Seeing Through Ensembles

Bröcker, Jochen
Max–Planck–Institut für Physik komplexer Systeme
Nöthnitzer Str. 38, 01187 Dresden, Germany
E-mail: broecker@pks.mpg.de

Smith, Leonard A.
London School of Economics, Centre for the Analysis of Time Series
Houghton Street, London WC2A 2AE, United Kingdom
E-mail: l.smith@lse.ac.uk

Abstract

Probability forecasts are becoming more and more common, as users facing decision problems under uncertainty are increasingly aware of the fact that probabilistic forecasts provide much more information than "deterministic" (point) forecasts. Studying probability forecasts is likewise appealing to mathematicians, as the calculus of probability offers an answer to every possible question a forecast user might have—at least in principle. The reality of ensemble forecasting however, especially in the weather and climate sector, looks a little bit less rosy. The epistemological challenges the reality of ensembles forecasting puts onto a mathematically minded person are considered. The procedures by which current operational centres produce ensemble weather forecasts are revisited, which exhibit a number of deficiencies, some being due to limited resources and some being due to inherent theoretical limitations of the problem. We are thus lead to contest the view that the resulting distributions should be considered probabilities, either in a frequentist's sense or a subjectivist's sense, as these forecast distributions neither coincide with observed frequencies nor could possibly represent any (reasonable) forecaster's opinion on future weather events. Nonetheless, ensembles have been shown to contain invaluable information. How can the view that current ensemble forecasts do not represent probabilities in any meaningful sense be reconciled with the claim that, as rational humans, we should be able to express our uncertainty concerning future events in the form of probabilities?

1 Introduction

In this paper we will often speak of forecasts, which seems to imply that what is to be forecast lies in the future, but what we have to say equally well applies to "now-casting" and reconstruction of past events. The events are nonetheless assumed to be generated by a dynamical process and hence have a temporal component, like weather events. The forecasts investigated

in this paper consist of probability assignments. These probabilities are generated by what will be referred to as a probabilistic forecasting system (PFS). This might consist not only of a model of the dynamics, but also of a network of measurement devices as well as a mechanism to assimilate the observations into the system.

The days when authors found it necessary to write introductory words or even a section defending probability forecasts against deterministic ones are over. The main arguments though are certainly worth being looked at again. It seems that two main types of arguments can be distinguished. The first one asserts that probability forecasts contain more decision relevant information than deterministic forecasts. It has more than amply been demonstrated that this is the case for currently operational medium range weather forecast systems. But this is not necessarily a compelling argument for probabilities. Decision relevant information might be produced and conveyed by other means, and we will discuss a simple example in Section 4. The second argument asserts that probabilities are necessary for the framework of decision theory to apply. Probabilities aid the user in calculating his or her exposure to risks. By minimising his expected risk he or she can arrive at an "optimal" decision. This argument might have been overrated. It has to be kept in mind that in order to be usefully employed in decision making, a probability has to satisfy certain requirements which go beyond just being a normalised density function. In Section 3, deficiencies of currently operational ensemble forecasting systems are discussed. The main point here is not that these systems fail to reproduce the true dynamics, but that they fail to represent our knowledge (and ignorance) of the problem adequately. Although this is to some extend unavoidable due to limited time and computational resources, it leads us to argue (in Section 4) that currently available "probability" forecasts do not sufficiently fulfil the requirements of probability theory. What remains (in our opinion) of the case for probabilities is outlined in Section 2.

2 Why we want to issue probability forecasts

Although the usefulness of currently operational probabilistic weather forecasts has amply been demonstrated, the case for probabilities as the ideal means to present forecast information is, at least in our opinion, less compelling than the amount of work devoted to probabilities in weather forecasting might suggest. This section will still argue for probabilities, but tries to do so using minimum amount of assumptions. The essential requirement is that we, the forecasters, want to give consistent advice to our customers. The math in this section closely follows [11], Section 8, although the interpretation is slightly different. As will be demonstrated, a certain form of consistency is ensured by basing this advice on probabilities. Beyond the

reasons stated in this section, we have not yet found any which favour probability over other concepts that convey the same amount of information to the forecast user.

