From Ensemble Forecasts to Predictive Distribution Functions

Jochen Bröcker*
Max–Planck Institut für Physik komplexer Systeme
01187 Dresden, Germany

—

Leonard A. Smith
Centre for the Analysis of Time Series
London School of Economics
London WC2A 2AE, United Kingdom

December 3, 2007

Abstract

The translation of an ensemble of model runs into a probability distribution is a common task in model based prediction. Common methods for such ensemble interpretations proceed as if verification and ensemble were draws from the same underlying distribution, an assumption not viable for most, if any, real–world ensembles. An alternative is to consider an ensemble as merely a source of information rather than possible scenarios of reality. This approach, which looks for maps between ensembles and probabilistic distributions, is investigated and extended. Common methods are revisited, and an improvement to standard kernel dressing, called "affine kernel dressing" (AKD), is introduced. AKD assumes an affine mapping between ensemble and verification; typically not acting on individual ensemble members but on the entire ensemble as a whole; the parameters of this mapping are determined in parallel with the other dressing parameters, including a weight assigned to the unconditioned (climatological) distribution. These amendments to standard kernel dressing, albeit simple, can improve performance significantly and are shown to be appropriate for both over– and under–dispersive ensembles, unlike standard kernel dressing which exacerbates over–dispersion. Studies are presented using operational numerical weather predictions for two locations and data from the Lorenz63 system, demonstrating both effectiveness given operational constraints and statistical significance given a large sample.

1 Introduction

Ensemble forecasts consist of several simulations of the future evolution of the dynamical process under concern (see e.g. Toth et al., 2003). In principle, en-

*Corresponding author. e–mail: broecker@pks.mpg.de

semble forecasts allow us to convey additional information on forecast uncertainty (Tennekes, 1988), which is invaluable for informed decision making (Taylor and Buizza, 2003; Richardson, 2003b,a; Roulston et al., 2003). In both scientific studies as well as practical applications, distribution functions are often more convenient to manipulate than a set of point values. The question then arises how to transform an ensemble into such a distribution function, a task often referred to as *statistical postprocessing of ensemble forecasts* in Wilks (2006); Wilks and Hamill (2007); Raftery et al. (2005) or *ensemble interpretation* in Jewson (2004a), the latter term being used in this paper. Any particular method for interpreting ensembles will be referred to as an ensemble interpretation method (other authors, e.g. Wilks, 2006; Wilks and Hamill, 2007, use the term ensemble–MOS method).

Ensemble interpretation methods generally differ due to the different families of distribution functions employed in building the ensemble interpretation and the way it is actually built. Both aspects are discussed in this paper. As to the different families of distribution functions, two particular approaches are considered here. The first one is referred to as *kernel dressing* and consists of replacing individual ensemble members by kernel functions. In the second approach, the ensemble is replaced by a parametrized distribution function, where the parameters of the distribution function have to be represented as functions of the original ensemble. This approach will be referred to as *distribution fit* or *DF interpretation*[1,2]. Both approaches typically involve parameters, which have to be determined.

Approaches to build the ensemble interpretation method differ in what the ensemble is taken to represent. In the simplest case, the ensemble is considered a collection of equally likely scenarios of reality, drawn from the same distribution as the verification (a *perfect ensemble*). This approach suggests that ensemble interpretation is accomplished by approximating this underlying distribution, for example by parametric estimation techniques (see e.g. Mood et al., 1974, Chapter VII) or kernel estimates (Silverman, 1986).

Although ensembles have been used to great effect even when assumed to be perfect (Wilks, 2002), we argue that a different paradigm is available which naturally includes the case where ensemble members and verifications do *not* share the same distribution. Nor need we assume that any one of the models in hand is true in any sense. Here we are interested in a distribution of the verification given the *information* contained in the ensemble. A formalism for constructing such distributions could take into account that ensembles and corresponding verifications are *not* draws from the same or at least fairly similar distributions, but entirely different ones.

This paradigm defines ensemble interpretation in a much broader sense than just interpolating a distribution function underlying the ensemble. In fact, there is no need to assume that ensembles are draws from distributions at all. As a simple example it will be demonstrated that a mere linear transformation of the ensemble already brings about a significant improvement in predictive perfor-

[1] In fact, kernel dressing and DF interpretation are not really distinct, as a sum of kernel functions can be interpreted as a special family of distribution functions, the centres of the kernel being part of the parameters. But when speaking of DF interpretations, we usually have somewhat more common families of distributions in mind, like Gaussian, Weibull or exponential distributions.

[2] The term distribution fitting is used by e.g. Wilks (2006).

mance of kernel dressing. Finding this linear transformation will be neither a preliminary nor a subsequent step to dressing, but integral part of it. Inasmuch as dressing involves finding unspecified parameters of the dressing method, we consider dressing a generalisation of statistical learning (Hastie et al., 2001).

The performance of forecast distributions is evaluated using *scoring rules* (Gneiting and Raftery, 2007; Selten, 1998). Some scores can be applied to the raw ensemble itself Gneiting and Raftery (e.g. CRPS–score 2007), while others can be applied to smoother probability assignments only, as provided for example by ensemble interpretation methods. Thus ensemble interpretations render the application of those scores to ensemble forecasts feasible. In this paper we focus attention on the *Ignorance* score (Roulston and Smith, 2002; Good, 1952). Strengths and weaknesses of this score are clarified as well.

Techniques for ensemble interpretation are the subject of Section 2, where state–of–the–art ensemble interpretation methods are revisited and a new *affine kernel dressing (AKD)* method is presented. A comparison of these ensemble interpretation methods in terms of their mathematical properties is subject to Section 3. Scoring rules are discussed briefly in Section 4, along with details of how to optimize the performance of ensemble interpretation methods, while questions of robust estimation and the value of blending in the climatological distribution are discussed in Section 5. In Section 6, we apply the ensemble interpretation techniques to temperature forecasts at London Heathrow and Heligoland (German Bight) as well as to the Lorenz63 system. The AKD method is shown to be capable of dealing with the imperfect ensembles more adequately than common ensemble interpretation methods in these cases. Furthermore, the Lorenz63 example demonstrates the insufficiency of Gaussian DF interpretations.

2 Interpreting Ensemble Forecasts

This section introduces a new dressing method referred to as *affine kernel dressing (AKD)* in the context of three well known methods, namely *Gaussian DF interpretation (GDF)*, *standard kernel dressing methods (SKD)*, and *bayesian model averaging (BMA)* (see e.g. Wilks, 2006; Wang and Bishop, 2004; Raftery et al., 2005; Roulston and Smith, 2003; Hoeting et al., 1999). We use the following notation throughout the paper. By

$$\mathbf{x} = [x_1, \ldots, x_d] \tag{1}$$

we denote an ensemble with d ensemble members. Typically, different ensemble members have different dynamical and statistical properties, depending on the ensemble generation scheme. In this paper though, we treat all ensemble members equally, or in other words, the ensemble interpretation methods considered in this paper do not depend on the ordering of the ensemble members. If some of the x_i need to be treated differently than others, for example if they come from different models[3], a superscript $x_i^{(J)}$ should be used. This case is to be distinguished from an ensemble in a higher dimensional space. Neither multi–model ensembles nor ensembles in high dimensional spaces are considered in this

[3]The unperturbed ensemble member (the "control") could be treated differently, which we will not do in this paper though.

paper. In general, the ensemble is a function of time, which we denote by $\mathbf{x}(t)$, while we write $y(t)$ for the *verification*, that is, the quantity to be forecast. The number of ensemble members d might even change over time. The ensemble has a mean and a variance, which are defined as

$$m(\mathbf{x}) = \frac{1}{d}\sum_i x_i, \tag{2}$$

$$v(\mathbf{x}) = \frac{1}{d}\sum_i (x_i - m(\mathbf{x}))^2, \tag{3}$$

respectively. Finally, $p(y; \mathbf{x}, \theta)$ is a probability density function derived from the ensemble $\mathbf{x}$, where θ denotes further parameters. In other words, $p(y; \mathbf{x}, \theta)$ denotes the interpreted ensemble as a probability density function, given the original ensemble. In fact, a probability density function need not be the goal, as will be discussed at the end of Section 4.

We first consider Gaussian DF interpretations (GDF), which can be written as

$$p(y; \mathbf{x}, \theta) := \frac{1}{\sqrt{\nu}} K(\frac{y-\mu}{\sqrt{\nu}}), \tag{4}$$

where K is a standard Gaussian density. Depending on the problem, other distributions can be more appropriate, for example Weybull or Γ–distributions. The parameters μ and $\sqrt{\nu}$ are the mean and the standard deviation of the distribution, respectively. Setting μ and $\sqrt{\nu}$ equal to the mean and the standard deviation of the ensemble is a possible choice (Wilks, 2002), but by doing so we would approximate the distribution of the ensemble, rather than the distribution of the verification given the ensemble, which is our goal. A conceptually different approach is to determine $\sqrt{\nu}$ and μ by functions of the ensemble and some free parameters θ, so that the DF interpretation shows good forecast performance. A variant of Gaussian DF interpretation following this philosophy was presented by Jewson (2004a,b), who suggested a mean μ and standard deviation $\sqrt{\nu}$ depending on the raw ensemble $\mathbf{x}$ as follows:

$$\mu = r_1 + r_2 \cdot m(\mathbf{x}), \tag{5}$$

$$\sqrt{\nu} = s_1 + s_2 \cdot \sqrt{v(\mathbf{x})}. \tag{6}$$

Thus $\sqrt{\nu}$ and μ are determined by linear functions of the standard deviations and the mean of the ensemble respectively. A very similar interpretation method was suggested by Gneiting et al. (2004), who replaced Equation (6) by $\nu = s_1 + s_2 \cdot v(\mathbf{x})$. The parameters $\theta = [r_1, r_2, s_1, s_2]$ are free parameters, for which $r_1 = 0$, $r_2 = 1$, $s_1 = 0$, $s_2 = 1$ are reasonable initial choices. The linear relationships in Equations (5) and (6) might be unable to cope with ensembles which are grossly different from the verification. The key insight of Jewson (2004a,b) and Gneiting et al. (2004) is that the parameters r_1, r_2, s_1, s_2 have to be determined according to forecast performance, rather than to represent the distribution of the ensemble members. Determining the parameters r_1, r_2, s_1, s_2 thus hinges on what counts as "good performance". Both the issue of finding the parameters as well as precise definitions of performance will be discussed in Section 4. This approach is distinctly different from for example Wilks (2002), where the probability distribution is fitted to the ensemble, without any reference to the verification.

