Diagnosis of Ensemble Forecasting Systems

Martin Leutbecher

ECMWF, Shinfield Park, Reading
RG2 9AX, United Kingdom

ABSTRACT

Ensemble forecasts should strive for quantifying forecast uncertainties as accurately as possible. Useful diagnostics should measure how well this goal has been achieved and they should identify aspects of an ensemble prediction system that require improvement.

First, an overview of standard diagnostics for ensemble forecasts is presented. Then, diagnostic methods for assessing variations of the second moment of the predicted distribution are discussed. Finally, a benchmark system for initial perturbations is described. It is a suboptimal ensemble forecasting system that uses flow-independent initial perturbations based on a sample of past short-range forecast errors.

1 Introduction

The value of reliable uncertainty information for forecasts of the atmosphere and ocean is getting increasingly recognized among the users of such forecasts. The need for uncertainty information, i.e. error bars, was anticipated more than two decades ago (Tennekes et al. 1987; Lewis 2005). Ensemble forecasting has been established operationally to satisfy this demand over 15 years ago (Molteni et al. 1996; Toth and Kalnay 1993). Since then, ensemble forecasting systems have benefitted from many improvements. These comprise better estimates of the initial state, higher accuracy in simulating dynamics and parametrized physical processes as well as improved representations of the sources of uncertainties. In addition, advances in computing technology permitted to increase the spatial resolution and the number of members in the ensembles which helped to further boost the usefulness of the uncertainty information provided by the ensembles.

The evaluation of ensemble forecasts differs from that of deterministic forecasts because the former provide a probability distribution that is not better known *a posteriori* than it was known *a priori* (Talagrand and Candille 2009). In a single case, this probability distribution is not even observable as there is just a single realisation from that distribution that materializes. In consequence, any meaningful evaluation has to accumulate forecast-verification pairs over sufficiently large samples in order to enable the comparison of like objects, i.e. predicted distributions and empirical distributions of observed or analysed values.

Diagnostic methods that aid forecast system development are the focus of this presentation. First, a brief overview of standard diagnostic methods for ensemble prediction will be given. These apply to any kind of probabilistic forecast. This topic is well covered in textbooks and many scientific publications. Thus, the overview emphasizes the essentials rather than attempts to be exhaustive. Then, diagnostic methods are discussed that aim specifically at the identification of deficiencies in ensemble prediction systems. Ultimately, diagnostic methods of this kind should provide guidance for refining the representation of initial uncertainty and model uncertainty in ensemble prediction systems.

One can envisage other aspects of diagnosis of ensemble forecasts beyond the scope of this presentation. For instance, one could use ensemble forecasts to understand the dynamical and physical processes gov-

erning the evolution of initially small perturbations. This could be useful for unravelling the dynamics of initial condition and model errors. Studying the origin of large forecast errors is another potentially interesting diagnostic field. Ensemble forecasts can provide plausible explanations of large forecast errors in terms of the amplifications of specific perturbations to the initial conditions or the model. For the sake of brevity, these aspects will not be covered here.

The outline of this contribution is as follows. Section 2 reviews a range of standard diagnostics for ensemble forecasts with some examples mainly from the ECMWF Ensemble Prediction System (EPS). Then, some recent work on diagnosing the spatio-temporal variations of the predicted variances is presented in Section 3. More generally speaking, this assesses the shape of the distribution rather than the quality of the distribution overall. To illustrate the ideas, examples from TIGGE[1] and the ECMWF EPS will be provided. Section 4 discusses the results of numerical experimentation designed to aid the diagnosis of deficiencies in existing ensemble forecasting systems. Discussion and Conclusions follow in Sections 5 and 6. Some mathematical derivations are provided in appendices.

2 Standard diagnostics

This section introduces a set of standard diagnostics that are commonly used to assess ensemble forecasts. To some extent this is a subjective selection that focusses on what is considered essential.

2.1 Model characteristics

The numerical model is the central component of any ensemble forecasting system. The skill of the ensemble relies to a large extent on the performance of this numerical model. The numerical model of the ensemble often differs in some aspects from the model used in the data assimilation system and for the single best guess forecast. For instance, the spatial resolution might be coarser and the integration time step might be longer in order to permit the timely production of several tens of ensemble members (Buizza et al. 2007). Furthermore, computationally less demanding options in the parameterizations may have been selected (Morcrette et al. 2008). Due to these differences, it is important to perform a similar range of diagnostics that are employed for the single best guess forecast. A minimalist approach should involve at least a comparison of forecast scores between the ensemble prediction model and the best guess model.

Another aspect that requires diagnostic is the climate of the numerical model. The systematic errors of the model climate should be small otherwise the ensemble cannot converge to a sample from the climatological distribution for long lead times. The members of the ensemble forecast are usually forecasts with a perturbed model (or even with different models). Therefore, it is also necessary to assess the realism of the climate of the perturbed model (or models). The diagnostic of the climate should include an analysis of errors in the mean and the variances but it may also involve more subtle aspects of the climate as in the example given below.

A detrimental effect of model tendency perturbations on the tail of the climatological distribution of precipitation was recently documented for the ECMWF EPS (Palmer et al. 2009). The parameterized tendencies of physical processes in the ECMWF EPS are stochastically perturbed. The original version of the scheme was developed by Buizza et al. (1999) and is commonly known as "stochastic physics". This version of the scheme increases the frequency of heavy precipitation events significantly compared to the unperturbed model. Precipitation exceeding 30 mm/6 h occurs about 3 times (1.5 times) as often in the perturbed model than in the unperturbed model in the tropics (extra-tropics). Despite this undesirable feature, the original stochastic scheme has a positive impact on the probabilistic skill of the EPS

[1] THORPEX Interactive Grand Global Ensemble

particularly in the tropics. A recent major revision of the scheme has reduced the frequency of heavy precipitation events significantly in spite of being more effective in generating ensemble spread than the original scheme (Palmer et al. 2009). Henceforth, this model perturbation scheme will be referred to as Stochastically Perturbed Parametrization Tendencies (SPPT).

2.2 Attributes of probabilistic forecasts

Determining the quality of forecasts — both probabilistic and deterministic — involves an analysis of the joint distribution of forecasts and observations (Murphy and Winkler 1987; Wilks 2006). What aspects of this joint distribution characterise good ensemble forecasts? The two basic properties that are needed for good probabilistic predictions are *reliability* and *resolution* (e.g. Candille and Talagrand 2005). Reliability refers to the statistical consistency between forecasted probabilities and the observations that are used for the validation of these forecasts. The statistical consistency is sometimes also referred to as calibration (Gneiting and Raftery 2007).

Let us consider a probabilistic prediction of a variable $\mathbf{x} \in \mathbb{R}^\ell$. In general, $\mathbf{x}$ can be the model state itself or any kind of projection of the model state in some lower-dimensional space. A probabilistic prediction is then a distribution F defined on $\mathbb{R}^\ell$. In loose terms, reliability is the property that the observations of $\mathbf{x}$ are distributed according to F, whenever F has been predicted. Most often, reliability is assessed for a sample consisting of a certain period, say one season, and over some region, say Europe. Thus, the reliability is measured in an average sense. However, one can interpret the *whenever* in the definition in a stricter sense and consider any sufficiently large subsample. The subsample has to be independent of observations valid at a time later than the time window used for estimating the initial conditions of the forecast. Demanding reliability in subsamples leads to a more rigorous definition of statistical consistency. We will return to this important variant of the diagnostic in Section 3.1.

The climatological distribution is at least in the average sense perfectly reliable. However, it does not constitute a skilful prediction as it is too broad. It lacks resolution which is the second property required in order to have a good probabilistic prediction. The property of resolution measures how different predicted distributions F_j with $F_j \neq F_k$ for $j \neq k$ sort the observations into distinct groups. A necessary condition for high resolution is sharpness, which is an attribute of the forecast alone. Sharpness describes the degree of localisation of the distribution in $\mathbb{R}^\ell$. Sharpness is also referred to as refinement (Murphy and Winkler 1987). Gneiting et al. formulated the goal of probabilistic forecasting to maximize sharpness subject to calibration (see references in Gneiting and Raftery 2007). A point distribution, corresponding to a deterministic forecast, is infinitely sharp. However, the sharper a distribution is, the more difficult it becomes to achieve reliability.

2.3 A zoo of measures

There are many different ways in which the joint distribution of observations and probabilistic forecasts can be summarized to obtain measures of the quality of the prediction. This leads to a zoo of measures. Now some of the common beasts in this zoo are introduced.

Practical limitations and the limited sample size of forecasts and associated observations imply that it is not possible to assess all aspects of a multivariate probabilistic prediction. Here, multivariate prediction means that we look at predictions of higher-dimensional variables $\mathbf{x} \in \mathbb{R}^\ell$ for $\ell > 1$. It is common practice to make some simplifications in order to proceed with the assessment of probabilistic forecasts. The majority of diagnostic work on ensemble forecasts has looked at univariate predictions ($\ell = 1$). For instance, the quality of probabilistic predictions of one variable, say 500 hPa geopotential or 850 hPa meridional wind component at single locations. Another common simplification is the assessment of binary events. Examples of such binary events are: Does a tropical cyclone affect a particular location

in a given time window? Does the temperature drop more than 8 K below the climatological average?

One of the basic measures that is an indicator of the overall quality of the predicted distribution is the skill of the ensemble mean. This skill can be measured for instance with the Anomaly Correlation Coefficient or with an RMS error. Ideally, the skill of the ensemble mean should be initially as high as that of an unperturbed forecast generated with the same model and then increase relative to the single unperturbed forecast as some aspects of the latter forecast become unpredictable and are filtered in the former (Leith 1974). For large lead times, the mean squared error of the ensemble mean should approach the climatological variance.

In a statistically consistent ensemble, the ensemble standard deviation should match the standard deviation of the ensemble mean error when a sufficiently large sample is considered to reliably estimate the error (see Sec. 2.5). Also, the ensemble standard deviation, often referred to simply as spread, is a measure of the sharpness of the forecast. The smaller the spread, the larger is the sharpness. Another measure of reliability that can be sensitive to higher moments of the distribution is the rank histogram (Hamill 2001, and references therein). The rank histogram tests whether the ensemble forecasts are statistically indistinguishable from the true state.

For binary events, the mean squared error of the predicted probability is a useful summary measure that assesses both reliabity and resolution. This measure is known as the Brier Score. It can be decomposed into a component that measures reliability and a component that measures resolution and a component that depends only on the distribution of the observed values. The Brier Score has been generalized to multiple categories and continuous scalar variables. These generalizations are arithmetic means of Brier Scores with varied thresholds for the event definitions and they are known as (Discrete) Ranked Probability Score and Continuous Ranked Probability Score, respectively. For the continuous case, the Ranked Probability Score measures the mean squared error of the cumulative distribution and it is identical to the mean absolute error for a deterministic forecast. There are also decompositions into reliability and resolution components for the Continuous and Discrete Ranked Probabability Scores (Hersbach 2000; Candille and Talagrand 2005, and references therein).

The area under the Relative Operating Characteristic (or ROC-area) is another summary measure that is used to assess the probabilistic prediction of binary events. The ROC-area is insensitive to reliability and quantifies the ability to discriminate events. A more general scoring rule that includes the ROC-area as special case is the two-alternative forced choice test (Mason and Weigel 2009).

The Logarithmic Score, also referred to as Ignorance Score, is a summary measure that can be used for binary events as well as multi-category events or density forecasts of a continuous variable. The Logarithmic score is defined as minus the logarithm of the predicted probability (or probability density in the continuous case) at the verifying value. It assesses both reliability and resolution.

The list of measures mentioned here is by no means exhaustive but it is deemed sufficiently comprehensive in order to permit a thorough assessment of ensemble forecasts.

2.4 Proper scores

The condensation of the information in the joint distribution of probabilistic forecasts and observations into summary measures can potentially provide misleading diagnostics when the focus is on the wrong kind of measures. However, it is possible to make some statements about the usefulness of summary measures independently of the actual distribution of forecasts and observations. Those scores that are *strictly proper* are superior in the sense that maximising such scores leads to the correct probability distribution. A rigorous mathematical definition of proper and strictly proper is given by Gneiting and Raftery (2007) together with theorems characterising properties of proper scores. The Logarithmic Score, the Brier Score and the Discrete and Continuous versions of the Ranked Probability Score are

examples of proper scores. Gneiting and Raftery (2007) also provide an example of how a score that is not proper can lead to a miscalibration of an ensemble forecast (their Fig. 3).

2.5 Spread-error relationship

A versatile diagnostic for ensemble forecasts that assesses reliability is the relationship between ensemble spread and the ensemble mean error (Talagrand et al. 1997). The relationship is usually quantified in terms of variances but other relationships may be useful too. Here, only variances will be discussed. Ideally, the diagnostic should be applied after bias-correcting the forecasts. In practice, it may be non-trivial to estimate the bias of the ensemble mean prior to the diagnostic. Often, the diagnostic is simply applied to direct model output. It is still useful for those cases where the random component of the forecast error is significantly larger than the mean forecast error.

Before discussing some examples, it is worth clarifying the dependence of the spread-error relationship on ensemble size. A necessary condition for statistical consistency is obtained by assuming a perfectly reliable member ensemble: The members $x_j, j = 1 \ldots M$ and the truth y are independent identically distributed random variables with mean μ and variance σ^2. In Appendix A, it is shown that the expected squared error of the ensemble mean is decreasing with ensemble size M according to

$$\mathbb{E}\left(\frac{1}{M}\sum_{j=1}^{M} x_j - y\right)^2 = \left(1+\frac{1}{M}\right)\sigma^2 . \tag{1}$$

Here, $\mathbb{E}$ denotes expectation. The error variance is larger than the distribution variance σ^2 as the sample mean deviates from the distribution mean. The expected ensemble variance also depends on ensemble size

$$\mathbb{E}\,\frac{1}{M}\sum_{j=1}^{M}\left(x_j - \frac{1}{M}\sum_{k=1}^{M} x_k\right)^2 = \left(1-\frac{1}{M}\right)\sigma^2 . \tag{2}$$

The sample variance estimate is smaller than the true variance as the deviations are computed with respect to the sample mean which is based on the members themselves. Due to the systematic underestimation of the true variance, this estimate is referred to as biased estimate of the variance (see e.g. Barlow 1989). For pedagogical reasons, the derivation is given in Appendix A.

