METEOROLOGICAL APPLICATIONS
Meteorol. Appl. (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/met.92

Communicating the value of probabilistic forecasts with weather roulette

Renate Hagedorn[a]* and Leonard A. Smith[b]
[a] *ECMWF, Shinfield Park, Reading RG2 9AX, UK*
[b] *Centre for the Analysis of Time Series (CATS), London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK*

ABSTRACT: In times of ever increasing financial constraints on public weather services it is of growing importance to communicate the value of their forecasts and products. While many diagnostic tools exist to evaluate forecast systems, intuitive diagnostics for communicating the skill of probabilistic forecasts are few. When the goal is communication with a non-expert audience it can be helpful to compare performance in more everyday terms than 'bits of information'. Ideally, of course, the method of presentation will be directly related to specific skill scores with known strengths and weaknesses.

This paper introduces Weather Roulette, a conceptual framework for evaluating probabilistic predictions where skill is quantified using an *effective daily interest rate*; it is straightforward to deploy, comes with a simple storyline and importantly is comprehensible and plausible for a non-expert audience. Two variants of Weather Roulette are presented, one of which directly reflects proper local skill scores. Weather Roulette contrasts the performance of two forecasting systems, one of which may be climatology. Several examples of its application to ECMWF forecasts are discussed illustrating this new tool as useful addition to the suite of available probabilistic scoring metrics. Copyright © 2008 Royal Meteorological Society

KEY WORDS diagnostic; forecast evaluation; value; score; communication; probability forecast; proper

Received 20 September 2007; Revised 17 May 2008; Accepted 24 June 2008

1. Introduction

Have you ever wondered how much your forecast is worth, or pondered the simpler question: how well a customer betting on your forecast would perform against someone using the forecast of a competitor? In times of ever increasing financial constraints on public services, including national and international meteorological services, such questions become increasingly popular with funding agencies and more forecast providers are confronted with them (Doswell and Brooks, 1998; Rosenfeld, 2000; Freebairn and Zillman, 2002a,b). This paper considers more accessible approaches to evaluating probabilistic forecasts, directly addressing the second question above.

Though the real value of a forecast indisputably depends on the specific application (Roebber and Bosart, 1996), it is nevertheless helpful for decision-makers to obtain an intuitive, yet relevant, quantitative indication of the relative skill of probabilistic forecast systems. Traditionally, relatively simple scoring metrics such as Anomaly Correlation Coefficients (ACC) for single member forecasts or Brier Skill Scores (BSS) for probabilistic forecasts are given to users for an overall assessment of the quality of a forecast system (Wilks, 2006a). More recently, other diagnostics have been developed to address specifically the potential economic benefits of a forecast system (e.g. the cost/loss approach of Richardson, 2000 or the relative income of Roulston *et al*., 2003). Such scoring metrics are used to focus not only on the general skill of the forecasts *per se*, but also reflect various aspects of the potential economic value of the forecast system. Even if such diagnostics are well understood in the scientific community they often provide little intuitive insight for a general audience and may prove ineffective in demonstrating to customers that your forecast is worth the money spent on it. For example, a recent review of the quality (fitness for purpose) of commercial weather forecasts in the United Kingdom has highlighted significant deficiencies in the methodologies and in the communication of forecast quality assessments (Mailier *et al*., 2006). This motivates exploration of new avenues towards simple, convincing and trustworthy tools to communicate the skill of probability forecasts to a wider audience. The target group here may be decision makers in funding agencies or commercial companies who cannot always be assumed expert on probabilistic verification, but who are likely to be educated in financial matters and interested in both profit making and value for money. We acknowledge that it might be disputed whether the strengths, weaknesses and implications of the majority of probabilistic diagnostics (including skill scores) used by the scientific community are well understood by most decision makers in the scientific community itself; the implications of statistical

* Correspondence to: Renate Hagedorn, ECMWF, Shinfield Park, Reading RG2 9AX, UK. E-mail: renate.hagedorn@ecmwf.int

uncertainty in the value of the diagnostic, and in particularly the ability to quantify the difference in utility implied by the diagnostics for two different systems are common topics of discussion.

The main objective of this paper is to present a new diagnostic tool, Weather Roulette, which will help forecast providers accomplish the tasks posed above while speaking in everyday metrics, like interest rates. Developed by Roulston and Smith, it was first introduced at the ECMWF Users Seminar in 2002. The development of Weather Roulette was inspired by the attempt to solve the dilemmas inherent in (1) developing/communicating scientifically proper scoring methods while (2) providing forecast assessment which is both comprehensible and plausible for a non-expert audience and/or review bodies. Weather Roulette provides a diagnostic tool to assess the relative value of two different probability forecasts, with a metric easy to understand for a wide range of forecast users who are not necessarily familiar with common probabilistic skill scores.

The definition of the Weather Roulette, including an examination of two of the many possible variants is given in Section 2. The dataset used in this study is described in Section 3. Demonstrations of how Weather Roulette works in practice follows in Section 4. The conclusions are summarized in Section 5.

2. Definition of weather roulette

The idea behind Weather Roulette is to literally assess the claim: 'I bet I can do better than your forecast!' Assuming this person is actually putting money behind their words (and that you are confident in your own forecast), certainly one's inclination is to accept the challenge. Weather Roulette aims to answer that question when the better probability forecast is desired; it can be recast in variants that evaluate situations in which other constraints lead to a different definition of 'better' as well.

Weather Roulette is most simply cast as a game between two players, let us call them Alice and Bette. Continuous, discrete or mixed probabilistic forecasts may be considered. For concreteness, Alice's and Bette's forecast systems will be tested on their ability to forecast in which of the five climatologically equally likely categories (normal, below-normal, above-normal, well-below-normal, well-above-normal) the 2 m temperature at London Heathrow will be 10 days ahead. Assuming Bette accepts Alice's challenge, she will 'gamble' her capital in Alice's weather casino. Alice offers odds for each of the five possible outcomes based on the probabilities given by her forecast system, or more generally, if N possible outcomes exist, the odds $o(i)$ in Alice's casino are set to:

$$o_{\mathrm{A}}(i) = \frac{1}{p_{\mathrm{A}}(i)} \tag{1}$$

with $i = 1, \ldots, N$, and $p_{\mathrm{A}}(i)$ the probability of the ith outcome assigned by Alice's forecast. The sum of probabilities over all possible outcomes must be one.

$$\sum_{i=1}^{N} p_{\mathrm{A}}(i) = 1 \tag{2}$$

We consider only probabilistic odds in this paper, specifically odds normalised so that their corresponding implied probabilities sum to one. The use of non-probabilistic odds allows forecasts to incorporate the effects of model inadequacy in their probabilistic forecasts (Smith, 2007; Judd 2008a, 2008b). For concreteness, we have assumed Kelly betting strategies (Kelly, 1956), where the predicted probability determines the fraction of the stake on each outcome, for all agents. This maintains the connection with Ignorance in the "fully proper" variant defined below. It is straightforward to evaluate other strategies which may prove more relevant in other cases.

In order to quantify the skill of her probability forecast system over Alice's, Bette uses the probabilities $p_{\mathrm{B}}(i)$ of her forecast system to distribute her initial capital c_0 on the possible outcomes:

$$s_{\mathrm{B}}(i) = p_{\mathrm{B}}(i) \times c_0 \tag{3}$$

with $s_{\mathrm{B}}(i)$ being the stake set by Bette on the outcome i. Bette's capital after establishing the outcome is the product of the odds on and the stake on the verifying outcome v:

$$c_1 = o_{\mathrm{A}}(v) \times s_{\mathrm{B}}(v) = \frac{p_{\mathrm{B}}(v)}{p_{\mathrm{A}}(v)} \times c_0 \tag{4}$$

That is, Bette receives a multiple of her total invested capital, with the multiplier (the return ratio below) defined by the quotient of her probability and Alice's probability on the verifying outcome. Bette's profit or loss (hereafter profit), f, is defined as:

$$f = c_1 - c_0 = (r - 1)c_0 \tag{5}$$

and r is her return ratio (hereafter return):

$$r = \frac{c_1}{c_0} = \frac{p_{\mathrm{B}}(v)}{p_{\mathrm{A}}(v)} \tag{6}$$

For probabilistic odds, the returns are entirely determined by the probabilities assigned to the verifying bin; in all cases the probabilities assigned to the remaining categories are irrelevant. As long as Bette's probability of the verifying category is greater than Alice's, she will get a profit on her invested capital.