An obvious shortcoming of Gaussian DF interpretation is that the shape of the dressed ensemble is invariably Gaussian. A more versatile method is provided by *kernel dressing*. Various versions of kernel dressing have been considered in the literature (Wilks, 2006; Roulston and Smith, 2003; Wang and Bishop, 2004; Raftery et al., 2005). A general way to present the kernel dressing approach reads as follows

$$p(y; \mathbf{x}, \theta) := \frac{1}{d\sigma} \sum_i K(\frac{y - ax_i - \omega}{\sigma}). \tag{7}$$

Hence, a kernel dressed ensemble is a sum of bumps, with one bump replacing each ensemble member. The shape of the bumps is determined by the *kernel* K. Each bump is centered at $ax_i + \omega$, where x_i is the i'th ensemble member. Thus a scales the ensemble, while ω acts as an offset. The width of each bump is determined by the *bandwidth* σ. As with GDF, a, σ and ω are quantities that might depend on the ensemble and on a parameter vector θ in a way we have to specify. Note that the bandwidth σ has to be positive. For simplicity, throughout this paper the kernel K will be a standard Gaussian density

$$K(\xi) := \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\xi^2). \tag{8}$$

Hence kernel dressing results in a sum of d Gaussians, in contrast to GDF, which gives a single Gaussian. Possible advantages of using different kernels with finite support like the Epanechnikov kernel (Silverman, 1986) are discussed in Section 5.

A wide variety of different kernels have been employed in similar or related circumstances (Roulston and Smith, 2003; Silverman, 1986). All results below apply to kernels which are normalized and positive and furthermore have mean zero and unit variance[4]. We remark that the Gaussian kernel employed here is furthermore symmetric, but this property is not used in this paper.

From the properties of the kernel immediately follows that the ensemble interpretation $p(y; \mathbf{x}, \theta)$ in Equation (7) is a positive and normalized probability density function. It is illustrative to compute the mean

$$\mu := \int y\, p(y; \ldots) \mathrm{d}y \tag{9}$$

and the variance

$$\nu := \int (y - \mu)^2\, p(y; \ldots) \mathrm{d}y \tag{10}$$

of the ensemble interpretation (Equation 7). We will now prove the following two identities on μ and ν, which we shall need later:

$$\mu = \omega + a\frac{1}{d}\sum_i x_i = \omega + am(\mathbf{x}), \tag{11}$$

$$\nu = \sigma^2 + a^2\frac{1}{d}\sum_i (x_i - m(\mathbf{x}))^2 = \sigma^2 + a^2 v(\mathbf{x}). \tag{12}$$

[4] As long as the kernel has a mean m and a variance s at all, we can always obtain mean zero and unit variance by using the kernel $\frac{1}{\sqrt{s}}K(\frac{\xi - m}{\sqrt{s}})$ instead of K. The Cauchy kernel provides an example of a kernel having neither a mean nor a variance.

The first Equation (11) states that the mean value of the ensemble interpretation is equal to the mean value $m(\mathbf{x})$ of the ensemble, scaled by the parameter a and shifted by the parameter ω. The second Equation (12) states that the variance of the ensemble interpretation is likewise equal to the variance $v(\mathbf{x})$ of the ensemble, scaled by the parameter a^2 and shifted by the parameter σ^2. Note however that a, σ and ω might depend on the ensemble as well, as mentioned above. To prove Equation (11), note that by substituting from Equation (7) into Equation (9) we get

$$\begin{aligned}
&\int y\, p(y;\ldots)\,\mathrm{d}y \\
&= \frac{1}{d\sigma}\sum_i \int yK(\frac{y-ax_i-\omega}{\sigma})\mathrm{d}y \\
&= \frac{1}{d}\sum_i \int (z+ax_i+\omega)K(z)\mathrm{d}z \\
&= a\frac{1}{d}\sum_i x_i + \omega \\
&= \omega + am(\mathbf{x}).
\end{aligned}$$

where we first substituted z for $\frac{y-ax_i-\omega}{\sigma}$, then used that the kernel is normalized and has zero mean and finally employed the definintion Equation 2 of the ensemble mean. To derive Equation (12), again substituting from Equation (7) we get along similar lines

$$\begin{aligned}
&\int y^2\, p(y;\ldots)\,\mathrm{d}y \\
&= \frac{1}{d\sigma}\sum_i \int y^2K(\frac{y-ax_i-\omega}{\sigma})\mathrm{d}y \\
&= \frac{1}{d}\sum_i \sigma^2 + (ax_i+\omega)^2 \\
&= \sigma^2 + \frac{1}{d}\sum_i (ax_i+\omega)^2. \qquad (13)
\end{aligned}$$

Furthermore, we expand

$$\begin{aligned}
&\frac{1}{d}\sum_i (ax_i - am(\mathbf{x}))^2 \\
&= \frac{1}{d}\sum_i (ax_i+\omega-am(\mathbf{x})-\omega)^2 \\
&= \frac{1}{d}\sum_i (ax_i+\omega)^2 - (\omega+am(\mathbf{x}))^2. \qquad (14)
\end{aligned}$$

Now employing Equations (13), (11) and then (14) we get

$$\begin{aligned}
\nu &= \int y^2 p(y;\ldots)\,\mathrm{d}y - \mu^2 \\
&= \sigma^2 + \frac{1}{d}\sum_i (ax_i+\omega)^2 - (\omega+am(\mathbf{x}))^2
\end{aligned}$$

$$= \sigma^2 + a^2 \frac{1}{d} \sum_i (x_i - m(\mathbf{x}))^2,$$

which establishes Equation (12). For constant a, σ, ω, these equations follow from Equations (4) and (7) in Raftery et al. (2005). All these identities are special instances of the well known fact that the overall variance of a model which is itself an average is given by the average of the individual variances plus the dispersion of the models.

The kernel dressing methods discussed in this paper (and in fact most other kernel dressing methods we know of) differ only in how the parameters σ, ω, and a are determined as functions of $\mathbf{x}$ and θ. For affine kernel dressing (AKD), σ and ω are set to

$$\omega = r_1 + r_2 \cdot m(\mathbf{x}), \tag{15}$$
$$\sigma^2 = h_{\mathcal{S}}^2 \cdot (s_1 + s_2 \cdot a^2 v(\mathbf{x})). \tag{16}$$

Here $h_{\mathcal{S}}$ is *Silverman's factor* (see Silverman, 1986)

$$h_{\mathcal{S}} = 0.5 \cdot (4/(3d))^{1/5},$$

the meaning of which will be explained below. Substituting Equation (15) for ω in (11) and Equation (16) for σ in (12) we get the relations

$$\mu = r_1 + (a + r_2) \cdot m(\mathbf{x}), \tag{17}$$
$$\nu = h_{\mathcal{S}}^2 s_1 + a^2 (h_{\mathcal{S}}^2 s_2 + 1) \cdot v(\mathbf{x}). \tag{18}$$

The dressing approach as presented in Equations (15,16) leaves the free parameter vector $\theta := [r_1, r_2, s_1, s_2, a]$ to be determined. There is a different way to write Equations (15) and (16) which reveals more about the structure of AKD and the role of Silverman's factor. Combining Equations (15) and (16), it is easy to see that the dressed ensemble Equation (7) reads as

$$p(y; \mathbf{x}, \theta) := \frac{1}{d\sigma} \sum_i K(\frac{y - z_i}{\sigma}), \tag{19}$$

where

$$z_i = a x_i + r_2 m(\mathbf{x}) + r_1 \tag{20}$$
$$\sigma^2 = h_{\mathcal{S}}^2 \cdot (s_1 + s_2 \cdot v(\mathbf{z})). \tag{21}$$

The relations (19), (20), and (21) allow for the interpretation of AKD as dressing the ensemble $\mathbf{z}$, which is obtained from the original ensemble $\mathbf{x}$ through the transformation in Equation (20). This transformation will henceforth be referred to as an *affine ensemble transform*. Hence also the name affine kernel dressing[5]. Further possible generalisations of dressing could be obtained by replacing the affine ensemble transform (i.e. Equ. 20) by more general ensemble transforms, which are discussed in Appendix B. Note that the affine ensemble transform acts on the ensemble as a whole and *cannot be represented as a function acting on each ensemble member individually*. We stress that the ensemble transformation (Equation (20)) as well as the dressing (Equation (19)) are both

[5]which should in fact be "affine ensemble transform kernel dressing"

integral parts of the entire method and should not be considered as separate steps. In other words, the parameters in Equations (19) and (20) will generally depend on each other.

From the theory of kernel density estimates (Silverman, 1986) we take the ansatz Equation (21) for the bandwith σ. In the highly idealized situation that the transformed ensemble $\mathbf{z}$ is gaussian and perfect, $\sigma^2 = h_S^2 \cdot v(\mathbf{z})$ is a close to optimal choice for the bandwidth. Although we do not assume $\mathbf{z}$ to be either gaussian or perfect, using Silverman's factor conveniently scales s_1 and s_2 to ranges around one.

In Section 6, affine kernel dressing will be compared to Gaussian dressing as well as a more standard version of kernel dressing, henceforth referred to as *standard kernel dressing (SKD)*, which obtains by setting $a = 1$, $r_2 = 0$, and $s_1 = 0$. That is, standard kernel dressing allows for a fixed offset r_1 to all ensemble members as well as a bandwidth correction factor s_2.

Another special case emerges by setting $r_2 = 0, s_2 = 0$. This ensemble interpretation method was studied by Wilks (2006), who introduced it as a special case of bayesian model averaging (BMA, Raftery et al., 2005). As was pointed out in Wilks (2006), the general BMA technique might be justified if the ensemble members are expected to have significantly different error statistics, as for example in ensembles of different numerical weather models. For the initial condition ensembles considered below however, the ensemble members are expected to have quite similar statistics, whence a general BMA approach would be overly complex.

3 Properties of AKD, SKD, and GDF

In this section a brief look is taken upon the advantages and shortcomings to be expected of the four dressing methods presented. It is plausible that any kernel dressing is better than Gaussian dressing if (but not only if) the ensemble $\mathbf{x}(t)$ and the verification $y(t)$ are independent draws from the same underlying distribution (perfect ensemble) and the ensemble is sufficiently large. The reason is that with increasing ensemble size (and suitable choice of the bandwidth σ), the kernel dressed ensemble will approach the underlying density. Although we did not venture to find a proof, analogy to density estimation problems (Silverman, 1986) suggests that a neccessary criterion would seem to be $\sigma(d) \to 0$ if the ensemble size d goes to infinity, but slow enough so that still $d \cdot \sigma(d) \to \infty$, that is, $\sigma(d)$ shrinks slower than d. This is expected, for example, in best member dressing (Roulston and Smith, 2003). Hence we would expect that, if the ensemble is perfect, yet not gaussian but, for example, bimodal (Smith, 1997, 2002), kernel dressing will eventually outperform Gaussian dressing. Even if the perfect ensemble is actually a draw from a gaussian, it is not clear that Gaussian dressing is better than kernel dressing, since the parameters ω and σ in Equations (5) and (6) still need to be estimated from the ensemble. It can be shown[6] that maximum likelihood estimates of these parameters are suboptimal, and a t–distribution should be used rather than a Gaussian (Johnson and Wichern, 1992). This effect is essentially due to the small ensemble size.