For reasonably large ensemble sizes, e.g. $M = 50$, the correction factors are close enough to one and can be neglected for most purposes. However, for smaller ensemble sizes, say $M \leq 20$, the correction factors should be accounted for. The necessary condition for an ensemble to be statistically consistent is

$$\frac{M}{M-1}\,\overline{\text{biased ensemble variance estimate}} = \frac{M}{M+1}\,\overline{\text{squared ensemble mean error}}, \tag{3}$$

As mentioned before, this assumes that the ensemble is unbiased.

There are different flavours of the spread-error relationship depending on how the average in Equation (3) is defined. The basic form of the spread-error relationship considers simple averages of the variances over a region and a set of forecasts. Several examples of these will be discussed below. The more sophisticated form of the spread-error relationship considers averages depending on the ensemble variance itself. This will be discussed in detail in Section 3.1.

As a first example, the comparison of ensemble spread and ensemble mean RMS error by Buizza et al. (2005) has been selected. Their Figure 5 shows spread and error of 500 hPa geopotential over the Northern Extra-tropics for the period May–June 2002. The diagnostic compares 10-member ensembles taken from the operational predictions of the Canadian Meterological Service, NCEP, and ECMWF. According to this study, none of the three systems is statistically consistent in terms of the average spread and ensemble RMS error at all lead times in the 10-day forecast range. All systems are overdispersive

initially and then become underdispersive later on. The ECMWF EPS exhibits the best overall agreement over the 10-day forecast range. However, the underdispersion of the ECMWF EPS towards the end of the forecast range is fairly moderate if the dependence of spread and error on ensemble size is accounted for using (3).

Since 2002, the EPS has evolved and the average spread-error relationship improved further. The two main steps towards achieving this were a resolution increase from T_L255 to T_L399 and subsequently a revision of the model physics in cycle 32r3. The change of the average spread-error relationship due to the revised model physics is discussed by Bechtold et al. (2008). Due to the revised physics being more active, the right level of spread in the medium-range could be achieved with a 30 % reduction of the initial perturbation amplitude. This reduction removed the initial overdispersion and led to a system with a significantly improved agreement between spread and error in the extra-tropics for 500 hPa geopotential (Fig. 12 in Bechtold et al. 2008).

The level of agreement between spread and error obviously depends on the sample size that is being considered. In order to estimate the uncertainty due to the finite sample size, one can compute confidence intervals of the difference between ensemble mean RMS error and the ensemble standard deviation using a bootstrap resampling technique. The temporal correlation of the difference between subsequent forecasts can be accounted for by resampling blocks of subsequent dates with the blocksize depending on the temporal correlation (Wilks 1997). An example of such confidence intervals can be found in the figure mentioned above.

In a statistically consistent ensemble, the ensemble spread will match the ensemble mean RMS error everywhere in phase space. Thus, the spread-error diagnostic can cover a broad range of aspects. Inconsistencies between spread and RMS error for particular regions or fields may help to identify lacking or misspecified sources of uncertainty. The link between the actual perturbations representing model uncertainty and initial uncertainty and the spread is expected to be strongest in the early forecast ranges as interactions of the perturbations during the model integration did not have much time to blur the various sources of uncertainty.

The tropics are an example of a region where the ECMWF EPS is still generally underdispersive. The recent revision of the Stochastically Perturbed Parametrization Tendency scheme has helped to improve the spread error relationship in this region (Palmer et al. 2009). The agreement between spread and error in the tropics will improve further with the introduction of perturbations from an ensemble of perturbed analyses (Buizza et al. 2008).

The spread-error diagnostic can be extended to also consider different spatial scales. Jung and Leutbecher (2008) use waveband filters to focus on planetary, synoptic and sub-synoptic scales. They showed that the overdispersion of the ECMWF EPS prior to model cycle 32r3 in the early forecast ranges is particularly prominent in the synoptic scales in the mid-latitudes.

3 Variations of the shape of the predicted distribution in space and time

The spread of an ensemble of forecasts varies in space and time. This is due to variations of initial perturbations, tendency perturbations and the modulation of perturbation growth by the flow. The diagnostics discussed so far did not specifically focus on this aspect. To fully assess an ensemble prediction system, the reliability of the spread variations needs to be quantified. A suitable diagnostic will be discussed in Section 3.1.

This will be followed by a description of diagnostics that assess the probabilistic skill of the variations of the shape of the probability distribution. In Section 3.2, the probabilistic skill of binary events defined with respect to an error climatology will be examined. In Section 3.3, the probabilistic skill of the

continuous probability distribution predicted by the ensemble, i.e. with flow-dependent variations of the shape of the distribution, is contrasted with the skill obtained from issuing a reference forecast dressed with a climatological distribution of its errors. This analysis will include a theoretical upper limit on the gain in skill due to reliable predictions of the variations of the spread.

3.1 Spread reliability

The largest variations in ensemble spread can be expected when local values are considered. Therefore, we consider the fields themselves rather than area-averaged values. From the fields for a fixed lead time and a set of start dates, we obtain a sample of the joint distribution of ensemble standard deviation and ensemble mean error. The *spread-reliability* is determined from the conditional distribution of the ensemble mean error for given ensemble standard deviation σ_{ens}. Such a kind of diagnostic was discussed earlier by e.g. Talagrand et al. (1997) and Leutbecher and Palmer (2008). Statistical consistency, i.e. reliability, requires that the standard deviation of this conditional distribution is equal to σ_{ens}, the value of the ensemble standard deviation on which the distribution of the error is conditioned. For small ensemble sizes, it will be necessary to account for the finiteness of the ensemble according to Eq. (3).

To compute the diagnostic, the sample is stratified by the predicted spread and partitioned into equally populated bins of increasing spread. Then, the standard deviation of the ensemble and the standard deviation of the ensemble mean error are computed for each bin. For variables with small bias compared to the random error, one can use the RMS error of the ensemble mean as proxy for its standard deviation. In the examples given below, the sample is split in 20 equally populated spread bins. The forecasts are truncated at wavenumber 63 and interpolated on a $2.5^\circ \times 2.5^\circ$ grid. The standard deviations and RMS errors are computed with cosine latitude weights, i.e. weights proportional to the area represented by each grid point.

Talagrand et al. (1997) and later Leutbecher et al. (2007) as well as Leutbecher and Palmer (2008) have discussed the spread-reliability of the ECMWF EPS. They concluded that the reliability of the spread is quite poor at early lead times but improves progressively as the forecast evolves. For instance, (Leutbecher et al. 2007, their Fig. 7) show the spread-reliability of 500 hPa geopotential in winter DJF06/07 for the Northern Mid-latitudes (35°–65°N). Initially, at a lead of 1–2 d, the EPS is quite overdispersive (underdispersive) for large (small) spread. However, at a lead time of 5 d, the relationship is close to ideal.

Now, differences in the spread-reliability among the four global ensembles from Canada, ECMWF, the Met Office, and NCEP will be discussed. The data for this comparison are available through the TIGGE project (Bougeault et al. 2009). Recent comparison studies (e.g. Park et al. 2008; Hagedorn et al. 2010) suggest that these four ensembles are the most skilful ensembles in TIGGE. While the latter studies have compared the average relationship between ensemble spread and ensemble mean RMS error, the detailed spread-reliability has not been presented yet. Here, the spread-reliability of 500 hPa geopotential height in the Northern Mid-latitudes (35°–65°N) is compared for winter DJF08/09. The data have been obtained by verifying direct model output with quasi-independent ERA-interim analyses (Simmons et al. 2007; Hagedorn et al. 2010).

Figure 1 shows the spread-reliability at 24 h and 48 h lead time. At a lead time of 24 h, the Canadian ensemble has the most reliable distribution of spread. The other three ensembles are clearly less reliable. They overpredict (underpredict) variance for large (small) spread. At a lead time of 48 h, the Canadian ensemble still exhibits the most reliable spread although a moderate underdispersion is present except for the largest spread classes. The spread-reliability of the other three ensembles improves from 24 to 48 hours. The Met Office and NCEP ensembles are also somewhat underdispersive for low and normal spread. The ECMWF ensemble still predicts too much variance for large spread and too little variance for small spread. The reliability of spread in the early lead times appears to be an area where the

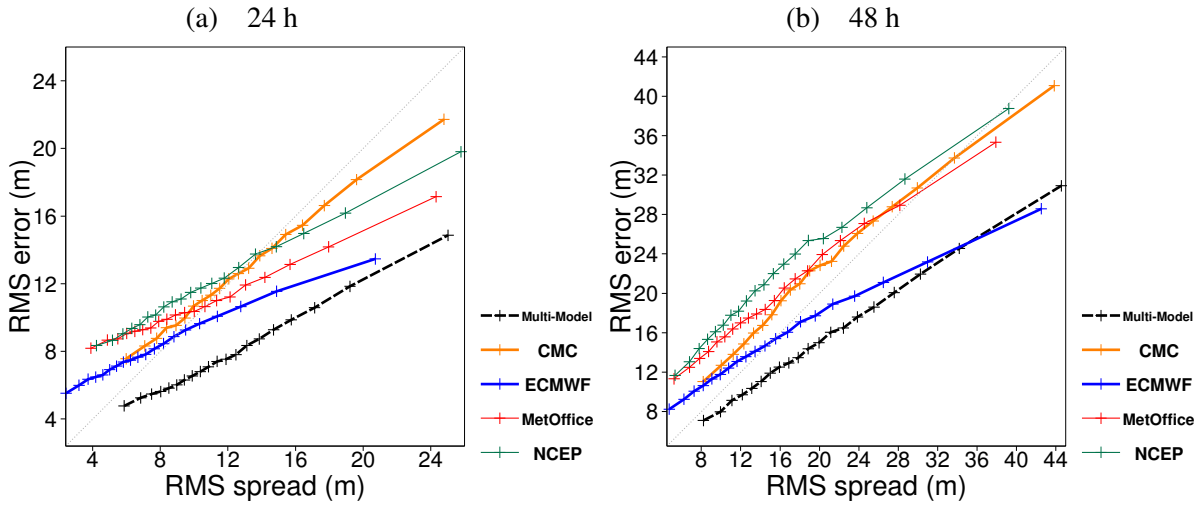

Figure 1: Spread-reliability of 500 hPa geopotential height at lead time 24 h (a) and 48 h (b) in the Northern Mid-latitudes (35°–65°N) for four ensembles in TIGGE: Canadian Met. Centre (CMC), ECMWF, Met Office, and NCEP. Multi-model refers to the combination of these four ensembles with equal weights. Period DJF08/09. Verification against ERA-interim analyses on a 2.5° × 2.5° grid.

ECMWF EPS is still quite suboptimal. Results by Leutbecher et al. (2007) give an early indication that the spread reliability can be improved through a revision of the initial perturbations that makes use of perturbations from an ensemble of analyses.

For completeness, the ensemble mean RMS error, the ensemble standard deviation and the Continuous Ranked Probability Skill Score (CRPSS) of the four ensembles are shown in Figs. 2 and 3. Although the Canadian ensemble has the most reliable spread distribution in the early lead times, it is not the ensemble with the highest probabilistic skill. This is probably due to the lack of resolution. The Canadian ensemble has the largest RMS error of the ensemble mean. Vice versa, the RMS error of the ECMWF ensemble mean is significantly smaller than that of the three other ensembles and it has a significant lead over the other three ensembles in terms of the CRPSS. Finally, it is worth noting that the multi-model ensemble consisting of the four ensembles compared here is quite overdispersive in terms of 500 hPa height up to a lead time of about 5 days.

3.2 Dichotomous events

Often, probabilistic forecasts are assessed by analysing their characteristics for the prediction of dichotomous events, i.e. the state can be described by a binary variable (0: event did not occur, 1: event occurred). For many diagnostic purposes, it is suitable to define events with respect to a climatological distribution. For instance, one can look at the probability that a variable exceeds a given quantile of the climatological distribution. Climatological distributions can be obtained from reanalyses or an archive of observations. A little care is required in constructing the climatological distribution in order to get a statistically homogeneous sample when data are spatially and temporally aggregated. Statistical homogeneity can be achieved if the climatological distribution resolves spatial and seasonal variations in the actual climate well. If the climatological distribution is too crude, one may diagnose fictitious skill (Hamill and Juras 2006).

In general, a probabilistic forecast starts with a tight distribution which gradually broadens and eventually converges to a distribution as wide as the climatological one. The probability of predicting an event defined with respect to the climate will be either in the vicinity of 0 or in the vicinity of 1 in most locations when the predicted distribution is much tighter than the climatological distribution (Fig. 4a). To

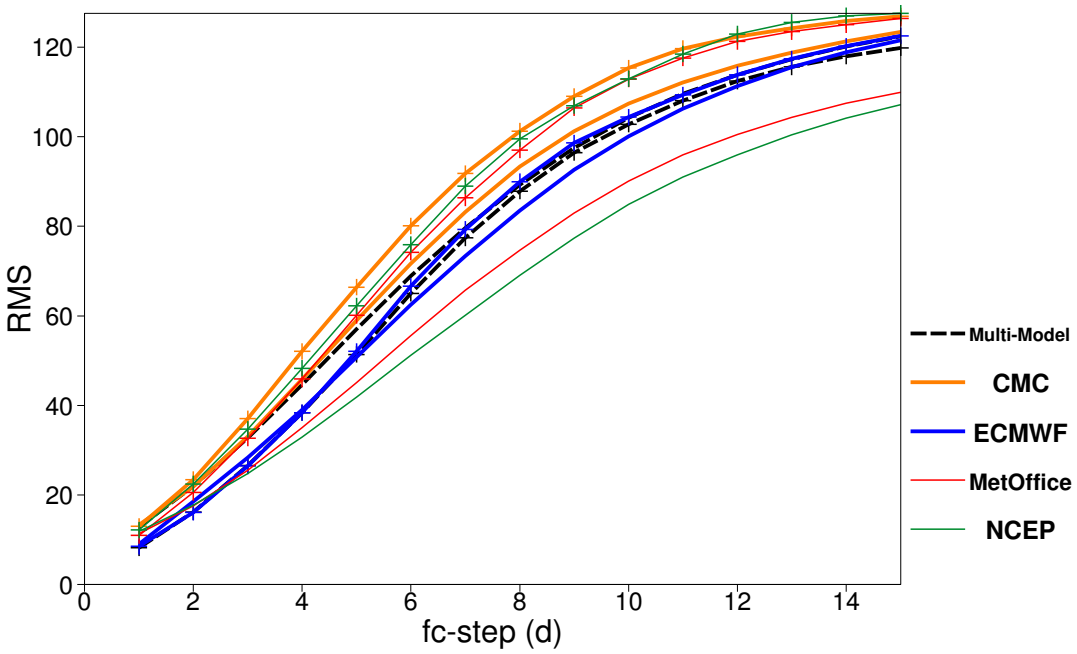

Figure 2: as previous Figure but for Ensemble standard deviation (no symbols) and ensemble mean RMS error (with symbols) versus lead time.