It is also evident that as the probability $p_{\mathrm{A}}(v)$ decreases Bette's profit increases. When Alice believed the verifying event to be extremely unlikely, she is exposed to the risk of extremely high losses. If Alice were to assign zero probability for the verifying event the pay-out would be infinite. Acting as if something unlikely was in fact impossible can lead to 'catastrophic' results in this scenario (as it can in aviation, or when crossing the street).

DOI: 10.1002/met

In the common (Monte Carlo) ensemble approach to forecasting, the model runs themselves do not provide a probability forecast: some type of ensemble interpretation must be applied. Following Roulston and Smith (2003) a 'dressing' approach is adopted both for the single-run high-resolution forecasts and for the ensemble forecasts. A number of such dressing methods exist and can be used. Wilks (2006b) provides one recent assessment of the performance of various methods, and it is noted that Weather Roulette can be used to contrast the skill of various ensemble interpretations (see also Bröcker and Smith, 2008). The focus of this paper is not on the performance of individual dressing techniques but rather on how Weather Roulette can be used to compare performance, and so we restrict ourselves to using the simple method of Gaussian ensemble dressing (Roulston and Smith, 2003), with the Gaussian dressing variance determined by minimizing the Ignorance Score in the training data used (Good, 1952; Roulston and Smith, 2002).

As the quality of a good probability forecast system cannot be adequately judged on a single forecast, Weather Roulette must be played for a number of rounds. Similar to traditional forecast verification in which forecasts are collected over certain areas and/or forecast cases, the Weather Roulette rounds can be realized through both playing for a number of successive forecasts (e.g. one whole season comprising 3 months of forecasts) and/or playing for a number of forecasts at different locations (London, Paris, Berlin,...). The skill of Bette's forecast in Alice's casino after M rounds can be quantified both by the arithmetic average of her profit f, and also by the geometric average of her return ratio r, which is defined as:

$$R = M\sqrt{\prod_{i=1}^{M} r_i} \tag{7}$$

When Weather Roulette is employed to diagnose the relative performance of two competing forecast systems, it might seem desirable for the diagnostic to give symmetrical results when exchanging the two systems in the assessment procedure. The question is: which symmetry? The symmetry may be either arithmetic (Alice and Bette's results merely change sign under exchange of forecast systems) or geometric (symmetry in returns). In terms of returns, we do not expect arithmetic symmetry, since in general

$$\frac{a}{b} \neq -\frac{b}{a} \tag{8}$$

taking a and b being placeholders for $p_{\mathrm{B}}(v_i)$ and $p_{\mathrm{A}}(v_i)$ respectively in Equation (7). Thus,

$$r_{\mathrm{B}} \neq -r_{\mathrm{A}} \tag{9}$$

Weather Roulette does, however, possess geometric symmetry (Section 2.1) and a correspondence with the information theoretical skill score called 'Ignorance' (Good, 1952; Roulston and Smith, 2003). Hence it will reflect the fact that that score is proper (Bröcker and Smith, 2007), and can be used for optimizing probability forecasts, as well as communicating their skill.

An alternative variant of Weather Roulette can be defined so as to obtain arithmetic symmetry (Section 2.2). More generally, Weather Roulette statistics can be computed in a number of ways, and the most appropriate choice may vary with the action to be taken. In different situations, the nature of the bets may change (each player may be fully invested on each round, or there may be a fixed stake placed on each round, or...), risk scenarios may vary (players may or may not follow a Kelly betting strategy (Kelly, 1956), which maximizes the expected growth rate), specific profit targets may exist, and so on. It is important to distinguish methods maximizing utility in a particular setting from scores used to tune a probabilistic forecast system. When maximizing utility in a specific scenario the variant selected needs to be relevant to the situation for which decision support is required and may in fact not correspond to a proper skill score if the aim is not an optimized probability forecast, particularly against an opponent with known weaknesses. When tuning a probability forecast system, scores need to be proper.

2.1. Fully proper

The first variant of the Weather Roulette considers the players as fully invested in each round and can be communicated as an effective interest rate. In this case, Weather Roulette starts with an arbitrary initial investment c_0 (e.g. 1 Euro), and after every round of betting the total wealth (both profit and initial capital) is fully re-invested. That is, under this varient, the capital c invested in every round is equal to the total Bette received in the previous round (as defined in Equation (4)); thus it is not constant but depends on the history:

$$\begin{aligned} c_t &= r_t \times c_{t-1} = r_t \times r_{t-1} \times c_{t-2} \\ &= r_t \times r_{t-1} \times \cdots \times r_1 \times c_0 \\ &= R^t \times c_0 \end{aligned} \tag{10}$$

Equation (10) shows that the capital at time t is equal to the product of the return ratio on each round times the initial capital. Considering the logarithm of R leads to the desired geometric symmetry since:

$$\log\left(\frac{a}{b}\right) = -\log\left(\frac{b}{a}\right) \tag{11}$$

demonstrating the symmetry between A and B's returns. Logarithms arise naturally, as we are multiplying the original stake by a number on each round; when logarithms are taken base 2 they can be immediately interpreted as bits of information. To facilitate a more intuitive interpretation, we evaluate the Weather Roulette diagnostic under the fully-proper variant as the *effective daily interest rate*, Д, achieved over M rounds. Specifically:

$$\text{Д} = R - 1 \tag{12}$$

DOI: 10.1002/met

One advantage of this variant of Weather Roulette is that the resulting Weather Roulette diagnostic, the effective interest rate, is a simple transformation of the Ignorance Score (Good, 1952; Roulston and Smith, 2002) which is a proper score (see Bröcker and Smith, 2007; Gneiting and Raftery, 2007). To be informative, of course, M must be sufficiently large for the result to be robust; we evaluate the robustness of a given result by bootstrap resampling, as discussed below. As outlined in more detail in Roulston and Smith (2002), the Ignorance Score is expressed in bits and can also be interpreted as wealth doubling rate, i.e. the number of bets expected before the gambler's initial wealth doubles. Since interest rates are more common than bits of information or doubling times, we express Ignorance as an effective interest rate; we believe this is a value more easily to understand and to relate to by the general public, but this simple numerical transformation makes the results more intuitive and does not affect important properties of the score.

2.2. Two-House Gaming

An alternative variant – 'Two-House Gaming' – achieves arithmetic symmetry under the condition of constant stakes. In this scenario, the Weather Roulette diagnostic is calculated not only from the profit Bette makes in Alice's casino, but also depends on the profit Alice would make in Bette's casino. Given the same (fixed) daily stake, the corresponding Weather Roulette diagnostic:

$$\Pi = \frac{1}{N}\sum_{i=1}^{N}(r_{\mathrm{B}} - r_{\mathrm{A}})_i = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{p_{\mathrm{B}}(v)}{p_{\mathrm{A}}(v)} - \frac{p_{\mathrm{A}}(v)}{p_{\mathrm{B}}(v)}\right)_i \tag{13}$$

is independent of the stake itself, Π has arithmetic symmetric under exchange of the forecast systems. When a forecast is used in a regulated system, it may prove rational to provide decision support which exploits known weaknesses of a competing system. Under variant one (fully proper), an optimal forecast will never appear to be beaten in the long run. Under variant two (two-house gaming) the optimal probability forecast may appear to perform worse against a third party than an alternative forecast which is 'sub-optimal' as a probability forecast *per se*. A simple illustration may help here: Carl's probability forecast system is optimal as it produces the probabilities with respect to which the verification is chosen, while Dave's probability forecast system is sub-optimal. If both play Alice within the two-house variant, Dave may indeed generate a greater profit (from Alice) in the long-run. Under the fully-proper variant, Carl's optimal forecast is always superior.

