RMetS
Royal Meteorological Society

Ensemble post-processing using member-by-member approaches: theoretical aspects

Bert Van Schaeybroeck* and Stéphane Vannitsem
Royal Meteorological Institute of Belgium (RMI), Brussels, Belgium

*Correspondence to: B. Van Schaeybroeck, Royal Meteorological Institute of Belgium, Ringlaan 3, B-1180 Brussels, Belgium.
E-mail: bertvs@meteo.be

Linear post-processing approaches are proposed and fundamental mechanisms are analyzed by which the probabilistic skill of an ensemble forecast can be improved. The ensemble mean of the corrected forecast is a linear function of the ensemble mean(s) of the predictor(s). Likewise, the ensemble spread of the corrected forecast depends linearly on that of the uncorrected forecast. The regression coefficients are obtained by maximizing the likelihood function for the error distribution. Comparing different calibration approaches on simple systems that exhibit chaotic features (the Kuramoto–Sivashinsky equation, the spatially extended Lorenz system), four correction mechanisms are identified: the ensemble-mean scaling and nudging using the predictor(s), and the ensemble-spread scaling and nudging. Ensemble-spread corrections turn out to yield improvement only when 'reliability' constraints are imposed on the corrected forecast. First of all *climatological reliability* is enforced and is satisfied when the total variability of the forecast is equal to the variability of the observations. Second, *ensemble reliability* or calibration of the ensembles is enforced such that the squared error of the ensemble mean coincides with the ensemble variance.

In terms of continuous ranked probability skill score, spread calibration provides much more gain in skill than the traditional ensemble-mean calibration and extends for lead times far beyond the error-doubling time. The skill performance is better than or as good as the benchmark calibration method which derives from statistical assumptions –non-homogeneous Gaussian regression. In addition to the member-by-member nature of the approach, benefits compared with the benchmark method can be pinpointed. In particular, although the post-processing methods are performed for each lead time, location and variable independently, they preserve the rank correlations and thus take dependencies across space, time, and different variables into account. In addition, higher-order ensemble moments like kurtosis and skewness correspond to those of the uncorrected forecasts.

Key Words: statistical post-processing; model output statistics; member-by-member approach

Received 5 December 2013; Revised 5 May 2014; Accepted 9 May 2014; Published online in Wiley Online Library 20 June 2014

1. Introduction

The atmosphere and its climate display the property of sensitivity to initial conditions which drastically limits their predictability horizon (Kalnay, 2002). Moreover, it is recognized that model errors also strongly degrade forecasts as a function of lead time (Nicolis *et al.*, 2009). The modern approach is to quantify these uncertainties using an ensemble of forecasts, each starting from different initial conditions, and/or with different model physics. In this way, a probabilistic forecast can be produced. The ensemble mean is the quantity usually disseminated while the ensemble spread is a measure of the flow-dependent forecast uncertainty. However, it is well-known that, for state-of-the-art weather forecasts, the uncertainty measure is not very accurate. Moreover, at the surface, experiments show that ensemble forecasts are consistently under-dispersive (or overconfident) for long lead times (Leutbecher and Palmer, 2008). This feature can be partly traced back to systematic errors, relevant to the model at hand, that could be partly corrected by calibration or post-processing.

A common method to calibrate (deterministic) forecasts is called Model Output Statistics (MOS) and is based on the statistical error features of past model output. The simplest approach applies ordinary least-squares regression to fit predictions to observations (Glahn and Lowry, 1972). Unfortunately the use of this approach implies a degradation of the variability at long lead times, an undesirable feature for ensemble forecasting (Wilks, 1995). In order to overcome this problem, new methods were introduced for scalar predictands based on different linear regression techniques (Vannitsem, 2009; Van Schaeybroeck and Vannitsem, 2011, 2012) but also for vector variables such as wind (Pinson, 2012). These will be referred to as 'member-by-member' (MBM) calibration methods in this work. On the other hand,

'statistical' calibration methods were also proposed. These are statistical in nature in the sense that they assume specific ensemble distributions and have predictive distributions as output, rather than an ensemble of discrete size. For example the logistic distribution has been successfully applied in the context of post-processing of precipitation forecasts (Wilks, 2009; Schmeits and Kok, 2010; Roulin and Vannitsem, 2012). For temperature, one of the most competitive approaches is the Non-homogeneous Gaussian Regression (NGR) (Gneiting *et al.*, 2005; Hagedorn *et al.*, 2008). NGR uses Gaussian predictive distributions with mean and spread that depend linearly on the corresponding quantities of the raw forecast. Also, NGR explicitly minimizes the associated continuous ranked probability score (CRPS) which is the squared difference between the cumulative distribution functions of the ensemble forecast and the observation integrated over all possible thresholds.

In practice, however, important side effects emerge when independently applying *statistical* post-processing methods at different stations, lead times or for multiple variables. The output of such a collection of statistical post-processing are independent predictive distributions which can be used to reconstruct ensembles by random sampling. However, strong correlations are likely to be present between the values of nearby stations, lead times and multiple variables of the *raw* forecast, which will be strongly reduced for a member reconstructed from the output of statistical calibration.

We illustrate this reduction of correlations in Figure 1. Consider in Figure 1(a) four ensemble members of the raw ECMWF ensemble forecast for 2 m temperature (2mT) and minimum temperature over the last 6 h (min$T$) on 21 June 2011 at Uccle, Belgium. For each ensemble member the 2mT (full lines) and min$T$ (dotted lines) have a strongly correlated temporal evolution. The presence of correlations is made more explicit in Figure 1(b) which displays the *raw* 2mT forecasts against the min$T$ forecasts for all 15 days of lead time at 0600 UTC and for all 51 members. Figure 1(c, d) show the same scatterplot but for calibrated forecasts–more specifically, the MBM correction method (CRPS MIN in Figure 1(c)) introduced in this article and a 51-member reconstructed ensemble forecast of the statistical calibration method NGR (Figure 1(d)). Even though the post-processing is performed with one predictor only and for 2mT and min$T$ independently, the Pearson correlation $\rho = 0.92$ among 2mT and min$T$ of the raw forecast is very close to the MBM method for which $\rho = 0.90$, while ρ is strongly reduced by NGR to a value of 0.15. In addition, the latter produces many physically inconsistent forecasts (red dots) where min$T$ is higher than 2mT.

Analogously to the correlations that exist among variables (pertaining to the same ensemble member), important information is contained in the spatial and temporal structures of the individual members. Calibration of the different sets (including different locations and lead times) independently using our approach does preserve the information as a consequence of the linear mapping of the individual ensemble members. Independent *statistical* calibration, on the other hand, yields independent predictive distributions from which a new ensemble is sampled, reducing correlations among the different sets compared with the raw forecast. Note that these problems of the statistical methods have been already recognized and form the subject of intense ongoing research (Schuhen *et al.*, 2012; Möller *et al.*, 2013; Schefzik *et al.*, 2013; Scheuerer and Büermann, 2014).

In this work, MBM (or 'deterministic') methods are developed for which each ensemble member is corrected individually by a linear mapping, thereby retaining rank correlations. Therefore each member retains to a large extent correlation structures in the case of multiple independent calibrations (as shown in Figure 1(c)). Moreover in terms of skill our MBM approach can be as high as NGR.

More specifically, the purpose is to correct the forecasts so that they respect two 'reliability' conditions. First of all, the distribution of errors of the ensemble mean should agree with the distribution

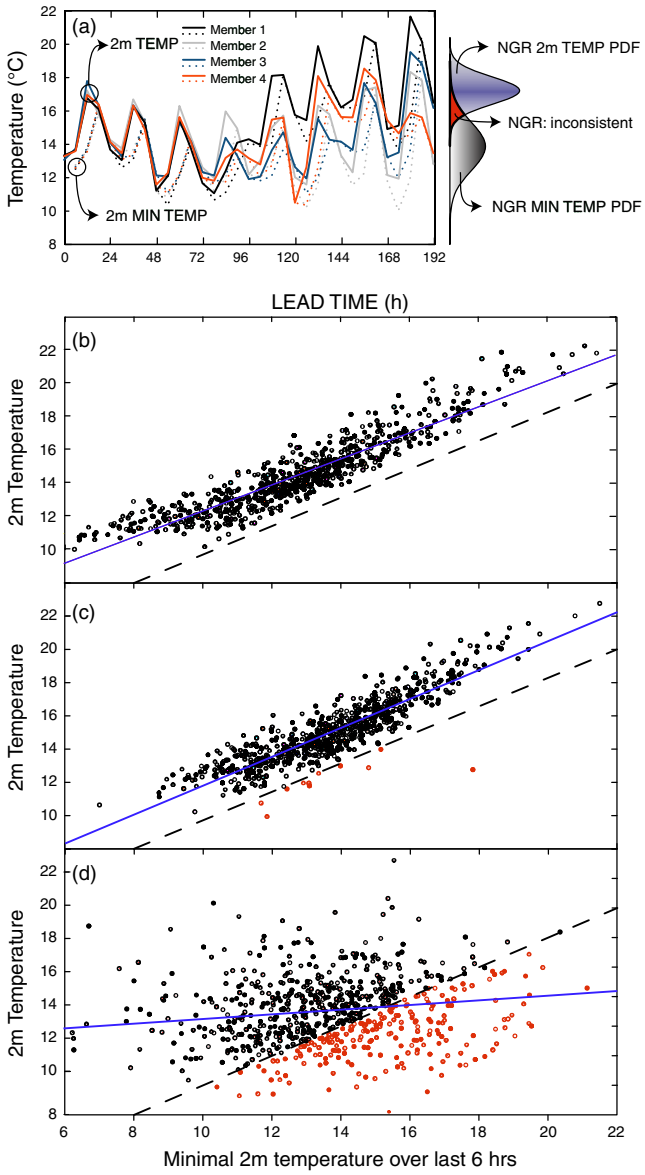

Figure 1. (a) Evolution of four members of the ECMWF ensemble forecast on 21 June 2011 at Uccle (Belgium) of 2 m temperature (2mT, full lines) and minimum 2 m temperature over the last 6 h (min$T$, dotted lines). For each member 2mT and min$T$ are strongly correlated. Also shown are the predictive densities of min$T$ and 2mT obtained with NGR calibration at lead time 192 h. Upon randomly sampling min$T$ and 2mT from these distributions, it may occur that 2mT is lower than min$T$, which is physically inconsistent. (b) Raw 2mT forecasts against min$T$ forecasts at 0600 UTC for lead times up to 15 days, for all 51 ensemble members. The solid line is a linear fit to the points. (c) is as (b), but obtained by post-processing using the member-by-member approach CRPS MIN. Red dots denote physically inconsistent forecasts. (d) is as (b), but shows ensemble members sampled from the predictive distributions obtained with NGR.

of the deviations of the ensemble members from the ensemble mean (Kharin and Zwiers, 2003; Johnson and Bowler, 2009; Glahn *et al.*, 2009). The second reliability condition concerns the climatological distribution of the forecasts which should agree with the distribution of all observations (Van Schaeybroeck and Vannitsem, 2011). This last property is called marginal calibration in Gneiting *et al.* (2007).