Gaussian dressing, on the other hand, is expected to beat standard kernel dressing when the ensemble $\mathbf{x}(t)$ is reasonably Gaussian but overdispersive, or

[6] Penzer, J., 2006. Personal communication.

in other words, the ensemble members are further away from each other than from the verification. Since σ^2 is positive, Equation (12) reflects the basic result (see e.g. Wilks, 2006) that the variance of the standard kernel dressed ensemble (i.e. if $a = 1$) is always larger than the variance of the raw ensemble, no matter how σ is determined. Affine kernel dressing, in contrast, allows for the variance of the dressed ensemble to be a linear function of the variance of the raw ensemble, a feature it shares with Gaussian dressing and BMA. In operational numerical weather prediction, the ensemble spread is typically too small *on average*, leading to convex Talagrand diagrams (Wilks, 1995; Hamill, 2001). Nevertheless, Equation (12) is a relation for each individual ensemble. Independent of whether the ensemble variance is too large or too small on average, affine kernel dressing allows for a more flexible relationship between the variance of the ensemble and the variance of the *dressed* ensemble than standard dressing in either case. A distinct advantage of AKD over BMA emerges from the relations (17, 18). For AKD, these two relations are independent. This would in principle permit to debiase the ensemble mean and simultaneously optimize the spread–skill relationship. The relations (5, 6) show that the same is true for GDF. For BMA though, $r_2 = 0$ and $s_2 = 0$, in which case the linear part in *both* the relations (17, 18) is determined by a. In other words, having debiased the ensemble, there remains little which can be done for a better spread–skill–relationship. As demonstrated in Section 6, AKD offers significant benefits when applied to numerical weather predictions for which the square error of the ensemble mean is not well represented by the ensemble variance. To the extent that it is Bayesian, BMA provides a principled framework for constructing probability forecasts. This comes with the cost of assuming that one of the models is true (Hoeting et al., 1999) or alternatively that the available model class admits a perfect model.

While all variants of kernel dressing borrow from and bear some resemblance to *Kernel Estimation*, a technique employed to estimate probability density functions (Silverman, 1986), we stress that kernel *dressing* (and in fact ensemble interpretation in general) rests on different assumptions than kernel *estimation*. The latter attempts to fit a probability density function to a single and unchanging archive of points. These points are simultaneously forecasts *and* verifications. Future points, although not expected to be equal to any point in the archive, are nevertheless assumed to be drawn from the same source. Thereby, in KE, the ensemble and the verification are draws from the desired distribution. For kernel *dressing* of ensemble forecast, there is but one verification for every ensemble, and typically, the verification is *not* drawn from the ensemble, that is, the ensemble is demonstrably *not* perfect. The improved dressing method as presented in Equations (19), (20), and (21) looks superficialy similar to a kernel estimator applied to the transformed ensemble z. It should be kept in mind though that eventually all parameters of kernel dressing are determined simultaneously and depend on each other, thus the ensemble transform (Equ. 20), the choice of the bandwidth (Equ. 21) and the dressing (Equ. 19) cannot be separated.

4 Scoring and Training

The ensemble interpretation methods presented in the Section 2 depend on the as yet unspecified parameters θ. We consider the problem of determining the parameters of ensemble interpretations to be similar to the learning problem of statistics (Vapnik, 1998; Hastie et al., 2001). In the latter problem, the objective is to fit a functional relationship between certain inputs and verifications, based on a *training set* of input–verification pairs. The functional relationship is picked from a range of functions or *model class* according to performance. An algorithmic procedure that tunes the parameters according to performance over a training set will be referred to as a *training* algorithm. At the core of most training algorithms lies an iterative procedure which optimizes the expected performance as a function of the parameters. A more classical term for training algorithm is "estimation technique". The difference is only a linguistic one, but estimation might imply the existence of a true parameter (like a physical quantity) that is to be estimated. The parameters of ensemble interpretation methods though need not have any physical interpretation, whence the term training algorithm seems more appropriate here.

When interpreting ensembles, the objective is to find a *probabilistic* relationship between the inputs and verifications, where the model class consists of sums of kernel functions, and the training set consists of ensemble–verification pairs (hence the training set is often referred to as *forecast archive*). The unspecified parameters should be determined solely by forecast performance, not by any *a priori* assumtions, like, for example, that the ensemble and the verification are draws form one and the same underlying distribution. This obviously involves finding appropriate performance measures or *scoring rules* for probabilistic forecasts, which we will turn to now.

A scoring rule is a function $S(p(y), Y)$, where $p(y)$ is a probability density and Y is the verification. In this paper, scoring rules are defined like cost functions: small scores indicate better forecast skill. For example, the *Ignorance Score* is defined by the scoring rule

$$S(p(y), Y) = -\log(p(Y))$$

The Ignorance score is related to the log–likelihood (Mood et al., 1974; Bröcker and Smith, 2007) and plays an importand role in gambling theory. Another interesting scoring rule (although not used in this paper) is the *Proper Linear Score*. It is defined as

$$S(p(y), Y) = \int p^2(z)\mathrm{d}z - 2p(Y). \tag{22}$$

It should be noted that the Ignorance depends only on the single number $p(Y)$, while the Proper Linear Score depends on the entire functional form of $p(y)$. This particular property of the Ignorance is called *locality*. Local scores are typically cheaper to evaluate than nonlocal scores. Computing functionals of the probability density (such as the integral in Equation 22) are often very costly. As noted by Gneiting et al. (2004), similar reasons have hampered the use of the CRPS–score.

It turns out that not all conceivable candidates for scoring rules yield useful scores. An indispensable property of scores is *propriety*. Roughly speaking, a

score is proper if $p(y)$ achieves an optimal (i.e. minimal) expected score whenever the verification is drawn from $p(y)$. A scoring rule is *strictly proper* if that happens *only* if the verification is drawn from $p(y)$. Propriety is a property only of the scoring rule itself. The Ignorance and the Proper Linear Score are proper (for a proof of this fact as well as a discussion of the notion of propriety see Bröcker and Smith, 2007). A general result due to Bernardo (1979) states that all smooth, proper and local scores are affine functions of the Ignorance. Proper scores in general have been characterized by Gneiting and Raftery (2007).

In evaluating forecast systems, one is not only concerned with a single probability density function $p(y)$ but rather with a sequence $p_n(x)$ of probability density functions and corresponding verifications Y_n which can be employed to estimate the performance of the forecast system, in other words, the expected score (with respect to a proper scoring rule S). To this end, define the *empirical score*

$$S_N := \frac{1}{N} \sum_{n=1}^{N} S(p_n(x), Y_n). \tag{23}$$

The empirical score values the average performance of the forecast system over all samples in the archive. In the case of dressed ensembles, the probability density functions are time depend through the ensemble $\mathbf{x}(n)$, that is $p_n(y) = p(y; \mathbf{x}(n), \theta)$, where θ denotes the ensemble interpretation parameters. In the case of affine kernel dressing for example, $\theta = [a, r_1, r_2, s_1, s_2]$. Replacing the expression for $p_n(y)$ in Equation (23) and using the Ignorance score we obtain

$$S_N(\theta) = \frac{1}{N} \sum_{n=1}^{N} -\log\left(p(Y_n, \mathbf{x}(n), \theta)\right) \tag{24}$$

In Equation (24) the empirical score of the ensemble (which essentially reflects the performance of the forecast system) can be regarded as a function of the free ensemble interpretation parameters θ. Minimizing the score (and thereby optimizing the performance of the dressed ensemble) with respect to the parameters θ provides a means to choose these parameters, i.e. a means of training, reminiscent of statistical learning. In statistical learning, a functional relationship is picked from a range of functions according to its performance, which is often (but not always) the quadratic error. In ensemble interpretation, a relationship between ensembles and probability density functions is picked from a range of functions according to performance, which in this paper is measured by the Ignorance score. The approach to minimize performance measures (such as the Ignorance score) to determine parameters of forecast interpretation methods for continuous events was, to our knowledge, first considered by Jewson (2004a,b) and (apparently independently) by Gneiting et al. (2004). In so far as minimizing the Ignorance can be considered as maximum likelihood, it is of course a very old concept.

A thorough theoretical investigation of the minimum–score training strategy and the properties of the obtained parameters would be invaluable, but is not subject to this paper. We used an optimisation algorithm that solves a sequence of constrained quadratic optimisation problems (Gill et al., 1982). Other options are the EM–algorithm, employed by Raftery et al. (2005). Both algorithms are only guaranteed to find local rather than global minima. We are ignorant as to whether the EM–algorithm could be applied to other scores, while preliminary

studies indicate that sequential quadratic optimisation works equally well with the proper linear score. The Ignorance of kernel dressing can display multiple minima with rather poor performance. Robust solutions with good performance however are obtained in practice by a regularisation strategy, discussed in Section 5, along with a careful initialisation of the minimisation algorithm. The results reported in this paper were obtained using the following methodology for finding the initial conditions. The mean of the dressed ensemble (as described by Equ. (5) in case of Gaussian DF interpretation resp. Equ. (11) in case of kernel dressing) is fitted to the verification in a mean square error sense. Then the *variance* of the dressed ensemble (as described by Equ. (6) in case of Gaussian DF interpretation resp. Equ. (12) in case of kernel dressing) should roughly correspond to the squared residuals of the fitted mean. Thus fitting the variance of the dressed ensemble to the squared residuals gives a further condition to find initialisation parameters. As it turns out, this allows for finding complete initial conditions for Gaussian DF interpretation and standard kernel dressing. For affine kernel dressing, this strategy leaves s_2 unspecified, which is set to 1. The structure of the problem as presented in Equations (20), (21), and (19) and the use of Silverman's factor guarantee that setting $s_2 = 1$ is a reasonable choice unless the transformed ensemble (Equ. 20) is extremely poor.