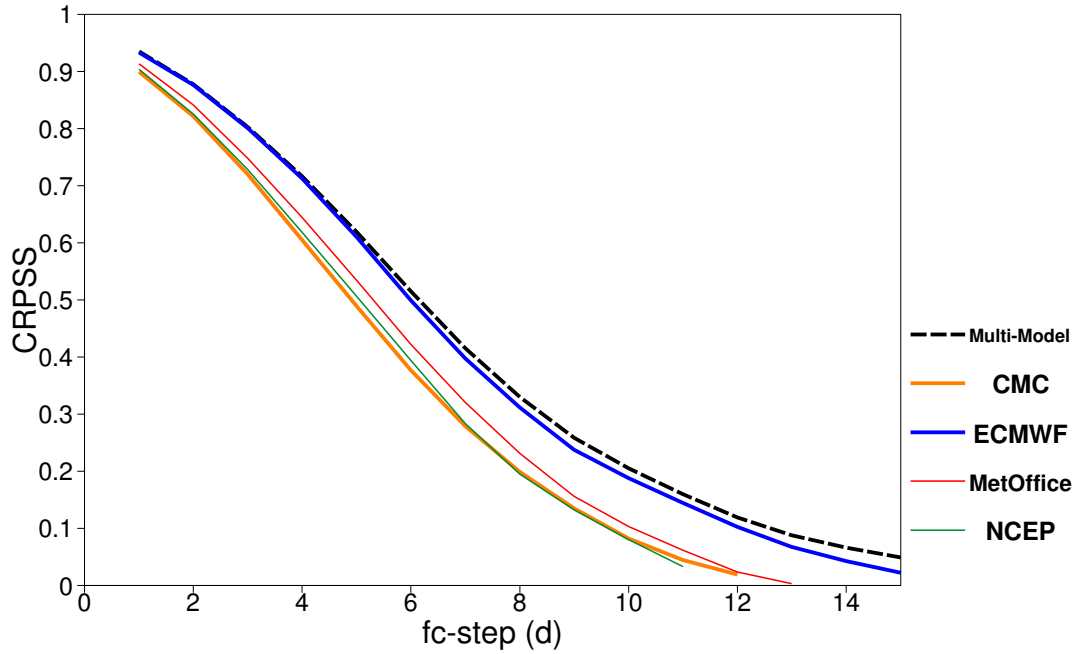

Figure 3: as previous Figure but for the Continuous Ranked Probability Skill Score. The skill score is computed with a climatological distribution based on ERA-40 analyses as reference.

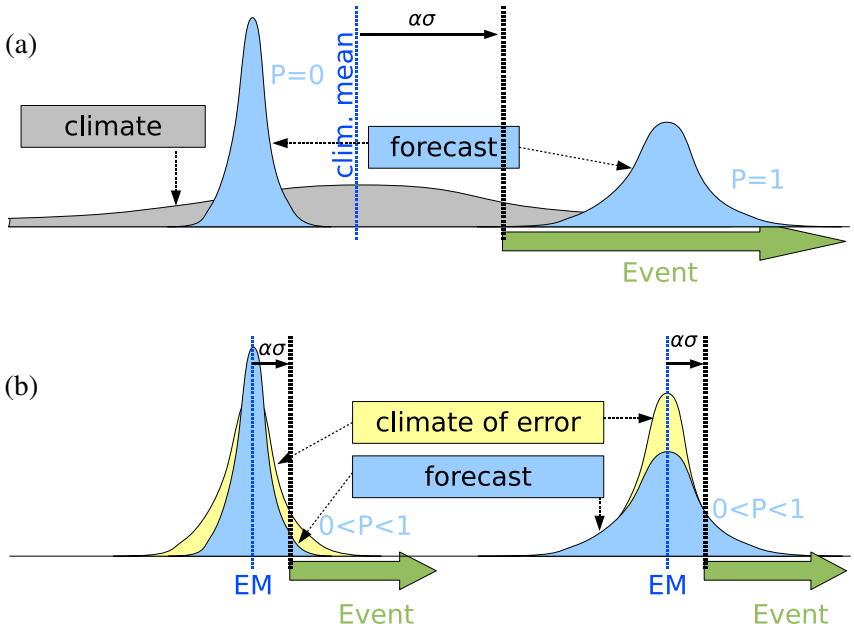

Figure 4: Event definition relative to climate (a) and relative to the ensemble mean (b). Two ensemble forecasts with a smaller than normal spread (left) and with a larger than normal spread (right) are shown as a probability density function (pdf, shaded in blue). Probabilities for the event (a) tend to be either 0 or 1 if the forecasted pdf is significantly tighter than the climate distribution as happens usually for the early lead times. This degeneracy is avoided in event definition (b) that uses the climate of the ensemble mean error (shaded in yellow) to set the position $\alpha\sigma_{\mathrm{err-em}}$ of the event threshold relative to the ensemble mean (EM). The variance of the error $\sigma^2_{\mathrm{err-em}}$ depends on lead time in contrast to the event definition in (a) based on the climatological variance σ^2_{clim} that is independent of lead time.

illustrate this, Fig. 5 shows probabilities for 850 hPa meridional wind from a 48-hour ECMWF ensemble forecast for exceeding a value of $\mu_{\mathrm{clim}} + \sigma_{\mathrm{clim}}$. Here, μ_{clim} and σ_{clim} denote the climatological mean and the climatological standard deviation, respectively. The climate is derived from ERA-40 analyses in the period 1979–2001 (Jung and Leutbecher 2008; Uppala et al. 2005). Thus, although the forecast is inherently probabilistic, diagnostics based on this kind of event will be almost equivalent to diagnostics of a deterministic forecast. With other words, diagnostics based on the climatological event definition are insensitive to the shape of the predicted distribution when short lead times are considered — except for the small fraction of the domain with probabilities deviating significantly from 0 and 1. However, assessing the distribution shape in the early part of the forecast is expected to be important for diagnosing deficiencies in the perturbation methodologies.

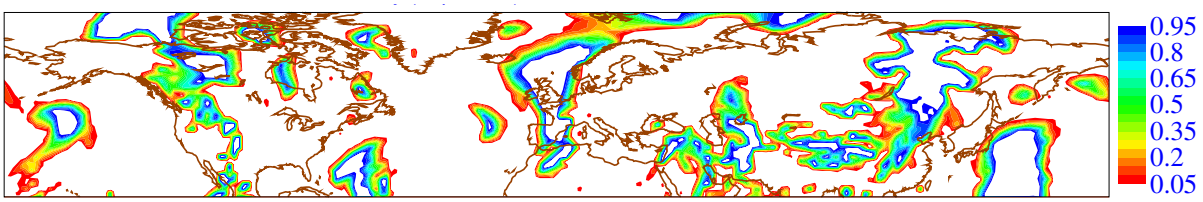

Figure 5: Probability to exceed $\mu_{\mathrm{clim}} + \sigma_{\mathrm{clim}}$ at a lead time of 48 h for 850 hPa meridional wind. ECMWF ensemble, valid at 0 UTC on 31 January 2009. Unshaded regions have a probability lower than 0.05 or larger than 0.95.

An alternative event definition is introduced now in order to assess the skill of ensembles in predicting the shape of the distribution including the early forecast ranges. The new event is defined relative to a reference forecast, which could be either the ensemble mean or an unperturbed control forecast. An archive of past errors of the reference forecast is used to define the climatological error distribution. Then, events of the kind

$$x > x_{\mathrm{ref}} + \alpha \sigma_{\mathrm{err-ref}} \tag{4}$$

are considered. As the standard deviation of the error grows with lead time, the event definition naturally scales with lead time and the degenerate situation of probabilities close to 0 and 1 is avoided (Fig. 4b). Here, the values x_{ref} and $\sigma_{\mathrm{err-ref}}$ denote the reference forecast and the climatological standard deviation of the error of the reference forecast, respectively. If such an event *occurs*, $-(x_{\mathrm{ref}} - x_o) > \alpha \sigma_{\mathrm{err-ref}}$ holds, i.e. minus the error of the reference forecasts exceeds a value of α standard deviations of the error climatology. One can consider different values of α, say, between -2 and $+2$.

The skill of such predictions can be evaluated with a probabilistic forecast given by a climatological distribution centred on the reference forecast. The climatological distribution consists of an estimated distribution of minus the error of the reference forecast in order to be consistent with the event definition (4). Here, we consider two kinds of error distributions: (i) a Gaussian error climatology given by $N(0, \sigma^2_{\mathrm{err-ref}})$ and (ii) an error climatology based on quantiles of the error anomalies $((x_{\mathrm{ref}} - x_o) - \mu_{\mathrm{err-ref}})$. The anomalies are with respect to the climatological mean error of the reference forecast $\mu_{\mathrm{err-ref}}$. Here, biases have not been corrected for. In order to account for biases, one needs to replace x_{ref} by $x_{\mathrm{ref}} - \mu_{\mathrm{err-ref}}$ in the event definition for "observed" events. The definition of "predicted" events remains unaltered assuming that the bias of the reference forecast is the same as that of the ensemble members.

Operational ECMWF reforecasts (Hagedorn et al. 2008; Hagedorn 2008) are used to define error climatologies for the control forecast and the ensemble mean. Results will be shown for the season DJF08/09. Reforecasts are available once weekly for the 18 preceding years. The reforecasts consist of four perturbed and one unperturbed control forecast. The ensemble mean is based on these five forecasts and is expected to be less accurate than the real-time ensemble mean of the EPS due to the small number of members in the reforecasts. The reforecasts are started from ERA-40 analyses for the considered season[2]. The error climate for a particular start date is computed from the 9 weeks of reforecasts centred on this day. Thus, $18 \times 9 = 162$ error fields are used to estimate the error climate. The computation of the errors is performed with ERA-interim analyses as proxy for the truth.

Figure 6 shows the probability of exceeding values defined relative to the ensemble mean ($x_{\mathrm{EM}} + \sigma_{\mathrm{err-EM}}$ and $x_{\mathrm{EM}} - \sigma_{\mathrm{err-EM}}$). For a Gaussian distribution, one would expect average values of 0.16 and 0.84 for the probabilities, respectively. The locations where the probabilities deviate significantly from 0 and 1 cover a much larger fraction of the domain than for the common event definition shown in Fig. 5.

The dependence of the error climatology on lead time yields a natural scaling of the event definition. This avoids the degeneracy of the probabilities towards small lead times and will allow a proper assessment of the probabilistic skill of the shape of the distribution even at early lead times. Another appealing aspect of the event definition based on the error climatology is the link between anomalies of the predicted probabilities and synoptic features in the flow. Figure 7 shows the 48-hour EPS probability of exceeding the ensemble mean by one standard deviation of the error climatology together with analysed fields of mean sea level pressure, 925 hPa equivalent potential temperature and 850 hPa wind in order to sketch the synoptic context. Generally speaking, areas of increased uncertainty tend to be associated with cyclonic features and fronts. The exception is a region in south-east Asia. The flow in this region is generally weak and due to the orography the 850 hPa level is close to (or even under) the surface. Thus, caution is required in interpreting the probability anomalies here.

The aim is now to quantify whether the uncertainty predicted by the ensemble has more skill than the

[2]Since March 2009, ERA-Interim analyses are used for initializing the reforecasts.

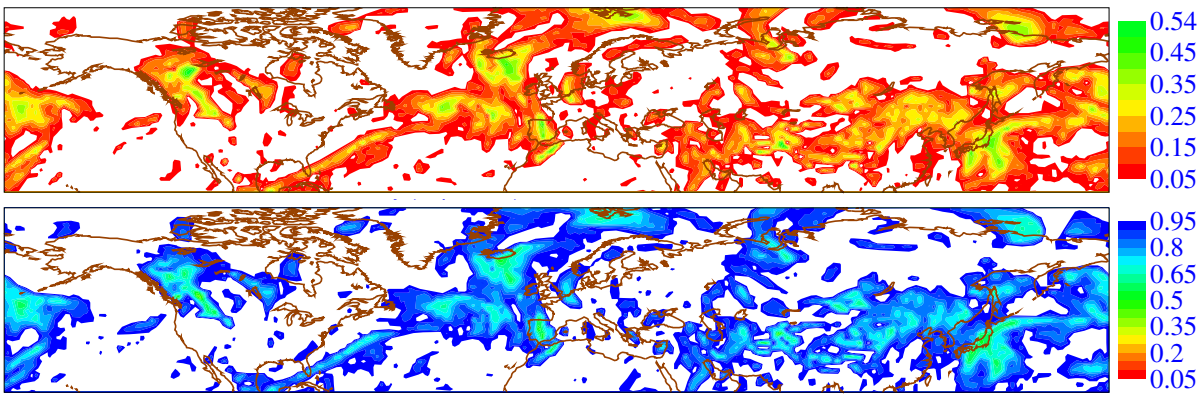

Figure 6: Probability to be larger than the ensemble mean by plus (top) and minus (bottom) one standard deviation of the climatological distribution of ensemble mean errors for 850 hPa meridional wind component at a lead of 48 h. ECMWF ensemble, valid at 0 UTC on 31 January 2009. Unshaded regions have a probability lower than 0.05 or larger than 0.95.