To this aim, a general framework is developed based on Lagrange multipliers and constrained maximum likelihood for imposing an arbitrary number of constraints. As it turns out, allowing the ensemble mean and spread to be expressed as linear functions of the uncorrected ones (as in NGR) are insufficient to decrease the CRPS at all lead times. Reliability constraints on the corrected forecast are the essential extra ingredient. Our theoretical findings are illustrated using the Kuramoto–Sivashinsky (KS) model.

We start in section 2 by outlining our general set-up of correcting forecasts and specifying the KS equation used to illustrate the findings. The new post-processing techniques are introduced in section 3, while a verification is performed in section 4. We come back to the issue of modification of correlation structures in section 5 and we present a preliminary test of the calibration techniques on real data in section 6. Finally, we conclude in section 7.

2. General set-up

2.1. Raw and corrected forecast

Consider the meteorological variable X for which N observations $(X_{\mathrm{O},1},\ldots,X_{\mathrm{O},N})$ are available. Corresponding to each observation n, the mth member of the ensemble forecast produces the values $(V^m_{1,n},\ldots,V^m_{P,n})$ for the P different meteorological variables or predictors. The first predictor V_1 is the one corresponding to the variable X and is also called the raw or uncorrected forecast. The ensemble-mean values are defined as $(\overline{V}_{1,n},\ldots,\overline{V}_{P,n})$. For the variable X, and for each member m of an ensemble n, a corrected forecast or predictand is constructed:

$$X^m_{\mathrm{C},n} = \alpha + \sum_{p=1}^{P} \beta_p \overline{V}_{p,n} + \tau_n \varepsilon^m_n. \tag{1}$$

Here regression coefficient α is the bias parameter while the coefficients $(\beta_1,\ldots,\beta_P)$ are the ensemble-mean scale parameters. Although in our verification only one predictor ($P = 1$) is used, multiple predictors can be used, provided care is taken to avoid overfitting. The parameter τ_n adjusts the spread of the new ensemble because the deviation from the ensemble mean is defined as $\varepsilon^m_n = V^m_{1,n} - \overline{V}_{1,n}$. A direct consequence of the form of Eq. (1) is that, in the case of one predictor (P = 1) and $\beta_1 \geq 0$, the correlation between the corrected ensemble mean and the observation is equal to the correlation between the uncorrected ensemble mean and the observation (Johnson and Bowler, 2009). The fact that τ_n depends on the ensemble index n comes from its dependence on ensemble spread as specified below. First, two spread measures, the ensemble standard deviation $\sigma_{\varepsilon,n}$ and the absolute-value spread δ_n, are defined as

$$\sigma^2_{\varepsilon,n} = \left\langle \left(V^m_{1,n} - \overline{V}_{1,n}\right)^2 \right\rangle_m, \tag{2a}$$

$$\delta_n = \left\langle \left\langle |V^{m_1}_{1,n} - V^{m_2}_{1,n}| \right\rangle_{m_1} \right\rangle_{m_2}. \tag{2b}$$

Here $\langle\cdot\rangle_m$ denotes the ensemble average. Note that the measure δ_n is of the Cramér–von Mises type (Baringhaus and Franz, 2004; Gneiting and Raftery, 2007). Depending on our choice of spread measure, τ_n is defined as

$$\tau^2_n = \gamma_1^2 + \gamma_2^2\, \sigma^{-2}_{\varepsilon,n}, \tag{3a}$$

$$\tau_n = \gamma_1 + \gamma_2\, \delta_n^{-1}. \tag{3b}$$

This implies that the ensemble spread measures of the corrected ensemble are given by

$$\sigma^2_{\mathrm{C},\varepsilon,n} = \gamma_1^2\, \sigma^2_{\varepsilon,n} + \gamma_2^2, \tag{4a}$$

$$\delta_{\mathrm{C},n} = \gamma_1\, \delta_n + \gamma_2. \tag{4b}$$

Here γ_1 is called the ensemble-spread scale parameter while γ_2 is the ensemble-spread nudge parameter. Scaling refers to inflation or deflation of the uncorrected quantity using a multiplicative factor. Nudging, on the other hand, refers to the fact that a quantity, here $\sigma_{\mathrm{C},\varepsilon,n}$ or $\delta_{\mathrm{C},n}$, becomes constant, i.e. here independent of $\sigma_{\varepsilon,n}$ or δ_n. It is therefore an additive correction. In terms of the parameters, this leads to

- *ensemble-mean scaling*: $\overline{X}_{\mathrm{C},n} \propto \beta_1 \overline{V}_{1,n}$,
- *ensemble-mean nudging*: $\overline{X}_{\mathrm{C},n} \approx \alpha$,
- *ensemble-spread scaling*: $\sigma_{\mathrm{C},\varepsilon,n} \propto \gamma_1 \sigma_{\varepsilon,n}$, or $\delta_{\mathrm{C},n} \propto \gamma_1 \delta_n$,
- *ensemble-spread nudging*: $\sigma_{\mathrm{C},\varepsilon,n} \approx \gamma_2$, or $\delta_{\mathrm{C},n} \approx \gamma_2$.

Both choices of Eq. (4) with non-zero values for γ_1 and γ_2 allow us to cover the crossover between the following two situations. First, at short lead time, the ensemble spread may be a reliable measure of skill up to a constant multiplicative factor owing to a systematic underdispersiveness. A good calibration scheme should be such that $\gamma_2 \approx 0$ and $\gamma_1 > 1$ (or $\gamma_1 < 1$, for overdispersion). Second, at long lead times, when no spread–skill relation exists, a good calibration method should set the spread to a constant and this is achieved using $\gamma_1 \approx 0$ and $\gamma_2 > 0$. Note that relation (4a) is also satisfied by the NGR approach (Gneiting *et al.*, 2005).

2.2. Maximum likelihood estimation (MLE)

The parameters $(\alpha, \boldsymbol{\beta}, \gamma_1, \gamma_2)$ are estimated by maximization of a likelihood function $\mathcal{L}_0$ associated with the ensemble-mean error distribution of the nth ensemble $\mathcal{P}_n$ (Wilks, 1995):

$$\ln \mathcal{L}_0 = \left\langle \ln\left\{\mathcal{P}_n\left(X_{\mathrm{O},n} - \overline{X}_{\mathrm{C},n}\right)\right\}\right\rangle_n, \tag{5}$$

where $\langle\cdot\rangle_n = (1/N)\sum_n \cdot$ denotes the average over all data points, all of which are assumed independent. Again, depending on the ensemble-spread measure, two choices for the error distribution are considered:

$$\mathcal{P}_n(y) = \frac{\mathrm{e}^{-y^2/(2\sigma^2_{\varepsilon,n})}}{\sqrt{2\pi\sigma^2_{\varepsilon,n}}} \sim \mathcal{N}(0,\sigma_{\varepsilon,n}), \tag{6a}$$

$$\mathcal{P}_n(y) = \frac{\mathrm{e}^{-y/\delta_n}}{\delta_n} \sim \mathcal{E}(0,\delta_n). \tag{6b}$$

The first choice for the normal distribution (Eq. (6a)) is the most natural since it allows analytic solutions and the reproduction of traditional post-processing approaches. As shown later, the exponential distribution (Eq. (6b)) is useful as it leads to more stable solutions of the MLE.

2.3. Constrained MLE

In order to obtain the best possible correction, one needs to enforce reliability constraints to our corrected forecast. From a climatological point of view, a reliable forecast is characterized by the fact that the distribution of all observations agrees with the distribution of all forecasts. Therefore, for bias-free forecasts, *climatological reliability* (CR) is defined as the equality of forecast variability σ^2_{C} with the variability of the observations σ^2_{O}:

$$\left\langle \left(X^m_{\mathrm{C},n} - \left\langle X^m_{\mathrm{C},n}\right\rangle_{m,n}\right)^2 \right\rangle_{m,n} = \left\langle \left(X_{\mathrm{O},n} - \left\langle X_{\mathrm{O},n}\right\rangle_n\right)^2 \right\rangle_n. \tag{7}$$

A reliable *probabilistic* forecast, on the other hand, is characterized by the fact that the observation may be considered as a member of the ensemble and therefore has the same statistical properties as the ensemble forecast. The conventional approach is to say that, for a bias-free forecast, the average ensemble variance $\left\langle \sigma^2_{\mathrm{C},\varepsilon,n}\right\rangle_n$ agrees with the mean squared forecast error. This is what is defined as *weak ensemble reliability* (WER):

$$\left\langle (\overline{X}_{\mathrm{C},n} - X_{\mathrm{O},n})^2 \right\rangle_n = \left\langle \sigma^2_{\mathrm{C},\varepsilon,n}\right\rangle_n. \tag{8}$$

Since ensemble spread may be strongly regime-dependent, a slightly different notion of ensemble reliability or ensemble calibration is defined as follows: *Strong ensemble reliability* (SER)

is satisfied when, for a bias-free forecast, the χ^2/N value (which is the standardized mean squared error of the ensemble mean) is equal to one. The standardization is done using the ensemble variance of the corresponding corrected forecast $\sigma^2_{\mathrm{C},\varepsilon,n}$. The condition for SER is therefore

$$\frac{\chi^2}{N} = \left\langle \frac{(\overline{X}_{\mathrm{C},n} - X_{\mathrm{O},n})^2}{\sigma^2_{\mathrm{C},\varepsilon,n}} \right\rangle_n = 1. \tag{9}$$

Note that the χ^2/N value is also called the reduced centred random variable (RCRV) as introduced in Candille *et al.* (2007).

The mathematical tool for imposing the constraints on the corrected forecast is a modified likelihood $\mathcal{L}$ with additional parameters η and μ. The constrained log-likelihood becomes

$$\ln \mathcal{L} = \ln \mathcal{L}_0 - \eta \left(1 - \frac{\sigma^2_{\mathrm{C}}}{\sigma^2_{\mathrm{O}}}\right)^2 - \mu \left(1 - \frac{\chi^2}{N}\right)^2. \tag{10}$$

In the case that one of the two conditions is not imposed, the corresponding parameter (η or μ) is set to zero. If one considers η and μ as variational parameters, they are in fact Lagrange multipliers and the two imposed conditions are 'hard' in the sense that they are exactly satisfied. However, in practice, it is often better to use 'soft' constraints by fixing η and μ to a value much larger than one. As will be discussed later and shown in Appendix C2, maximizing some hard-constrained likelihoods may not yield any solution. Unless mentioned otherwise, we further use soft constraints with $\eta = \mu = 10^3$. Optimization of these parameters could be performed but will be strongly system-dependent. Furthermore for the system under consideration, no substantial quantitative changes are expected. Note also that no constraint on the mean is necessary since our forecasts are to a large extent bias-free after maximization of the log-likelihood with respect to the bias-adjustment parameter α. Moreover, when assuming the normal distribution Eq. (6a) for the error with a variance that is independent of the ensemble, the corrected forecast is exactly bias-free.

In summary, maximization of the first term of Eq. (10) forces the corrected ensemble means $\overline{X}_{\mathrm{C},n}$ to be as close as possible to the observations $X_{\mathrm{O},n}$, ensuring a maximal resolution. The second and third term are introduced to enforce climatological and ensemble reliability, respectively.

Note that, for a forecast which satisfies both WER and CR constraints, the correlation between the observation and the ensemble mean equals the correlation between the ensemble members and the ensemble mean. This feature is proven in Appendix A and may be seen as another measure of reliability.