Suppose the highly idealised situation where the weather has only two states, rain and sunshine. We give (or sell) decision support to two customers A and B with weather dependent business. Depending on what decisions they take in advance and whether it will rain or not, they will incur losses (or wins, which we will count as negative losses). Hence the losses $L_A(a, y)$ of customer A (resp. $L_B(a, y)$ of customer B) are functions of the act a taken and the weather y which actually obtains. An example which assumes that both customers can act in only two ways is summarised in the two panels of the following table:

Cust.A	Rain	Sunshine
$a = 1$	3	1
$a = 2$	0	2

Cust.B	Rain	Sunshine
$a = 1$	1	1
$a = 2$	0	4

As forecasters, we are supposed to aid the two customers in reaching a decision. It should be intuitively clear that telling customer A to choose $a = 1$ while telling customer B to choose $a = 2$ (a decision which we will denote by $[1, 2]$) is bad advice. The total loss for both customers is 5 for sunshine and 3 for rain, which is in any case inferior to decision $[2, 1]$, which would give accumulated losses of 3 for sunshine and 1 for rain, respectively. The customers could pool these losses and arrange for pay-outs which at any rate would leave them better off than with decision $[1, 2]$. Such a strategy would not require any knowledge of the weather but only of the loss structure. Another way to see the deficiency of decision $[1, 2]$ is to think of the customers A and B being in fact only one customer with two businesses. We (the forecasters) would face embarrassment as soon as the customer behind the straw men A and B reveals his true identity.

The problem in the above mentioned example is that the decision $[1, 2]$ is *dominated* by decision $[2, 1]$ in that the overall loss incurred is smaller for the latter, no matter if it is rainy or sunny. It seems to be a reasonable minimum requirement that forecasters should only support decisions which at least are not dominated by any other decisions. A sufficient criterion for a decision a not to be dominated by another is that it is supported by a probability p, by which we mean that

$$pL(a, 1) + (1 - p)L(a, 0) \leq pL(\alpha, 1) + (1 - p)L(\alpha, 0)$$

for any other decision α (here 1 might stand for "rain" and 0 for "sunshine"). In other words, a decision a is supported by a probability p if it minimises the expected loss with respect to p. Indeed, if a is dominated by another decision b, then $L(a, 1) > L(b, 1)$ and also $L(a, 0) > L(b, 0)$, whence for all p

$$pL(a, 1) + (1 - p)L(a, 0) > pL(b, 1) + (1 - p)L(b, 0).$$

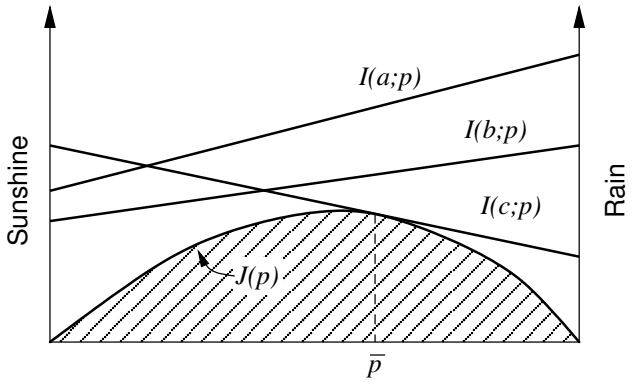

Figure 1: Possible decisions (a, b, c) and corresponding affine functions.