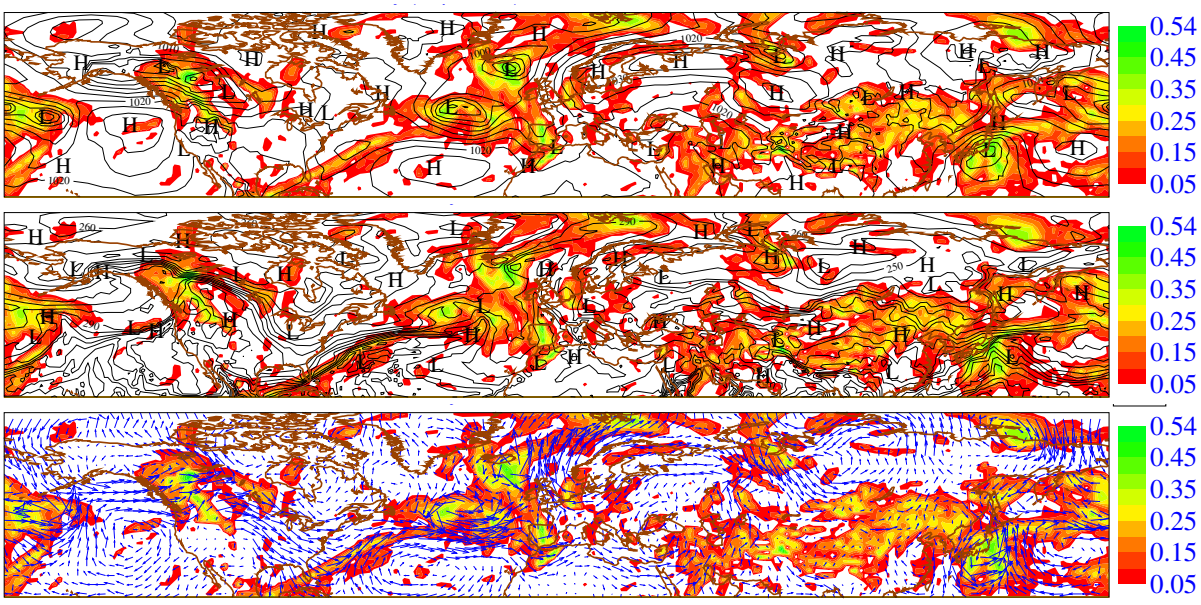

Figure 7: as top panel in previous Figure with analysed synoptic features overlaid: mean sea level pressure (top), equivalent potential temperature at 925 hPa (middle) and wind at 850 hPa (bottom).

uncertainty provided by the error climate. The skill of predicting events relative to the control forecast and the ensemble mean has been evaluated with the Brier Score, the ROC-area and the Ignorance Score (Logarithmic Score). Let p and o denote the predicted probability and the observed value ($o = 1$ if the event occured 0 otherwise), respectively. Then, the Brier Score is given by

$$\mathrm{BS} = \overline{(p-o)^2}. \tag{5}$$

The ROC-area is given by

$$\int_0^1 H \, \mathrm{d}F \in [0.5, 1], \tag{6}$$

where $H = \frac{\text{number correct fc}}{\text{number of occurrences}}$ and $F = \frac{\text{number of false alarms}}{\text{number of nonoccurrences}}$ denote Hit Rate and False Alarm Rate, respectively. They are a function of the probability at which one decides to predict the event.

Here, the Ignorance Score is defined as

$$-\overline{(o \log(p^{(T)}) + (1-o)\log(1-p^{(T)}))}, \tag{7}$$

where the forecasted probability, when n members of an M-member ensemble predict the event, is given by

$$p^{(T)}(n) = \frac{n+2/3}{M+4/3} \in \left[\frac{2}{3M+4}, \frac{3M+2}{3M+4}\right]. \tag{8}$$

These probabilities are known as Tukey's plotting positions (Wilks 2006). The plotting positions are an improved way of estimating probabilities from a finite ensemble without using a sophisticated calibration technique. They take into account the sampling uncertainty and account for the fact that the actual probability is larger 0 and smaller 1 even if no member or all members predict that the event is going to happen. Tukey's plotting position are a simple way of heeding Cromwell's advice that one should never issue statements with probability 0 or 1. As the Ignorance Score is defined with the logarithm of the probability the use of the standard probabilities $p = n/M \in [0, 1]$ would lead to an ill-defined measure.

Figure 8 shows the Ignorance score for the new type of event and the threshold $\alpha = +1$. For the event defined relative to the control forecast, the EPS is clearly more skilful than the forecasts based on climatological error distributions; the difference in skill grows continuously with lead time until about 10 d. In contrast, for events relative to the ensemble mean, the difference in skill between EPS and the climatological error distributions does not get as large. The largest gap is reached at intermediate lead times. For lead times larger than 10 d, the EPS is only as good as the ensemble mean with the Gaussian error climatology. Results for the Brier Score are qualitatively similar (not shown). For the longer (shorter) lead times > 7 d (< 7 d), the Gaussian error climatology is more (less) skilful than the quantile-based error climatology. Note, however, that this difference in skill between the two climatologies varies with the magnitude of the threshold. For larger anomalies, $|\alpha| = 2$, the Gaussian error climate is more skilfull for all lead times. Presumably, this is due to the fact that the quantile based climate is noisier for the tails of the distribution due to the moderate sample size of $18 \times 9 = 162$ used to estimate the climatology.

The area under the Relative Operating Characteristic for the error-based events is shown in Fig. 9. The ROC-area for the event relative to the ensemble mean peaks at a lead time of 4 d while the ROC-area for the event relative to the control forecast increases steadily up to a lead of $\approx 6-10$ d. The areas under the ROC for all examined thresholds, $\alpha = \pm 1, \pm 2$ are qualitatively similar. The ROC-area for all ensemble mean based events peak between 2.5 and 4 d. This peak at intermediate lead times is consistent with the results for Ignorance Score and Brier Score. The initial increase of either the ROC-area or the Ignorance and Brier Score relative to the skill of the climatological error distribution is expected to be due to the initial improvement of the spread-reliability. Finally, it is worth mentioning that qualitatively similar results have been obtained for another season (MAM2009) and for other fields (temperature at 850 hPa and geopotential at 500 hPa).

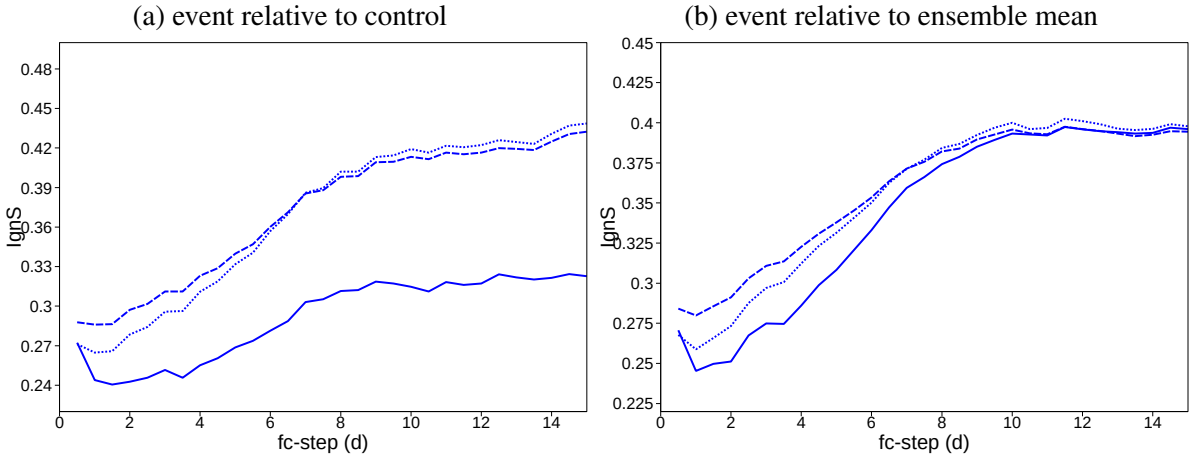

Figure 8: Ignorance Score for events defined relative to control forecast (a) and relative to ensemble mean (b). Scores for the EPS (solid), the Gaussian climatological error distribution (dashed) and the quantile-based climatological error distribution (dotted). Meridional wind component at 850 hPa in the Northern-Midlatitudes (35°-65°N), DJF2008/9. The event is for exceeding the reference forecast by one standard deviation of the error of the reference forecast.

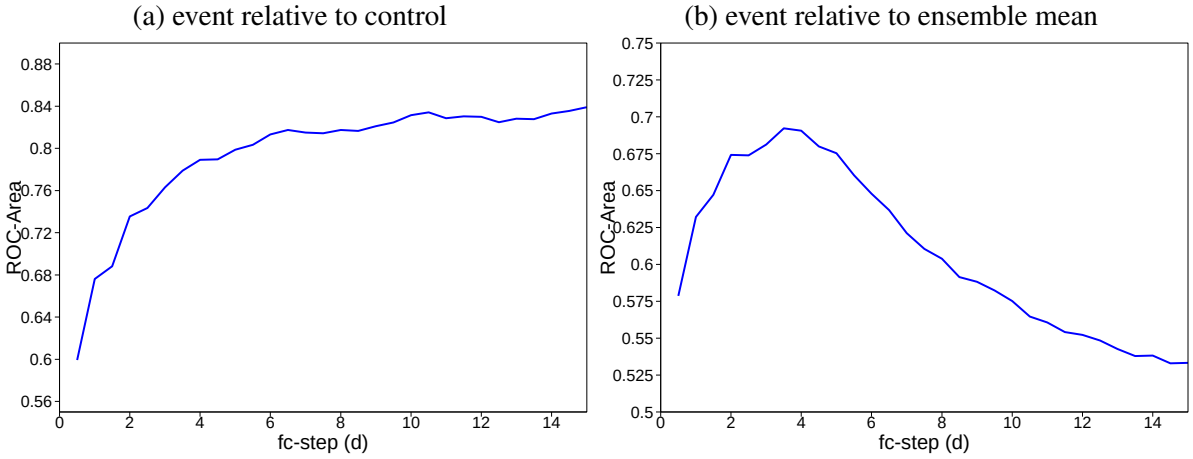

Figure 9: as previous Figure but for the area under the Relative Operating Characteristic.

3.3 Continuous distributions

The goal of this section is to evaluate variations of the shape of the predicted distribution for continuous scalar variables. The evaluation uses two different summary measures that assess both reliability and resolution and which are proper scores. The first metric is the Continuous Ranked Probability Score. It is defined as

$$\mathrm{CRPS} = \int_{-\infty}^{+\infty} (P_f(x) - H(x-y))^2 \, \mathrm{d}x, \tag{9}$$

where P_f, H and y denote the forecasted cumulative distribution, the Heavyside step function and the observed or analysed value of the variable, respectively. The step function is the cumulative distribution of the value that materializes.

The second metric is the Continuous Ignorance Score (or Logarithmic Score) given by

$$\mathrm{CIgnS} = -\log p_f(y). \tag{10}$$

It requires that the prediction is a continuous probability density function (pdf) p_f rather than a set of discrete values as provided by the raw ensemble output. There are many ways in which an ensemble forecast can be transformed into a continuous pdf (e.g. Silverman 1986). As the focus will be on the variance of the distribution, it seems appropriate to start with one of the simplest procedures possible. Here, we will evaluate the Continuous Ignorance Score by using a Gaussian distribution with the mean and variance given by the values predicted by the ensemble $p_f = N(\mu_f, \sigma_f^2)$. With this ansatz, the Continuous Ignorance Score is given by

$$\mathrm{CIgnS} = \frac{(y-\mu_f)^2}{2\sigma_f^2} + \log\left(\sigma_f\sqrt{2\pi}\right). \tag{11}$$

The first term on the right hand side is proportional to the squared ensemble mean error normalized by the ensemble variance. The value $\frac{(y-\mu_f)}{\sigma_f}$ can also be interpreted as a reduced centred random variable (RCRV) of the verifying value y. The mean and variance of the RCRV should be 0 and 1, respectively (Candille et al. 2007). The second logarithmic term penalises large spread. Without it, the lowest Ignorance score would be achieved with an infinitely large spread.

Now, the distribution predicted by the ensemble will be compared with that given by reference forecasts with a static Gaussian distribution. As in the previous section, the unperturbed control forecast and the ensemble mean are the reference forecasts. The variance of the static Gaussian distribution is set to the climatological variance of the error of the respective reference forecast. Before results for the ECMWF EPS will be presented, it is instructive to look at an idealized situation for which analytical results can be obtained.

3.4 An idealized heteroscedastic model

It is traditional in the verification of probabilistic forecasts to define skill scores so that a skill of zero implies a forecast as good as the climatological distribution and a skill of one implies a perfect deterministic forecast. However, even the initial state from which forecasts are started will never be perfect. This is a direct consequence of the nonlinear chaotic dynamics of the atmosphere and the finite number of imperfect observations available to estimate the initial state and the imperfections of the forecast model. These facts will always impose a lower limit on the average forecast error variance $\overline{v} \equiv \overline{\sigma^2} > 0$.

What we will consider now is the definition of a perfect probabilistic forecast under the constraint that the average variance of the error of the reference forecast is fixed. For the unperturbed control forecast, this constraint will be satisfied to the extent that a particular forecast system consisting of a forecast model, a set of observations and an assimilation system is considered. For the ensemble mean this is

a simplification. Changes in the representation of initial and model uncertainties can also affect the accuracy of the ensemble mean.

For the derivations that follow now, it is useful to consider a fixed lead time of the forecast. The label t will be used to refer to different *valid times* of the forecast. Two hypothetical limiting cases will be considered for a given reference forecast x_{ref}. Let us consider the probabilistic forecast with a perfect static distribution defined as

$$p(x,t) = p_s(x - x_{\mathrm{ref}}(t)). \tag{12}$$

Here, the position of the pdf varies with the reference forecast $x_{\mathrm{ref}}(t)$ but the pdf p_s is either constant or only seasonally varying. This forecast will be referred to as the *perfect static forecast*, henceforth. The attribute *perfect* refers to the assumption that the pdf p_s is statistically consistent with the error of the reference forecast in an average sense over all t, i.e. the empirical distribution of minus the error of the reference forecast should converge towards p_s for large samples.

The other limiting case is the *perfect dynamic forecast*. It is given by

$$p(x,t) = p_d(x - x_{\mathrm{ref}}(t), t). \tag{13}$$

Here, the pdf p_d varies explicitly with t. It is assumed to be statistically consistent with the error of the reference forecast in a stricter sense: For any subsample (which may be conditioned on p_d, or even on any information available to us up to the time the forecast is started), the empirical distributions of minus the error of the reference forecast converge to the subsample mean of p_d.

For the idealized example considered here, we assume that the reference forecasts are unbiased so that $\int x p_s \,\mathrm{d}x = \int x p_d \,\mathrm{d}x = 0$. Furthermore, we constrain both the perfect static forecast and the perfect dynamic forecast to have the same average variance

$$\int x^2 p_s \,\mathrm{d}x = \overline{v} \quad \text{and} \tag{14}$$

$$\int x^2 \overline{p_d} \,\mathrm{d}x = \overline{v}. \tag{15}$$

The overline refers to an average over the valid times.

Let us now consider the particular case where the reference forecast is the mean μ_t of a Gaussian distribution and the true state is distributed according to

$$y \sim N(\mu_t, \sigma_t^2). \tag{16}$$

In this situation, the perfect dynamic forecast is $p_d = N(\mu_t, \sigma_t^2)$ and the perfect static forecast is $p_s = N(\mu_t, \overline{\sigma^2})$ with $\overline{\sigma^2} \equiv \mathbb{E}_t \sigma_t^2$.