While both variants of Weather Roulette may sometimes lead to similar results, they need not. Depending on the task for which decision support is required, there may be cases where the desired action is not to produce the ideal probability forecast but, for example, to outperform an opponent with particular weaknesses. If, however, the goal is to evaluate a probability forecast *per se*, as it is for example when constructing a particular forecast system, one should consider only variants that correspond to proper scores, such as the fully-proper variant above. After giving one example comparing the results of both Weather Roulette variants, the 'fully proper' variant of Weather Roulette is chosen for the remaining examples.

3. Data

Forecasts from both the high-resolution model (HRES) and ensemble prediction system (EPS) of the European Centre for Medium-Range Weather Forecasts (ECMWF) are used in this study. The dataset comprises daily operational forecasts from 1 March 2001 until 28 February 2007, that is 6 years of forecasts for all four seasons March/April/May (MAM), June/July/August (JJA), September/October/November (SON), and December/January/February (DJF). The first 5 years are used as training data for the dressing procedure, and the last year is exclusively used to evaluate the models out-of-sample. The forecast start time is 1200 UTC and all lead times from 24, 48,. . ., 240 h, i.e. 1-day to 10-day forecasts are studied.

The demonstration of Weather Roulette is performed for 100 weather stations mainly located in Europe (Figure 1). Note that, obviously, these results at these stations on a given forecast need not be independent, indeed they are expected to display striking interdependence at long lead times. Following standard procedures, the global forecast fields are interpolated to these locations using a 12-point interpolation scheme (White, 2003), and synoptic observations from the World Meteorological Organization (WMO) Global Telecommunication System (GTS) are used as verification. Mainly forecasts of 2 m temperature are studied, but a comparison of the results for other parameters like Mean Sea Level Pressure (MSLP) and 10 m wind speed (10 FF) is performed as well.

All datasets used (forecasts and observations) are converted to anomalies with respect to the observed climate at the respective station. The daily climate at the station locations is calculated from observations of the 25 year period 1982–2006 and a $\pm$20-day window around the target date. Selecting the size of this window is another potential application for Weather Roulette.

4. Applications

To demonstrate Weather Roulette, this section gives a simple example of how it can work in practice. This is followed by further examples of investigations that can be performed with Weather Roulette diagnostics. The aim here is not to optimize the forecast system but to illustrate the manner in which weather roulette can be used to both optimize and quantify the utility of such optimization. Additional results using Weather Roulette to optimize and

DOI: 10.1002/met

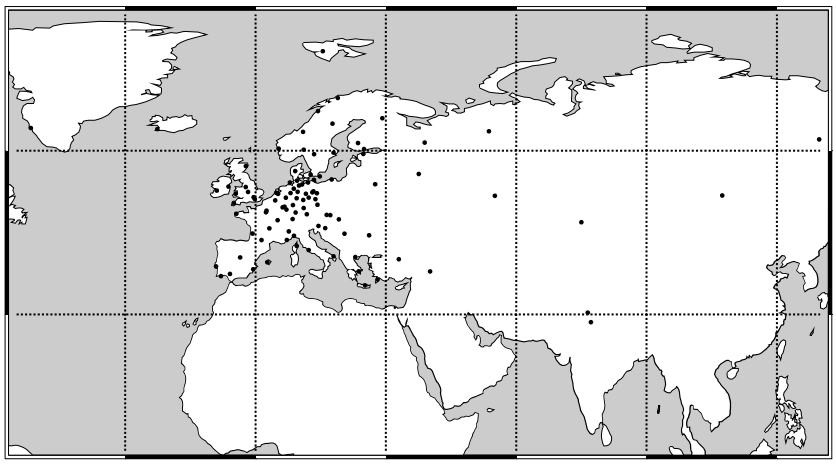

Figure 1. Location of 100 stations used for Weather Roulette application.

evaluate, and incorporating some of the suggestions made below, will be presented elsewhere.

4.1. The basic concept

Imagine the funding agency of your weather forecast centre states that your forecast is not worth the money spent on it, and just using climatology as forecast would be as good as (or even more cost effective) than funding the rather expensive production of your forecast. In such a situation, offering to play Weather Roulette with the funding agency might be of help to convince them otherwise or even could help you to fund your activities.

Consider Weather Roulette with forecasts targeting which of five possible categories (well-below normal, below normal, normal, above normal, well-above normal) the 2 m temperature at London-Heathrow will fall, evaluating the systems over a 4 week period (28 rounds) and predicting with 3 days lead every day of the period 11 May until 7 June 2006 (Figure 2(a)).

In the 'Two-House Gaming' variant, both parties distribute every day a fixed amount, for example $c = 1000$ Euros, according to their own forecasts on the five possible categories. The funding agency, betting on climatology, assigns equal probabilities of $p_{\mathrm{A}} = 0.2$ to every possible category, thus they distribute an equal amount of 200 Euros on each category on each day. In contrast to that, you distribute your 1000 Euros proportionally to the predicted probability of each bin according to your forecast, in our case according to the dressed ECMWF EPS forecast probabilities p_{EPS}. Comparing the probabilities assigned to the verifying category by the climatological forecast (Figure 2(b) open diamonds) and the dressed EPS (Figure 2(b) filled squares) gives a first hint of the quality of the two forecast systems; note that in by far the majority of cases, the probability the EPS assigns to the verifying category is significantly greater than its (constant) climatological probability The monetary reward for playing the EPS forecast against climatology is shown in Figure 2(c). This accumulated profit, expressed in units of the daily stake, increases for the EPS player as long as the EPS probability of the verifying bin is above the climatological probability. At the end of the 4 weeks you have an accumulated profit of more than 30 times the daily stake, i.e. the total pay-out the agency's casino made to you was more than 30 000 Euros. On average over the whole period you got more than 1000 Euros every day, that is more than 100% return on the daily investment at stake! Under the 'fully proper' variant, an initial investment on day 1 of just a 1 Euro yields, accumulated capital after 28 days of 729 416 Euros (Figure 2(d)) which corresponds to an effective daily interest rate of $\sim$62% per day. (That is, 1.62^{28} is about 730 000.) The profits achieved under these two different variants are not easily compared as they reflect very different scenarios of compounded interest and a sum of daily profit.

Weather Roulette allows the comparison of the temporal development of the achieved return, and Figure 2(c) and (d) convey the same message. Namely that – probably not surprisingly – even under the simple dressing scheme used here the 3-day forecast is vastly superior to betting on climatology. As mentioned above, the choice of which variant to apply depends on the individual decision that is to be informed; for the remaining examples we consider only the 'fully proper' variant.

Taking into account the decreasing skill of weather forecasts with increasing lead time, we expect a lower return with longer lead time forecasts. Indeed the results of Weather Roulette over the same period and at the same location as above but with the 10-day forecast, reveal the reduced value of the forecasts compared to climatology (Figure 3). At longer lead times the probabilities of the verifying category are similar for both climatology and EPS forecasts, and more often p_{EPS} is even lower than p_{HRES}, i.e. the EPS player has a loss on some days (about 50% of the 10-day forecasts show a loss for the EPS). Due to the quality of good forecasts, however, the effective daily interest rate reveals the EPS to have greater skill: overall return is positive.