For the following considerations, the reader might find Figure 1 useful. For any decision a, we can consider the functions

$$I(a;p) := pL(a,1) + (1-p)L(a,0)$$

which are linear in p (three of those are plotted in Figure 1). They define a concave function

$$J(p) := \min_a I(a;p).$$

We can say that a sufficient criterion for a decision c not to be dominated by another is that there is a probability $\bar{p}$ so that the linear function $I(c;p)$ is a line of support of $J(p)$ at $\bar{p}$ (cf. Figure 1). The converse is true only if we assume the following:

Condition (C): Every linear function that is nowhere smaller than $J(p)$ can be represented as $I(a;p)$, that is, corresponds to a decision.

Suppose that a certain a is not supported by a p_a (for example $I(a;p)$ in Figure 1). This means that it is possible to draw a line between $I(a;p)$ and the function $J(p)$ (for example $I(b;p)$ in Figure 1). Because of condition (C), this line represents a possible decision which dominates $I(a;p)$.

In the preceding example of customers A and B, we were concerned with two decision problems described by two loss functions L_A and L_B. Now any pair of decisions $[a_A, a_B]$ made by the customers A and B, respectively, can be considered as a compound decision a for the loss function $L := L_A + L_B$. This decision corresponds to the linear function

$$I(a;p) := I_A(a_A;p) + I_B(a_B;p).$$

As was argued before, the decision $a = [a_A, a_B]$ is not dominated by any other decision if there is a p_a so that $I(a;p)$ is a line of support of the concave function

$$J(p) := \min_a I(a;p) = J_A(p) + J_B(p)$$

at p_a. In particular, this is the case if a_A and a_B are supported by the *same* p_a. However, the latter is also necessary if we again assume condition (C) for both customers. Suppose that the decision $a = [a_A, a_B]$ is not supported by a p_a. Hence a line $\lambda(p)$ parallel to $I(a, p)$ can be found which is a line of support of $J(p)$ at a certain p_λ. Any such line can be represented as a sum of two lines supporting $J_A(p)$ and $J_B(p)$ respectively at the same point p_λ. As was assumed in (C), both lines correspond to decisions which the customers have at their disposal, and the combined decision (which corresponds to the line $\lambda(p)$) dominates a.

To summarise this section, it has been demonstrated that supporting decisions by probabilities provides a sufficient safeguard against incoherent decisions, especially in more complex decision problems. Under condition (C), which requires a certain richness of decisions available to the customer, probabilities even naturally come about if incoherent decisions are to be avoided. The presented arguments though by no means assert that probabilities are a necessary or natural or even practical way to convey forecast information to end users or to provide decision support.

3 Typical Deficiencies of Probabilistic Dynamical Models

In this section some typical deficiencies of probabilistic forecasting systems (PFS) in general and ensemble forecasting systems in particular will be revisited. The deficiencies considered could be labelled either *a priori* or *a posteriori*. Any modelling process inevitably involves approximating or simplifying our understanding of the problem, often due to limited computational resources. Hence the PFS is known to be deficient *a priori* or before the model has actually been put to a test. In other words, *a priori* deficiencies are those that would go away if the model faithfully represented the forecaster's knowledge about the process under concern. For example, the fact that global circulation models use only finite grid resolutions could be labelled an *a priori* deficiency. It is known that processes on a sub-grid scale matter for the overall dynamics of the weather [15, 10]. However, *a priori* deficiencies are not limited to inadequate representations of the laws governing the phenomenon under concern. Another important reason for *a priori* deficiencies is inadequate representation of the laws of mathematics and especially probability theory. A particularly interesting example is the problem of assimilating observations into the system. In order to make forecasts, the model has to be initialised at a suitable initial condition, depending on the history of observations. These observations are usually corrupted by noise, whence the initial condition is in fact a distribution of initial conditions. Calculating this distribution is often referred to as *data assimilation* in geophysics [9], while the term *filtering* is used in the engi-

neering and stochastics literature [6, 8]. Data assimilation is known to be a formidable problem. Ideally, data assimilation should yield the conditional probability of the model's state given the past history of observation. According to the laws of probability, this problem has a mathematically well defined solution called the *optimal filter*, but calculating it typically involves solving an infinitely large set of dynamical equations (see, for example [1]). In other words, proper data assimilation is an infinitely complex problem.[1] Obviously, any operational PFS needs to compromise here. Finally, it should be noted that ensembles are in some sense but a simplified way to issue probability assignments. Often this is referred to as sampling error, although it is doubtful whether currently operational ensembles are even only a sample from the probability distribution the forecaster would issue if he could. Hence this not–sampling from the desired probability distribution could be seen as yet another *a priori* deficiency.