2.4. *Benchmark: non-homogeneous Gaussian regression*

The MBM approach introduced will be compared with NGR, considered here as a benchmark. It has been shown (Hagedorn *et al.*, 2008; Vannitsem and Hagedorn, 2011) that NGR is a state-of-the-art method, at least for Gaussian-like distributed variables such as surface temperature. Wilks (2006) confirmed that NGR is in many situations the best post-processing method available, in the context of the Lorenz 96 model. However, as already alluded to in the Introduction, cross-correlation properties could be considerably affected when multiple variables, locations and lead times are considered. We therefore think that a MBM approach as presented here is worth using, provided a skill can be comparable with that of the NGR method.

The NGR approach minimizes the CRPS score subject to the assumption that the predictive distribution is normally distributed. An analytical form can be obtained for the CRPS (Gneiting *et al.*, 2005):

$$\mathrm{CRPS} = \left\langle \sigma_{\mathrm{C},\varepsilon,n} \left[z_n \{2\Phi(z_n) - 1\} + 2\phi(z_n) - \pi^{-1/2} \right] \right\rangle_n,$$

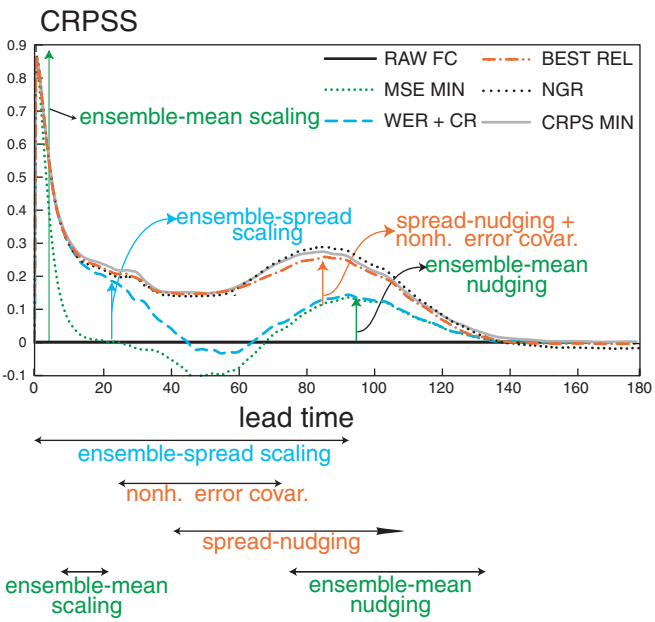

Figure 2. The CRPSS as defined in Eq. (12) as a function of lead time for the different calibration methods. Below the figure, the different time-scales associated with different types of corrections are indicated. The characteristics of the correction schemes are given in Table 1. This figure is obtained using the KS equation with a biased model error parameter $\nu' = \nu + 0.001$.

where ϕ is the normal probability density function (PDF), Φ its cumulative distribution function (CDF), z_n the standardized error $= (X_{\mathrm{O},n} - \overline{X}_{\mathrm{C},n})/\sigma_{\mathrm{C},\varepsilon,n}$, and both $\sigma_{\mathrm{C},\varepsilon,n}$ and $\overline{X}_{\mathrm{C},n}$ are related to the regression parameters $(\alpha, \boldsymbol{\beta}, \gamma_1, \gamma_2)$ as before.

2.5. *Illustration based on test case*

Our main findings are illustrated based on a spatially extended system, the KS equation (Manneville, 1990) which pertains to a class of partial differential equations describing the flow dynamics in the vicinity of a convective instability threshold of two-dimensional stationary rolls. The dynamic equation reads:

$$\partial_t \psi = -\nu\psi - \partial_x^2\psi - \partial_x^4\psi - 2\psi\partial_x\psi\,, \tag{11}$$

where $\psi(x,t)$ is the convective velocity and $\nu = 0.01$ is a small damping parameter. For the numerical integration, we use a semi-implicit Adams–Bashford Cranck–Nicholson scheme with a time step of 0.1 time units and 256 spatial grid points of 0.5 space units and periodic boundary conditions. The forecasts extend up to 200 units of time. The properties of ordinary least-squares (OLS) post-processing have also been investigated for this system in Vannitsem and Nicolis (2008). With this choice of parameters, the system exhibits chaotic behaviour. The error-doubling time associated with the largest Lyapunov exponent λ_{L} is equal to $\ln(2)/\lambda_{\mathrm{L}} = 7.54$ time units. For generating the observations, we assume a slightly biased parameter $\nu' = \nu + 0.001$ from the one used for generating the model data. The system is perfectly reliable at time zero in the sense that the observation is randomly sampled from the normally distributed forecast ensemble with standard deviation of 10^{-4} at each grid point. Both the training and verification sets include 200 ensembles of 50 members each.

3. The different practical correction methods

To evaluate the approach proposed in sections 2.1–2.3 using the test case (section 2.5), a hierarchy of correction methods are now detailed and developed, and a summary of their main characteristics is provided in Table 1. In this table, for each calibration method, we denote the use of the regression parameters α, β_1, γ_1, γ_2 by a checkmark (✓). These coefficients are determined independently for all lead times and spatial points.

Table 1. Schematic representation of the characteristics of the post-processing schemes. For the calibration methods indicated in the first column, the regression parameters α, β_1, γ_1, γ_2 (section 2.1) are fixed to a certain value (0 or 1) or are variable regression parameters (✓). It is also indicated whether the different reliability constraints are satisfied: climatological reliability (CR; Eq. (7)), weak ensemble reliability (WER; Eq. (8)) and strong ensemble reliability (SER; Eq. (9)). When exactly satisfied, it is denoted by ✓, while ± means overall satisfaction in numerical experiments. Note that, for CRPS MIN and NGR, the choice of error distribution $\mathcal{P}_n$ is not applicable (n/a).

Method	Abbreviation	α	β_1	CR	γ_1	WER	SER	γ_2	Error distrib. $\mathcal{P}_n$ of ensemble n	Exact solution	Comput. load
Normal-error minimization	MSE MIN	✓	✓	–	1	–	–	0	$\mathcal{N}\left(0, \langle\sigma^2_{\varepsilon,n}\rangle_n\right)$	✓	Low
Weak ensemble-reliability and climatological-reliability constraint	WER + CR	✓	✓	✓	✓	✓	–	0	$\mathcal{N}\left(0, \gamma_1^2\langle\sigma^2_{\varepsilon,n}\rangle_n\right)$	✓	Low
Exponential-error and non-homogeneous error variance	BEST REL	✓	✓	✓	✓	✓	±	✓	$\mathcal{E}(0, \delta_{C,n})$	–	Higher
Member-by-member CRPS minimization	CRPS MIN	✓	✓	±	✓	±	±	✓	n/a	–	Highest
Non-homogeneous Gaussian regression	NGR	✓	✓	–	✓	±	±	✓	n/a	–	Higher

The different methods are compared by displaying in Figure 2 the continuous ranked probability skill score (CRPSS) as a function of lead time. Denoting the CRPS of the corrected and uncorrected forecasts by $\mathrm{CRPS_{corr}}$ and $\mathrm{CRPS_{unc}}$, respectively, the associated skill score is

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS_{corr}}}{\mathrm{CRPS_{unc}}}. \tag{12}$$

A forecast better (worse) than the raw forecast has positive (negative) skill and a perfect forecast corresponds to a skill of one.

Different time-scales associated with the different correction mechanisms which were defined in section 2.1, are indicated below Figure 2. The scales reflect the time intervals during which skill increase is observed, by comparison of the different methods.

3.1. *Ensemble-mean correction*

We define now the *ensemble-mean correction* or MSE MIN. This method adjusts the ensemble mean using one or more predictors, the first one, V_1, being the uncorrected forecast. However, MSE MIN leaves the deviation of the members from the ensemble mean untouched. Therefore we set $\gamma_1 = 1$ and $\gamma_2 = 0$, constraints which are later relaxed upon introduction of the other correction methods. The MSE MIN forecast is therefore:

$$X^m_{\mathrm{C},n} = \alpha + \sum_p^P \beta_p \overline{V}_{p,n} + \varepsilon^m_n. \tag{13}$$

The log-likelihood function associated with a normal error distribution with ensemble-independent variance $\mathcal{N}(0, \langle\sigma^2_{\varepsilon,n}\rangle_n)$ reads:

$$\ln \mathcal{L}_0 = -\frac{N}{2}\ln\left(2\pi\langle\sigma^2_{\varepsilon,n}\rangle_n\right) - \frac{\left\langle(\overline{X}_{\mathrm{C},n} - X_{\mathrm{O},n})^2\right\rangle_n}{2\langle\sigma^2_{\varepsilon,n}\rangle_n}. \tag{14}$$

Note that the second term on the right-hand side is proportional to the mean squared error. The maximization of the likelihood function leads to the well-known result of OLS regression:

$$\boldsymbol{\beta} = \boldsymbol{\sigma}^2_{\mathrm{O}\overline{V}}\boldsymbol{\sigma}^{-2}_{\overline{V}}, \tag{15a}$$

$$\alpha = \langle \mathbf{X}_{\mathrm{O}}\rangle_n - \left\langle\boldsymbol{\beta}\overline{\mathbf{V}}\right\rangle_n, \tag{15b}$$

with $\mathbf{X}_{\mathrm{O}} = (X_{\mathrm{O},1}, \ldots, X_{\mathrm{O},N})$ the vector of all N observations, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$ the vector of P regression parameters and $\overline{\mathbf{V}}$ the matrix of size $P \times N$ containing all ensemble-mean predictors. The expression $\langle \boldsymbol{A}\rangle_n$ denotes averaging over all N vector elements of $\boldsymbol{A}$. The $P \times P$ covariance matrix $\boldsymbol{\sigma}^2_{\overline{V}}$ and the vector $\boldsymbol{\sigma}^2_{\mathrm{O}\overline{V}}$ of size P that contain the elements (p_1 and p_2 are between 1 and P) are defined as:

$$\sigma^2_{\overline{V},p_1,p_2} = \left\langle\left(\overline{V}_{p_1,n_1} - \langle \boldsymbol{V}_{p_1}\rangle_n\right)\left(\overline{V}_{p_2,n_1} - \langle \boldsymbol{V}_{p_2}\rangle_n\right)\right\rangle_{n_1}, \tag{16a}$$

$$\sigma^2_{\mathrm{O}\overline{V},p_1} = \left\langle\left(X_{\mathrm{O},n_1} - \langle \boldsymbol{X}_{\mathrm{O}}\rangle_n\right)\left(\overline{V}_{p_1,n_1} - \langle \boldsymbol{V}_{p_1}\rangle_n\right)\right\rangle_{n_1}. \tag{16b}$$

Note that the solution of Eq. (15) does not depend on the average ensemble variance $\langle\sigma^2_{\varepsilon,n}\rangle_n$.

Figure 2 indicates that MSE MIN has a period ($t < 20$) of improved skill compared with the raw forecast due to ensemble-mean scaling. This period is of the order of the error-doubling time $\ln(2)/\lambda_{\mathrm{L}} = 7.54$ time units. On the other hand, for long lead times ($t > 70$), ensemble-mean nudging induces skill gain: the bare ensemble mean and observations are nearly decorrelated such that skill is gained by setting the ensemble mean to a constant. However, an intermediate period ($20 < t < 70$) of negative skill is encountered due to reduced forecast variability caused by the ensemble-mean nudging and thus a reduction of climatological reliability. This is illustrated in Figure 3 where the ratio of forecast variance with respect to the observational variance is displayed. The MSE MIN forecast exhibits a clear variance dip at intermediate times.

