The Weather Forecaster as an 'Intuitive Statistician'

Anders Persson, Swedish Meteorological Society, Sweden

Introduction

In May 2011, as I was about to retire, Erik Andersson, Head of the ECMWF Meteorological Division, wanted me to re-write their User Guide. Instead of the customary two-thirds of its content devoted to the excellent ECMWF forecast system and only one-third about how to make the best use of its products, he wanted the reverse.

This challenge forced me to really take on the question I had been pondering since the start of my meteorological career more than 40 years before: "What does a good weather forecaster do?"

During the work it become gradually clear to me that skilful weather forecasters must not only have a good grasp of the physics and dynamics of the atmosphere, numerical models and ensemble system, but must also have the ability to quickly draw conclusions from a wide selection of, often contradictory, information, an ability which is in the literature referred to as 'intuitive statistical thinking'.

I will illustrate this with a common weather situation: the drop in wintertime temperature after the lifting of low clouds.

An Illustrative Example – the Dispersion of Stratus

Consider the following forecast: 'The 2 metre temperature of +3°C with low stratus will drop to -5°C if the clouds disperse.' If this is predicted by a computer, the human forecasters might be challenged to justify their existence by 'adding value', i.e improving on the computer's deterministic forecast.

It is not a trivial matter to calculate the accumulated effects of radiative cooling (depending on time of day and year, moisture and stratification), heating from the ground, the effect of wind shear and turbulence etc plus knowing to what degree these phenomena are already well described by the model. With this challenge the forecasters are, in my view, lured into competing with the deterministic NWP on its own conditions.

But improving on the deterministic NWP is not the only way the forecasters can modify the -5°C forecast. *They can for example express doubts that the clouds will disperse at all!*

Assume that they arrive to the conclusion that the clouds are slightly more likely to stay than to disperse. If they can roughly quantify their judgement into a probability of 40% for clearing and 60% for remaining overcast, the 'best' all purpose temperature forecast is neither +3°C or -5°C but the weighted average ±0°C.

But the forecasters can do more. They know that the public generally is more sensitive to a drop in the temperature than it remaining unchanged. Consequently, if they are wrong, it is for most purposes better to be on the cold side than on the mild. This insight might make the forecasters tweet the forecast to -1°C or even -2°C.

But of course the best solution is to be able to tell the clients in probabilistic terms that although it is most likely (60%) that the clouds will stay with +3°C there is a substantial possibility (40%) they will break up and lead to a temperature drop to -5°C.

These three different deliberations are made, not on a physical but on a statistical basis.

a) The ±0°C forecast was made to minimize the expected statistical error and the general 'pain' the public would suffer. This assumes a symmetric penalty function, that errors in the positive direction are as bad or harmful as those in the negative direction.

b) The -1°C or -2°C forecasts, however, assume an asymmetric penalty function where a positive error (forecast is too warm) is deemed as being more harmful than a negative error (forecast is too cold). Note that all these three forecasts ±0°C , -1°C or -

2°C will probably not occur and if they do, only for a few minutes in case of a clearing. Still they are 'the best'.

c) Finally, to express uncertainty verbally or numerically (though explicit use of probabilities) really demands that forecasters are more open and developed in their use of statistics, and their understanding of these methods becomes more crucial.

When the forecasters make these deliberations they might take a lot of information into consideration, such as deterministic forecasts from more than one model, ensemble systems, statistical interpretation schemes etc. For short range forecasts they might also have to consider newly arrived information from observing stations, satellite and radar. Add to this ideas and suggestions from the forecasting team as a whole.

The final forecasts are less products of physical insights as of an intuitive, ad hoc weighting together of often contradictory information where the forecasters act like 'intuitive statisticians' without being really aware of it.

Statistics – the Science of Uncertainty

That weather forecasters act like 'intuitive statisticians' should not come as a surprise. Weather forecasting has been non-perfect since its start 150 years ago. The interesting thing is why this has to be stressed. Perhaps their one-sided deterministic Newtonian physical education is to blame.

1. Weather forecasting is intrinsically a statistical, or rather probabilistic problem. Prior to Edward Lorenz's discovery of the 'Butterfly Effect' there was among meteorologists only a qualitative understanding of the role small changes in the initial conditions could affect forecasts.

2. To quantify and communicate this uncertainty increases the value of the forecasts. This has been known since the early 20th century when Cleveland Abbe in the US and later Anders Ångström in Sweden promoted probability forecasting. What has muddied the water was the erroneous perception that probabilities or uncertainty information served to 'cover the backs' of the forecasters.

3. Indeed, it is in uncertain weather situations that the forecasters really have a chance to show their skill. Experience shows that it is when the forecasters communicate their uncertainty in an active, clever way that they get appreciation from the public. The ability to turn a potential weakness, non-perfect and uncertain forecasts, into an advantage could be more explored by weather forecasters.

The main political or psychological conceptual problem is not that 'probabilities are difficult to understand' but that many users of weather forecasts want them in a deterministic, categorical form. This essentially means that the forecasters make the decisions for them, thus relieving them from any controversial responsibility for 'wrong' decisions[1].