Next, Weather Roulette is demonstrated in a somewhat more interesting case: that of comparing the probabilistic forecast performance of two competing forecast systems

DOI: 10.1002/met

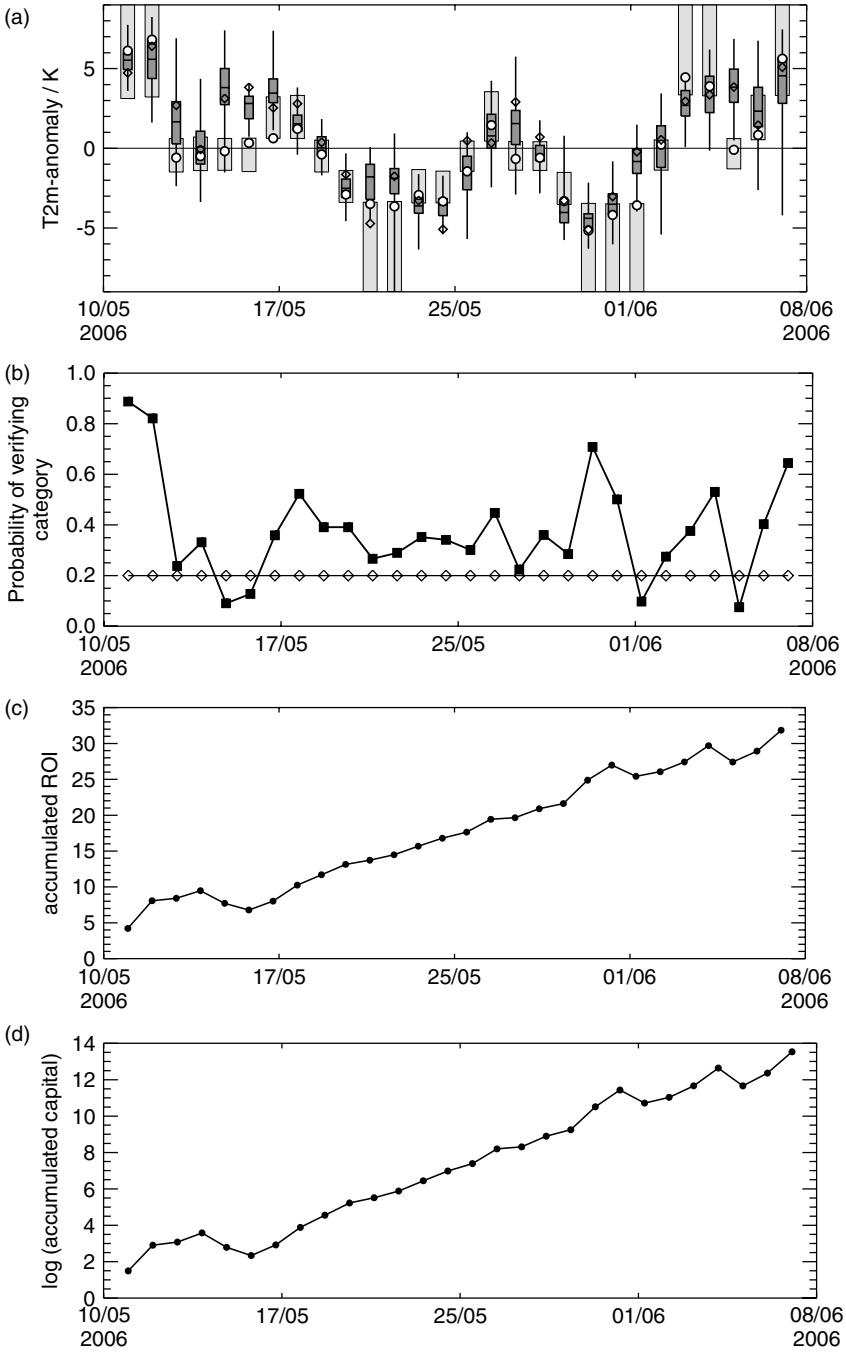

Figure 2. Illustration of the results of Weather Roulette for predicting with a 3-day lead time in which quintile-category the 2 m temperatures at London-Heathrow will fall during the period 11 May–7 June 2006. (a) 3-day EPS forecasts are symbolized with dark-grey box-and-whisker symbols, 3-day HRES forecasts are denoted by open diamonds, and the corresponding observations are depicted by open circles, with the verifying categories marked in addition as light-grey box; (b) probabilities of the verifying categories as predicted by climatology (open diamonds) and the dressed EPS forecasts (filled squares); (c) resulting accumulated profit (in units of daily stake) when playing the dressed EPS forecasts against climatology under the 'Two-House Gaming' variants. Positive values indicate winnings (e.g. the value 2 corresponding to an additional return of twice the daily investment on top of the investment.) A value of 0 indicates neutral return, i.e. no gain and no loss; (d) logarithm of the return on investment when playing the dressed EPS forecasts against climatology under the 'fully proper' variants.

each more skillful than climatology. Imagine a weather forecast centre runs more than one forecast system and is asked to decide which of the two is more valuable, if due to budget constraints one of the systems has to be abandoned. In the real world such important decisions will also depend on other factors as for example the cost of running the forecast system. Nevertheless we believe Weather Roulette diagnostics provide intuitive insight; an effective daily interest rate may be considered more intuitive than bits (or nats) of information. In this case the 'fully proper' game as above is played, but now replacing climatology with a second forecast system. In this case, the ECMWF high-resolution forecasts (HRES) are used for this. In Figure 4 Weather Roulette highlights that using these dressing schemes at short lead times, such as 3 days, the dressed HRES forecasts are quite

DOI: 10.1002/met

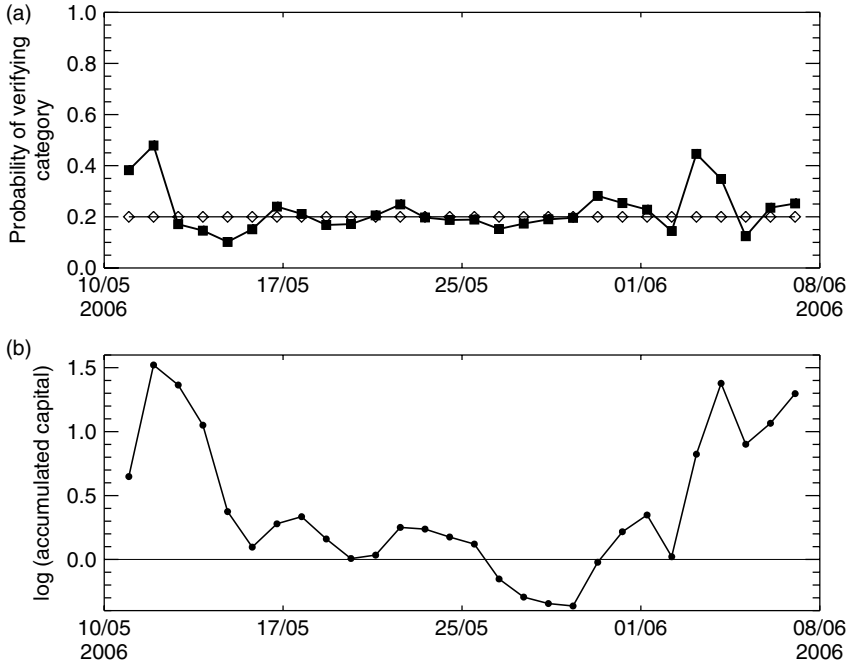

Figure 3. Illustration of the results of Weather Roulette for predicting with a 10-day lead time in which quintile-category the 2 m temperatures at London-Heathrow will fall during the period 11 May–7 June 2006. (a) Probabilities of the verifying categories as predicted by climatology (open diamonds) and the dressed EPS forecasts (filled squares); (b) logarithm of the accumulated capital when playing the dressed EPS forecasts against climatology under the 'fully proper' variant.

competitive to the dressed EPS forecasts, while for longer lead times such as the 10-day forecasts, the EPS comfortably achieves clear winnings (Figure 5). This result suggests that the higher resolution forecast can be particularly valuable in the early forecast range, whereas in the later forecast range the flow-dependent uncertainty information of the EPS easily outperforms the high-resolution advantage. Forecasts, and results of actions based on them, will depend on the dressing scheme used to interpret the ensemble of simulations. Of course, a reliable statement on the overall value of one forecast system compared to another certainly cannot be based on just one forecast location and a short period of 4 weeks of forecasts. One can begin to test whether or not this result is robust by examining the results for a larger number of stations and forecast cases have been aggregated.