5 Robustness Issues

Obtaining robust estimates for the parameters of ensemble interpretation methods can be difficult, especially if forecast busts are numerous or when the ensemble is small. This problem is often traced back to the empirical score showing a large variance. Recently, several authors (Gneiting and Raftery, 2007; Selten, 1998) criticized the Ignorance for being particularly prone to large variation. The Ignorance is a quite unforgiving score in that it extremely severly penalizes low probability assignment to verifications that actually obtain. Indeed, assigning vanishing probability to a verification yields an Ignorance of infinity. Even if the assigned probabilities are never exactly zero, a few "bad forecasts" can render the variance of the empirical Ignorance undesirably large, resulting in parameters obviously useless (this may be a positive attribute in decision support). It should be noted that the Ignorance has a clear interpretation in terms of gambling returns (Roulston and Smith, 2002; Good, 1952; Kelly, 1956). Under a certain betting scenario ("Kelly Betting", Kelly, 1956) the Ignorance describes the rate at which the forecaster's fortune increases with time. The properties of the Ignorance hence can be defended as representing properties of a game. Furthermore, large variations in the empirical score are *always* to be expected if the forecasts are poor and should adequately be dealt with, especially as the score might not even be a matter of choice. So how can large variations in the empirical score be avoided?

It was suggested by Gneiting and Raftery (2007) that the summands in Equation (24) could be censored, that is, a certain percentage of the data could be rejected as outliers. Another option could be to use a truncated logarithm, which would be reminiscent of ϵ–insensitive loss functions in regression (Vapnik, 1998). This seems inadvisible in cases where such "outliers" have a firm physical interpretation and are expected to become more relevant in the future dynamics, for example in seasonal forecasting. These and other means to com-

bat the influence of outliers on the score and subsequently the parameters are often refered to as *regularisation*. It has to be kept in mind though that the Ignorance (or whichever score is employed) is used both to train the ensemble interpretation parameters and also to evaluate the interpreted forecast. During *training*, any kind of regularisation is permissible and even recommended. For *evaluation* however, censoring or truncating of the score would require it to be re–interpreted. Important properties and interpretations of the score might not hold for the regularized score. For example, common interpretations of the Ignorance in terms of gambling return rates cease to apply if the sum in Equation (24) is censored, which essentially would be tantamount to canceling the highest winnings and to default on the worst bankruptcies. In practice, certain scoring procedures (e.g. sailing, ski jumping or ice skating) actually allow to retrospectively discount the worst results (sometimes requiring the best results to be cancelled too), but this is certainly not the case in "games" as for example casinos, energy markets or air traffic control. Hence in general it seems to depend on the particular problem whether a censored (or truncated) score is an appropriate measure of forecast performance.

In situations where a regularisation of the problem is neccesary during training, but where the problem statement does not allow for any censoring or otherwise altering of the score, it seems inevitable to apply a slightly different (i.e. regularized resp. not regularized) scoring methodology during the training (resp. evaluation) period. In this paper, the logarithm was effectively truncated by replacing all p_n which were equal to zero (up to numerical precision) by the smallest nonzero p_n. For evaluation though, the Ignorance was neither censored or truncated. Such discrepancies (which are inherent to all regularisation approaches) might seem disturbing at first sight. Currently we lack a full theoretical justification of this approach, but as an *ad hoc* scheme we found it to give superior results, presumably because of smaller variance in the dressing parameters.

To account for forecast failures during evaluation, the dressed ensemble was blended with an estimate of the climatology of the verification, thereby circumventing the problem of large variances in the empirical score. For a finite ensemble size, this is justifiable even in the case of a perfect ensemble. More specifically, let $p_n(y)$ be the interpreted ensemble and $q(y)$ be an estimate of the climatology of the verification. We use a mixture of both, like

$$r_n(y) := \alpha p_n(y) + (1-\alpha)q(y), \qquad 0 \leq \alpha \leq 1. \tag{25}$$

as the forecast distribution. The *weight* α is determined so as to minimize the Ignorance (i.e. to optimize the performance) of the combination, and hence must be involved in the optimisation. The resulting probability assigned to a verification Y is never smaller than $(1-\alpha)q(Y)$. The effect therefore is that a small, yet nonvanishing probability is assigned to the verification, as long as the latter does not fall outside the range of the data record employed to estimate the climatology. Forecast performance is often stated in relation to the performance of climatology as a reference. This means that the (mean of the) *difference* in performance between $p_n(y)$ and the climatology $q(y)$ is reported. Thus the climatology acts as a reference forecast, itself yielding a score of zero.

In case of the Ignorance, this can be written as

$$S_N[p] - S_N[q] := \frac{1}{N} \sum - \log \left(\frac{r_n(Y_n)}{q(Y_n)} \right). \tag{26}$$

Replacing $r_n(y)$ from Equation 25 we get for every summand

$$\begin{aligned} \frac{r_n(Y_n)}{q(Y_n)} &= \frac{\alpha p_n(Y_n) + (1-\alpha) q(Y_n)}{q(Y_n)} \\ &= \alpha \frac{p_n(Y_n)}{q(Y_n)} + (1-\alpha) \\ &\geq (1-\alpha), \end{aligned}$$

from which we can conclude

$$-\log\left(r_n(Y_n)\right) \leq -\log\left(q(Y_n)\right) - \log(1-\alpha).$$

Hence the empirical Ignorance of a forecast combined with climatology relative to climatology is never larger (i.e. worse) than $-\log(1-\alpha)$. Blending in climatology thus acts as a hedge against forecast busts. Another way to interpret a blend with climatology is to play cancelling bets. The Ignorance of a forecast *relative* to climatology describes the rate at which the forecasters fortune increases in a betting scenario where the odds are set according to climatology.[7] Mixing in a proportion $1-\alpha$ of climatology hence is equivalent to staking a proportion α of the fortune according to the forecast and a proportion $1-\alpha$ according to the odds given, which guarantees a certain return of at least a proportion $1-\alpha$ of the stake. The forecaster thus avoids being infinitely worse off than the house.

Only few forecast busts are sufficient to render a good climatology worth being blended with the forecast proper. As an example, Figure 1 shows $-\log(r_n(Y_n))$, combined with climatology, versus $-\log(q(Y_n))$, that is the climatology itself. The ensemble forecast was from ECMWF's medium range 10 day 51 member ensemble predicion system. The lead time was ten days. The weight assigned to the climatology is $1-\alpha = 0.051$. It is obvious from the plot that $-\log(r_n(Y_n))$ is never larger than $-\log(q(Y_n)) - \log(1-\alpha) = -\log(q(Y_n)) + 2.97$ at *every* verification (not just in the mean). Figure 2 shows the weight assigned to the climatology over lead time. The uncertainty bars display variations of the weight estimate obtained through cross–validation (see Appendix A).

It might prove difficult to determine α robustly, as the optimal combination of α and kernel bandwith (i.e. σ in Equation 21) for the training set might be a local (and very poor) minimum by suggesting a very wide bandwith to compensate for forecast busts, instead of employing the climatology for that purpose. This could be adressed by using a kernel function with a limited domain (like the quadratic Epanechnikov kernel, see e.g. Silverman, 1986), which yields infinite Ignorance for all points outside its domain. Alternatively, large kernel bandwidths σ could be penalized. Taking into account the finite ensemble size and probably a known rate of forecast busts, it should even be possible to derive an upper bound on α (i.e. a lower bound on the weight assigned to

[7] Or alternatively, relative Ignorance between two forecasts A and B describes the rate at which the fortune of forecaster A exceeds that of forecaster B.

climatology). The suggested precautions where however not neccessary for the data sets considered in this paper (see Section 6).

Another interesting interpretation of the weight $1 - \alpha$ assigned to the climatology could be to quantify of belief in or uncertainty of our forecast. The question arises if and how uncertainty of probabilistic forecasts could be quantified more generally, for example if the climatology is unknown or is known to be changing. If the predictive distribution is interpreted as a probability, we are now speaking about assigning an uncertainty to what is already a probability, thus introducing the idea of *second order probabilities*, that is quantifying statements like "the probability that it rains tomorrow at London Heathrow is evenly distributed between 10% and 20%". Second order probabilities lead to odds forecasts[8], that is forecasts with a total mass larger than one, the excess representing uncertainty in the forecast (Smith, 2007). Although it is not yet clear how uncertainty in probabilistic forecasts in general or odds in particular could be assigned or used, such a framework requires ensemble interpretation methods that focus on information content in the ensembles to hand, while the assumption that the resulting predictive distributions can be interpreted or acted on as if they are (decision relevant) probability distributions has to be dropped.

6 Comparative Studies

This section analyzes the performance of standard kernel dressing (SKD), affine kernel dressing (AKD), Gaussian DF interpretation (GDF). Shortcomings of SKD and GDF, which originally motivated the developement of AKD, are illustrated. AKD was compared to BMA too, albeit less comprehensively. All ensemble interpretation methods were blended with climatology, with the exception of Gaussian DF interpretation (GDF). AKD is shown to be superior to all other methods for the problems considered. As far as we are aware, previous implementations of BMA do not blend in climatology, leading to significantly larger variations in performance and often inferior skill.

Results are presented for three different data sets. The first and second data sets constist of forecasts of the two metre temperature at London Heathrow Airport (WMO station Nr.03772) and Heligoland, German Bight (WMO station Nr.10015), respectively. The forecasts consist of ECMWF's 51 member ensemble (as for Figures 1 and 2). The verifications consist of station data, kindly provided by ECMWF as well. Forecasts were available for the years 2001–2005, featuring lead times from one to ten days. Verifications were available as far back as 1981. The years 1981–2000 were used to build a climatology. For any given day, the climatology is calculated only from data falling into the same annual period, defined by a window of ± 20 days. Hence the climatology depends as well on the season. All data verified at noon. The results for the weather data are shown in Figures 3 and 4 and are discussed below.

The third data set was generated using the Lorenz63 system (Lorenz, 1963). The ensemble, comprising 50 members, was generated from observations of the full state of the system, corrupted with 15dB noise[9]. The sampling interval was

[8] Judd, K., 2006. Personal communication

[9] The dB scala measures the ratio between the variances of two signals. A signal to noise ratio of d dB indicates that $d = 10 \cdot \log_{10}(\frac{v_s}{v_n})$, where v_s (resp. v_n) is the variance of the clean

0.05. For data assimilation, a variant of the indistinguishable states importance sampler (Judd and Smith, 2001) was employed. Data assimilation is necessary here, since we have but noisy measurements of the true underlying state of the system. Although ensembles could also be generated by perturbing the true initial condition, this option would of course not be available in real applications. Hence, using a data assimilation scheme corresponds much more to realistic circumstances. Forecasts were considered at ten lead times $[0.1, 0.2, \ldots, 1]$. The same model was used to generate both forecasts and the verifications. Moreover, the verifications formed a single trajectory. In general, the AKD significantly outperforms SKD and GDF, especially for the Lorenz63 system. The AKD method also appears to be the most robust method among the three, in the sense that the performance of AKD showed the least variability. The results for the Lorenz63 system are shown in Figure 5, and are discussed below.

