 1991: The relationship between spread and forecast error in extended-range forecasts. *J. Climate*, **4**, 733–742, doi:10.1175/1520-0442(1991)004<0733:TRBSAF>2.0.CO;2.

Benedetti, R., 2010: Scoring rules for forecast verification. *Mon. Wea. Rev.*, **138**, 203–211, doi:10.1175/2009MWR2945.1.

Bentzien, S., and P. Friederichs, 2014: Decomposition and graphical portrayal of the quantile score. *Quart. J. Roy. Meteor. Soc.*, **140**, 1924–1934, doi:10.1002/qj.2284.

Box, E. P., J. S. Hunter, and W. G. Hunter, 2005: *Statistics for Experimenters: Design, Innovation, and Discovery*. 2nd ed. Wiley-Interscience, 664 pp.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **125**, 99–119, doi:10.1175/1520-0493(1997)125<0099:PFSOEP>2.0.CO;2.

——, M. Leutbecher, L. Isaksen, and J. Haseler, 2010: Combined use of EDA- and SV-based perturbations in the EPS. *ECMWF Newsletter*, No. 123, ECMWF, Reading, United Kingdom, 22–28.

Candille, G., C. Côté, P. L. Houtekamer, and G. Pellerin, 2007: Verification of an ensemble prediction system against observations. *Mon. Wea. Rev.*, **135**, 2688–2699, doi:10.1175/MWR3414.1.

Chatterjee, S., and A. S. Hadi, 2006: *Regression Analysis by Example*. 4th ed. Wiley Series in Probability and Statistics, J. Wiley and Sons, 375 pp.

Christensen, H. M., I. M. Moroz, and T. N. Palmer, 2014: Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. *Quart. J. Roy. Meteor. Soc.*, **141**, 538–549, doi:10.1002/qj.2375.

Draper, N. R., and H. Smith, 1998: *Applied Regression Analysis*. J. Wiley and Sons, 706 pp.

Eckel, F. A., M. S. Allen, and M. C. Sittel, 2012: Estimation of ambiguity in ensemble forecasts. *Wea. Forecasting*, **27**, 50–69, doi:10.1175/WAF-D-11-00015.1.

Ehrendorfer, M., 1997: Predicting the uncertainty of numerical weather forecasts: A review. *Meteor. Z.*, **6**, 147–183.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, doi:10.1198/016214506000001437.

——, ——, A. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, doi:10.1175/MWR2904.1.

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, doi:10.1111/j.1467-9868.2007.00587.x.

Grimit, E. P., and C. F. Mass, 2007: Measuring the ensemble spread–error relationship with a probabilistic approach: Stochastic ensemble results. *Mon. Wea. Rev.*, **135**, 203–221, doi:10.1175/MWR3262.1.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hopson, T. M., 2014: Assessing the ensemble spread–error relationship. *Mon. Wea. Rev.*, **142**, 1125–1142, doi:10.1175/MWR-D-12-00111.1.

Houtekamer, P. L., 1993: Global and local skill forecasts. *Mon. Wea. Rev.*, **121**, 1834–1846, doi:10.1175/1520-0493(1993)121<1834:GALSF>2.0.CO;2.

Johnson, C., and N. Bowler, 2009: On the reliability and calibration of ensemble forecasts. *Mon. Wea. Rev.*, **137**, 1717, doi:10.1175/2009MWR2715.1.

Kolczynski, W. C., D. R. Stauffer, S. E. Haupt, N. S. Altman, and A. Deng, 2011: Investigation of ensemble variance as a measure of true forecast variance. *Mon. Wea. Rev.*, **139**, 3954–3963, doi:10.1175/MWR-D-10-05081.1.

Kumar, A., P. Peitao, and C. Mingyue, 2014: Is there a relationship between potential and actual skill? *Mon. Wea. Rev.*, **142**, 2220–2227, doi:10.1175/MWR-D-13-00287.1.

Leutbecher, M., 2009: Diagnosis of ensemble forecasting systems. *Seminar on Diagnosis of Forecasting and Data Assimilation Systems*, ECMWF, Reading, United Kingdom, 235–266.

——, and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, doi:10.1016/j.jcp.2007.02.014.

Lorenz, E. N., 1996: Predictability—A problem partly solved. *Proc. Seminar on Predictability*, Vol. 1, Reading, United Kingdom, ECMWF, 1–18.

Nicolis, C., R. A. P. Perdigao, and S. Vannitsem, 2009: Dynamics of prediction errors under the combined effect of initial condition and model errors. *J. Atmos. Sci.*, **66**, 766–778, doi:10.1175/2008JAS2781.1.

Raftery, A. E., F. Balabdaoui, T. Gneiting, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, doi:10.1175/MWR2906.1.

Roulin, E., and S. Vannitsem, 2012: Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Mon. Wea. Rev.*, **140**, 874–888, doi:10.1175/MWR-D-11-00062.1.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660, doi:10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

Scherrer, S. C., C. Appenzeller, P. Eckert, and D. Cattani, 2004: Analysis of the spread–skill relations using the ECMWF Ensemble Prediction System over Europe. *Wea. Forecasting*, **19**, 552–565, doi:10.1175/1520-0434(2004)019<0552:AOTSRU>2.0.CO;2.

Toth, Z., Y. Zhu, and T. Marchok, 2001: The use of ensembles to identify forecasts with small and large uncertainty. *Wea. Forecasting*, **16**, 463–477, doi:10.1175/1520-0434(2001)016<0463:TUOETI>2.0.CO;2.

——, O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. Jolliffe and D. B. Stephenson, Eds., J. Wiley and Sons, 137–163.

Van Schaeybroeck, B., and S. Vannitsem, 2011: Post-processing through linear regression. *Nonlinear Processes Geophys.*, **18**, 147–160, doi:10.5194/npg-18-147-2011.

——, and ——, 2015: Ensemble post-processing using member-by-member approaches: Theoretical aspects. *Quart. J. Roy. Meteor. Soc.*, **141**, 807–818, doi:10.1002/qj.2397.

Veenhuis, B. A., 2013: Spread calibration of ensemble MOS forecasts. *Mon. Wea. Rev.*, **141**, 2467–2482, doi:10.1175/MWR-D-12-00191.1.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158, doi:10.1175/1520-0469(2003)060<1140:ACOBAE>2.0.CO;2.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.*, **126**, 3292–3302, doi:10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, doi:10.1002/met.134.

——, 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.

Yeo, I.-K., and R. A. Johnson, 2000: A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954–959, doi:10.1093/biomet/87.4.954.

Ziehmann, C., 2001: Skill prediction of local weather forecasts based on the ECMWF ensemble. *Nonlinear Processes Geophys.*, **8**, 419–428, doi:10.5194/npg-8-419-2001.