One could try to avoid the variance dip of the MSE MIN forecast by imposing the climatological reliability constraint. In Appendix B the solutions for such correction scheme are presented. The green line in Figure 4 shows that, although the skill gain due to the ensemble-mean scaling is preserved, the gain due to ensemble-mean nudging is lost. This is because the CR constraint eliminates the decreased variability of the MSE MIN method by increasing the ensemble-mean variability, which implies an increase of the ensemble-mean error. In the next subsection we allow for an increase of the ensemble spread to satisfy the CR constraint while preserving the reduced ensemble-mean error of MSE MIN.

3.2. *Ensemble-spread scaling (WER + CR)*

The detrimental features of reduced total variability of MSE MIN, and the underdispersion of ensemble forecasts could be overcome by allowing for a spread correction. As a first approach, a spread-scaling factor γ_1 only, is introduced while $\gamma_2 = 0$ in Eq. (3).

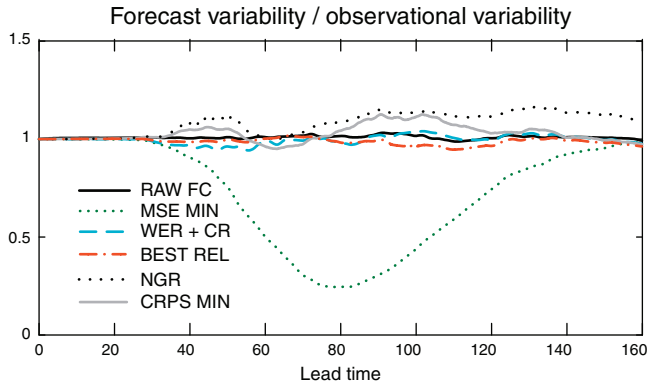

Figure 3. Ratio of the forecast variability to the observational variability (σ_C^2/σ_O^2) as a function of lead time for the different calibration methods.

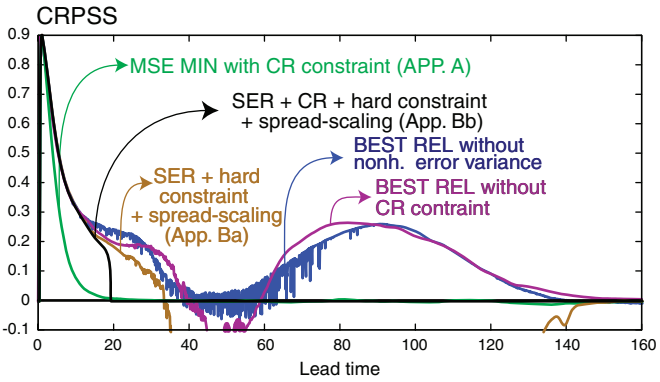

Figure 4. The CRPSS as defined in Eq. (12) as a function of lead time for calibration methods which were tested but yield to insufficient skill gain.

Therefore

$$X^m_{C,n} = \alpha + \sum_{p=1}^{P} \beta_p \overline{V}_{p,n} + \gamma_1 \varepsilon^m_n \,. \tag{17}$$

The log-likelihood of the normal distribution is then readily obtained by replacing $\langle \sigma^2_{\varepsilon,n} \rangle_n$ with $\gamma_1^2 \langle \sigma^2_{\varepsilon,n} \rangle_n$ in Eq. (14). Again the error variance is homogeneous in the sense that it is independent of the specific ensemble forecast. The MLE solutions for α and $\boldsymbol{\beta}$ are as in Eq. (15) and the spread-scale parameter is given by

$$\gamma_1^2 = \langle \sigma^2_{\varepsilon,n} \rangle_n^{-1} \left\{ \sigma^2_O - \boldsymbol{\sigma}^2_{O\overline{V}} \boldsymbol{\sigma}^{-2}_{\overline{V}} (\boldsymbol{\sigma}^2_{O\overline{V}})^T \right\}. \tag{18}$$

It can be checked that this calibrated forecast satisfies exactly the WER and CR conditions of Eqs (7) and (8). Therefore this approach is called the WER + CR approach. As is clear from Figure 2, the skill gain of ensemble-spread scaling is large compared with the other correction mechanisms and extends for a considerable time range of around 100 time units. This time interval indicated below Figure 2 by 'ensemble-spread scaling' is the interval during which the CRPSS value of WER + CR exceeds that of MSE MIN.

In the case where one predictor is used ($P = 1$), the results already obtained by several authors (Kharin and Zwiers, 2003; Johnson and Bowler, 2009) are recovered for the regression coefficients:

$$\beta_1 = \frac{\rho_{O\overline{V}} \sigma_O}{\sigma_{\overline{V}}}, \tag{19a}$$

$$\gamma_1^2 = \frac{\sigma^2_O}{\langle \sigma^2_{\varepsilon,n} \rangle_n} \left(1 - \rho^2_{O\overline{V}} \right), \tag{19b}$$

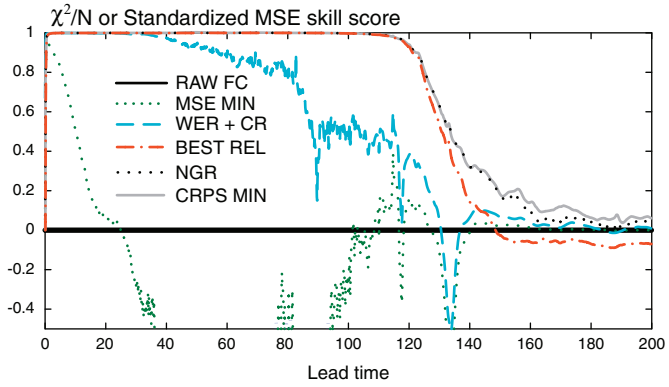

Figure 5. The skill associated with the χ^2/N score (Eq. (9)) or reduced centred random variable (RCRV) as a function of lead time for the different calibration methods.

where $\rho_{O\overline{V}}$ is the correlation between uncorrected ensemble mean and observation. Note that Johnson and Bowler (2009), rather than minimizing the MSE or the likelihood as done here, asked for the equality of correlation between the truth and the corrected ensemble mean with the correlation between the corrected ensemble members and corrected ensemble mean, but this leads to the same answer (Appendix A). Our solution can be considered as an extension for any arbitrary number of predictors.

Figure 5 shows how well WER + CR satisfies the strong ensemble-reliability constraint, as indicated by the χ^2/N skill score. The expression for this skill score is obtained by replacing CRPS in Eq. (12) with χ^2/N as given in Eq. (9). Compared with the uncorrected and the MSE MIN approaches, WER + CR performs reasonably well, but clearly even better results are obtained using the approaches introduced in the next section.

3.3. *Ensemble-spread nudging*

Although large skill improvements are obtained by ensemble-spread scaling, Figure 2 indicates that the WER + CR approach also has a short time period of negative skill (which becomes longer upon decrease of the model error; section 4.4). The skill on this time-scale can be improved using different approaches. Let us now enable spread nudging with a non-zero value of γ_2 in Eq. (3). Again this is implemented for all lead times and spatial points. This is analogous to the situation of ensemble-mean nudging. If, for some period of time, there is hardly any spread–skill relation, a good calibration scheme would set the spread to a constant ($\gamma_1 \approx 0$) and thereby increase the probabilistic skill. To ensure climatological and ensemble reliability of the corrected forecasts, it is necessary to constrain explicitly using the full constrained likelihood function, Eq. (10), allowing non-zero values for α, β_1, γ_1 and γ_2. The results are shown in Figure 4 (blue curve), and, although better in skill than the WER + CR method, strong fluctuations appear probably caused by the occurrence of degenerate solutions of the MLE. As the stability of the regression parameters is an important requirement for a good calibration method, we prefer to discard this approach.

3.4. *Non-homogeneous error variance and SER + CR constraint: BEST REL method*

For all calibration methods considered so far, the error variance as present in the ensemble-mean error distribution $\mathcal{P}$ was assumed to be homogeneous or independent of the ensembles (in other words $\mathcal{P}_n = \mathcal{P}$). For ensemble forecasts, the assumption of a constant error variance (made for OLS regression and used for all the above-introduced approaches) can be readily improved. For a skilful ensemble forecast, the variance of the error on the ensemble mean may be estimated by the ensemble variance.

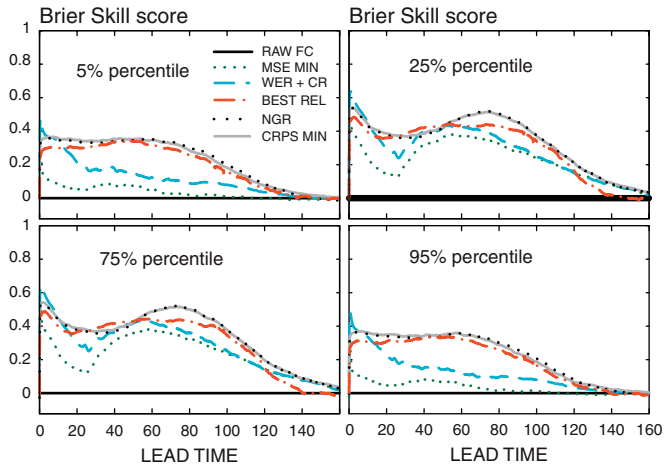

Figure 6. The Brier skill scores for different thresholds (5, 25, 75 and 95% percentiles of the climatological distribution) as a function of lead time for the different calibration methods.

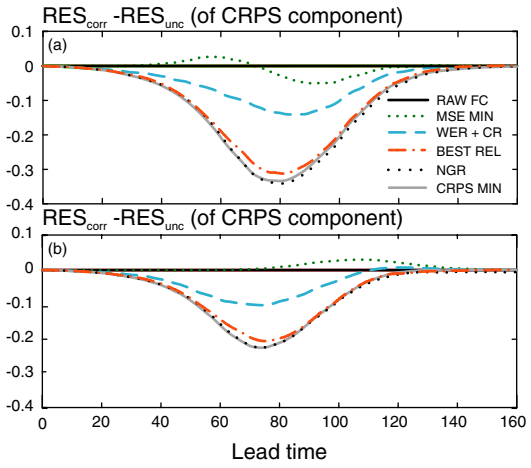

Figure 7. Change of reliability (a) $\mathrm{REL}_{\mathrm{corr}} - \mathrm{REL}_{\mathrm{unc}}$ and (b) resolution $\mathrm{RES}_{\mathrm{corr}} - \mathrm{RES}_{\mathrm{unc}}$ as a function of lead time for the different calibration methods. Eq. (21) gives the definitions of REL and RES.

In general a training set contains ensembles of both large and small spread. Analogous to the improvement of weighted least-squares regression upon ordinary least-squares regression, errors should be weighted in case prior information concerning their statistical properties is available. Upon disposal of an ensemble forecast, the best guess concerning how far the observation may be from the ensemble mean is the spread of the ensemble itself.