In order to proceed further, the expected values of the CRPS and the CIgnS need to be known. In Appendix B, it is shown that the expected CRPS for predicted pdf $N(\mu_t, \sigma_f^2)$ and truth y distributed according to (16) is given by

$$\mathbb{E}_y \,\mathrm{CRPS}(N(\mu_t, \sigma_f^2), y) = \frac{\sigma_t}{\sqrt{\pi}} \left[-\frac{\sigma_f}{\sigma_t} + \sqrt{2 + 2\sigma_f^2/\sigma_t^2} \right] \tag{17}$$

Figure 10 shows how the expected CRPS varies as function of the ratio of the predicted standard deviation σ_f and the true standard deviation σ_t. As the CRPS is a proper score, the minimum is attained for $\sigma_f = \sigma_t$.

The expected value of the Continuous Ignorance Score follows directly from (11). It is given by

$$\mathbb{E}_y \,\mathrm{CIgnS}(N(\mu_t, \sigma_f^2), y) = \frac{1}{2} \left[\ln(2\pi\sigma_f^2) + (\sigma_t/\sigma_f)^2 \right] \tag{18}$$

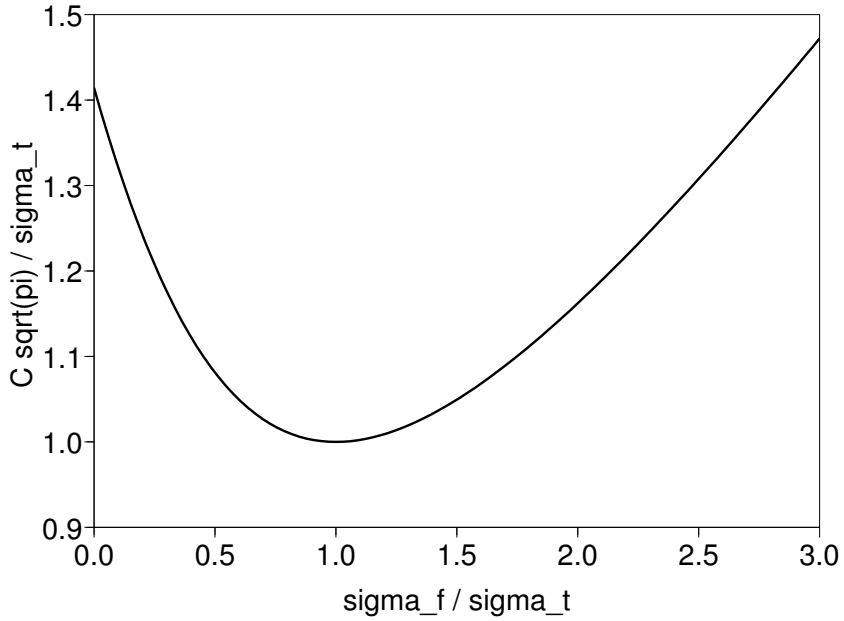

Figure 10: Expected CRPS for a Gaussian prediction with standard deviation σ_f and a truth distributed according to a Gaussian distribution with standard deviation σ_t. The CRPS has been normalized with $\sigma_t/\sqrt{\pi}$, which is the expected CRPS for $\sigma_f = \sigma_t$.

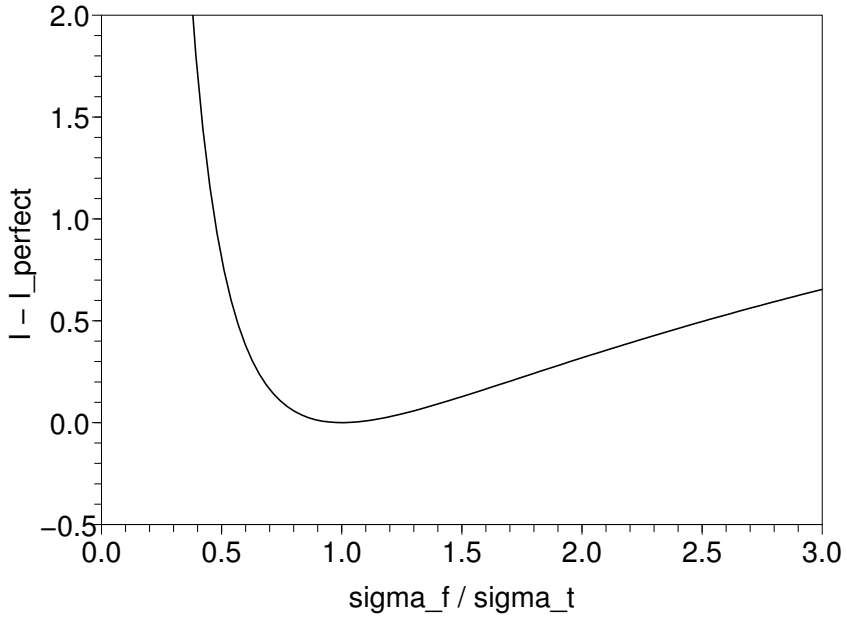

Figure 11: As previous Figure but for the difference of the Ignorance score for the predicted distribution with variance σ_f^2 and the Ignorance attained when the predicted distribution has the true variance σ_t^2.

The Ignorance score penalises underdispersion significantly stronger than the CRPS (Fig. 11). However, being a proper score, the Ignorance score also is smallest for the correct level of variance ($\sigma_f = \sigma_t$).

The degree by which the scores for the perfect dynamic forecast and the perfect static forecast differ is a function of the distribution of the variance of the error $v_t = \sigma_t^2$. For an arbitrary distribution of variance of the error of μ_t, given by the pdf $f(v)$, the expected values of the CRPS for the perfect static forecast and the perfect dynamic forecast are

$$\mathscr{C}_s \equiv \mathbb{E}_v\, \mathbb{E}_y\, CRPS(N(\mu_t,\overline{v}),y) = \int f(v)\frac{\sqrt{v}}{\sqrt{\pi}}\left[-\sqrt{\overline{v}/v}+\sqrt{2(1+\overline{v}/v)}\right]\mathrm{d}v = \\ -\sqrt{\overline{v}/\pi}+\sqrt{2/\pi}\int f(v)\sqrt{v+\overline{v}}\,\mathrm{d}v \quad (19)$$

and

$$\mathscr{C}_d \equiv \mathbb{E}_v\, \mathbb{E}_y\, CRPS(N(\mu_t,v),y) = \int f(v)\frac{\sqrt{v}}{\sqrt{\pi}}\left[-\sqrt{v/v}+\sqrt{2(1+v/v)}\right]\mathrm{d}v = \overline{\sqrt{v/\pi}}, \quad (20)$$

respectively. The expected Continuous Ignorance Scores for an arbitrary variance distribution $f(v)$ are given by

$$\mathscr{I}_s = \int \frac{1}{2}\left[\log(2\pi\overline{v})+\frac{v}{\overline{v}}\right]f(v)\,\mathrm{d}v = \frac{1}{2}\log(2\pi\overline{v})+\frac{1}{2} \quad \text{and} \quad (21)$$

$$\mathscr{I}_d = \int \frac{1}{2}\left[\log(2\pi v)+\frac{v}{v}\right]f(v)\,\mathrm{d}v = \frac{1}{2}\overline{\log(2\pi v)}+\frac{1}{2} \quad (22)$$

for the two kinds of forecasts.

As an example, two simple variance distributions will be considered: A continuous uniform distribution of variance between a lower value v_1 and and upper value v_2 and a discrete uniform distribution which selects the values $v_t = v_1$ and $v_t = v_2$ with equal probability (Fig. 12). Thus, $f(v) = (v_2 - v_1)^{-1}$ for $v \in [v_1, v_2]$ and 0 otherwise in the continuous case. In the discrete case, we have $f(v) = \frac{1}{2}\delta(v-v_1)+\frac{1}{2}\delta(v-v_2)$. Here, δ denotes the Dirac distribution. The mean variance for both distributions is $\overline{v} = (v_1+v_2)/2$. We will use the dimensionless parameter

$$\Delta_v = (v_2 - v_1)/(2\overline{v}) \ \in [0,1] \quad (23)$$

to characterize the width of the variance distributions in both situations. Evaluating the integrals in (19) and (20) for the continuous uniform distribution yields

$$\mathscr{C}_s = \sqrt{\overline{v}/\pi}\left(\frac{4}{3\Delta_v}\left[(1+\Delta_v/2)^{3/2}-(1-\Delta_v/2)^{3/2}\right]-1\right) \quad (24)$$

$$\mathscr{C}_d = \sqrt{\overline{v}/\pi}\,\frac{1}{3\Delta_v}\left[(1+\Delta_v)^{3/2}-(1-\Delta_v)^{3/2}\right] \quad (25)$$

for the expected CRPS of the perfect static forecast and the perfect dynamic forecast, respectively. For the discrete uniform variance distribution, one obtains

$$\mathscr{C}_s = \sqrt{\overline{v}/\pi}\left(\sqrt{1+\Delta_v/2}+\sqrt{1-\Delta_v/2}-1\right) \quad (26)$$

$$\mathscr{C}_d = \sqrt{\overline{v}/\pi}\left(\sqrt{1+\Delta_v}+\sqrt{1-\Delta_v}\right)/2. \quad (27)$$

The skill of the perfect dynamic forecast relative to the perfect static forecast is given by $1-\mathscr{C}_d/\mathscr{C}_s$. The ratio $\mathscr{C}_d/\mathscr{C}_s$ decreases monotonically with increasing width Δ_v of the variance distribution (Fig. 13). For the discrete variance distribution, the ratio is smaller than for the continuous variance distribution as the typical error in predicting the mean variance $\overline{v}$ instead of the actual variance v_t is significantly larger for the discrete variance distribution. The perfect dynamic forecast is always better than the static forecast

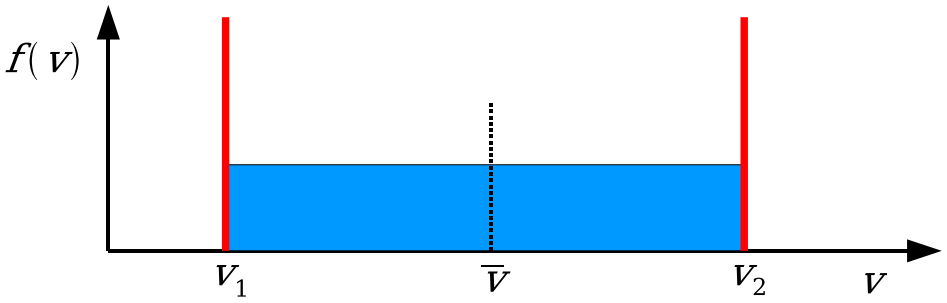

Figure 12: Idealized distributions of the true variance of the error of the distribution mean.

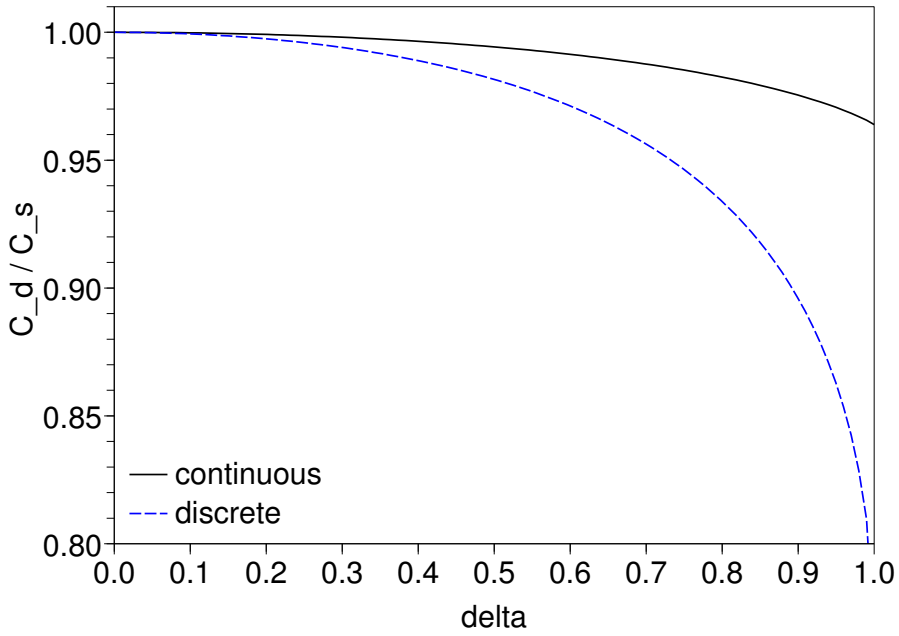

Figure 13: Ratio between expected CRPS of the perfect dynamic forecast and the expected CRPS of the perfect static forecast as function of the dimensionless width Δ_v of the variance distribution.

except for the trivial case $\Delta_v = 0$. It is instructive to convert the increase in probabilistic skill due to the use of a dynamically varying distribution instead of a static distribution into a lead time gain. Let us assume that the standard deviation grows exponentially with lead time t_{lead} as $\exp(t_{\text{lead}}/\tau)$, where τ is the e-folding time scale of the error. Then, the expected CRPS for both types of forecast will also grow exponentially with the same e-folding time scale according to (17). Thus, the gain in lead time is given by

$$\Delta t_{\text{lead}} = -\tau \log\left(\mathscr{C}_d/\mathscr{C}_s\right) \tag{28}$$

Simmons and Hollingsworth (2002) estimated error doubling times for 500 hPa height around 1.4 d in the Northern Extra-tropics ($\tau = 48$ h). For wide variance distributions, i.e. $\Delta_v \approx 1$, the reduction in the CRPS from the perfect static forecast to the perfect dynamic forecast amounts to 3% (20%) for the continuous (discrete) uniform variance distribution. The corresponding lead time gains are 1.5 h and 11 h for the continuous variance distribution and the discrete variance distribution, respectively. The variance distribution predicted by a real ensemble will be better approximated by the continuous distribution than the discrete distribution. Therefore, only a modest gain in lead time of less than 2 h can be expected by representing the dynamic variations of the variance perfectly instead of using a perfect static distribution.

Now, we continue the example and look at the Continuous Ignorance Score. The score for the perfect static forecast is given by (21). Averaging of the logarithm of the variance in (22) gives the score for the perfect dynamic forecast. For the continuous variance distribution, one obtains

$$\mathscr{I}_d = \mathscr{I}_s - \frac{1}{2} + \frac{1}{4\Delta_v}\log\left(\frac{1+\Delta_v}{1-\Delta_v}\right) + \frac{1}{4}\log\left(1-\Delta_v^2\right) \tag{29}$$

and the score for the discrete variance distribution is given by

$$\mathscr{I}_d = \mathscr{I}_s + \frac{1}{4}\log\left(1-\Delta_v^2\right). \tag{30}$$

The skill of one forecast with respect to another in terms of ignorance is measured by the difference of the ignorances (Roulston and Smith 2002). Figure 14 shows the difference in Ignorance scores between the static and dynamic forecast as function of the width Δ_v of the variance distribution. For wide distributions $\Delta_v \approx 1$, the difference in the Ignorance Score amounts to values of about 0.15 (0.5) for the continuous (discrete) variance distribution.