4.2. Aggregation of results

The effective daily interest rate provides summary assessment of the skill of a forecast system over a certain period of time; the value will vary with the lead time of the forecasts. Hence, the д Weather Roulette diagnostic, giving the effective daily interest rate under the 'fully proper' variant, is displayed as a function of lead time in Figure 6. The average results for the 100 considered locations are bootstrapped in time. That is, the Weather Roulette diagnostic д (Equation (12)) is recalculated 1000 times, each time consisting of M randomly chosen dates (or rounds) of the period under consideration. The box-and-whisker symbols mark the 1, 25, 50, 75, and 99 percentiles of the bootstrap resampled results.

The final results for the months MAM 2006 with dressed EPS forecasts *versus* climatology show in Figure 6(a) illustrates how Weather Roulette clearly reveals that the EPS is more valuable than climatology throughout the whole forecast range; an effective daily interest rate near 90% in the short-range stresses that this really is a significant (large) advantage. Using the dressed high-resolution forecasts instead (Figure 6(b)) Weather Roulette reveals similar returns for the very early forecast range, but from day-8 forecasts onwards the high-resolution forecasts are unable to beat climatology. The direct comparison of Gaussian dressed EPS *versus* HRES forecasts (Figure 6(c)) demonstrates more clearly the similar performance of HRES and EPS forecasts in the early forecast range but also the increasing value of the EPS forecasts with longer lead times. These results suggest that, if you had the choice to use the EPS or the HRES forecasts in a Weather Roulette casino, you might choose to set your bets according to the HRES forecast for the early forecast ranges but for the longer lead times you should definitely use the Gaussian dressed EPS. Both forecast systems, however, seem to add value beyond climatology, one might want to consider whether it is possible to optimize the return when combining both forecast systems; any decision to do this would, of course, depend on their relative cost. Alternatively, the dressing algorithm used above treats the ensemble control run and perturbations as exchangeable, whereas one expects that, in the short range at least, they are not: the EPS might significantly at perform the HRES even in the short run using a dressing scheme that exploited this fact, and treated

DOI: 10.1002/met

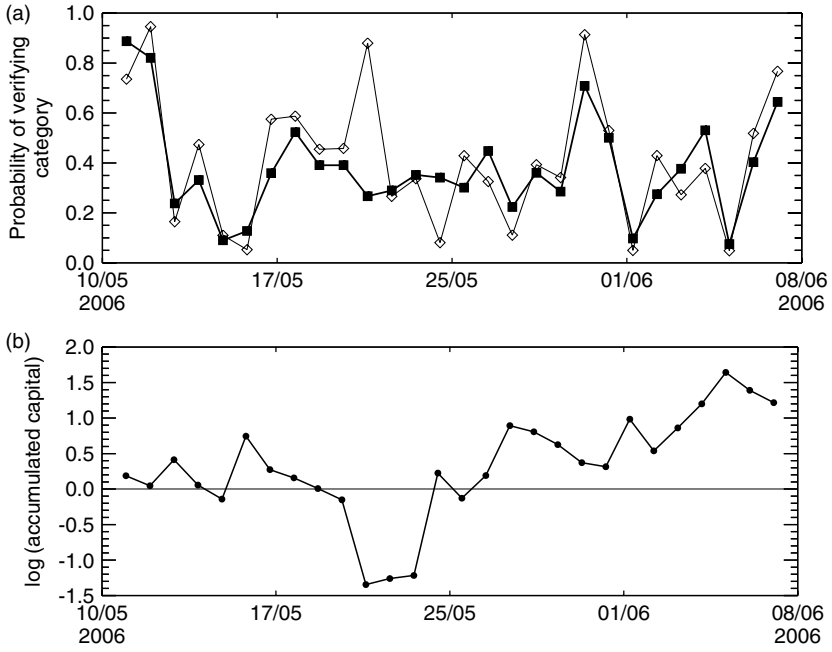

Figure 4. Illustration of the results of Weather Roulette for predicting with a 3-day lead time in which quintile-category the 2 m temperatures at London-Heathrow will fall during the period 11 May–7 June 2006. (a) Probabilities of the verifying categories as predicted by the dressed HRES forecast (open diamonds) and the dressed EPS forecasts (filled squares); (b) logarithm of the accumulated capital when playing the dressed EPS against the HRES forecast under the 'fully proper' variant.

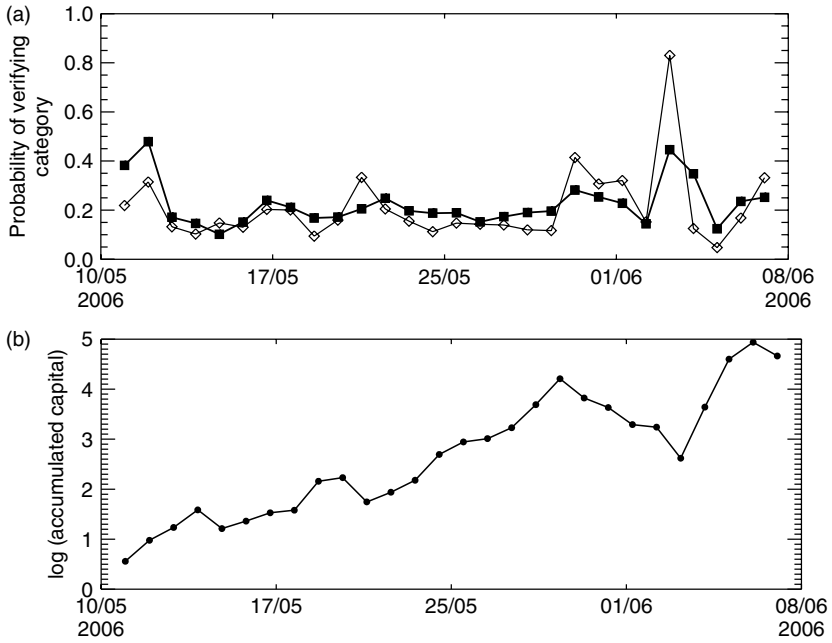

Figure 5. Illustration of the results of Weather Roulette for predicting with a 10-day lead time in which quintile-category the 2 m temperatures at London-Heathrow will fall during the period 11 May–7 June 2006. (a) Probabilities of the verifying categories as predicted by the dressed HRES forecast (open diamonds) and the dressed EPS forecasts (filled squares); (b) logarithm of the accumulated capital when playing the dressed EPS against the HRES forecast under the 'fully proper' variant.

the control member differently? Again, Weather Roulette provides the framework to evaluate such conjectures.