Here stable and improved calibration schemes are obtained by considering non-homogeneous error variances in the likelihood function. More specifically, for both choices of the error probability function $\mathcal{P}_n$ (Eq. (6)) the ensemble errors are divided by their corrected-ensemble spreads which are given by Eq. (4). The best results are obtained with the exponential distribution $\mathcal{E}(0, \delta_{C,n})$ for which the spread is given by Eq. (4b).

Finally, the BEST REL approach is defined by using the full constrained likelihood function Eq. (10) with SER and CR conditions *including* non-homogeneous error variances. Again the variational parameters are α, β_1, γ_1 and γ_2. Figure 2 shows that BEST REL is clearly one of the best methods, comparable in skill with the NGR method. Both SER and CR are necessary conditions for obtaining good skill with the BEST REL. The purple line in Figure 4 shows the CRPSS values of a calibration method which is exactly the same as the BEST REL approach but without the CR constraint. These results are far worse than those of BEST REL in Figure 2. Although this approach is optimized in order to improve the CRPSS, it also perform well for events with different thresholds. This is illustrated in Figure 6 where the Brier skill scores are shown for different thresholds, corresponding to the 5, 25, 75 and 95% percentiles of the climatological distribution. Apparently the BEST REL approach is distinctly better than WER + CR at the most extreme percentiles.

Note also that exact solutions exist for the case of normally distributed errors with non-homogeneous error variance without ensemble-spread nudging, but with hard SER and CR constraints. These are outlined in Appendix C. Figure 4 shows that enforcing SER only (brown line) is not worthwile and the approach with both enforced (black line) has only solutions up to a certain lead time (here $t \approx 20$). This advocates the use of soft rather than hard constraints.

3.5. *CRPS minimization*

The last method is again a MBM method which, analogous to NGR, explicitly minimizes the CRPS, but without any assumption on the distribution. The CRPS corresponding to the observations $X_{\mathrm{O},n}$ $(n = 1, \ldots, N)$ and the corrected-forecast members $X^m_{\mathrm{C},n}$ $(m = 1, \ldots, M)$ can be written as (Gneiting and Raftery, 2007)

$$\mathrm{CRPS} = \left\langle \left\langle |X^m_{\mathrm{C},n} - X_{\mathrm{O},n}| \right\rangle_m - \frac{\delta_n}{2} \right\rangle_n . \tag{20}$$

The corrected forecast Eq. (1) is used with the absolute-value spread given in Eq. (3b). Note that, given the ensemble spreads δ_n of the uncorrected forecast, the right-hand side of Eq. (20) consists of $N \times M$ terms due to the averaging over all N data points and M ensemble members. Therefore the minimization of this CRPS expression is computationally intensive which makes this method mostly suitable for small ensemble sizes M or small sample sizes N.

4. Verification

As already mentioned in the Introduction, for a perfect ensemble forecast, the ensemble-mean error distribution should be the same as the distribution of ensemble members around the ensemble mean. It is now assessed how well the ER constraint is satisfied, and to what extent moments higher than two of the distributions agree using the different calibration approaches presented in section 3. The influence of the model, the model error amplitude, the initial condition error and the ensemble size on the calibration techniques are also discussed.

4.1. *Reliability component of CRPS*

In Hersbach (2000) it was shown that the CRPS can be decomposed into the following components:

$$\mathrm{CRPS} = \mathrm{REL} + \mathrm{UNC} - \mathrm{RES}\,, \tag{21}$$

where REL stands for reliability, UNC for uncertainty and RES for resolution. The reliability term expresses how well the predicted probability for a certain event matches the corresponding observed frequency. The resolution, on the other hand, expresses how different the probabilistic forecasts are from the climatological distribution, overall. Since UNC depends on the observations only, the interesting quantities are the reliability and resolution. Figure 7(a) displays the gain in reliability compared with the uncorrected forecast, that is $\mathrm{REL}_{\mathrm{corr}} - \mathrm{REL}_{\mathrm{unc}}$. Positive values of this difference indicate a deterioration of the reliability. Except for the MSE MIN method, all our methods improve the reliability of the uncorrected forecast. Interestingly, ensemble-spread scaling as featured in the WER + CR approach substantially increases the reliability with respect to the spread-scaling approach MSE MIN. A comparable additional skill gain is obtained using

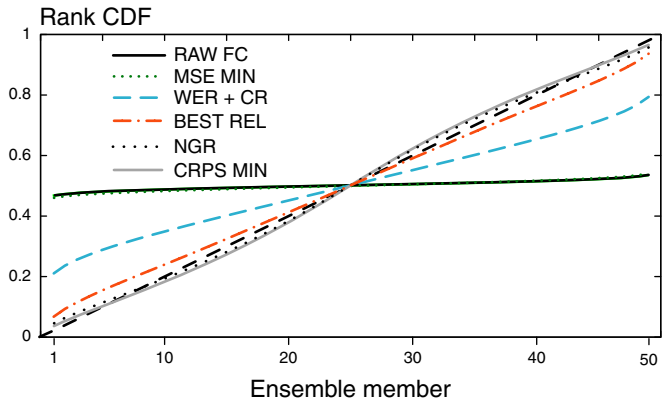

Figure 8. The cumulative rank distribution as a function of the rank for the different calibration methods. This diagram is obtained by pooling the rank distributions of all lead times before 50 units of time. The rank CDF of a perfect forecast would lie on the diagonal (thin dashed) line.

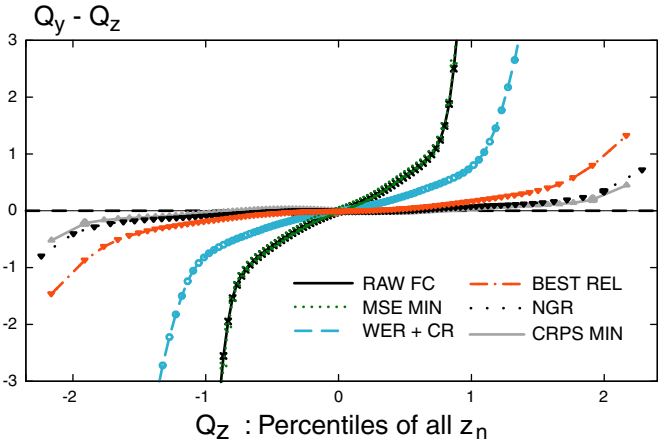

Figure 9. The difference of percentiles $Q_y - Q_z$ against the percentiles Q_z associated with the distributions of all standardized errors $z_n = (X_{O,n} - \overline{X}_{C,n})/\sigma_{C,n}$ and the standardized ensemble members $y_n^m = (X_{C,n}^m - \overline{X}_{C,n})/\sigma_{C,n}$. All lead times are pooled. For a perfectly reliable forecast $Q_y - Q_z$ should be zero for all percentiles.

ensemble-spread nudging imposed by BEST REL, NGR and CRPS MIN. Although there is a net gain in skill, the improvement in terms of reliability by ensemble-spread correction is at the expense of a loss in resolution, as shown in Figure 7(b) where all methods, except for MSE MIN clearly degrade the resolution with respect to the uncorrected forecast. The loss is highest for the most reliable methods NGR and CRPS MIN.

4.2. *Reliability in terms of cumulative rank distribution*

Figure 8 displays the CDF of the rank probability distributions (or integrated rank histogram), obtained by pooling all times up to $t = 50$. The rank CDF for a perfectly reliable ensemble forecast would be on the diagonal (dashed black line). Clearly this is far from being realized for the uncorrected and MSE MIN forecasts. The NGR and CRPS MIN forecasts approach very well perfect reliability while BEST REL is close to it.

4.3. *Higher moments of the ensemble distribution*

Finally note that all ensembles obtained using a MBM correction scheme preserve (normalized) ensemble moments like skewness and kurtosis exactly. This can be checked using the expression of the corrected forecast Eq. (1). On the other hand, statistical methods such as NGR produce ensembles with constant skewness and kurtosis and are highly sensitive to outliers even though the extreme event may be very unlikely and would be discarded by forecasters. Moreover the calibrated forecasts based on MBM approaches BEST REL and CRPS MIN display an error distribution which is well represented by the ensemble distributions. In order to verify this, we compare in Figure 9 the percentiles Q_z of the distribution of all standardized errors $z_n = (X_{O,n} - \overline{X}_{C,n})/\sigma_{C,n}$ (x-axis) with the percentiles Q_y of the distribution of all standardized ensemble members $y_n^m = (X_{C,n}^m - \overline{X}_{C,n})/\sigma_{C,n}$. All values are obtained by pooling all lead times. For a perfectly reliable forecast, the difference $Q_y - Q_z$ would be zero. Due to the absence of bias, the signs of the quantiles of z_n and y_n^m seem to agree well for all forecasts however, due to the underdispersiveness, the standardized errors z_n of the raw forecast and of the MSE MIN method are generally much larger than the ensemble spread. In agreement with our previous findings, the spread-scaling approach WER + CR improves upon MSE MIN but BEST REL, NGR and CRPS MIN are much better. Note that CRPS MIN is slightly better than NGR in the sense that its associated quantile differences $Q_y - Q_z$ are close to zero but also in the sense that the quantiles Q_z themselves are closer to zero.

4.4. *Sensitivity to model, model error amplitude and ensemble size*

The sensitivity of the post-processing methods to changes of the amplitude of the model error and the initial-condition error have been tested. Figure 10 shows CRPSS as a function of lead time for a small ($\nu' = \nu + 0.0001$) and a large model error ($\nu' = \nu + 0.005$). Note that the inflation of the initial-condition spread is qualitatively equivalent to a decrease of model error, and vice versa (not shown). In this case the best methods (BEST REL, NGR and CRPS MIN) retain their positive skill and the efficiency hierarchy of post-processing methods remains unaffected. The effect of a reduction of the model error (or increase of initial-condition error) is a reduction of the skill gain and in particular for the ensemble-mean nudging in the time interval between 60 and 140 time units. This can be understood by the fact that, with a larger ensemble spread, the uncorrected ensemble will probe the entire attractor at an earlier stage thereby setting the ensemble mean equal to the climatological mean. In other words, the ensemble mean of the uncorrected forecast will already be nudged to the climatological mean.

Analogously to Figure 2, Figure 11 displays the CRPSS but for the single-scale Lorenz 96 system (Lorenz, 1996) instead of the KS equations. Again it seems that our conclusions are generic in the sense that the efficiency hierarchy of post-processing methods is preserved. Tests with reduced ensemble sizes for calibration training have also been performed. Remarkably, even with only four members, skill changes are small for all presented methods, except at long lead times ($t > 140$) for which all skills become slightly negative.

5. Conservation of correlation structure

As illustrated in the Introduction, the main benefit of using a MBM post-processing method is that, when independent post-processing is performed on different spatial points, lead times or variables, the structural information that was present in the uncorrected forecast is largely preserved. This is a direct consequence of Eq. (1) which imposes a strong relation between the MBM-corrected forecast and the raw forecast. Note that this preservation is not a consequence of the use of more than one predictor. Randomly reconstructed ensembles based on the predictive distributions of statistical post-processing methods such as NGR destroy to a large extent the correlation structure.