Again, it is informative to relate the change in scores to a lead time gain. It follows from (21) that the Ignorance Score of the static forecast grows with lead time as $\mathscr{I}_s = \text{const.} + t_{\text{lead}}/\tau$ for the assumed exponential growth of the standard deviation $\sqrt{v} \propto \exp(t_{\text{lead}}/\tau)$. Thus, a change in Ignorance Score of $\Delta\mathscr{I}$ converts to a lead time gain according to

$$\Delta t_{\text{lead}} = \tau\,\Delta\mathscr{I}. \tag{31}$$

For a value of $\tau = 48$ h, the reductions of the Ignorance Score of 0.15 (0.5) imply lead time gains of about 7 h (24 h) for the case of a wide continuous (discrete) uniform variance distribution.

The idealised considerations indicate that the Continuous Ignorance Score is a much more sensitive measure of variations in the shape of the pdf than the Continuous Ranked Probability Score. Now, we will explore the results for the ECMWF Ensemble Prediction System.

3.5 Results for the EPS

The Continuous Ranked Probability Score and the Continuous Ignorance Score have been computed for the operational EPS for two seasons (DJF08/09, MAM09) and for three variables (500 hPa geopotential,

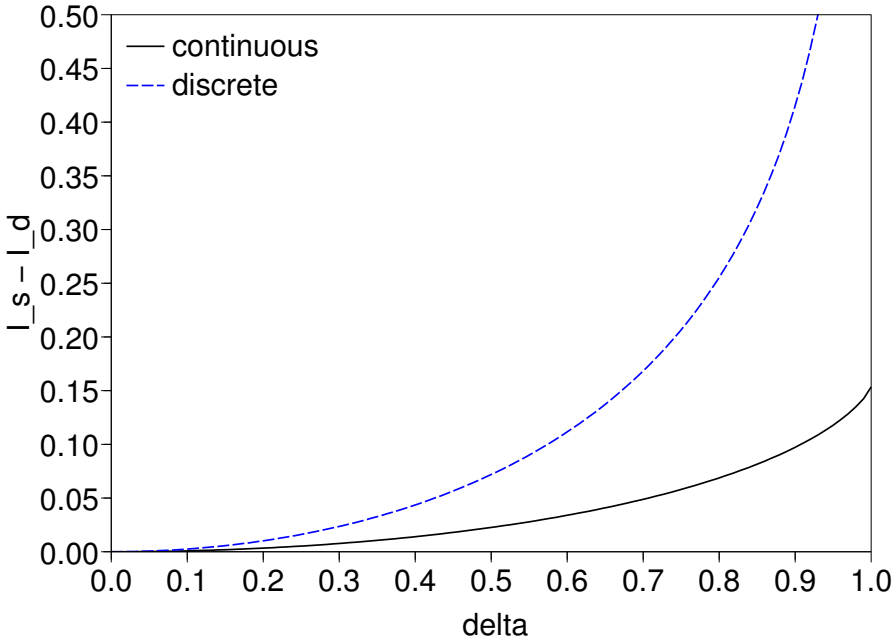

Figure 14: Difference between expected Continuous Ignorance Score (CIgnS) of the perfect static forecast and the expected CIgnS of the perfect dynamic forecast as function of the dimensionless width Δ_v of the variance distribution $f(v)$.

	control		ens. mean	
	D+2	D+5	D+2	D+5
CRPS	4	21	1.5	4
CIgnS	0	30	0	11

Table 1: Gain in lead time (h) due to using the EPS instead of a static distribution centred on the control forecast (columns 2 and 3) and centred on the ensemble mean (columns 4 and 5). 850 hPa meridional wind component in the Northern Mid-latitudes in DJF08/09.

850 hPa temperature and meridional wind). In order to define the static forecasts centred on the control and the ensemble mean, error variances from the reforecasts are used (see Section 3.2 for details). Due to the remaining inevitable inconsistencies between the reforecasts and the real time forecasts, the static forecast is likely to be suboptimal.

The Continuous Ranked Probability Score of the EPS and the static Gaussian distribution centred on the two reference forecasts are compared in Fig. 15. By construction, the EPS ensemble mean is identical to the control forecast at initial time. Therefore, the scores based on the static Gaussian distributions centred on the two forecasts are equal initially. Already at a lead time of 2 d, however, the probabilistic forecasts based on the static Gaussian centred on the ensemble mean is noticably better than the Gaussian centred on the control forecast. The gap between the two forecasts progressively widens with lead time reaching a difference equivalent to 17 h at a lead time of about 5 d. The EPS is more skilful than the static Gaussian centred on the ensemble mean. Initially, however, the difference in the Continuous Ranked Probability Score (CRPS) is marginal. At a lead time of about 5 d, the EPS is 4 h more skilful than the static Gaussian centred on the ensemble mean. Here, the CRPS for the EPS is computed with the empirical distribution function obtained from the 51 forecasts. Note that the results are not altered if the CRPS of the EPS is computed from the Gaussian centred on the ensemble mean with the variance predicted by the EPS (not shown).

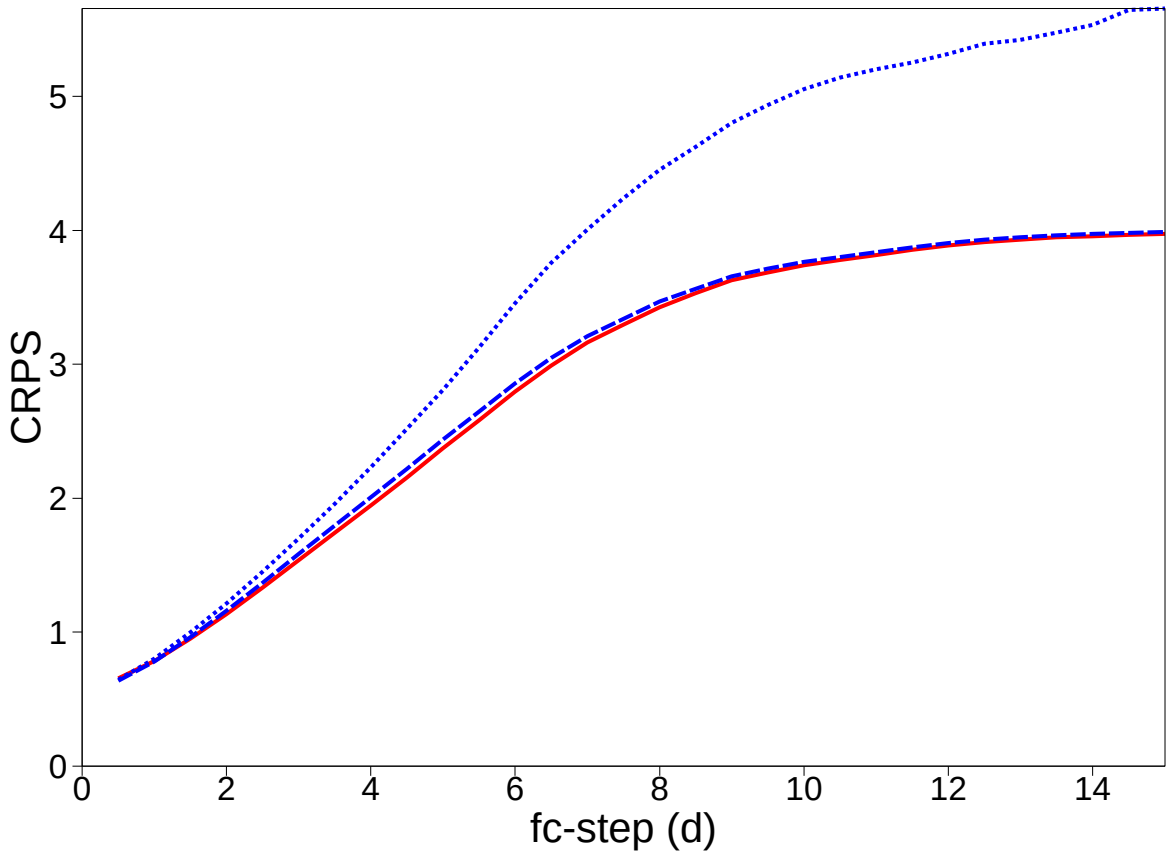

Figure 15: Continuous Ranked Probability Score for the EPS (red-solid) and a static Gaussian distributions centred on the control forecast (blue-dotted) and centred on the ensemble mean (blue-dashed) for the 850 hPa meridional wind component in the Northern Mid-latitudes (35°–65°N) in the season DJF08/09. The variance of the static Gaussian distribution has been estimated from the variance of the respective errors in the 18-year reforecast dataset.

	D+2	D+5
τ	83	176
$\Delta t_{\text{lead}}[\text{CRPS}, \Delta_v = 1.0]$	2.5	5.3
$\Delta t_{\text{lead}}[\text{CRPS}, \Delta_v = 0.8]$	1.5	3.1
$\Delta t_{\text{lead}}[\text{CIgnS}, \Delta_v = 1.0]$	12	26
$\Delta t_{\text{lead}}[\text{CIgnS}, \Delta_v = 0.8]$	5.7	12

Table 2: Variance growth time-scale τ for the error of the Ensemble Mean and predicted lead time differences between the perfect static forecast and the perfect dynamic forecast for 850 hPa meridional wind component in the Northern Mid-latitudes in DJF08/09. All values in hours. Estimates are given for the width $\Delta_v = 0.8$ and 1.0 of the variance distribution and are based on the continuous uniform distribution.

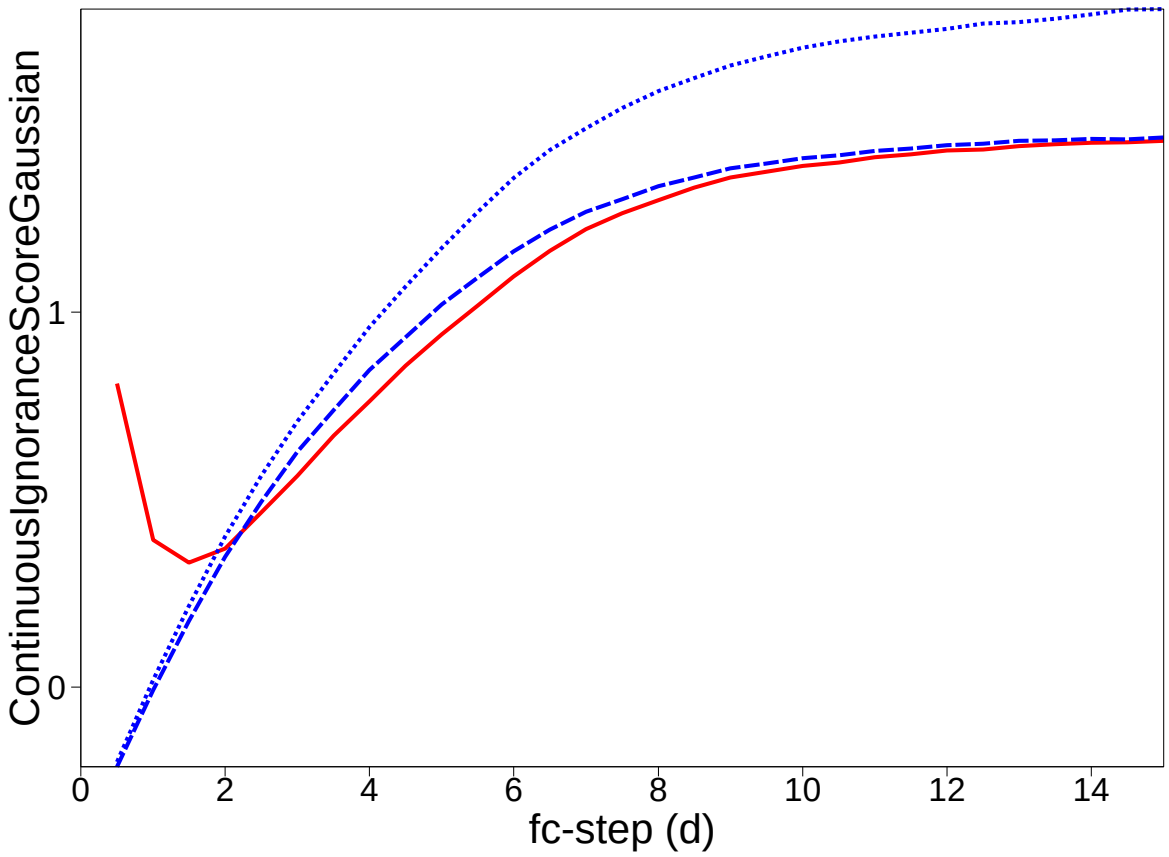

Figure 16: as previous Figure but for the Continuous Ignorance Score.

The corresponding results for the Continuous Ignorance Score (CIgnS) are shown in Figure 16. The two reference forecasts differ qualitatively in the same way as for the CRPS. The Ensemble Mean based static forecast is better than the forecast based on the Control from a lead time of 1 d onwards. The difference in skill amounts to 4 h (24 h) at a lead time of 2 d (5 d). The EPS is significantly worse than the reference forecasts up to a lead time of 2 d. It is expected that this can be attributed to deficiencies in the representation of initial uncertainties. We will return to this aspect in Section 4. For longer lead times, the variations in the distribution predicted by the EPS are clearly useful as the EPS has a significant lead over the static Gaussian centred on the ensemble mean. The EPS is about 11 h more skilful at a lead time of 5 d. For even longer lead times, the difference between EPS and static forecast decreases again. This is to be expected as the variance of EPS converges towards the climatological variance for very long lead times and variations of the width of the predicted distribution become small. In summary, the situation for the CIgnS is qualitatively similar to the one for the CRPS except for the early forecast ranges where the static distributions are superior to the distribution based on the EPS. Quantitatively, the Ignorance score is much more sensitive to spread variations than the CRPS.