4.3. Value of combined forecasts

Finding optimal weights for the different forecasts systems is a significant task when designing combined forecast systems. While there is no operationally relevant theory for an 'optimal' approach, a number of methods exist to determine useful weights, examples include Bayesian Model Averaging (Raftery *et al*., 2005), using an analytical equation to maximize the Brier Skill Score (Rodwell, 2005), and variations on kernel dressing (Roulston and Smith, 2002, 2003; Bröcker and Smith, 2008). Model weights used in the current paper are calculated through minimizing the Ignorance score in the training

DOI: 10.1002/met

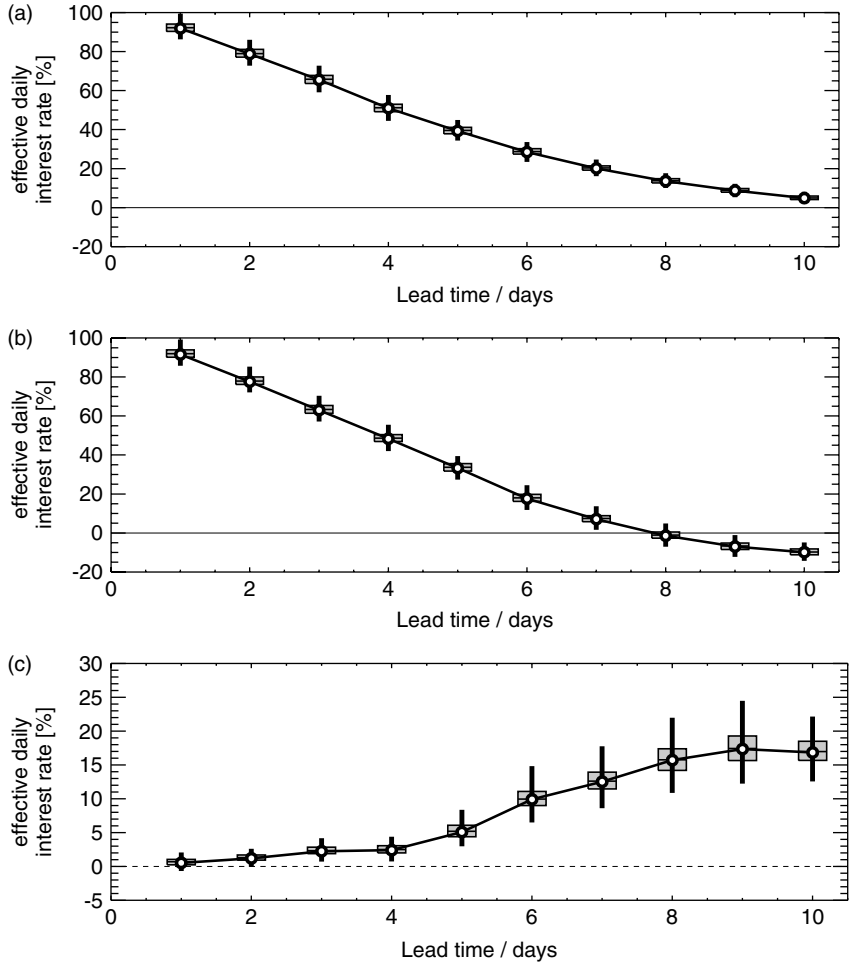

Figure 6. Contrasting two forecasts with effective daily interest rate as a function of lead time. Weather Roulette diagnostic Д aggregated over 100 locations and for the 3 month period of March–May 2006, shown for lead times from 1-day to 10-day forecasts. (a) Dressed EPS playing against climatology, (b) dressed HRES playing against climatology, and (c) dressed EPS playing against dressed HRES forecasts (please note changed scale of y-axis). Box-and-whisker symbols indicate 1, 25, 50, 75, and 99 percentiles of time-bootstrapped results.

period. The return achieved when combining the Gaussian dressed EPS and HRES forecasts and playing against the Gaussian dressed EPS only forecasts, is shown in Figure 7(a). It is obvious that – in particular for the early forecast ranges – the combined system yields to a better return, i.e. is more valuable than the Gaussian dressed EPS probabilities on its own. As expected – considering the results from Figure 6 – both EPS and HRES are assigned similar weights in the early forecast range, but with constantly increasing weight for the EPS forecasts with increasing lead times (Figure 7(b)).

Weather Roulette can be used to illustrate, as in Figure 8, that blending climatological information into the combined forecast system leads to an increased effective interest rates at longer lead times (Bröcker and Smith, 2008). The blending weights for the combined prediction system, consisting of EPS and HRES forecasts as well as climatological information, show that climatological information is hardly used in the early forecast range, but that for longer forecast ranges the climatological component can have significant weight (Figure 8(b)).

4.4. Diagnostic of different seasons and parameters

Weather Roulette was used above to consider forecasts during one season (MAM 2006) and it is advisable to consider a broader assessment including other seasons and/or target forecast variables. Contrasting the skill of the dressed EPS forecasts against that of the dressed high-resolution forecasts for all four seasons (Figure 9(a)–(d)), Weather Roulette reveals variability in the timing of the cross-over point from HRES to EPS superiority. Though the general feature that the value of EPS forecasts increases with time is valid for all seasons, it is also apparent that in autumn 2006 the high-resolution forecasts are much more valuable (in particular in the first few forecast days) compared to the other seasons, when playing the EPS forecasts lead to positive returns already very early in the forecast range. Of course, we expect predictability to vary with the state of the flow (Wilks, 2006a; Ziehmann *et al*., 1999), and this result might reflect the fact that the dressed EPS forecasts contain flow-dependent uncertainty information which is not available in the dressed HRES forecasts and might

DOI: 10.1002/met

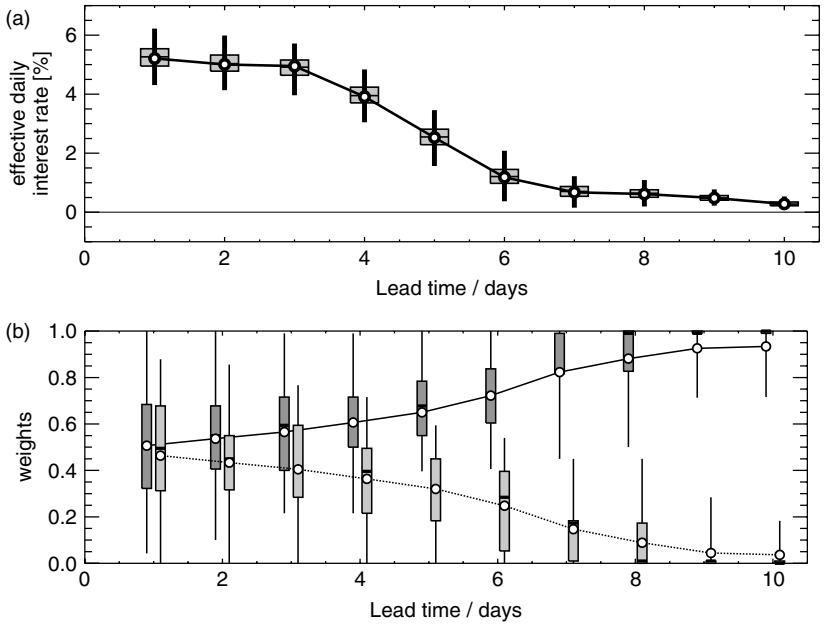

Figure 7. (a) Weather Roulette diagnostic Д when evaluating the return of a combined prediction system consisting of the dressed EPS and HRES forecasts playing against the dressed EPS only forecasts. The diagnostics are aggregated over 100 locations and for the 3 month period of March–May 2006, shown for lead times from 1-day to 10-day forecasts. (b) Weights given to the dressed EPS (dark-grey, left box-and-whisker symbols, solid line) and HRES forecasts (right, grey box-and-whisker symbols, dotted line).