Figure 3.a shows the performance in terms of Ignorance of AKD relative to climatology for the London Heathrow data set. The x–axis shows the lead time. The uncertainty bars (in fat line style) mark the $10\% - 90\%$–range obtained from a tenfold cross validation. The thin line shows the Ignorance of the out–of–train output. The corresponding thin confidence bars show the $\pm 2\sigma$–range (see Appendix A). Cross validation is known to have a large variance (Hastie et al., 2001), while the variance of the out–of–train output (see Appendix A) on the other hand tends to be too small. In any case, AKD gives a significantly higher skill than the climatology under both validation methods. In order to compare the performance of AKD, SKD, and GDF interpretation, we plot the difference of the Ignorance (Equ. 37) directly, rather than leave it to the reader to compare performances across multiple graphs, allowing confidence bars of the *relative performance*, as the uncertainty in the relative performance does not follow from the uncertainties of the absolute performances (see Appendix A). The axis scaling has been set so as to allow for easy comparison across different graphs.

Figure 3.b shows the performance of GDF versus AKD. The out–of–train confidence bars overlap the zero line slightly for lead time 24, 48 and 72 hours, but sees AKD significantly ahead of GDF beyond lead time 72h. The cross–validation assessment indicates essentially the same, the bars being wider though.

Figure 3.c shows the performance of SKD versus AKD for London Heathrow. Up to lead time 120 hours, the AKD method outperforms SKD substantially, at least according to out–of–train calculation. For higher lead times, AKD still appears to be better for a large fraction of cross validation runs.

Figure 3.d shows the performance of SKD versus GDF. From lead time 96 hours onwards, the two are essentially similar. The potential advantage of SKD when dealing with strongly nongaussian ensembles seems to play little role for temperature at lead times up to 100 hours.

The comparison between BMA and AKD (Fig. 3.f) remains somewhat inconclusive, although AKD is certainly better than BMA for medium and larger lead times. In terms of out–of–train performance, AKD is significantly better than BMA. Note that our implementation of BMA includes blending with the climatology. This blending is not a common part of BMA, and some Bayesians might object to it on principle, but it allows for a better comparison between BMA and AKD. Without climatology, BMA shows considerably larger variation

signal (resp. the noise)

in performance (not shown).

The findings for Heligoland (Fig. 4) are very similar to the results obtained for London Heathrow, a notable exception being that AKD wins over GDF by an even wider margin. Furthermore, the superior performance of AKD over BMA occurs for higher lead times when compared to London Heathrow (cf. Fig. 3.f with Fig. 3.f).

It is interesting to look at the nongaussianity of the ensemble for these two datasets, especially in connection to the performance of AKD versus GDF (Figures 3.b and 4.b), as we expect AKD to outperform GDF if the ensembles deviate from gaussianity. As a measure of nongaussianity, we employ the kurtosis of the ensemble, that is the centered moment of fourth order,

$$k(\mathbf{x}) = \frac{1}{d}\sum (x_i - m(\mathbf{x}))^4,$$

where $m(\mathbf{x})$ is, as before, the ensemble mean. For Gaussian distributions, the fourth centered moment is expected to be three times the variance, hence we expect for Gaussian ensembles

$$\kappa(\mathbf{x}) := \frac{k(\mathbf{x})}{3v(\mathbf{x})} - 1 \approx 0.$$

The distribution of this statistic κ for Gaussian ensembles can be simulated through bootstrapping and subsequently compared with the distribution of κ for the actual ensembles. In Figures 6 and 7, the $10\% - 90\%$–range of the actual κ's is indicated by a black bar, for London Heathrow and Heligoland, respectively. The y–axis is calibrated in terms of quantiles of κ for gaussian ensembles. If the actual ensembles were gaussian, all bars should extend from 0.1 to 0.9[10]. It emerges that at both locations the ensembles tend to be particularly nongaussian at lead times around 96 hours. Interestingly, for larger lead times at London Heathrow, the κ statistic indicates again a more Gaussian ensemble. For Heligoland, the ensembles are also particularly nongaussian at lead times around 96 hours, but contrary to London Heathrow, the ensembles stay fairly nongaussian out to lead time 240 hours. This provides a possible explanation for the better performance of AKD in relation to GDF at Heligoland. It is worth noting that the better performance of AKD versus GDF furthermore indicates that the nongaussian ensembles carry information beyond the second moment. The AKD interpretation outperforms GDF not only because the ensembles are nongaussian, but because this nongaussianity actually carries information.

As to the reasons why AKD outperforms the other discussed methods, further investigation is necessary. There is some evidence though that the mechanisms discussed in Section 3 are in fact responsible. We investigated the parameters for both BMA and AKD for London Heathrow at lead time 120h. Note that AKD is particularly strong here, and that the ensembles are particularly non–gaussian. The parameters were substituted into Equations (17) and (18). For AKD, these relations read

$$\mu = 0.0 + 0.99\, m(\mathbf{x}), \tag{27}$$

$$\nu = 1.93 + 0.53\, v(\mathbf{x}). \tag{28}$$

[10]The scale of the y–axis is not linear in p but in $\log\left(\frac{p}{1-p}\right)$. For small (resp. large p), this renders the plot effectively logarithmic in p (resp. $1-p$).

For BMA, these relations read

$$\mu = 0.003 + 1.0\, m(\mathbf{x}), \tag{29}$$

$$\nu = 0.17 + 1.0\, v(\mathbf{x}). \tag{30}$$

The cross–validation approach (see Appendix A) yields an uncertainty of less than 10^{-3} for all these coefficients. Since Equations (27) and (29) agree to a high degree, AKD and BMA always have very similar means. The Equations (28) and (30) though differ. As was mentioned already in Section 3, for BMA the slope of the variance relation (Equ. 30) is always the square of the slope of the mean relation (Equ. 29), whence it is impossible for BMA to have mean and variance relations like Equations (27) and (28). It appears though that the variance relation of AKD (Equ. 28) gives the better performance. It is interesting to note that the two variance relations intersect at $v(\mathbf{x}) = 3.74$, as this is almost exactly the temporal average of $v(\mathbf{x})$, which is 3.76. This means that on *average* over time, BMA and AKD feature the same variance (3.93), which is in fact the ensemble variance, slightly inflated. For individual ensembles though, their variances generally differ. In particular, the variations of the variance (i.e. the variance of ν) is larger for BMA than for AKD. The lead times 48h and 216h (for Heathrow and Heligoland) were investigated along the same lines, with similar findings. Finally, we would like to mention that for AKD, BMA and SKD, the weight assigned to climatology behaves roughly as in Figure 2.

The experiments carried out using the Lorenz63 data confirm the general picture already obtained from the weather data experiments, thereby confirming that any positive results are not only due to limited counting statistics. AKD is the best performing and most robust method. The performance of AKD versus climatology is shown in Figure 5.a. AKD and SKD perform roughly equal (Fig. 5.b). We suspect that this is due to the high quality of the ensemble. If the ensembles were either over or underdispersive, we would expect AKD to perform better than SKD. Talagrand diagrams (not shown) however indicate that the ensembles are very reliable (i.e. neither over nor underdispersive), which explains the similar performance of both AKD and SKD. Inspection of the AKD models (not shown) indicate that the parameter a (see Equs. 20 and 21) is close to one, in particular for small lead times, rendering AKD and SKD essentially equal. Kernel dressing (i.e. AKD and SKD) significantly outperform GDF for higher lead times (Figs. 5.c and d). A main reason for this is certainly the increasingly nongaussian ensembles for higher lead times, as is obvious from a plot of the κ–statistic (Fig. 8). Again, by comparing the variance of the performance across different graphs, it can be concluded that AKD features not only the best, but also the most robust performance.

7 Conclusion

There is valuable information in ensemble weather forecasts; extracting this information requires interpreting the ensemble. Comparing different methods for interpreting ensembles shows that the affine kernel dressing technique introduced in this paper yields promising results for operational temperature forecasts using the ECMWF ensemble; its strengths are also illustrated in the context of perfect model and large forecast–verification archive with the Lorenz 63

system. In terms of the ignorance score, affine kernel dressing outperforms the other methods in all cases considered; both cross validation and out–of–train evaluation confirm the results. The importance of blending climatology into the probability distribution function is shown.

Our approach aims at extracting information from an ensemble without making assumptions regarding the perfection of the model or the ensemble. There is no assumption that the verification represents "just another draw" from the distribution that generated the ensemble, nor any assumption that the model class available admits a "true" model. There is abundant evidence that such assumptions are not justified in operational forecast systems. We furthermore touch on the question of whether or not probability distributions functions are indeed the best representation of the valuable information contained in these systems.

To the extent that operational forecasts are made to be used, the ensemble interpretation is a critical component contributing to the value of an ensemble prediction system. By aiming merely to extract information from the model simulations and other available distributions (for example climatology), affine kernel dressing has been shown to improve this critical component, and may contribute to enhancing the value of ensemble–based prediction, particularly in applications like weather forecasting at all lead times.

8 Acknowledgements

We are grateful to Renate Hagedorn and ECMWF for providing ensemble forecast as well as station data for London Heathrow and Heligoland. The forecasts and verifications for the Lorenz63 system were provided by Hailiang Du. We gratefully acknowledge fruitful discussions with Jeremy Penzer, Antje Weisheimer, and Liam Clarke. Questions and suggestions by two anonymous referees led to further improvements of the manuscript.

A On Out–Of–Train Evaluation and Cross Validation

Performance evaluation of forecast systems aims to provide a sound estimate of the *future* or *out of sample* performance, or more specifically on data the forecast system will encounter while in operation. Estimating the the performance on data which was already used to build or select the forecast system or any parts of it, including the ensemble interpretation methods, is likely to give overoptimistic results. Ideally, the ensemble interpretation methods are trained on one part of the available data, while the other part is left aside as test data. To get reliable estimates of the out of sample performance, the test data set has to be sufficiently large. But typically, as the total amount of data available is already limited, we cannot afford to sacrifice large proportions of the data for out of sample performance assessment, as a small training set is expected to provide inferior parameter values. We apparently face the problem of having either unrealistic parameters or unreliable estimates of performance.