It is interesting to compare the differences in lead time between EPS and the static Gaussian centred on the ensemble mean with the estimates provided by equations (28) and (31). For this, we assume that the variance distribution can be reasonably well approximated by the uniform variance distribution with a dimensionless width Δ_v close to 1. The variance growth time scale τ is estimated from the ensemble spread using centred finite differences at ± 12 h of the lead time. The lead time differences between the perfect dynamic forecast and the perfect static forecast estimated from (28) and (31) are given in Table 2 for $\Delta_v = 0.8$ and 1. For the CRPS, the estimates agree well with the observed values at lead times of 2 d and 5 d. For the CIgnS, the values agree also well at a lead time of 5 d and using the value of $\Delta_v = 0.8$. However, for the earlier lead time, the predicted gain of ≥ 6 h is much larger than the observed value of about 0 h. It is expected that this discrepancy is due to the fact that the ensemble spread is too unreliable in the early forecast ranges. In the future, one could extend the idealized model of Section 3.4 to unreliable predictions of uncertainty to assess this hypothesis.

4 Diagnostic numerical experimentation

Diagnostic work does not need to be limited to the analysis of already existing numerical experiments. Sometimes, a deeper understanding can be gained from analysing new sets of experiments that are specifically designed to answer a particular question. Experiments that will be discussed now are motivated by the following question: To what extent is an ensemble using a flow-dependent representation of initial uncertainties, say for instance, singular vector initial perturbations, more skilful than an ensemble using a climatological representation of initial uncertainty. In other words, the climatological representation of initial uncertainty provides a simple benchmark for assessing more sophisticated representations of initial uncertainty.

Mureau et al. (1993) compared ensembles using singular vector based initial perturbations with ensembles using short-range forecast errors as initial perturbations. The latter have been constructed from orthonormalized 6-hour forecast "errors" collected over 30 days prior to the start time of the ensemble. Although these perturbations are not a climatological representation, they are at least partially independent of the flow of the day. A T63 version of the ECMWF model was used and the initial state was estimated with an Optimum Interpolation scheme. The two different sets of initial perturbations were scaled to have the same variance initially. Mureau et al. concluded that the singular vector based initial perturbations led to superior probabilistic forecasts.

In a more recent study, Magnusson et al. (2009) compared the skill of three ensembles which used (i) singular vectors, (ii), the Ensemble Transform method and (iii) scaled differences of randomly selected atmospheric states, which they refer to as Random Field Perturbations. While, (i) and (ii) are flow-

dependent representations of initial uncertainty, (iii) can be viewed as a simple flow-independent initial uncertainty representation. The experiments were performed with the ECMWF ensemble at a resolution of T_L255 with 40 levels. The amplitude of the initial perturbations in configurations (ii) and (iii) was set to a value that leads to the same level of ensemble spread as in (i) at day 3 for 500 hPa geopotential in the Northern Extra-tropics. As a consequence, the ensemble with Random Field perturbations is quite overdispersive initially. However, the probabilistic skill in the medium-range is close to the skill of the ensembles (i) and (ii) that have flow-dependent initial uncertainty representations.

The structural characteristics of the Random Field perturbations are the same as those of the full fields. Therefore, the Random Fields cannot be expected to have the same structure as analysis errors. For instance, current estimates of the variance distribution across different spatial scales show that analysis errors have a significantly larger (smaller) proportion of variance at sub-synoptic (planetary and synoptic) scales than the full fields (e.g. Errico et al. 2007). Estimates of short-range forecast errors may provide a useful alternative. As the idealized predictability experiments by Tribbia and Baumhefner (2004) indicate, 1-day forecast errors still retain to a large degree the flatness of the variance spectrum of analysis errors. Obviously, it is impossible to get samples of the true 1-day forecast error. Nevertheless, with a modern numerical weather prediction system with a good data assimilation scheme and a good global observational coverage, forecast-analysis differences may be able to provide a reasonable proxy for short-range forecast errors which in turn may then be a reasonable proxy for analysis errors. The good observational coverage is crucial because otherwise the forecast-analysis difference is small in unobserved regions despite the fact that the analysis error can be large.

Now, results of recent experiments are summarized which compare an ensemble using the operational singular vector configuration with an ensemble with initial uncertainty represented by a sample of the random component of 24-hour forecast error estimates. Apart from the initial conditions, the experiments both use the ECMWF Ensemble Prediction System at a spatial resolution of $T_L255L62$. The model version is labelled cycle 32r3 and is known to be more active than previous versions due to changes in the physical parameterisations (Bechtold et al. 2008). The diagnostics are based on 50 start dates in the period November '07 to February '08. The 24-hour forecast errors are used unscaled as this yields an ensemble with a similar mid-latitude spread at a lead of 3 days in terms of 500 hPa geopotential as the singular vector ensemble. However, this comes at the price of initial overdispersion of about 35% at a lead of 1 d. In addition, individual members in the forecast error ensemble have localized large amplitude initial perturbations in some cases. The independence of the initial perturbations from the actual flow can result in locally unrealistic perturbations in some members. For instance, a perturbation representing a small-scale low several hectopascal deep can be placed within an anticyclonic region. Such features are undesirable in an operational system. However, we need not be overly concerned with this aspect as we are simply interested in the sensitivity of the overall ensemble skill to the use of an independent set of initial perturbations. It is expected that the overall skill is not too sensitive to the presence of unrealistic localized perturbations in a few members.

In general terms, the ensemble using forecast error perturbations appears to be at least as skilful as the ensemble using singular vector perturbations except for the aspect of the initial overdispersion and the spurious initial perturbations in some members. The largest differences between the two experiments are evident during the first couple of days. Generally, skill scores converge at later lead times. A detailed comparison of the two experiments will be reported separately. Here, some results for the 850 hPa meridional wind component at a lead time of 48 h are discussed. Figure 17 shows the ensemble standard deviation and the ensemble mean RMS error. The ensemble standard deviation is quite similar in the mid-latitude storm tracks in the two experiments. Elsewhere, in particular in the low latitudes and the polar regions, the forecast error ensemble has a significantly larger spread. The ensemble mean RMS errors of the two experiments are rather similar. The overall agreement between RMS error and spread is better in the forecast error ensemble. The difference in probabilistic skill as quantified by the CRPS is shown in Fig. 18 for the same variable and lead time. Negative values (forecast error

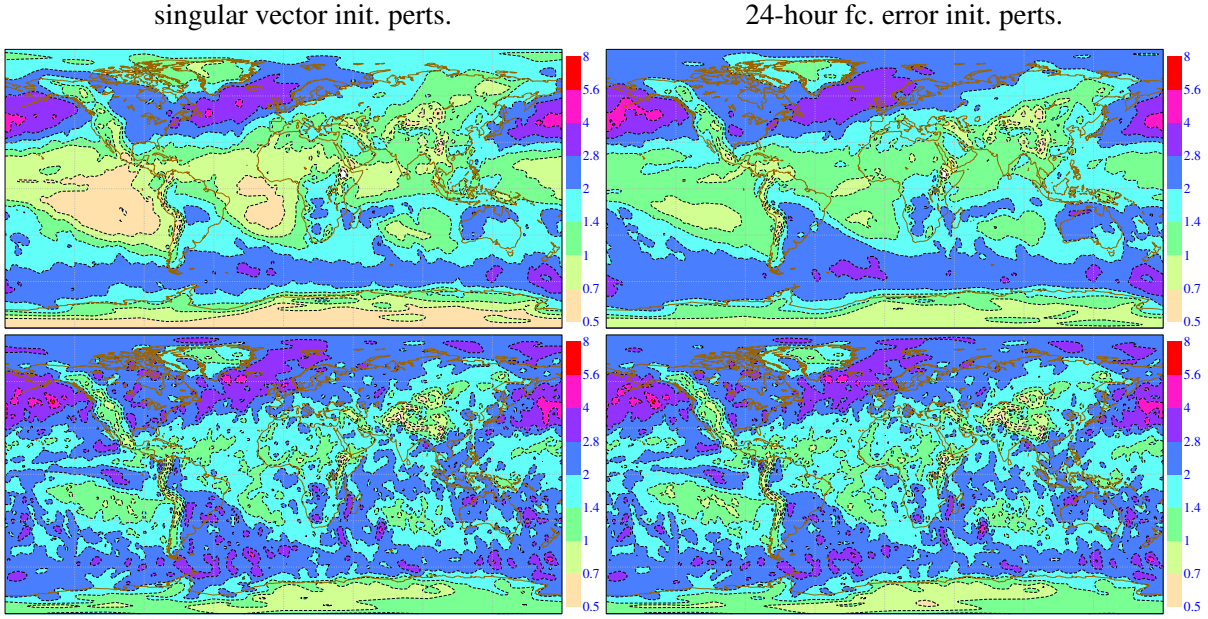

Figure 17: Ensemble standard deviation (top) and ensemble mean RMS error (bottom) at 48 h lead time for the meridional wind component ($m\,s^{-1}$) at 850 hPa.

ensemble better than singular vector ensemble) prevail outside the storm tracks. This is consistent with the better agreement of spread and ensemble mean error in the forecast error ensemble. These results are also consistent with the results shown in the previous section which showed that a climatological error distribution centred on the ensemble mean can provide a probabilistic forecast as skilful or even superior (depending on the measure) to the EPS in the early forecast ranges.

Representing initial uncertainties by samples of past forecast errors is clearly suboptimal as variations of uncertainty depending on the flow and the observation usage are ignored. The fact that the ensemble with forecast error perturbations is as good as or superior (depending on variable, lead time and region) as the ensemble using the operational singular vector initial perturbations indicates that the latter are also a suboptimal set of initial perturbations. The reliable prediction of the temporal and spatial variations of initial uncertainties is, at least in principle, possible through ensemble data assimilation techniques. This statement is corroborated by the good spread-reliability in the early forecast ranges of the Canadian Ensemble, which uses an Ensemble Kalman Filter (Section 3.1). Work on implementing an ensemble of 4D-Var assimilations with perturbations to observations and the forecast model is in progress at ECMWF (Isaksen et al. 2007). For the EPS, it is planned to replace the perturbations based on evolved singular vectors with perturbations from this 4D-Var ensemble (Buizza et al. 2008). Diagnostic comparisons with the benchmark system are expected to continue to aid the development of further improved representations of flow-dependent variations of initial uncertainty.

5 Discussion

Diagnostics in the early forecast ranges are likely to be the most useful in order to determine what aspects in the representation of initial uncertainties are deficient as interactions between different regions and variables as well as the presence of model uncertainties had less time to complicate the picture. However, uncertainties of the data used for verifying the ensemble forecasts may become an issue when early lead times are considered. Two aspects are relevant here. Firstly, forecast error and analysis error will be correlated to some degree in the early ranges if the forecast is evaluated with an analysis obtained from

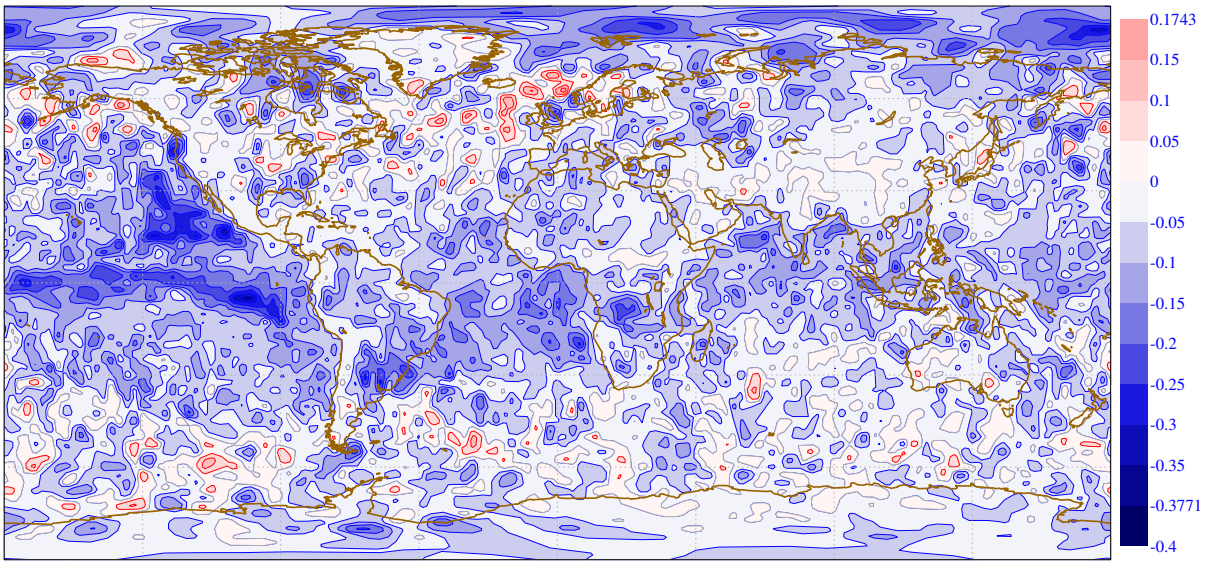

Figure 18: Difference ($m\,s^{-1}$) between CRPS of ensemble using short-range forecast errors and CRPS of ensemble using singular vectors at 48 h lead time for the meridional wind component at 850 hPa.

assimilation cycles that depend on the analysis from which the ensemble has been initialized. This aspect can be addressed at least partially by using an independent analysis for the forecast evaluation. Secondly, the ensemble aims to represent the distribution of truth. As any verifying data deviates from truth by some error ε, the ensemble needs to be transformed in order to predict the distribution of the verifying data. This can be achieved through addition of noise to the ensemble. The noise should be distributed according to the estimated characteristics of the distribution of errors ε (Saetra et al. 2004). Alternative approaches to account for uncertainty in the verifying data are discussed by Bowler (2006) and by Candille and Talagrand (2008).

It is an open question whether it is easier to account for uncertainties of the verifying data when using observations or when using analyses. On the one hand, analysis errors are affected by model errors and are difficult to estimate. Furthermore, the diagnostic risks to become incestuous if analysis error estimates are required that are also used in defining the initial perturbations of the ensemble. On the other hand, observation errors will need to account for the instrument error and the error of representativeness. The latter depends on the model and is difficult to estimate. Furthermore, observation errors tend to be larger than typical analysis errors and observational coverage is very inhomogeneous. As ensemble methods are refined for shorter lead time forecasts, the topic of accounting for uncertainty of the verification data is likely to receive more attention in the future.

This overview has entirely focussed on diagnosing univariate aspects of the ensemble forecasts. However, some applications of ensembles depend on multivariate aspects of the pdf. It is conceivable that a probabilistic forecast is reliable independently for variable A and variable B but not for predicting the joint distribution of A and B. An example may be an application depending on wind speed, temperature and humidity at a given location. Another example is the prediction of flow-dependent background error covariances for use in data assimilation schemes. For this purpose, it would be natural to verify the correlations in addition to the predicted variances.