Figure 12 shows the Pearson correlation between the KS order parameters $\psi(x)$ and $\psi(x+1)$ as a function of lead time. Note that, due to the translational symmetry of the KS system, all values

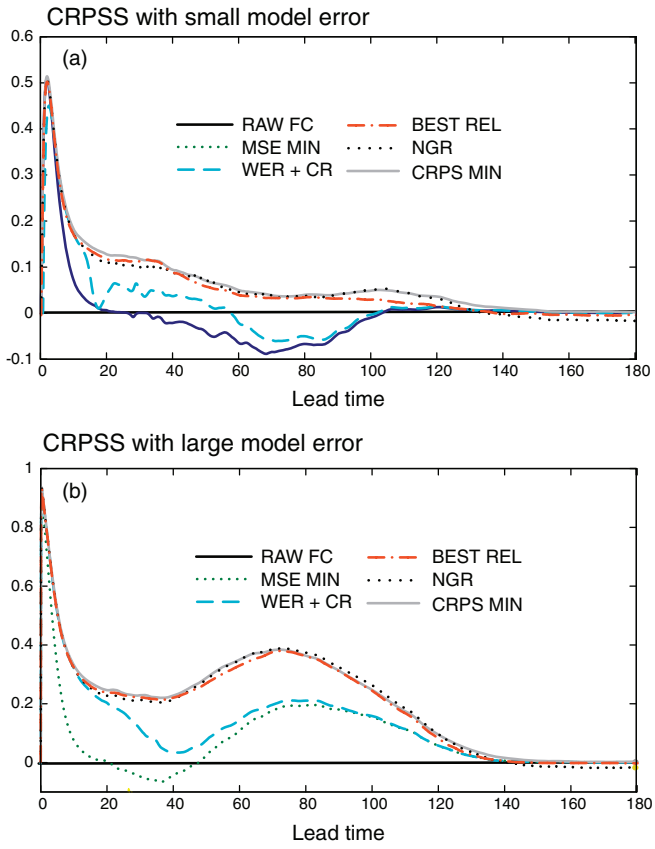

Figure 10. As Figure 2, but for (a) a smaller model error ($\nu' = \nu + 0.0001$), and (b) a larger model error ($\nu' = \nu + 0.005$).

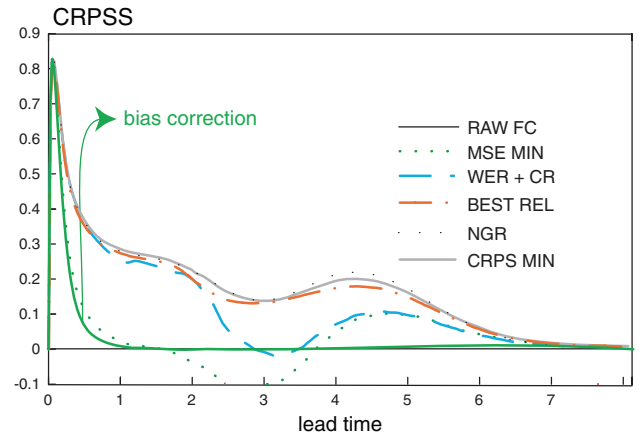

Figure 11. As Figure 2, but for the single-scale Lorenz 96 system. Both the training and verification sets include 2000 ensembles each of 100 members. 36 grid points and a forcing parameter $F = 10$ are used, while for generating the observations $F' = F + 0.01$ is used. The error-doubling time associated with the largest Lyapunov exponent λ_L is equal to $\ln(2)/\lambda_L = 0.3$ time units. At time zero, the observation is randomly sampled from the normally distributed forecast ensemble which has a standard deviation of 10^{-4} at each grid point. The bias correction which, contrary to the KS system, introduces a substantial correction at short time-scales ($t < 1$), is also plotted.

x are equivalent. Clearly, the correlation of all the MBM methods is quite close to that of the raw forecast and deviates weakly for lead times between 60 and 120 time units. On the other hand, for the NGR method the correlation becomes small after 100 time units. This can be understood from the fact that, at long lead times, both $\psi(x)$ and $\psi(x+1)$ are random and independent samples from a normal distribution, the mean and variance of which are very close to the climatological values.

For short and long lead times, the spatial correlation structures of the MBM approaches are exactly equal to those of the raw

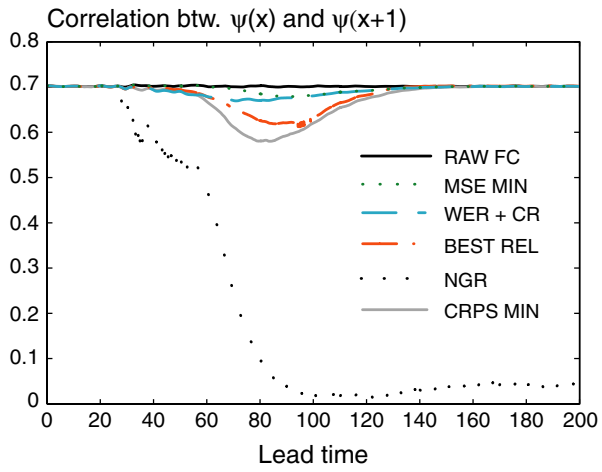

Figure 12. Pearson correlation between the KS order parameters $\psi(x)$ and $\psi(x+1)$ against lead time for the different calibration methods. All methods except NGR are close to the correlation of the raw forecast for all lead times.

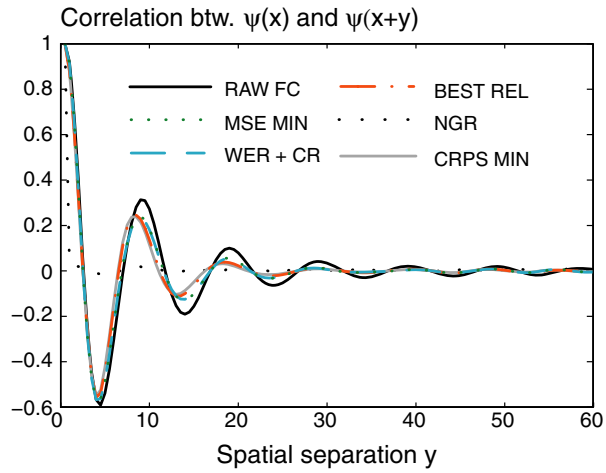

Figure 13. Pearson correlation between the KS order parameters $\psi(x)$ and $\psi(x+y)$ against spatial distance y for the different calibration methods at lead time equal to 100 time units. All methods except for NGR keep information on the correlation structure.

forecast (not shown). They deviate slightly for intermediate lead times between 60 and 120 time units, as we illustrate for lead time 100 in Figure 13. Clearly, all but the NGR method reproduce well the correlation structure of the raw forecast between $\psi(x)$ and $\psi(x+y)$ as a function of y.

Note that, due to a particular symmetry of the system, the correlation between $\psi(x)$ and $\psi(x+y)$ is equal to that between $\psi(x)$ and $\psi(x-y)$. In addition, due to the Wiener–Khinchin theorem, the power spectra of the MBM methods are still close to the ones of the uncorrected forecast. The power spectrum of the NGR-corrected forecast, on the other hand, is white at long lead times.

The structure of forecast *uncertainty* among the different spatial locations is also of interest. This can be measured by the correlation between the deviations from the ensemble mean at location x and the one at $x+1$. This correlation closely follows the one shown in Figure 12 for all calibration techniques except for NGR, for which it is identically zero. Another measure is the correlation between the error (compared with the observation) at location x and the one at location $x+1$. Again, except for a convergence to half of its initial value for NGR, this correlation behaves identically to the one given in Figure 12.

6. Application

Results of a preliminary test on the European Centre for Medium-range Weather Forecasts (ECMWF) forecast for Belgium are

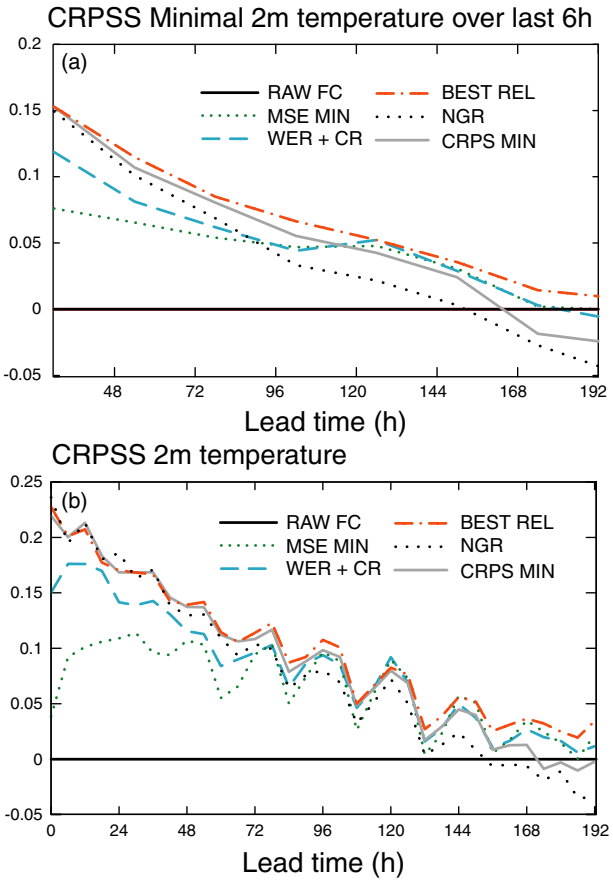

Figure 14. CRPSS against lead time for (a) minimum 2 m temperature over the last 6 h (min$T$) and (b) 2 m temperature (2mT) using different corrections of the ECMWF ensemble forecast. Between the lead times, in (a) there are 24 h intervals starting at 30 h, while in (b) there are 6 h intervals starting at 0 h.

shown in Figure 14 where the new calibration methods are compared with the benchmark NGR method. Two-metre temperature (2mT) and minimum temperature over the last 6 h (min$T$) are validated over 1 year (June 2011–June 2012) of the 51-member ECMWF Ensemble Prediction System (EPS) forecast and averaged over seventeen SYNOP stations in Belgium. For training, the five-member ECMWF EPS hindcasts were used. All our post-processing methods were independently applied on 2mT and min$T$, both of which are model variables of the EPS. Since observations of min$T$ are only available once daily at 0600 UTC, only these post-processed values are shown. On the other hand, for 2mT results are shown every 6 h. The conclusion drawn from our simple models, indicating that WER + CR improves upon MSE MIN, extends to real applications. It is clear also that the best new methods (BEST REL and CRPS MIN) provide more skill than the benchmark NGR method. Moreover, the NGR approach suffers from loss of correlations and consistency issues as highlighted in Figure 1 and explained in the Introduction. Such problems are avoided using a MBM calibration method. Interestingly, the spread calibration gives improvements during the first week although the training set includes only five hindcast ensemble members. Note that post-processing with both 2mT and min$T$ as predictors might lead to overfitting due to their strong correlations.