A way around this apparent *circulus vitiosus* is *cross validation* (see e.g. Hastie et al., 2001). The price to be paid though is having to train the ensemble

interpretation method a number of times rather than only once. More specifically, cross validation works as follows. The training set $T = \{(\mathbf{x}(n), Y_n), n = 1 \dots N\}$ is partitioned into J partitions T_j of equal lenght N/J. Let $\theta^{(j)}$ be the parameter vector obtained by training the ensemble interpretation method on $T \setminus T_j$, that is the training set without partition j. The score S_j for this particular $\theta^{(j)}$ is evaluated *only* on T_j (i.e. the data that had been left out for finding $\theta^{(j)}$) and is given by

$$S_j := \frac{J}{N} \sum_{n \in T_j} -\log\left(p(Y_n; \mathbf{x}(n), \theta^{(j)})\right). \tag{31}$$

The mean of all S_j is called the *cross validation estimate* of the score

$$S_{\mathrm{CV}} := \frac{1}{J} \sum_j S_j. \tag{32}$$

The standard error of S_{CV} can be estimated thus

$$\Delta S_{\mathrm{CV}} := \sqrt{\frac{1}{J(J-1)} \sum_j S_j^2 - S_{\mathrm{CV}}^2}. \tag{33}$$

In similar fasion, quantiles of the S_j can be computed to give confidence intervals for the score. In the figures of Section 6, we plotted the median score along with the $10\% - 90\%$–range as confidence bars. In Hastie et al. (2001), using the standard error is recommended, but this gives obscure results if the distribution of the S_j is rather non–gaussian.

Another way to estimate the likely variations of the score, referred to as the *out–of–train* estimate, works as follows. Using the parameters $\theta^{(j)}$ obtained through cross validation, we can compute the *out–of–train* output by

$$\pi_n := p(Y_n; \mathbf{x}(n), \theta^{(j_n)}), \tag{34}$$

where j_n denotes the index of the partition containing $(Y_n, \mathbf{x}(n))$. Recall that the sample $(Y_n, \mathbf{x}(n))$ was not used during the training of the particular parameter $\theta^{(j_n)}$. Using the out–of–train output, the expected score

$$S_{\mathrm{OOT}} := \frac{1}{N} \sum_n -\log(\pi_n), \tag{35}$$

and its standard error

$$\Delta S_{\mathrm{OOT}} := \sqrt{\frac{1}{N(N-1)} \sum_n \log(\pi_n)^2 - S_{\mathrm{OOT}}^2} \tag{36}$$

can be computed. An easy calculation (comparing Equations (34,35) with (31,32)) reveals that S_{OOT} is actually equal to S_{CV}. The standard errors however generally differ. Since it does not make sense to compute quantiles for the out–of–train method, we used $\pm 2 \cdot \Delta S_{\mathrm{OOT}}$ confidence intervals. It is hard to say which of the two methods is to be preferred, whence we used both for performance assessment. The CV–method explicitely takes into account model variations, but as the individual CV–partitions are shorter than the training set, the model

variations are likely to be over–estimated. The OOT–technique uses the entire data set to estimate the variations, but both model variations as well as performance variations are compounded. Furthermore, the individual outputs π_n are assumed to be independent, an idealisation that leads to underestimation of the variations.

It is often neccessary to consider the *improvement* of the Ignorance obtained by $p_n(y)$ over $q(y)$. This improvement is naturally measured by the increase in Ignorance (also often referred to as the *relative* Ignorance of $p_n(y)$ with respect to $q(y)$)

$$S_p - S_q = \frac{1}{N} \sum_{n=1}^{N} - \log\left(p_n(Y_n)\right) + \log\left(q(Y_n)\right). \tag{37}$$

This quantity, as the estimate of the Ignorance proper, carries an uncertainty. It is important to realize that there is no simple relationship between the uncertainty in $S_p - S_q$ and the individual uncertainties in S_p and S_q, since both are highly dependent. In other words, the standard error of $S_p - S_q$ is *not* in any simple way related to the individual standard errors of S_p and S_q. To estimate the standard error of relative ignorances through either cross–validation or out–of–train technique, the Equation (33) (respectively (36)) has to be applied to the differences in the performance S_j between the forecasts on each partition (respectively the differences of $-\log(\pi_n)$).

All performance plots (Figures 3 to 5) show relative Ignorance (either with respect to another forecast or with respect to climatology). The cross validation estimates are plotted in fat line style, while the out–of–train estimates are in thin line style. As noted above, cross validation and out–of–train differ only in their estimates of the standard error.

B On Ensemble Transforms

In this paper we considered the interpretation of ensembles as the problem of finding a map from a series of ensembles onto a series of distribution functions for a corresponding series of verifications. The method of affine kernel dressing provides a special class of such mappings by combining a simple kernel estimator (Equation 19) with what we termed an affine ensemble transform (Equation 20). This idea could be generalized by using different ensemble transforms, probably involving nonlinear elements. A particular linear ensemble transform was used in this paper, and there is the possibility that the concept is of wider applicability in postprocessing ensemble forecasts. To this end, ensemble transforms need to be properly understood and classified first. At this point, we are not even sure if the ensemble transform used in this paper is the most general linear ensemble transform. In this appendix, some neccessary conditions will be formulated that we deem general ensemble transforms should obey and are hopefully sufficient for a conclusive analysis of the aforementioned question.

The key property of an ensemble, which distinguishes it from a vector, is that it is still considered the same ensemble if some members are interchanged either across parts of or the entire ensemble. For example, although the 50 perturbed members of the ECMWF ensemble are distinguishable by the initial perturbations used to compute them, they can be considered indistinguishable for the purpose of many applications. For the numerical studies in Section 6, even

the unperturbed ("control") forecast was considered indistinguishable from the perturbed ensemble members. Such an ensemble of mutually interchangeable members will be called a *pure* ensemble. Ensembles consisting of a *collection* of pure ensembles (say, if we combine pure ensembles produced by different models) might be called *compound* ensembles. All ensemble interpretation methods studied in this paper tread the ensembles as pure, as they are invariant to any permutation of the ensemble members.

An ensemble transform f is defined simply as a mapping between ensembles (not neccessarily having the same number of members). They key property of a (pure) ensemble, namely that the ordering of the ensemble members is irrelevant, imposes certain restrictions on f, which we are going to formulate. Let (as before) $\mathbf{x} = [x_1 \dots x_d]$ be the original ensemble (consisting of d members) and

$$\mathbf{z} = f(\mathbf{x})$$

be the transformed ensemble (of d' members). If we now permute the elements in $\mathbf{x}$, then $\mathbf{z}$ must remain the *same ensemble*, which means, as we have seen, that at most some permutation of the elements of $\mathbf{z}$ should take place. In other words, if π denotes a permutation of d elements and $\pi\mathbf{x}$ denotes the permuted original ensemble, there must be a permutation κ of the members of the transformed ensemble $\mathbf{z}$ so that

$$\kappa\mathbf{z} = f(\pi\mathbf{x}) \tag{38}$$

holds.

The permutation κ so obtained obviously depends on π, or in other words, the relation (38) defines a mapping $\kappa(\pi)$ between permutations. If ι is the *identity*, that is the permutation of d elements that actually keeps all elements the same, then likewise $\kappa(\iota)$ is the identity permutation (of d' elements). This relation can (with a slight abuse of notation) be written as

$$\kappa(\iota) = \iota. \tag{39}$$

Furthermore, if π_1, π_2 are two permutations, a third permutation $\pi_1 \circ \pi_2$ arises through composition of π_1, π_2. It follows immediately from Equation (38) that

$$\kappa(\pi_1 \circ \pi_2) = \kappa(\pi_1) \circ \kappa(\pi_2). \tag{40}$$

Properties (39) and (40) state that any ensemble transform gives rise to a *representation* κ of the group of permutations of d symbols in the group of permutations of d' symbols.

The transformed ensemble $\mathbf{z}$ in Equation (38) is not neccessarily a pure ensemble though, but it might be possible to split the members of $\mathbf{z}$ into two sub–ensembles, $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2]$, so that for any permutation π, the corresponding permutation $\kappa(\pi)$ permutes in fact only the members of $\mathbf{z}_1$ and $\mathbf{z}_2$ among each other, but does not interchange members of $\mathbf{z}_1$ with members of $\mathbf{z}_2$. If this is the case, we have created a compound ensemble consisting of (at least) two pure ensembles. In order to exclude this behaviour we have to require that for any two indices i, j in the set $[1 \dots d']$, there is at least one permutation π so that $\kappa(\pi)$ permutes i into j. Groups of permutations with this property are called *transitive*. Hence the conclusion of this appendix can be summarized thus:

Via Equation (38), an ensemble transform induces a transitive representation of the group of permutations of d symbols in the group of permutations of d' symbols.

Transitive representations of the permutation groups have been widely studied and classified. Hence by means of group theory (Weyl, 1946) it should be possible to address questions like whether the affine ensemble transform as presented in Equation (20) is the most general class of ensemble transforms which can be obtained by linear operations.

References

J. M. Bernardo. Expected information as expected utility. *Annals of Statistics*, 7(3):686–690, 1979.

Jochen Bröcker and Leonard A. Smith. Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22(2):382–388, 2007.

P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, London, 1982.

T. Gneiting, A. Raftery, A. H. Westveld III, and T. Goldmann. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 2004.

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

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society*, XIV(1): 107–114, 1952.

Thomas M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560, 2001.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, first edition, 2001.

Jennifer A. Hoeting, David Madigan, Adrian Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

Stephen Jewson. Comparing the ensemble mean and the ensemble standard deviation as inputs for probabilistic medium–range temperature forecasts. *arXiv:physics*, 2004a.

Stephen Jewson. Moment based methods for ensemble assessment and calibration. *arXiv:physics*, 2004b.

Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, Englewood Cliffs, New Jersey, 3 edition, 1992.

Ian T. Jolliffe and David B. Stephenson, editors. *Forecast Verification; A practicioner's Guide in Athmospheric Science*. John Wiley & Sons, Ltd., 2003.

Kevin Judd and Leonard A. Smith. Indistinguishable states i: the perfect model scenario. *Physica D*, 151:125–141, 2001.

J. L. Kelly, Jr. A new interpretation of information rate. *Bell System Technical Journal*, 35, 1956.

Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 1963.

Alexander M. Mood, Franklin A. Graybill, and Duane C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill Series in Probability and Statistics. McGraw-Hill, 1974.

Adrian E. Raftery, Tilman Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 2005.

David S. Richardson. Predictability and economic value. Technical report, European Centre for Medium Range Weather Forecast, 2003a.

David S. Richardson. Economic value and skill. In Jolliffe and Stephenson (2003), chapter 8.