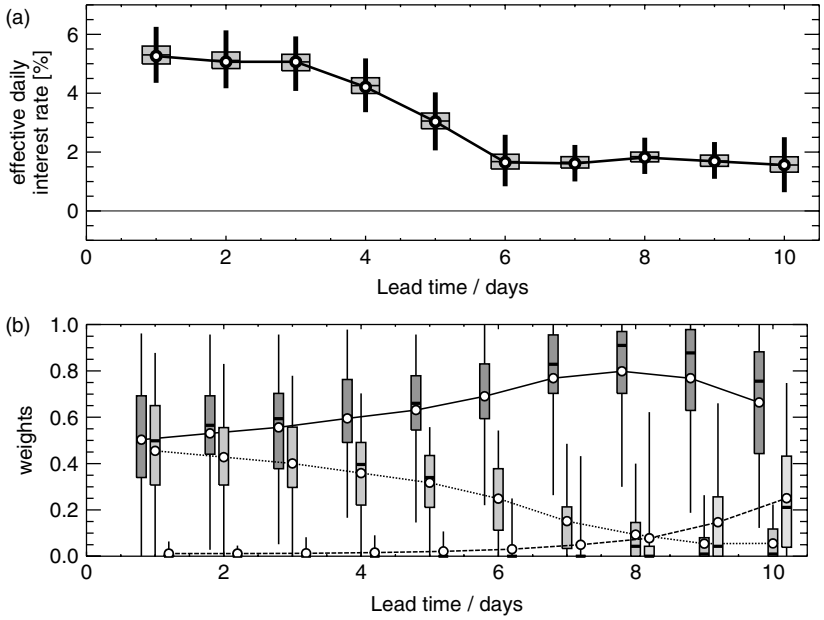

Figure 8. (a) Weather Roulette diagnostic Д when evaluating the return of a combined prediction system consisting of the dressed EPS and HRES forecasts plus climatological information, playing against the dressed EPS only forecasts. The diagnostics are aggregated over 100 locations and for the 3 month period of March–May 2006, shown for lead times from 1-day to 10-day forecasts. (b) Weights given to the dressed EPS (dark-grey, left box-and-whisker symbols, solid line) and HRES forecasts (right, grey box-and-whisker symbols, dotted line). The weights given to the climatological information are marked by the third, light-grey box-and-whisker symbols (dashed line).

have been of particular value in all seasons except this particular autumn. Alternatively it may reflect variations between model versions or ensemble formation systems deployed during this period. Examining day to day performance, this seems to be a period when the EPS was performing unusually poorly.

In order to assess how the findings above depend on the target variable forecast, the Weather Roulette diagnostic is also computed for predictions of mean sea level pressure and 10 m wind speed, and then compared to the 2 m temperature results (Figure 10). For the case of MSLP predictions the cross-over point between HRES and EPS forecasts is moved to longer lead times, that is the dressed EPS forecasts achieve a positive return only from day 8 onwards. On the other hand, for 10 m wind speed predictions EPS forecasts have a positive return over the whole forecast period. A possible explanation for the differences in the results depending on the

DOI: 10.1002/met

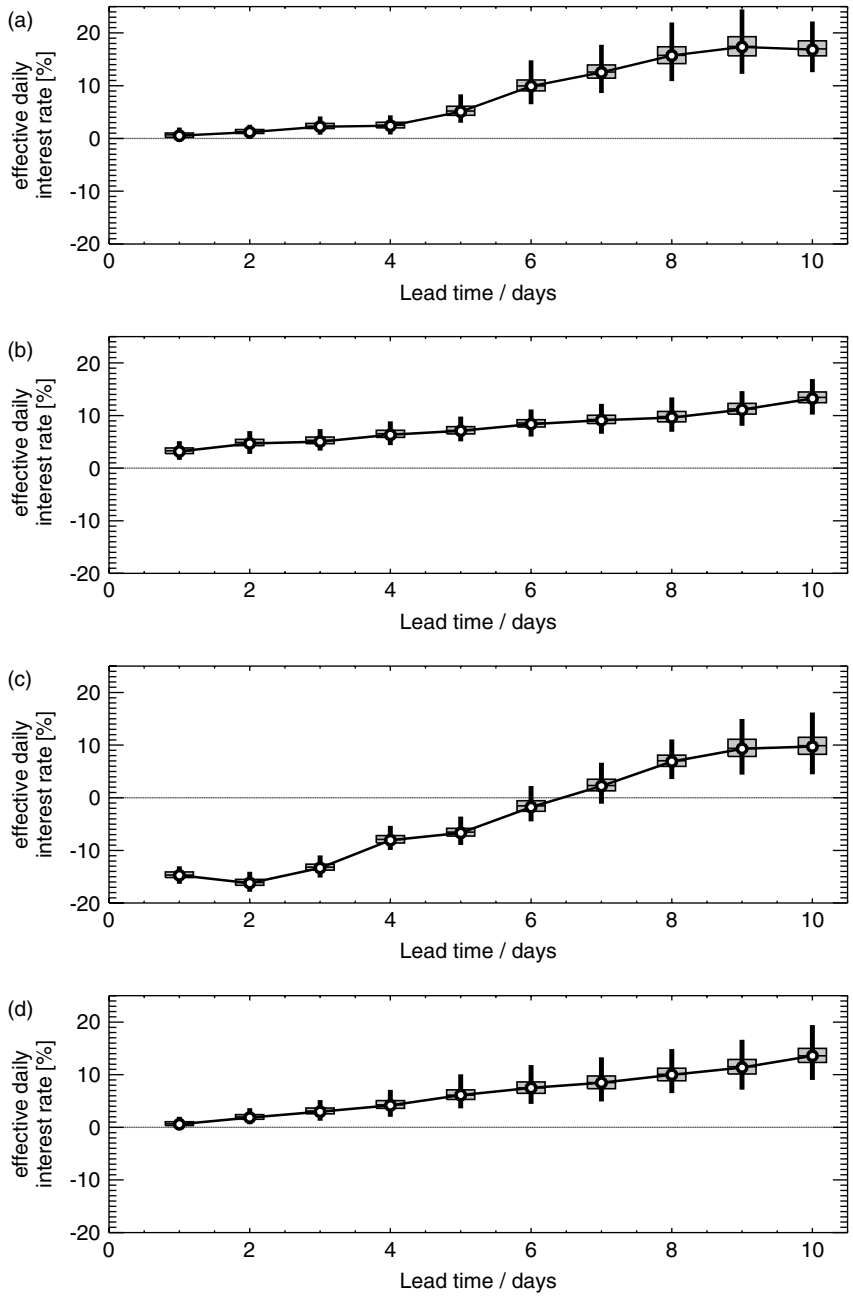

Figure 9. Weather Roulette diagnostic Д aggregated over 100 locations, as in Figure 6(c), but here shown for all four seasons. (a) MAM 2006, (b) JJA 2006, (c) SON 2006, and (d) DJF 2006/07.

forecast parameter considered lies in the fact that the Weather Roulette diagnostic reflects the potential cost of extremely bad forecasts probability forecasts (as when a low probability event is forecast as impossible, know as a violation of Cromwell's Rule). Hence, for relatively well predicted target like MSLP, neither the dressed HRES forecast system nor the dressed EPS is likely to predict as very low probability for the verifying bin; neither can then expect spectacular returns no matter how good a probability forecast it may be; nevertheless, in the long run, a better probability forecast will be identified as such. In such cases, one might consider increasing the number of bins (or evaluating continuous densities), yet if only a small number of bins is relevant to the decision maker, then it just may be the case that the two approaches are indistinguishable when the forecast-verification archive is small. On the other hand, in the case of a target which is generally forecast poorly such as 10 m wind speed, Weather Roulette allows reliable uncertainty information contained in the dressed EPS forecasts to stand out clearly; the EPS outperforms the probability forecasts from the dressed high-resolution forecasts significantly. Note that Weather Roulette evaluates the probability forecast system as a whole, so that it can also be used to compare different dressing methods, which may prove more appropriate to particular forecast targets.