This is, fortunately or unfortunately, not only a problem in weather forecasting but in all straights of life, according a wide range of current literature. They are aimed at economists, medical doctors, politicians, military commanders etc all involved in decision making under uncertainty. Among these Daniel Kahneman's bestseller 'Thinking Fast and Slow' stands out. It is not only written by a Nobel Prize Laureate, it is also a good read. He presents a number of case studies which, although not of meteorological nature, still can be 'translated' into meteorology.

His discussion on 'fast' and 'slow' thinking (System 1 and System 2) explains, at least to me, why a lot of the meteorological educational activities I have been involved in or subjected to, don't seem to have left much impression in the operational activity. Perhaps this is because most of the teaching has been designed to serve our 'slow' system 2 (to pass exams and conduct scientific investigations) and not so much the 'fast' System1 (quick decision making in the heat of operational duties).

The Problem with Probabilities

Any teaching of uncertainty has to involve the concept of probability. The problem here is that it

1. Specialist meteorologists, such as military or marine weather forecasters, know their customers sufficiently well to be able to essentially make the decisions for them by presenting their forecasts in a confident deterministic way.

emerged quite late in the human history (in gambling in the 16th century, in science around 1800). Then there are at least three different types of probabilities:

a) the classical, for example the chance of getting a tail when tossing a coin or a '6' when tossing a die

b) the frequentist, for example the number of rainy days over 30 years which define the probability of rain for a certain location

c) the subjective or Bayesian probability where we try to estimate the chance that our football team will win, being 2-1 down with 15 minutes remaining.

We will meet all three categories in weather forecasting; in particular the Bayesian varity since purely subjectively estimated probabilities as well as those from the ensemble system belong to this group. Probabilities from MOS (Model Output Statistics) belong to the frequentist category since they are based on observed relations between past forecasts and verifications. Knowledge about the classical definition is useful when we want to understand the problems with combining or splitting up probabilities.

A Five-point Programme

Having identified weather forecasters as 'intuitive statisticians' allows us to translate the advice given in the above mentioned literature into forecasting meteorology, formulated as a five point programme:

1. Combat over-confidence
2. Do not underestimate the power of randomness
3. Pitfalls estimating probabilities
4. How to present probabilities
5. How to make decisions from probabilities

For each section and sub-section a brief introduction from the non-meteorological world will be followed by examples from operational weather forecasting.

Combat overconfidence

Over-confidence is a global human phenomenon or weakness. It stems mainly from

a) Conclusions from too small samples. Commonly three months of NWP output (ran twice a day yielding 180 forecasts) is regarded as an adequate sample size. But this is only true if the forecasts are mutually independent. The nature of data assimilation, using a "first guess" then only partly modified by observation, creates dependence between successive analyses and thereby forecasts. To minimize this dependence one might select only every 5th NWP run. In other words, results from a 180 case sample is as representative as one of a 36 case sample.

b) The confirmation bias: There is a human tendency to look only for arguments which support one's opinion. It is of course equally important to search for contradictory arguments. Another human weakness is to become more stubborn when challenged by counter arguments, when the most logical reaction would be to hold on to the opinion, but increase one's uncertainty about its validity.

c) The selection bias: Another common error is to terminate an investigation, for example concerning the quality of two different forecast methods, when the results seem to confirm the 'desired' conclusion. A similar error is to discard 'bad' data or 'outliers' too superficially.

d) Lack of knowledge: If you have never heard about tsunamis you confidently stay on the beach when the sea water is rapidly receding. If you are not aware of the existence of katabatic winds in mountains you might underestimate the chances of strong winds during a high pressure regime. Not knowing that the Coriolis Effect also has a strong impact on sea breezes will affect your sailing forecast.

e) Lack of imagination: The 'normalcy bias' assumes that since a certain type of events has not happened before, it will not happen in the future either. The atmosphere normally behaves in a 'familiar fashion', but the exceptions are numerous. The ensemble forecasts are not able to include all possible synoptic scenarios. Because of the limited number (N) of members, a fraction 2/N of verifying observations has to be outside the spread.

Our quest for certainty, leading to over-confidence, also makes us underestimate the power of purely random effects.

Do not underestimate the power of randomness

Randomness is often seen as something rather harmless, some 'noise' that will even out in the long term. As will be shown below, random effects can in meteorology yield quite convincing patterns and spurious correlations. In meteorology we encounter the power of randomness in at least five situations:

a) Conclusions based on too few samples: False regularities will of course more easily appear in a small data sample. Since our memory is short we tend to remember weather cases only a few weeks back, during which time false regularities or apparent systematic errors might easily appear.

b) Conditional sampling: If we evaluate the forecasts with respect to predicted anomalies, such as heavy rain, cut-offs west of Portugal or tropical cyclones we will notice an increasing degree of over-forecasting with increasing lead time. This does not necessarily constitute any systematic error but non-systematic errors due to the lack of predictive skill. If we on the other hand evaluate the forecasts after occurred anomalies, the opposite will appear to be true: an apparent increasing under-forecasting with lead time. Only if the number of forecast and observed anomalies over a specific time period differ in number (irrespective if they verified or not) does this indicate a model systematic error.