M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(130):1653–1660, 2002.

M. S. Roulston and L. A. Smith. Combining dynamical and statistical ensembles. *Tellus A*, 55, 2003.

Mark S. Roulston, Daniel T. Kaplan, Jost von Hardenberg, and Leonard A. Smith. Using medium range weather forecasts to improve the value of wind energy producton. *Renewable Energy*, 28, 2003.

Reinhard Selten. Axiomatic characterisation of the quadratic scoring rule. *Experimental Economics*, 1:43–62, 1998.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, first edition, 1986.

Leonard A. Smith. Predictability and chaos. In J. Holton, J. Pyle, and J. Curry, editors, *Encyclopedia of Atmospheric Sciences*, pages 1777–1785. Academic Press, 2002.

Leonard A. Smith. *A Very Short Introduction to Chaos*. Oxford University Press, 2007.

Leonard A. Smith. The maintenance of uncertainty. In *Proc International School of Physics "Enrico Fermi"*, volume CXXXIII, pages 177–246, Bologna, Italy, 1997. Società Italiana di Fisica.

J. W. Taylor and R. Buizza. Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, 19:57–70, 2003.

Hendrik Tennekes. The outlook: Scattered showers. *Bulletin of the American Meteorological Society*, 69(4), 1988.

Zoltan Toth, Olivier Talagrand, Guillem Candille, and Yuejian Zhu. Probability and ensemble forecasts. In Jolliffe and Stephenson (2003), chapter 7.

Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., New York, 1998.

Xuguang Wang and Craig H. Bishop. Improvement of ensemble reliability with a new dressing kernel. *Quarterly Journal of the Royal Meteorological Society*, 131, 2004.

Hermann Weyl. *The Classical Groups*. Princeton Mathematical Series. Princeton University Press, second edition, 1946.

Daniel S. Wilks. Smoothing forecast ensembles with fitted probability distributions. *Quarterly Journal of the Royal Meteorological Society*, 128, 2002.

Daniel S. Wilks. Comparison of ensemble–MOS methods in the Lorenz'96 setting. *Meteorological Applications*, 13, 2006.

Daniel S. Wilks. *Statistical Methods in the Athmospheric Sciences*, volume 59 of *International Geophysics Series*. Academic Press, first edition, 1995.

Daniel S. Wilks and Thomas M. Hamill. Comparison of ensemble–MOS methods using gfs reforecasts. *Monthly Weather Review*, (in press), 2007.

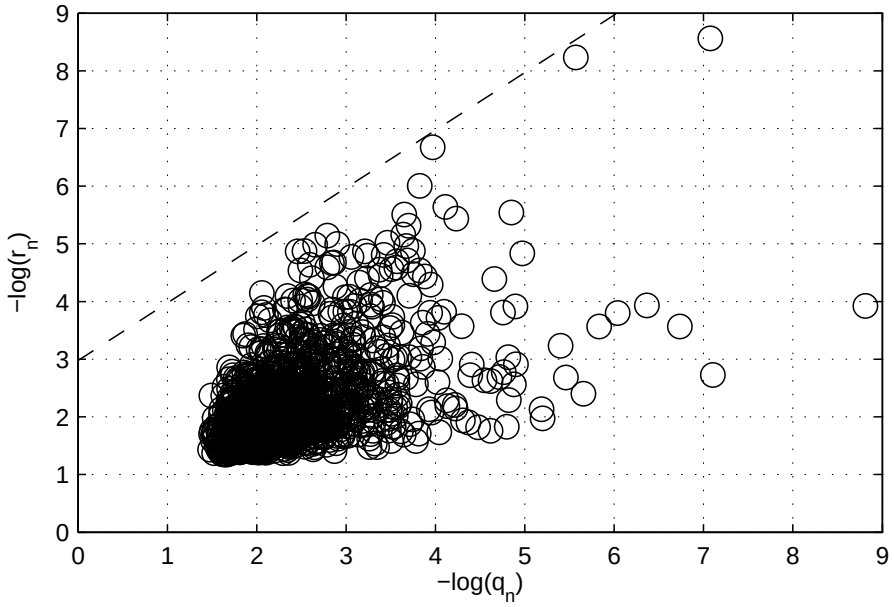

Figure 1: The Ignorance $-\log(r_n(Y_n))$ (ECMWF ensemble and climatology) vs $-\log(q(Y_n))$ (only climatology) for temperature at Heligoland, German Bight (WMO 10015), lead time ten days. The dressing method here is AKD. Obviously, $-\log(r_n(Y_n))$ is never larger than $-\log(q(Y_n)) - \log(1-\alpha)$. The weight assigned to the climatology is $1-\alpha = 0.051$, whence $-\log(1-\alpha) \approx 3$.

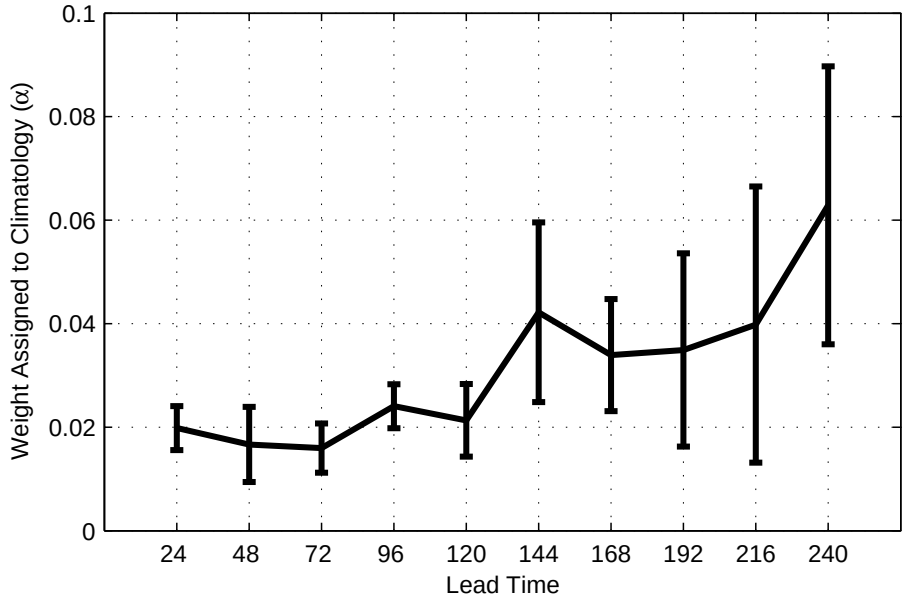

Figure 2: The weight assigned to the climatology over lead time. The rest is as in Figure 1. The uncertainty bars display variations of the weight estimate obtained through cross–validation (see Appendix A).

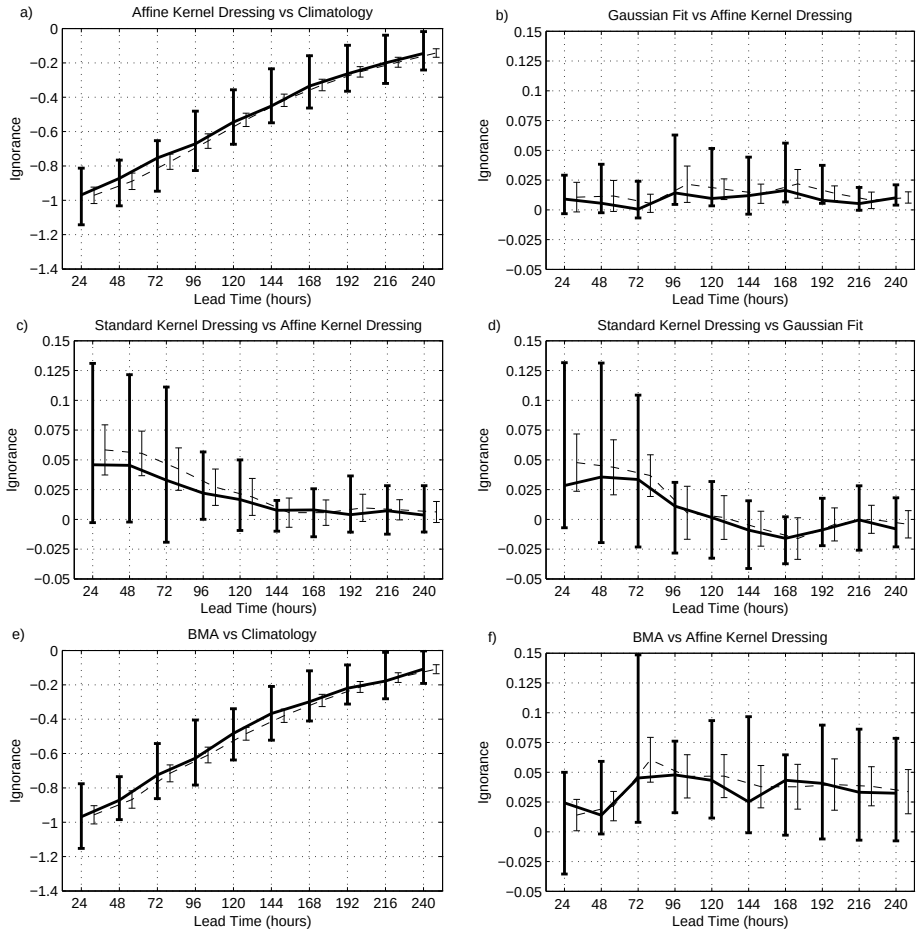

Figure 3: The relative Ignorance of the investigated ensemble interpretation methods and climatology for London Heathrow over lead time. The fat uncertainty bars are from tenfold cross validation ($10\% - 90\%$–range). The thin uncertainty bars correspond to the out–of–train performance ($\pm 2\sigma$–range).