By *a posteriori* deficiencies we denote disagreement between forecasts made by the PFS and actual observations. These deficiencies evidently require the PFS to have been compared to observations. For example, in order for the forecasts to actually bear any connection with reality, the issued probabilities should agree (up to expected sampling error) with actual observed frequencies. To give a rather simple example, suppose the problem is to forecast (on a day–by–day basis) the occurrence of rain in London. Every day, our PFS issues a probability p_n, where n enumerates the days. Let us fix an interval $B \subset [0, 1]$ and consider the collection I_B of all days where p_n falls into B, that is

$$I_B := \{n; p_n \in B\}. \tag{1}$$

Considering the observed frequency f_B of rain over all days in I_B, we should expect f_B to fall into B as well, or more precisely, we should expect f_B to be equal to the mean value of the set $\{p_n; n \in I_B\}$. This property is often referred to as *reliability* or alternatively the forecasts are called *calibrated* [14]. Reliability generalises to more complicated PFS thus: Let the variable to be forecast be denoted by Y_n and the range of values of Y, the observation space, by E. Examples are an only finite set of possible outcomes (only two in the case of the previous rain/no-rain example) or the real line. Suppose at time n, the PFS issues a probability assignment P_n over E, which can be a probability density function, a cumulative distribution function or even a measure. This P_n subsumises our current information on Y_n. Reliability (or calibration) means that P_n describes the distribution of

[1] It should be said that there are notable exceptions to this statement, namely linear systems with Gaussian uncertainties. In this situation, the Kalman Filter provides a complete solution to the optimal filter. In nonlinear situations though, the Kalman Filter fails to apply, and with very few exceptions, optimal filtering is an infinite dimensional problem.

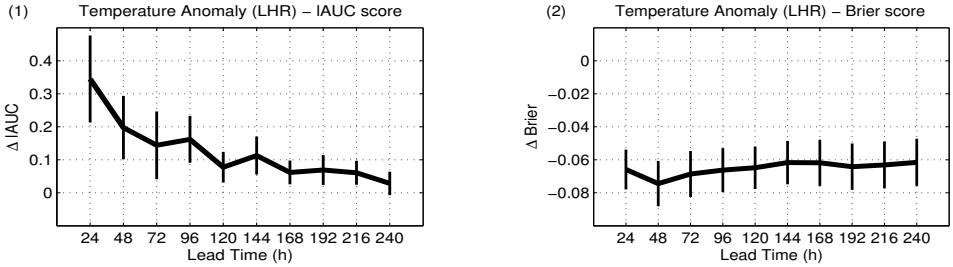

Figure 2: Relative skill of two forecasts for 2m–temperature anomaly at London Heathrow, noon. Left panel: logarithmic AUC score, right panel: Brier score.

Y conditioned on the currently available information, or in symbols[2]

$$\text{Prob}(Y_n|P_n) = P_n. \tag{2}$$

As checking reliability involves the estimation of a rather complicated conditional probability on the left hand side of Equation (2), various simplified measures of reliability have been proposed in the literature (PIT [2], Talagrand diagrams [14], minimum spanning tree histograms [4]) which generally form but necessary conditions for Equation (2). Using these measures, it has amply been demonstrated that operational ensemble forecasting systems are not reliable. In the next section we will discuss a very simple experiment that yields further evidence to that, but at the same time demonstrate the immense potential value of the output of operational ensemble forecasting systems.