DOI: 10.1002/met

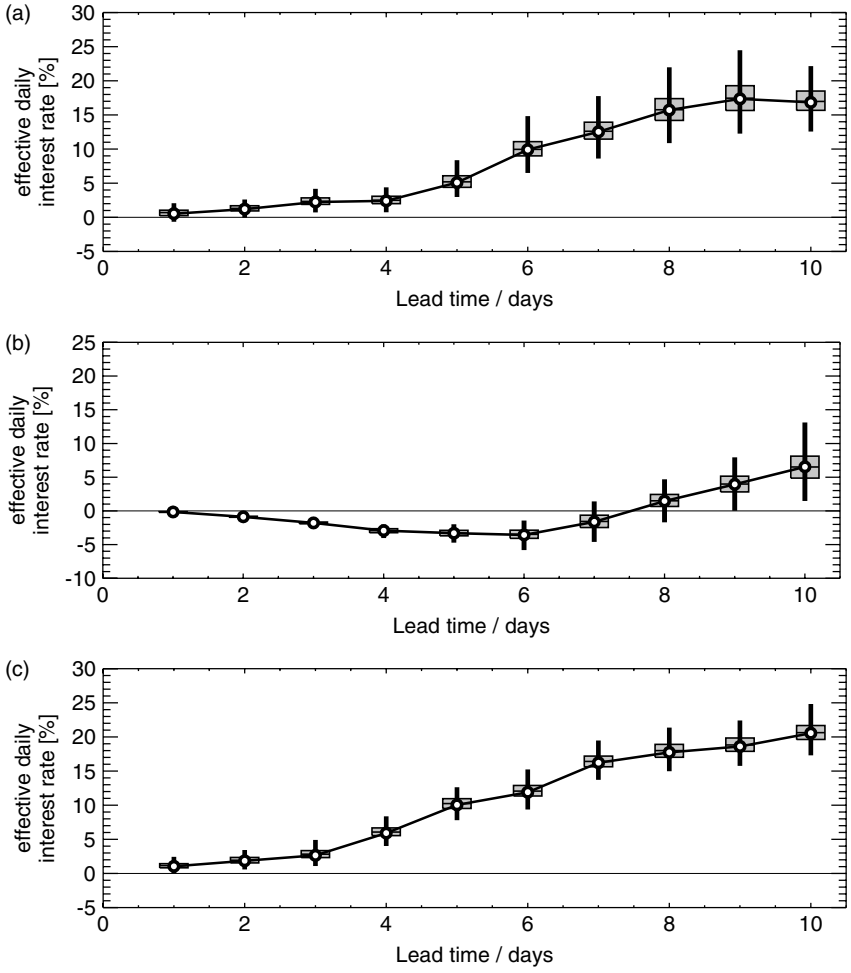

Figure 10. Weather Roulette diagnostic Д aggregated over 100 locations, as in Figure 6(c), but here shown for different parameters forecasted. (a) 2 m temperature, (b) mean sea level pressure, and (c) 10 m wind speed.

5. Summary

A new approach for communicating the performance and relative value of probabilistic forecast systems has been introduced. Key advantages of Weather Roulette are (i) that it provides a simple and easy to understand framework for evaluating probabilistic forecasts, (ii) that the evaluation can be expressed as an effective interest rate, which is more intuitive than some other methods, (iii) that the evaluation is related to a proper score when desired and (iv) that it allows one to assess the actual (monetary) value of a forecast system when it is used in a simplified but realizable form of decision support. This last fact could prove useful in demonstrating the value of a forecast system to an audience not familiar with probabilistic diagnostic tools. The 'fully proper' variant used in most examples in this paper, is connected to the Ignorance score, (which is the only proper local score for continuous probability forecasts). Thus probability forecast systems can be safely optimized using this variant. The 'two-house' variant does not have this connection and so should not be used for tuning probability forecast systems, but might prove relevant to evaluating schemes aiming to beat a particular opponent rather than evaluate a probability forecast system as such.

Weather Roulette diagnostics can be used to evaluate the performance of different forecast systems and/or different dressing techniques and/or the combination of prediction systems or even defining a better climatological distribution. Particular aspects of the quality of a forecast system for different parameters, seasons and locations are easily explored. Examples of such studies are given above, and it is hoped that future diagnostic work on evaluating probabilistic forecast systems will find Weather Roulette of use. In fact, it is now planned to incorporate Weather Roulette diagnostics routinely as a performance measure for ECMWF operational forecasts. In addition to the evaluation for local station data, the diagnostic will be extended to also include the evaluation of global and/or regional fields, and will be used to communicate the relative skill of probability forecasts, hopefully broadening the justification for their production and use. A direct comparison of operational probability forecasts from different operational centres at a number of specific sites would be of great interest. It is hoped that Weather Roulette finds widespread application in easing the communication of skill and the fundamental value of probabilistic forecasts.

DOI: 10.1002/met

Acknowledgements

This work was supported by the NERC NAPSTER project. The authors are grateful for the ideas and insights of M Roulston, J Bröcker, L Clarke and K Judd, as well as the reviewers.

References

Bröcker J, Smith LA. 2007. Scoring probabilistic forecasts: on the importance of being proper. *Weather and Forecasting* **22**: 382–388.

Bröcker J, Smith LA. 2008. From ensemble forecasts to predictive distribution functions. *Tellus* **60**(4): 663–678.

Doswell CA, Brooks HE. 1998. Budget-cutting and the value of weather forecasting. *Weather and Forecasting* **13**: 206–212.

Freebairn JW, Zillman JW. 2002a. Economic benefits of meteorological services. *Meteorological Applications* **9**: 33–44.

Freebairn JW, Zillman JW. 2002b. Funding meteorological services. *Meteorological Applications* **9**: 45–54.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**(477): 359–378.

Good IJ. 1952. Rational decisions. *Journal of the Royal Statistical Society* **14**: 107–114.

Judd K. 2008a. Forecasting with imperfect models. *Physica D* **237**(2): 216–232.

Judd K. 2008b. Non-probabilistic odds and forecasting with imperfect models. *Physica D* submitted.

Kelly J. 1956. A new interpretation of information rate. *Bell Systems Technical Journal* **35**: 916–926.

Mailier PJ, Jolliffe IT, Stephenson DB. 2006. Quality of weather forecasts. Online publication of the *Royal Meteorological Society*, available at: http://www.subscriptions.rmets.org/pdf/fqp_report.pdf.

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* **133**: 1155–1174.

Richardson DS. 2000. Skill and economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society* **126**: 649–668.

Rodwell MJ. 2005. Comparing and combining deterministic and ensemble forecasts: how to predict rainfall occurrence better. *ECMWF Newsletter* **106**: 17–23.

Roebber PJ, Bosart LF. 1996. The complex relationship between forecast skill and forecast value: a real-world comparison. *Weather and Forecasting* **11**: 544–559.

Rosenfeld J. 2000. Do we need the national weather service? *Scientific American Presents: Weather* **11**(1): 28–31.

Roulston MS, Kaplan DT, Hardenberg J, Smith LA. 2003. Using medium-range weather forecasts to improve the value of wind energy production. *Renewable Energy* **28**(4): 585–602.

Roulston MS, Smith LA. 2002. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review* **130**: 1653–1660.

Roulston MS, Smith LA. 2003. Combining dynamical and statistical ensembles. *Tellus Series A-Dynamic Meteorology and Oceanography* **55**: 16–30.

Smith LA. 2007. *Chaos, A Very Short Introduction*. Oxford University Press: Oxford; 180.

White PW (ed.). 2003. IFS documentation, Part VI: Technical and Computational Procedures (CY25R1), Available online: http://www.ecmwf.int/research/ifsdocs/CY25r1/pdf_files/Technical.pdf.

Wilks D. 2006a. *Statistical Methods in the Atmospheric Sciences*, 2nd edn. Academic Press: New York, 2nd edition, 627, see chapter 7 (255–332).

Wilks D. 2006b. Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications* **13**: 243–256.

Ziehmann C, Smith LA, Kuths J. 2000. Prediction of probability. *Phys. Lett. A.* **271**(4): 237–251.

DOI: 10.1002/met