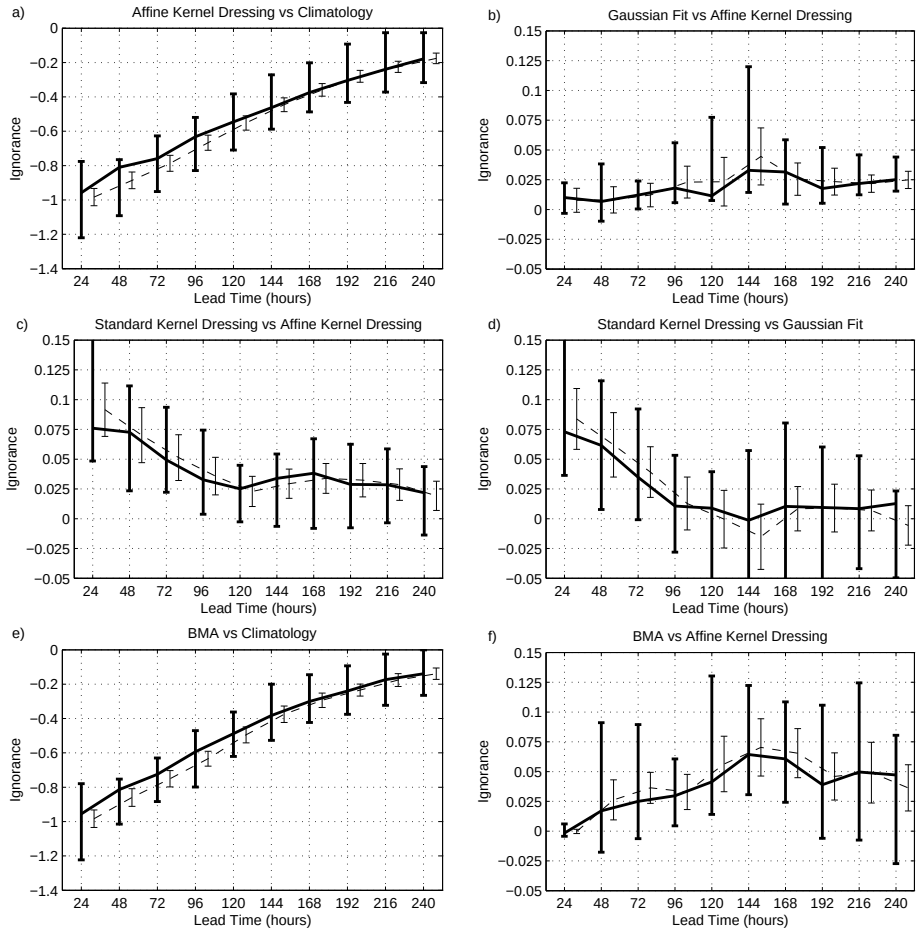

Figure 4: As in Figure 3, but for Heligoland.

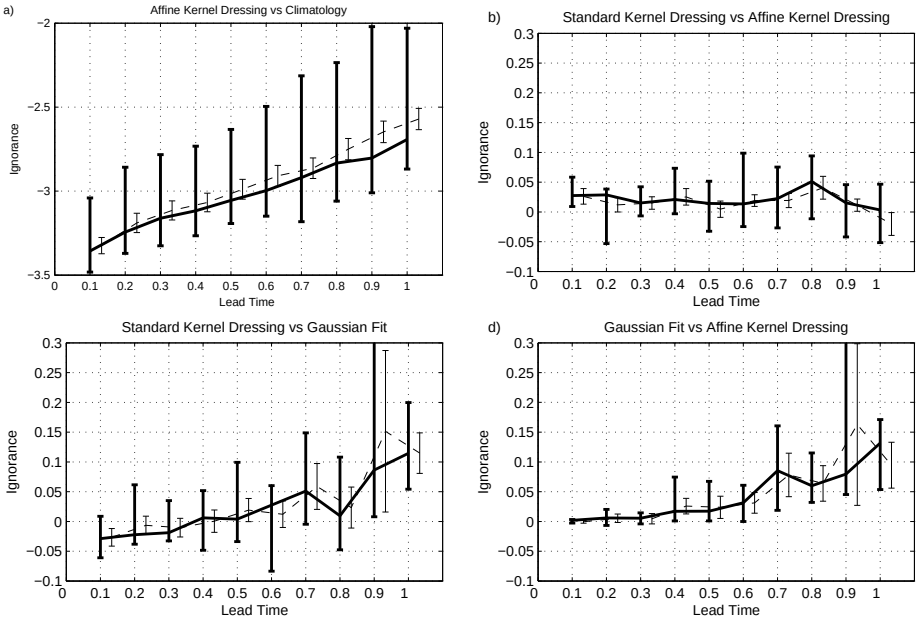

Figure 5: The relative Ignorance of the investigated ensemble interpretation methods for the Lorenz63 data set over lead time. Uncertainty bars are as in Figure 3.

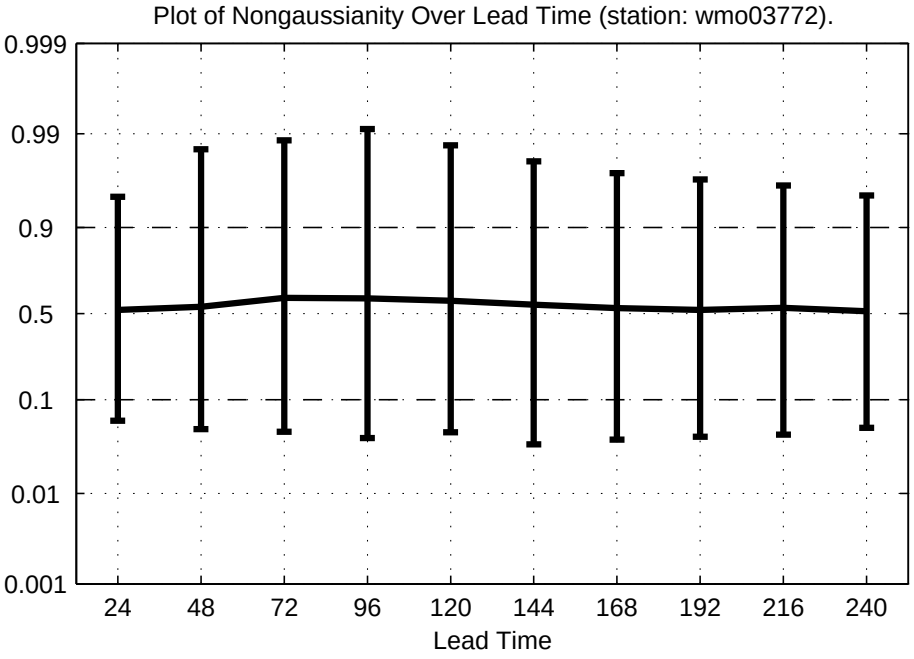

Figure 6: The $10\% - 90\%$–range of the κ–statistic for London Heathrow. The y–axis is calibrated in terms of quantiles of the κ–statistic for Gaussian ensembles and plot on $\log\left(\frac{p}{1-p}\right)$–scale.

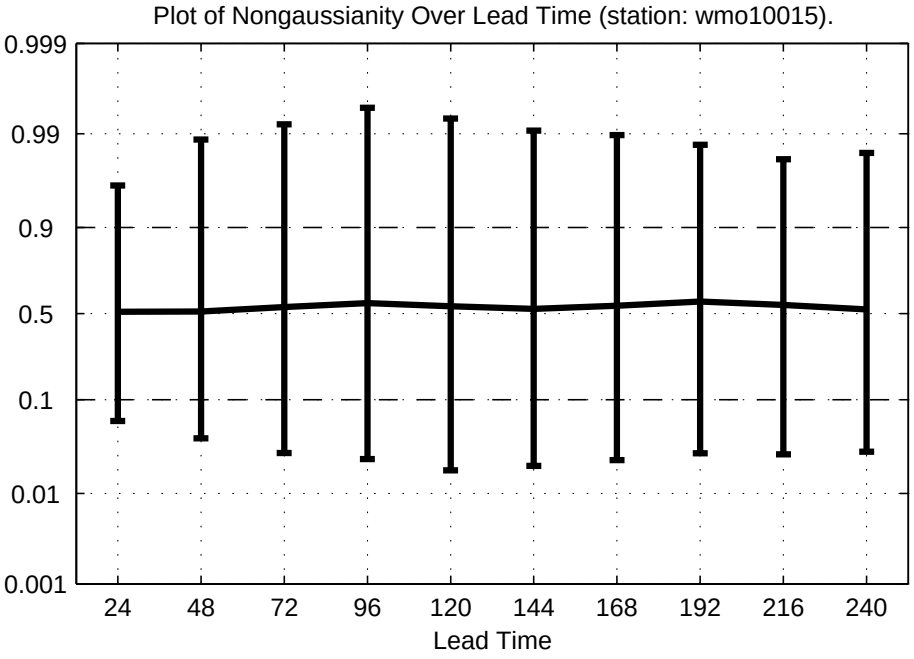

Figure 7: The $10\% - 90\%$–range of the κ–statistic for Heligoland. The y–axis is calibrated in terms of quantiles of the κ–statistic for Gaussian ensembles and plot on $\log\left(\frac{p}{1-p}\right)$–scale.

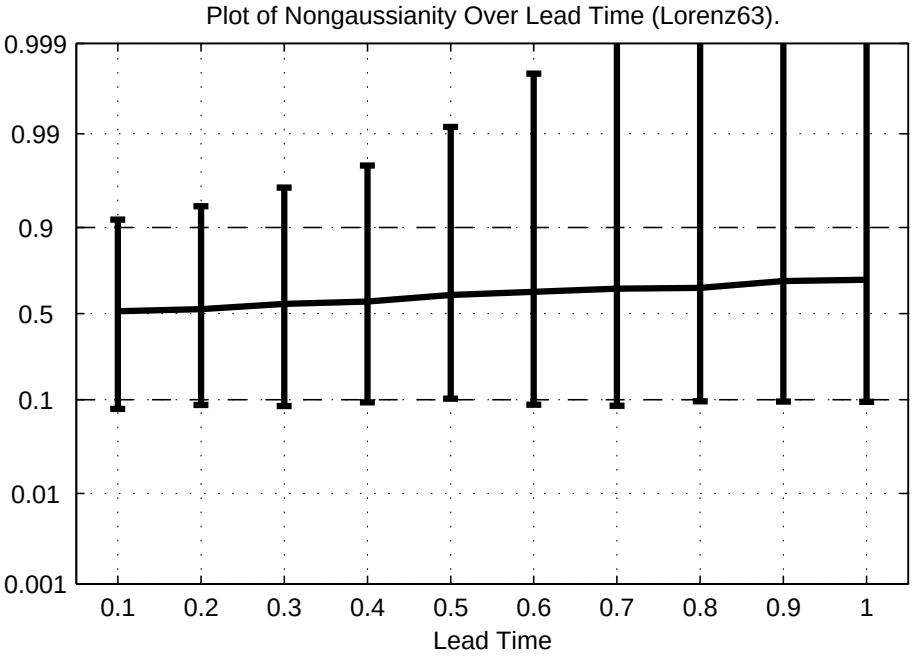

Figure 8: The $10\% - 90\%$–range of the κ–statistic for Lorenz63. The y–axis is calibrated in terms of quantiles of the κ–statistic for Gaussian ensembles and plot on $\log\left(\frac{p}{1-p}\right)$–scale.