6 Conclusions

Among the many measures available for evaluating ensemble forecasts, the assessment of the joint distribution of ensemble variance and the ensemble mean error remains an indispensable basic tool. In the ECMWF EPS, recent advances in modeling the variability of the atmosphere through improvements in the model physics and spatial resolution have led to a very good agreement between the mean ensemble variance and the variance of the ensemble mean error for the large-scale flow in the mid-latitudes across the whole reange of lead times up to 15 days. However, the variations of ensemble variance exhibit some degree of unreliability at lead times of up to two days. A comparison of the EPS with three other very skilful ensembles in the TIGGE archive indicates that the Canadian ensemble appears to have the most consistent spatio-temporal distribution of ensemble variance in the early lead times. Yet, the Canadian ensemble is not the most skilful in terms of probabilistic skill as it has less resolution than the EPS.

By combining probabilistic scores and a climatology of errors of the ensemble mean (or the control forecast), the skill of variations in the shape of the predicted pdf can be evaluated. This has been demonstrated for the pdf of continuous variables and for binary events. The binary events are defined relative to a reference forecast (ensemble mean or control) instead of a climatological mean in order to focus on the shape of the pdf. The distance of the event threshold from the reference forecast is scaled with the climatological standard deviation of the error. This procedure provides a natural scaling with lead time and avoids the degeneracy of probabilities observed for short lead times. To assess the prediction of continuous variables, two proper scores, the Continuous Ranked Probability Score and the Continuous Ignorance Score (the latter defined for the Gaussian distribution with mean and variance predicted by the ensemble) have been used. Empirical evidence from diagnosis of the ECMWF EPS as well as an idealized heteroscedastic model based on predictions with Gaussian distributions show that the Continuous Ignorance Score is a more sensitive measure for assessing the variations in pdf shape than the Continuous Ranked Probability Score. The idealized heteroscedastic Gaussian model can be used to provide quantitatively useful estimates of the difference in skill between the EPS and a static error distribution centred on the ensemble mean for lead times were the EPS has a reliable spread distribution (say ≥ 3 d).

In order to assess the sensitivity of ensemble prediction systems to the specification of initial perturbations, a benchmark system has been developed that uses a sample of the random component of past short-range forecast errors to perturb the initial conditions. The performance of an ensemble with short-range forecast error initial perturbations has been compared to an ensemble using singular vector perturbations in experiments with the ECMWF forecast system. Despite some obvious deficiencies in the benchmark system in terms of initial overdispersion and unrealistic perturbations due to the flow-independence, the benchmark ensemble is as skilful or more skilful than the ensemble using the singular vector perturbations. This indicates the potential for further improvements of the ECMWF ensemble by advancing the methodology for initial perturbations.

Acknowledgements

I would like to thank Renate Hagedorn for her comments on an earlier version of this manuscript and for providing the TIGGE verification data. I am grateful to Linus Magnusson for his contributions in studying benchmark systems for initial perturbations. Furthermore, I appeciate a discussion with Hans Hersbach on the static distribution centred on the control which will minimize the CRPS. Discussions that took place in the THORPEX mailing list on Theoretical Aspects of Ensemble Prediction, in particular the contributions by Olivier Talagrand and Frederic Atger, also helped preparing this presentation.

References

Barlow, R. J., 1989: *A Guide to the Use of Statistical Methods in the Physical Sciences*. Wiley.

Bechtold, P., M. Köhler, T. Jung, F. Doblas-Reyes, M. Leutbecher, M. J. Rodwell, F. Vitart, and G. Balsamo, 2008: Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales. *Q. J. R. Meteorol. Soc.*, **134**, 1337–1351.

Bougeault, P., Z. Toth, C. Bishop, B. Brown, D. Burridge, D. H. Chen, B. Ebert, M. Fuentes, T. M. Hamill, K. Mylne, J. Nicolau, T. Paccagnella, Y.-Y. Park, D. Parsons, B. Raoult, D. Schuster, P. S. Dias, R. Swinbank, Y. Takeuchi, W. Tennant, L. Wilson, and S. Worley, 2009: The THORPEX interactive grand global ensemble (TIGGE). *Bull. Am. Meteor. Soc.* (conditionally accepted).

Bowler, N. E., 2006: Explicitly accounting for observation error in categorical verification of forecasts. *Mon. Wea. Rev.*, **134**, 1600–1606.

Buizza, R., J.-R. Bidlot, N. Wedi, M. Fuentes, M. Hamrud, G. Holt, and F. Vitart, 2007: The new ECMWF VAREPS (variable resolution ensemble prediction system). *Q. J. R. Meteorol. Soc.*, **133**, 681–695.

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097.

Buizza, R., M. Leutbecher, and L. Isaksen, 2008: Potential use of an ensemble of analyses in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **134**, 2051–2066.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **125**, 2887–2908.

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.

Candille, G. and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Q. J. R. Meteorol. Soc.*, **131**, 2131–2150.

Candille, G. and O. Talagrand, 2008: Impact of observational error on the validation of ensemble prediction systems. *Q. J. R. Meteorol. Soc.*, **134**, 959–971.

Errico, R. M., R. Yang, M. Masutani, and J. S. Woollen, 2007: The estimation of analysis error characteristics using an observation system simulation experiment. *Meteorol. Z.*, **16**, 1–14.

Gneiting, T. and A. E. Raftery, 2007: Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

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

Hagedorn, R., 2008: Using the ECMWF reforecast dataset to calibrate EPS forecasts. *ECMWF Newsletter*, **117**, 8–13.

Hagedorn, R., R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2010: Comparing TIGGE multi-model forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Mon. Wea. Rev. (in review)*.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble forecasts. Part I: 2-metre temperature. *Mon. Wea. Rev.*, **136**, 2608–2619.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560.

Hamill, T. M. and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Q. J. R. Meteorol. Soc.*, **132**, 2905–2923.

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

Isaksen, L., M. Fisher, and J. Berner, 2007: Use of analysis ensembles in estimating flow-dependent background error variances. In *Flow-dependent aspects of data assimilation*, Workshop Proceedings, ECMWF, Shinfield Park, Reading, UK, 65–86. Available at www.ecmwf.int/publications/library/ecpublications/_pdf/workshop/2007/Data_assimilation.

Jung, T. and M. Leutbecher, 2008: Scale-dependent verification of ensemble forecasts. *Q. J. R. Meteorol. Soc.*, **134**, 973–984.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.

Leutbecher, M., R. Buizza, and L. Isaksen, 2007: Ensemble forecasting and flow-dependent estimates of initial uncertainty. In *Flow-dependent aspects of data assimilation*, Workshop Proceedings, ECMWF, Shinfield Park, Reading, UK, 185–201. Available at www.ecmwf.int/publications/library/ecpublications/_pdf/workshop/2007/Data_assimilation.

Leutbecher, M. and T. N. Palmer, 2008: Ensemble forecasting. *J. Comp. Phys.*, **227**, 3515–3539.

Lewis, J. M., 2005: Roots of ensemble forecasting. *Mon. Wea. Rev.*, **133**, 1865–1885.

Magnusson, L., J. Nycander, and E. Källén, 2009: Flow-dependent versus flow-independent initial perturbations for ensemble prediction. *Tellus*, **61A**, 194–209.

Mason, S. J. and A. P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.*, **137**, 331–349.

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

Morcrette, J.-J., G. Mozdzynski, and M. Leutbecher, 2008: A reduced radiation grid for the ECMWF Integrated Forecasting System. *Mon. Wea. Rev.*, **136**, 4760–4772.

Mureau, R., F. Molteni, and T. N. Palmer, 1993: Ensemble prediction using dynamically conditioned perturbations. *Q. J. R. Meteorol. Soc.*, **119**, 299–323.

Murphy, A. and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338.

Palmer, T., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. Technical Report Tech. Memo. 598, ECMWF, Reading, UK.

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

Roulston, M. S. and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660.

Saetra, Ø., H. Hersbach, J.-R. Bidlot, and D. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Wea. Rev.*, **132**, 1487–1501.

Silverman, B., 1986: *Density estimation for statistics and data analysis*. Chapman & Hall/CRC.

Simmons, A., S. Uppala, D. Dee, and S. Kobayashi, 2007: Era-interim: New ECMWF reanalysis products from 1989 onwards. *ECMWF newsletter*, **110**, 25–35.

Simmons, A. J. and A. Hollingsworth, 2002: Some aspects of the improvement in skill of numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **128**, 647–677.

Talagrand, O. and G. Candille, 2009: Verification of ensemble systems. In *Diagnostics of data assimilation system performance*, Workshop, ECMWF, Reading, UK.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. In *Proc. of Workshop on Predictability*, ECMWF, Reading, UK, 1–25.

Tennekes, H., A. P. M. Baede, and J. D. Opsteegh, 1987: Forecasting forecast skill. In *Proc. of Workshop on Predictability in the Medium and Extended Range*, ECMWF, Reading, UK, 277–302.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Am. Meteor. Soc.*, **74**, 2317–2330.

Tribbia, J. J. and D. P. Baumhefner, 2004: Scale interactions and atmospheric predictability: An updated perspective. *Mon. Wea. Rev.*, **132**, 703–713.

Uppala, S. M., P. W. Kållberg, A. J. Simmons, and Coauthors, 2005: The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.*, **131**, 2961–3012.

Wilks, D. S., 1997: Resampling hypothesis tests for autocorrelated fields. *J. Clim.*, **10**, 65–82.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed., Academic Press.

A Expected ensemble variance and variance of ensemble mean error

Here, the dependence of the ensemble mean RMS error and the ensemble variance estimate on ensemble size are derived. Let, y and $x_j, j = 1,\ldots,M$ denote independent, identically distributed (i.i.d.) random variables with mean μ and variance σ^2. Then, $y' \equiv y - \mu$ and $x'_j \equiv x_j - \mu$ are also i.i.d. with mean 0 and variance σ^2. The expected RMS error of the ensemble mean is given by

$$\begin{aligned}\mathbb{E}\left(\frac{1}{M}\sum_{j=1}^{M} x_j - y\right)^2 &= \mathbb{E}\left(\frac{1}{M}\sum_{j=1}^{M} x'_j - y'\right)^2 = \mathbb{E}\left(\frac{1}{M}\sum_{j=1}^{M}(x'_j - y')\right)^2 = \\ &= \frac{1}{M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}\mathbb{E}(x'_j - y')(x'_k - y') = \frac{1}{M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}\mathbb{E}(x'_j x'_k - y' x'_k - y' x'_j + y' y') \\ &\underset{(iid)}{=} \frac{1}{M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}(\delta_{jk} - 0 - 0 + 1)\sigma^2 = \left(1 + \frac{1}{M}\right)\sigma^2. \quad (32)\end{aligned}$$

Here, δ_{jk} denotes Kronecker's delta: For $j = k$, $\delta_{jk} = 1$ and 0 otherwise.

Likewise, the expected ensemble variance can be obtained as

$$\mathbb{E}\,\frac{1}{M}\sum_{j=1}^{M}\left(x_j-\frac{1}{M}\sum_{k=1}^{M}x_k\right)^2=\frac{1}{M^3}\sum_{j=1}^{M}\mathbb{E}\left(\sum_{k=1}^{M}(x_j-x_k)\right)^2=\frac{1}{M^3}\sum_{j=1}^{M}\mathbb{E}\left(\sum_{k=1}^{M}(x'_j-x'_k)\right)^2$$
$$=\frac{1}{M^3}\sum_{j=1}^{M}\sum_{k=1}^{M}\sum_{l=1}^{M}\mathbb{E}(x'_j-x'_k)(x'_j-x'_l)=\frac{1}{M^3}\sum_{j=1}^{M}\sum_{k=1}^{M}\sum_{l=1}^{M}\mathbb{E}\left(x'^2_j-x'_jx'_l-x'_kx'_j+x'_kx'_l\right)$$
$$\underset{(iid)}{=}\frac{1}{M^3}\sum_{j=1}^{M}\sum_{k=1}^{M}\sum_{l=1}^{M}\left(1-\delta_{jl}-\delta_{kj}+\delta_{kl}\right)\sigma^2=\frac{1}{M^3}\left(M^3-M^2-M^2+M^2\right)\sigma^2=\left(1-\frac{1}{M}\right)\sigma^2. \quad (33)$$

B Expected Continuous Ranked Probability Score for Gaussian distributions

Here, equation (17) for the expected value of the CRPS is derived. Gneiting et al. (2005) obtained the CRPS for verifying datum x when the prediction is a Gaussian distribution (their Eq. (5)). Note, that the CDF of the Gaussian distribution with mean 0 and variance 1 evaluated at x can be expressed as $\frac{1}{2}+\frac{1}{2}\Phi(x/\sqrt{2})$, where

$$\Phi(x)=\frac{2}{\sqrt{\pi}}\int_0^x\exp\left(-t^2\right)\mathrm{d}t \quad (34)$$

denotes the error function. Using this relationship, Eq. (5) of Gneiting et al. (2005) can be written as

$$\mathrm{CRPS}(N(\mu,\sigma^2),x)=\frac{\sigma}{\sqrt{\pi}}\left[-1+\sqrt{\pi}\frac{x-\mu}{\sigma}\,\Phi\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)+\sqrt{2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right]. \quad (35)$$

Now, the expected value of the CRPS for Gaussian truth and Gaussian forecast can be computed. Without loss of generality, we can assume $\mu=0$. Let the prediction be $N(0,\sigma_f^2)$ and the truth be distributed according to $N(0,\sigma_t^2)$.

$$\mathbb{E}\,\mathrm{CRPS}(N(0,\sigma_f^2),x_t)=$$
$$\frac{1}{\sigma_t\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left(-\frac{z^2}{2\sigma_t^2}\right)\frac{\sigma_f}{\sqrt{\pi}}\left[-1+\sqrt{\pi}\frac{z}{\sigma_f}\Phi\left(\frac{z}{\sqrt{2}\sigma_f}\right)+\sqrt{2}\exp\left(-\frac{z^2}{2\sigma_f^2}\right)\right]\mathrm{d}z \quad (36)$$

The three terms in the square bracket give rise to three definite integrals. The first and last term in square brackets are straigtforward to integrate because they involve only Gaussian densities. The integral due to the second term in square brackets can be expressed as

$$\int z\exp(-b^2z^2)\,\Phi(az)\,\mathrm{d}z=\frac{1}{2b^2}\left(\frac{a}{[a^2+b^2]^{1/2}}\Phi([a^2+b^2]^{1/2}z)-\exp(-b^2z^2)\Phi(az)\right), \quad (37)$$

where $a^{-2}=2\sigma_f^2$ and $b^{-2}=2\sigma_t^2$. This concludes the derivation of (17).