7. Conclusion

A set of post-processing techniques based on linear regression enforcing the forecast to become climatologically reliable (CR) and/or to have reliable ensembles (ER), have been introduced. The methods have been tested extensively using a low-order chaotic system with model error, and preliminary results on operational weather forecasts are presented. For lead times much longer than the error-doubling time, the constrained approaches are superior to the uncorrected forecast and the forecast obtained by an unconstrained maximum likelihood estimation. Moreover, the use of the ensemble spread for estimating the error variance of the ensemble mean leads to a better calibration and avoids the undercorrection of ensembles with small spreads. Apart from the fact that the best corrected forecasts satisfy both constraints, their CRPS scores are at the same level as that of the NGR-corrected forecast which, by construction, minimizes the CRPS. Moreover, the reliable forecast preserves not only the higher moments from the uncorrected forecast (kurtosis and skewness), but also the rank structure of the ensemble and, in a large part, the spatio-temporal coherence of the forecasts. Calibration results of the ECMWF forecast for Belgium even indicate slightly superior scores for the MBM approach compared with the NGR approach. An extensive analysis of these forecasts will be reported soon.

Appendix A

Equality of correlations

In this Appendix we prove that, for a forecast that satisfies both WER and CR constraints, the correlation between the observation and the ensemble mean equals the correlation between the ensemble members and the ensemble mean. Without loss of generality, we assume the forecast and the observation to have a zero mean. We first write the weak ensemble reliability constraint Eq. (8):

$$\left\langle (\overline{X}_{\mathrm{C},n} - X_{\mathrm{O},n})^2 \right\rangle_n = \left\langle \left(\overline{X}_{\mathrm{C},n} - X^m_{\mathrm{C},n}\right)^2 \right\rangle_{m,n}. \tag{A1}$$

Expanding the squares, we obtain:

$$\sigma^2_{\mathrm{O}} - 2\left\langle \overline{X}_{\mathrm{C},n} X_{\mathrm{O},n} \right\rangle_n = \sigma^2_{\mathrm{C}} - 2\left\langle \overline{X}_{\mathrm{C},n} X^m_{\mathrm{C},n} \right\rangle_{m,n}. \tag{A2}$$

We continue by using the CR constraint Eq. (7) and by dividing by $\sigma_{\mathrm{O}} = \sigma_{\mathrm{C}}$ and by the standard deviation of the ensemble mean $\sigma_{\overline{X}_{\mathrm{C}}}$:

$$\frac{\left\langle \overline{X}_{\mathrm{C},n} X_{\mathrm{O},n} \right\rangle_n}{\sigma_{\overline{X}_{\mathrm{C}}}\sigma_{\mathrm{O}}} = \frac{\left\langle \overline{X}_{\mathrm{C},n} X^m_{\mathrm{C},n} \right\rangle_{m,n}}{\sigma_{\overline{X}_{\mathrm{C}}}\sigma_{\mathrm{C}}}. \tag{A3}$$

This expresses the equality of the correlations.

Appendix B

MSE MIN with CR constraint

Here exact results for the regression coefficients are provided for the MSE MIN with the CR constraint. Consider the corrected forecast for each member m of an ensemble n (we assume $\tau_n = 1$):

$$X^m_{\mathrm{C},n} = \alpha + \sum_p^P \beta_p \overline{V}_{p,n} + \varepsilon^m_n. \tag{B1}$$

The log-likelihood with the CR constraint reads (Van Schaeybroeck and Vannitsem, 2013):

$$\ln \mathcal{L} = \ln \mathcal{L}_0 - \eta \left(1 - \frac{\sigma^2_{\mathrm{O}}}{\sigma^2_{\mathrm{C}}}\right)^2. \tag{B2}$$

Maximizing this constrained likelihood with respect to α, $\boldsymbol{\beta}$ and η, one straightforwardly arrives at:

$$\boldsymbol{\beta} = \frac{\boldsymbol{\sigma}^2_{\mathrm{O}\overline{V}} \boldsymbol{\sigma}^{-2}_{\overline{V}} \sqrt{\sigma^2_{\mathrm{O}} - \langle \sigma^2_{\varepsilon,n} \rangle_n}}{\sqrt{\boldsymbol{\sigma}^2_{\mathrm{O}\overline{V}} \boldsymbol{\sigma}^{-2}_{\overline{V}} (\boldsymbol{\sigma}^2_{\mathrm{O}\overline{V}})^T}}, \tag{B3a}$$

$$\alpha = \langle \boldsymbol{X}_{\mathrm{O}} \rangle_n - \langle \boldsymbol{\beta} \overline{\boldsymbol{V}} \rangle_n, \tag{B3b}$$

with $\boldsymbol{X}_O = (X_{O,1}, \ldots, X_{O,N})$ the vector of all N observations, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$ the vector of P regression parameters and $\overline{\boldsymbol{V}}$ the matrix of size $P \times N$ containing all ensemble-mean predictors. The expression $\langle \boldsymbol{A} \rangle_n$ denotes an averaging over all N vector elements of $\boldsymbol{A}$ and $\boldsymbol{\beta}\overline{\boldsymbol{V}}$ corresponds to the summation of Eq. (B1). The covariance matrix $\boldsymbol{\sigma}^2_{\overline{V}}$ and the vector $\boldsymbol{\sigma}^2_{\mathrm{O}\overline{V}}$ have been defined in Eq. (16). For a deterministic forecast the average ensemble variance $\langle \sigma^2_{\varepsilon,n} \rangle_n$ is zero and Eq. (B3) correspond to the known solution for Error-in-Variables MOS (EVMOS; Vannitsem, 2009; Van Schaeybroeck and Vannitsem, 2011, 2012). Therefore EVMOS can be interpreted as an approach that minimizes the MSE of the ensemble mean subject to the CR constraint. Note that other methods exist that satisfy the CR constraint. Examples include the geometric-mean regression and the time-dependent Tikhonov regression (Van Schaeybroeck and Vannitsem, 2011).

Appendix C

Hard-constrained MLE with non-homogeneous error variance

In this Appendix exact solutions of the hard-constrained MLE are provided with a normal error distribution with non-homogeneous or ensemble-*dependent* variance $\mathcal{N}(0, \sigma^2_{\varepsilon,n})$. These methods were already partly discussed by Van Schaeybroeck and Vannitsem (2013). Let us first define the standardized variables:

$$\widetilde{V}^m_{p,n} = \frac{V^m_{p,n}}{\sigma_{\varepsilon,n}}, \quad \widetilde{X}^m_{\mathrm{O},n} = \frac{X^m_{\mathrm{O},n}}{\sigma_{\varepsilon,n}}, \quad \widetilde{X}^m_{\mathrm{C},n} = \frac{X^m_{\mathrm{C},n}}{\sigma_{\varepsilon,n}}. \tag{C1}$$

Using this notation we can redefine the strong ensemble reliability constraint of Eq. (9):

$$\left\langle \left(\widetilde{\overline{X}}_{\mathrm{C},n} - \widetilde{X}_{\mathrm{O},n} \right)^2 \right\rangle_n = 1. \tag{C2}$$

Two types of 'debiasing' can be defined: the conventional definition implies $\langle \boldsymbol{X}_\mathrm{C} \rangle_n = \langle \boldsymbol{X}_\mathrm{O} \rangle_n$ but a standardized debiasing $\langle \widetilde{\boldsymbol{X}}_\mathrm{C} \rangle_n = \langle \widetilde{\boldsymbol{X}}_\mathrm{O} \rangle_n$ can also be proposed. Two debiasing coefficients α_1 and α_2 can then be introduced such that

$$X^m_{\mathrm{C},n} = \alpha_1 + \alpha_2 \sigma_{\varepsilon,n} + \sum_p^P \beta_p \overline{V}_{p,n} + \gamma_1 \varepsilon^m_n. \tag{C3}$$

We proceed with a corrected forecast of this form, thus allowing for spread scaling.

C1. Unconstrained approach with non-homogeneous error variance

Minimization of the likelihood function Eq. (5) of the corrected forecast Eq. (C3) associated with a normal ensemble-mean error distribution with respect to α_2, $\boldsymbol{\beta}$ and γ_1 and assuming $\alpha_1 = 0$ yields

$$\boldsymbol{\beta} = \widetilde{\boldsymbol{\sigma}}^2_{\mathrm{O}\overline{V}} \widetilde{\boldsymbol{\sigma}}^{-2}_{\overline{V}}, \tag{C4a}$$

$$\alpha_2 = \langle \widetilde{\boldsymbol{X}}_\mathrm{O} \rangle_n - \langle \boldsymbol{\beta} \widetilde{\overline{\boldsymbol{V}}} \rangle_n, \tag{C4b}$$

$$\gamma_1^2 = \widetilde{\sigma}^2_\mathrm{O} - \widetilde{\boldsymbol{\sigma}}^2_{\mathrm{O}\overline{V}} \widetilde{\boldsymbol{\sigma}}^{-2}_{\overline{V}} \left(\widetilde{\boldsymbol{\sigma}}^2_{\mathrm{O}\overline{V}} \right)^\mathrm{T}, \tag{C4c}$$

where the covariances $\widetilde{\boldsymbol{\sigma}}_{\mathrm{O}\overline{V}}$, $\widetilde{\boldsymbol{\sigma}}_{\overline{V}}$ and $\widetilde{\sigma}_\mathrm{O}$ are the same as $\boldsymbol{\sigma}_{\mathrm{O}\overline{V}}$, $\boldsymbol{\sigma}_{\overline{V}}$ and σ_O (Eq. 16) but for the standardized variables. Importantly, the strong ensemble reliability (SER) criterion is satisfied.

C2. SER and CR with hard constraints

Both SER and CR are now enforced allowing for ensemble-spread nudging and scaling. Differentiation with respect to all parameters leads to the equations:

$$\sigma^2_\mathrm{O} - \boldsymbol{\beta} \boldsymbol{\sigma}^2_{\overline{V}} \boldsymbol{\beta}^\mathrm{T} = \langle \sigma^2_{\varepsilon,n} \rangle_n \left(\boldsymbol{\beta} \widetilde{\boldsymbol{\sigma}}^2_{\overline{V}} \boldsymbol{\beta}^\mathrm{T} - 2 \widetilde{\boldsymbol{\sigma}}^2_{\mathrm{O}\overline{V}} \boldsymbol{\beta}^\mathrm{T} + \widetilde{\sigma}^2_\mathrm{O} \right), \tag{C5a}$$

$$\boldsymbol{\beta} = \widetilde{\boldsymbol{\sigma}}^2_{\mathrm{O}\overline{V}} \left(\frac{\widetilde{\boldsymbol{\sigma}}^2_{\overline{V}} - 2\eta \boldsymbol{\sigma}^2_{\overline{V}}}{\langle \sigma^2_{\varepsilon,n} \rangle_n (1 - 2\eta)} \right)^{-1}, \tag{C5b}$$

$$\gamma_1^2 = \langle \sigma^2_{\varepsilon,n} \rangle_n^{-1} \left(\sigma^2_\mathrm{O} - \boldsymbol{\beta} \boldsymbol{\sigma}^2_{\overline{V}} \boldsymbol{\beta}^\mathrm{T} \right), \tag{C5c}$$

$$\begin{aligned} \alpha_1 = & \left(1 - \langle \sigma_{\varepsilon,n} \rangle_n \langle \sigma^{-1}_{\varepsilon,n} \rangle_n \right)^{-1} \\ & \times \left[\langle \boldsymbol{X}_\mathrm{O} \rangle_n - \langle \sigma_{\varepsilon,n} \rangle_n \langle \widetilde{\boldsymbol{X}}_\mathrm{O} \rangle_n - \boldsymbol{\beta} \left(\langle \overline{\boldsymbol{V}} \rangle_n - \langle \widetilde{\overline{\boldsymbol{V}}} \rangle_n \langle \sigma_{\varepsilon,n} \rangle_n \right) \right], \end{aligned} \tag{C5d}$$

$$\begin{aligned} \alpha_2 = & \left(\langle \sigma_{\varepsilon,n} \rangle_n \langle \sigma^{-1}_{\varepsilon,n} \rangle_n - 1 \right)^{-1} \\ & \times \left[\langle \boldsymbol{X}_\mathrm{O} \rangle_n \langle \sigma^{-1}_{\varepsilon,n} \rangle_n - \langle \widetilde{\boldsymbol{X}}_\mathrm{O} \rangle_n - \boldsymbol{\beta} \left(\langle \overline{\boldsymbol{V}} \rangle_n \langle \sigma^{-1}_{\varepsilon,n} \rangle_n - \langle \widetilde{\overline{\boldsymbol{V}}} \rangle_n \right) \right]. \end{aligned} \tag{C5e}$$