4 The Interpretation of Probabilistic Forecasts

This section starts with the discussion of a very simple experiment which will demonstrate (yet again) that the output of operational ensemble forecasting systems does not reach its full potential if employed directly as probability forecasts. Figure 2 shows the relative skill of probabilistic forecasts for temperature anomalies at London Heathrow weather station. The event to be forecast is whether the temperature exceeds a certain normal provided by a third order harmonic polynomial fitted to the archive of observations. Two types of forecasts are generated using the ECMWF global medium range ensembles, consisting of 51 members. The first simply consists of the fraction of ensemble members that exceed the normal. In other words, if

[2]A somewhat different formulation of reliability (and in fact of the entire forecasting problem) has been given in [2]. We conjecture that the formulation though is entirely equivalent to ours if the quantity $G_n(x)$ in [3], "the distribution from which Nature draws", is replaced by $\text{Prob}(Y_n = x|P_n)$.

$\mathbf{x}_n = [x_n^{(1)} \dots x_n^{(51)}]$ is the ensemble at day n, the first forecast investigated is given by

$$p_n := \frac{1}{51} \sum_i \{x_n^{(i)} \geq \nu_n\} \tag{3}$$

where ν_n is the normal. This approach will be referred to as *counting approach*. The second forecast was generated using the ansatz

$$q_n := f(\mathbf{x}_n; \theta) \tag{4}$$

where f is a linear combination of 15 different statistics generated from the ensemble, such as the mean, standard deviation, interquartile range etc. The coefficients θ where determined using regularised (*ridge*) regression (see [5]). This approach will be referred to as *regression approach* (cf [13], where logistic instead of linear regression was used). The two forecasts where then compared in terms of the difference of their respective scores, where the forecast obtained through counting was subtracted from the regression approach. Two scores were considered, namely the logarithmic AUC score (Figure 2, first graph) and the Brier score (Figure 2, second graph). For a discussion of these scores see e.g. [14]. Both indicate a significant superiority of the regression approach (note that the log-AUC score is positively oriented, i.e. a higher score indicates a better forecast, while the Brier score is negatively oriented). The obvious conclusion is that the ensemble contains more information about the problem than what is revealed by the simple counting approach, or in other words, considering the ensemble a bunch of equally likely scenarios of the future weather under-exploits it.

Both because of the preceding example and what has been discussed in Section 3 it seems at least questionable whether the pure ensemble should be termed a probabili*ty* forecast. First note that there are (at least) two distinct interpretations of probability, namely the frequentist's and the subjectivist's view, although both agree on Kolmogorov's axioms as the proper mathematical formalism for probabilities. The frequentist's view essentially asserts that probabilities are limiting observed frequencies of repetitive independent trials. The notion of probability only applies in situations where limiting observed frequencies exist (or at least are a reasonable idealisation of the true circumstances). In other words, a probabilistic forecast is a probabili*ty* forecast only if it is reliable. As was already mentioned, currently operational weather forecasts are not reliable. In particular, failure of the counting approach demonstrates that the ensemble is not a sample from a reliable forecast distribution. Hence, the ensemble fails to get the frequentist's seal of approval. It might be argued that what the forecast needs is recalibration, thereby ensuring a better agreement with observed frequencies (and hence, probabilities). This is certainly a reasonable approach for problems as simple as the one presented above, and the mentioned regression

approach can be understood as a means to recalibrate the forecast. Recalibrating more complicated forecasts though quickly becomes a formidable program, invariably suffering from lack of data. In weather forecasting, large recurrence times and frequent model changes already hamper the calibration of one dimensional continuous forecasts, and estimating higher dimensional conditional probabilities seems utterly impossible. Note that, in some sense, the model changes even on a daily basis, as the configuration of the observation network can differ quite drastically from day to day. As a summary, it seems quite hard to thoroughly carry out the frequentist's program in weather forecasting. The situation of the frequentist appears to be even more challenging conceptually (to say the least) for seasonal or climate forecast. It seems quite hard to interpret a probability assignment for the global mean temperature in 2050 (be it conditioned on a certain CO_2–scenario or not) in a frequentist's sense.

Another important interpretation of probability is the personalist approach [12, 7]. Personalist views hold that probability represents the degree of someone's belief that a certain proposition (for example that it will rain tomorrow) holds true. That the laws of probability hold for personal probabilities can be deduced from certain assumptions on the individuum being "rational" (according to some axioms of rationality). This does by no means assume the person to be omniscient or entirely logical. Two people, equipped with the same body of evidence, might still hold different personalist probabilities on whether it will rain tomorrow. Personalist views on probability aim at providing a framework of consistent behaviour in the face of uncertainty. Consequently, various approaches to defining personal probabilities proceed along similar lines as presented in Section 2 of this paper (see [11]). Above the frequentist's interpretation, the personalist view has the advantage of being still well defined even under the above mentioned circumstances, which seems to suggest this interpretation as the more suitable for weather and especially climate forecasting. The personalist approach to probability does not give any kind of guidance though as to how possible inconsistencies in the decision making process might be removed. The theory only guarantees consistency of decisions if the laws of probability are obeyed. In Section 3 however, we saw that doing so can be of enormous, if not insurmountable difficulty in practice. Current data assimilation systems work with a plethora of approximations, resulting in inconsistent probabilities. Concessions due to limited resources are inconsistent with rational behaviour, as required by the personalist view on probability. The personalist view does not take into account that scrutinising all possible courses of action in advance might require prohibitively large amounts of time and computer power. In a particular situation it is of course almost always possible to devise a probabilistic model simple enough to allow for exact mathematical solutions and hence fully consistent probability assignments, but usually at the cost of discarding essential knowledge about the phenomenon under

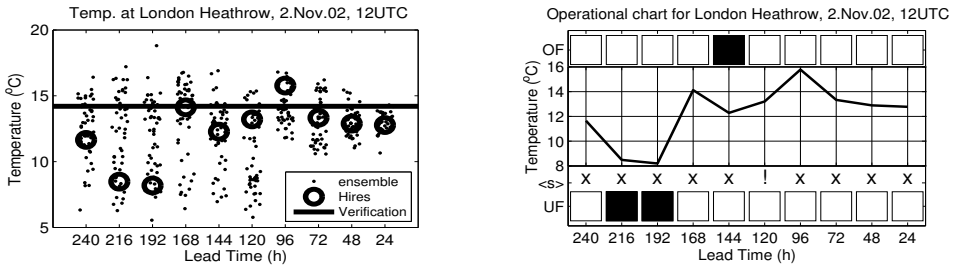

Figure 3: Left panel: Ensemble forecasts for 2m–temperature at London Heathrow for 2. Nov. 2002, noon and ECMWF's deterministic forecast (circles). The verification is plotted as a straight line. Right panel: Forecast summary (for details see text).

concern. We might for example propose that the temperature at London Heathrow Airport at noon follows a white noise process (thereby obviating any need for data assimilation). Such a model would be a valid personalistic model, although an utterly useless one.

To finish this section, a simple example will be discussed how ensemble forecasts can convey useful information without using probabilities at all. The left panel of Figure 3 shows a set of ensemble forecasts for 2m–temperature at London Heathrow Airport. The ensembles possess different lead times (shown on the x–axis), but all of them verify on 2. November 2002 at noon. To be able to distinguish between the individual ensemble members they have been plotted with a slight jitter along the abscissa. The circles represent ECMWF's deterministic forecast, generated by using a model with a very high resolution. Finally, the verification (i.e. the actual temperature) was about 14^oC and is plotted as a straight line. Examining this figure closely, a couple of interesting facts emerge. The deterministic forecast first underestimates the temperature until it jumps to roughly the correct value at lead time 168h. The ensemble though seems to anticipate the correct temperature well before that, as a large fraction of ensemble members are close to the correct value even at 240h. Hence, the error in the deterministic forecast could have been anticipated. It is fair to say though that at lead time 120h, a large fraction of ensemble members indicate a temperature of about 8^oC, whence the deterministic forecast is expected to overestimate the temperature while it is in fact quite correct. A comparison with other dates would furthermore reveal that the ensemble spread at lead time 120h is unusually large, thereby indicating the whole synoptic situation to be uncertain; a fact that is probably of interest to the user. A possible way to summarise the mentioned features is presented in Figure 3, right panel. The deterministic forecast is plotted, while black squares in the row above (resp. below) the axes box indicate likely over-forecasting (resp. under-forecasting) of the deterministic forecast. Furthermore, an exclamation mark in the row