In principle these equations can be solved by substituting Eq. (C5b) into Eq. (C5a) and solving for η and then substituting the result into the subsequent equations. An exact result for the case of one predictor ($P = 1$) has been obtained:

$$\begin{aligned} \beta_1 = & \left[\sigma^2_{\overline{V}} + \widetilde{\sigma}^2_{\overline{V}} \langle \sigma^2_{\varepsilon,n} \rangle_n \right]^{-1} \left\{ \widetilde{\sigma}^2_{\mathrm{O}\overline{V}} \langle \sigma^2_{\varepsilon,n} \rangle_n \right. \\ & \left. + \sqrt{\widetilde{\sigma}^4_{\mathrm{O}\overline{V}} \langle \sigma^2_{\varepsilon,n} \rangle^2_n + \left(\sigma^2_{\overline{V}} + \widetilde{\sigma}^2_{\overline{V}} \langle \sigma^2_{\varepsilon,n} \rangle_n \right) \left(\sigma^2_\mathrm{O} - \widetilde{\sigma}^2_\mathrm{O} \langle \sigma^2_{\varepsilon,n} \rangle_n \right)} \right\}, \end{aligned} \tag{C6a}$$

$$\begin{aligned} \gamma_1^2 = & \frac{\sigma^2_\mathrm{O} \widetilde{\sigma}^2_{\overline{V}} + \sigma^2_{\overline{V}} \widetilde{\sigma}^2_\mathrm{O}}{\sigma^2_{\overline{V}} + \widetilde{\sigma}^2_{\overline{V}} \langle \sigma^2_{\varepsilon,n} \rangle_n} - \frac{2 \sigma^2_{\overline{V}} \langle \sigma^2_{\varepsilon,n} \rangle_n \widetilde{\sigma}^4_{\mathrm{O}\overline{V}}}{\left(\sigma^2_{\overline{V}} + \widetilde{\sigma}^2_{\overline{V}} \langle \sigma^2_{\varepsilon,n} \rangle_n \right)^2} \\ & - \frac{2 \sqrt{\widetilde{\sigma}^4_{\mathrm{O}\overline{V}} \langle \sigma^2_{\varepsilon,n} \rangle^4_n + \left(\sigma^2_{\overline{V}} + \widetilde{\sigma}^2_{\overline{V}} \langle \sigma^2_{\varepsilon,n} \rangle_n \right) \left(\sigma^2_\mathrm{O} - \widetilde{\sigma}^2_\mathrm{O} \langle \sigma^2_{\varepsilon,n} \rangle_n \right)}}{\sigma^{-2}_{\overline{V}} \widetilde{\sigma}^{-2}_{\mathrm{O}\overline{V}} \left(\sigma^2_{\overline{V}} + \widetilde{\sigma}^2_{\overline{V}} \langle \sigma^2_{\varepsilon,n} \rangle_n \right)^2}. \end{aligned} \tag{C6b}$$

The coefficients α_1 and α_2 are given in Eqs (C5d) and (C5e). Note that solutions only exist when both the right-hand side of Eq. (C6b) and the quantity under the square root are positive. For chaotic systems this condition is not satisfied at long lead times when forecast and observations are decorrelated since the argument under the square root is approximately equal to

$$\left(\sigma^2_{\overline{V}} + \widetilde{\sigma}^2_{\overline{V}} \langle \sigma^2_{\varepsilon,n} \rangle_n \right) \sigma^2_\mathrm{O} \left(1 - \left\langle \sigma^{-2}_{\varepsilon,n} \right\rangle_n \langle \sigma^2_{\varepsilon,n} \rangle_n \right).$$

At long lead times, the distribution of the ensemble variance $\sigma^2_{\varepsilon,n}$ is peaked and, in the absence of a strong skewness, the peak is located at the mean value $\left\langle \sigma^2_{\varepsilon,n} \right\rangle_n$. An expansion of $\sigma^{-2}_{\varepsilon,n}$ around this value amounts to

$$\left\langle \sigma^{-2}_{\varepsilon,n} \right\rangle_n \approx \left\langle \sigma^2_{\varepsilon,n} \right\rangle_n^{-1} + \left\langle \sigma^2_{\varepsilon,n} \right\rangle_n^{-3} \left\langle \left(\sigma^2_{\varepsilon,n} - \left\langle \sigma_{\varepsilon,n} \right\rangle_n \right)^2 \right\rangle_n.$$

This implies that $\left\langle \sigma^{-2}_{\varepsilon,n} \right\rangle_n \left\langle \sigma^2_{\varepsilon,n} \right\rangle_n > 1$ and the argument under the square root becomes negative and both β_1 and γ_1 are ill-defined. The case of a positively skewed distribution for the ensemble variance distribution can be modelled by assuming $\ln\left(\sigma^2_{\varepsilon,n}\right)$ to be normally distributed, or $\ln\left(\sigma^2_{\varepsilon,n}\right) \sim \mathcal{N}(a, b^2)$. It is then readily found that $\left\langle \sigma^{-2}_{\varepsilon,n} \right\rangle_n \left\langle \sigma^2_{\varepsilon,n} \right\rangle_n = \mathrm{e}^{b^2} > 1$. Another argument considers the SER constraint $\chi^2/N = 1$, which, at long lead times and together with the CR constraint, becomes

$$2\sigma^2_{\overline{X}_\mathrm{C}} = \left\langle \sigma^{-2}_{\varepsilon,n} \right\rangle_n^{-1} - \langle \sigma^2_{\varepsilon,n} \rangle_n.$$

Again, when $\left\langle \sigma^{-2}_{\varepsilon,n} \right\rangle_n \langle \sigma^2_{\varepsilon,n} \rangle_n > 1$, the SER constraint cannot be satisfied.

Acknowledgements

This work has benefited from useful discussions with Michael Scheuerer, Emmanuel Roulin, Pascal Mailier, Martin Leutbecher, and Olivier Talagrand. The remarks of Tilmann Gneiting, Lesley De Cruz, Maurice Schmeits, Kees Kok and two anonymous referees strongly improved the manuscript. This work is partially supported by the Belgian Federal Science Policy office under contract SD/CA/04A.

References

Baringhaus L, Franz C. 2004. On a new multivariate two-sample test. *J. Multivariate Anal.* **88**: 190–206.

Candille G, Côté C, Houtekamer PL, Pellerin G. 2007. Verification of an ensemble prediction system against observations. *Mon. Weather Rev.* **135**: 2688–2699.

Glahn HR, Lowry DA. 1972. The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.* **11**: 1203–1211.

Glahn B, Peroutka M, Wiedenfeld J, Wagner J, Zylstra G, Schuknecht B, Jackson B. 2009. MOS uncertainty estimates in an ensemble framework. *Mon. Weather Rev.* **137**: 246–268.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Assoc.* **102**: 359–378.

Gneiting T, Raftery AE, Westveld A, Goldman T. 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133**: 1098–1118.

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc: Ser. B: Stat. Methodol.* **69**: 243–268.

Hagedorn R, Hamill TM, Whitaker JS. 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Weather Rev.* **136**: 2608–2619.

Hersbach H. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting* **15**: 559–570.

Johnson C, Bowler N. 2009. On the reliability and calibration of ensemble forecasts. *Mon. Weather Rev.* **137**: 1717–1720.

Kalnay E. 2002. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press: Cambridge, UK.

Kharin VV, Zwiers FW. 2003. Improved seasonal probability forecasts. *J. Clim.* **16**: 1684–1701.

Leutbecher M, Palmer TN. 2008. Ensemble forecasting. *J. Comput. Phys.* **227**: 3515–3539.

Lorenz EN. 1996. Predictability –a problem partly solved. In *Proceedings of Seminar on Predictability*, Vol. 1: 1–18. ECMWF: Reading, UK.

Manneville P. 1990. *Dissipative Structures and Weak Turbulence*. Academic Press: London.

Möller A, Lenkoski A, Thorarinsdottir TL. 2013. Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Q. J. R. Meteorol. Soc.* **139**: 982–991.

Nicolis C, Perdigao RAP, Vannitsem S. 2009. Dynamics of prediction errors under the combined effect of initial condition and model errors. *J. Atmos. Sci.* **66**: 766–778.

Pinson P. 2012. Adaptive calibration of (u, v) wind ensemble forecasts. *Q. J. R. Meteorol. Soc.* **138**: 1273–1284.

Roulin E, Vannitsem S. 2012. Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Mon. Weather Rev.* **140**: 874–888.

Schefzik R, Thorarinsdottir TL, Gneiting T. 2013. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statist. Sci.* **28**: 465–660.

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

Schmeits MJ, Kok KJ. 2010. A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Mon. Weather Rev.* **138**: 4199–4211.

Schuhen N, Thorarinsdottir TL, Gneiting T. 2012. Ensemble model output statistics for wind vectors. *Mon. Weather Rev.* **140**: 3204–3219.

Vannitsem S. 2009. A unified linear Model Output Statistics scheme for both deterministic and ensemble forecasts. *Q. J. R. Meteorol. Soc.* **135**: 1801–1815.

Vannitsem S, Hagedorn R. 2011. Ensemble forecast post-processing over Belgium: Comparison of deterministic-like and ensemble regression methods. *Meteorol. Appl.* **18**: 94–104.

Vannitsem S, Nicolis C. 2008. Dynamical properties of model output statistics forecasts. *Mon. Weather Rev.* **136**: 405–419.

Van Schaeybroeck B, Vannitsem S. 2011. Post-processing through linear regression. *Nonlinear Proc. Geophys.* **18**: 147–160.

Van Schaeybroeck B, Vannitsem S. 2012. Toward Post-Processing Ensemble Forecasts Based on Hindcasts. Scientific et Technical Publications No. 61. Royal Meteorological Institute: Brussels.

Van Schaeybroeck B, Vannitsem S. 2013. Reliable probabilities through statistical post-processing of ensemble forecasts. In *Proceedings of the European Conference on Complex Systems*, Gilbert T, Kirkilionis M, Nicolis G. (eds.): 347–352. Springer: Berlin.

Wilks DS. 1995. *Statistical Methods in the Atmospheric Sciences*. Academic Press: London.

Wilks DS. 2006. Comparison of ensemble-MOS methods in the Lorenz 96 setting. *Meteorol. Appl.* **13**: 243–256.

Wilks DS. 2009. Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorol. Appl.* **16**: 361–368.