labelled $< s >$ indicates an unusually large ensemble spread. Further symbols could for example indicate when the ensemble shows bi-modality, which seems to be present here for lead times 144h and 120h. Although we do not deny that probabilities are implicit in this example as well (and even should be, as was argued in Section 2), it shows that probabilities do not have to be explicit in decision support. An enlighted user, if sufficiently instructed, is able to form his own opinion about how to use the forecast product, or, as a personalist would put it, form his own personal probabilities given the provided information.

5 Conclusion

It was discussed how a small set of assumptions in principle compels the forecaster to issue probability forecast. On the other hand, issuing probabilities which are acceptable by either a frequentist or a personalist meets with considerable difficulty. This is due to a number of typical deficiencies in probabilistic forecasting systems. The problem is not our limited understanding of the phenomenon, but our inability to express our understanding in the form of forecasts, mainly due to restrictions in time and computational resources. We noted that issuing the information at hand through probabilities is but one possibility, and a probability–free way of summarising an ensemble forecast was presented. It should be obvious though from this discussion that either a modified interpretation of the probability concept or something different is needed, most likely obtained by relaxing the requirements of probability somewhat. We are not only unable to say exactly what will happen, but that we are even unable to say exactly what we think will happen. This somehow has to be taken into account, lest the forecast user be mislead about what the forecasts actually mean.

References

[1] Marco Ferrante and W. J. Runggaldier. On necessary conditions for the existence of finite dimensional filters in discrete time. *Systems and Control Letters*, (14):63–69, 1990.

[2] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration, and sharpness. Technical report, Department of Statistics, University of Washington, 2005.

[3] Tilmann Gneiting and Adrian E. Raftery. Weather forecasting with ensemble methods. *Science*, 310:248–249, 2005.

[4] J.A. Hansen and L.A. Smith. Extending the limits of forecast verification with the minimum spanning tree. *Monthly Weather Review*, 132(6), 2004.

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

[6] Jazwinsky. *Stochastic Processes and Filtering Theory*, volume 64 of *Mathematics in Science and Engineering*. Academic Press, 1970.

[7] Richard Jeffrey. *Subjective Probability*. Cambridge University Press, first edition, 2004.

[8] Gopinath Kallianpur. *Stochastic Filtering Theory*. Number 13 in Applications of Mathematics. Springer Verlag, 1980.

[9] Eugenia Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, first edition, 2001.

[10] Tim N. Palmer. A nonlinear dynamical perspective on model error: a proposal for non–local stochastic–dynamic parametrisation in weather and climate prediction models. *Quarterly Journal of the Royal Meteorological Society*, 127, 2001.

[11] Leonard J. Savage. Elicitation of personal probabilities and expectation. *Journal of the American Statistical Association*, 66(336), 1971.

[12] Leonard J. Savage. *Foundations of Statistics*. Dover Publications Inc, New York, 1972.

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

[14] Daniel S. Wilks. *Statistical Methods in the Athmospheric Sciences*, volume 59 of *International Geophysics Series*. Academic Press, second edition, 2006.

[15] Paul D. Williams. Modelling climate change: the role of unresolved processes. *Philosophical Transactions of The Royal Society*, 363:2931–2946, 2005.