c) The regression to the Mean Effect: This powerful statistical artefact, discovered by the British 19th century scientist Francis Galton, causes misinterpretation of verification statistics during persistent anomalous periods. With increasing lead time the forecasts' skill decreases and the forecasts start to scatter more and more around the climatological mean. This is reflected in an increasing mean error which might be mistaken for a model drift.

d) The helpful Regression to the Mean Effect: In cases of large uncertainties a forecast based on climatological averages can serve as a default. This is what aviation forecasters make use of in their 2-hour TREND forecasts when some extreme weather has unexpectedly appeared at the airport. So do medium range forecasters when faced with an extreme development in the ECMWF deterministic forecast beyond the first 3-4 days. So does the ensemble system in particular during the last 5-8 days in the extended 15 day epsgrams: a smooth gravitation back towards the (model) climate.

e) Other biases: The 'publication bias', also called 'the desk drawer effect', occurs when negative results are not publicised. This increases the risk that positive results, arrived to by pure luck, are accepted. When other scientists try to confirm the result and fail, they might wrongly suspect manipulation or fraud. The Slutsky-Yule Effect creates convincing, but spurious periodic variations in time series when subjected to running averages.

These two sections have pointed out two main sources of misjudgement of probabilities: overconfidence and underestimating the power of randomness. These two also figure among numerous other pitfalls when estimating probabilities.

Possible pitfalls trying to estimate probabilities

a) The halo effect: We encounter the 'halo effect' when a certain NWP model for non-rational reasons is given unduly high weight. Before ECMWF came into being weather forecasters in Norway tended to favour the American model, the Danish forecasters the British and the Finnish forecasters the German model, apparently echoing their countries' allegiances during the Second World War. Consequently forecasters at SMHI, in 'neutral Sweden', looked keenly at all three models.

b) The primacy effect: The order under which information arrives might have impact on the user. If output from model A arrived earlier than from model B of equal quality, the forecasters would put more weight to A than B. See also the 'confirmation bias' above.

c) The availability error: Verifications of manual thunderstorm predictions show that the +24 hour forecasts are better than the +12 hour ones, counter to what could be expected. When the +24 hour forecast is made in the afternoon, for the next day's afternoon, the daily cycle makes local thunderstorms 'available' on the maps or radar screens. In the morning, 12 hours later, there are rarely thunderstorms 'available' when the +12 h thunderstorm forecasts are made, just mist or perhaps fog patches.

d) The mean and the variance: Although the ECMWF model, presently at least, is 'the best' glob-

al NWP model statistically, doesn't mean that it is 'the best' every day. The day-to-day variation in skill is much larger than the average difference of the mean skill of other global models. An optimal approach is therefore to consider all NWP models.

e) Correlation between models: The heuristic weight we might put on the NWP models should not only depend on their average skill but also on similarities in their characteristics reflected in their mutual error correlation. Assume for the sake of discussion that the ECMWF operational model, its EPS Control and the UK Met Office global model are equally skilful. This would still motivate a weight › 33% on the UK model and ‹ 33% for each of the EC models, since they are more similar than each of them with the UK model.

f) The representativity error: It is a common human mistake to confuse what is 'probable' with what is 'typical'. In a lottery it is a common mistake to believe that a sequence 853347 is more random and more likely to win than the sequence 111111 just because the former looks more random. The representative error might partly explain why meteorologists tend to favour detailed high-resolution NWP forecasts with realistic looking but unreliable details than forecasts where these details have been smoothed out or removed. The latter will no longer look like 'typical' images of the atmospheric flow pattern.

g) Consistency and forecast skill: We tend to trust a person who doesn't change his mind too often, but is 'consistent'. This does not apply to NWP where the correlation between the spread between consecutive forecasts and skill of the most recent forecast is small if not zero. Any connection ought to be between the spread and the mean of the forecasts.

h) The outcome, memory or hindsight bias: These artefacts reflect the unconscious selection of information, either because it is very recent or because knowledge about the final outcome would favour information that supports this development. After an unexpected storm it is of course easy to find exactly those pieces of evidence (an isolated observation, an individual forecast or an odd ensemble member) that would indicate that the storm would develop.

i) The bandwagon effect: The tendency to do or believe things because many other people do or believe the same. This is important at weather conferences where a minority view might not change the deterministic part of the forecast but modify its probabilistic part.

j) Deceptive consistency: In cases where the last three deterministic forecasts are 'jumpy' we tend to trust a sequence where the last two agree rather than a sequence where the first and the last agree, but the one in between differ ('flip-flop'). But that that is to ignore the importance of their mutual correlation. If two persons have the same opinion we regard it as more significant if they do not know each other than if they are twins. The first forecast might be the on average least skilful but on the other hand less correlated with the most recent. Their agreement therefore carries more weight than the agreement between the two consecutive forecasts.

The role of the 'intuitively statistical forecasters' could very well end here. But their skill is also needed to communicate their results.

Communicating probabilities

The best indication that probability information has been understood by the receivers is that right conclusions have been drawn.

a) Odds, intervals or verbal communication? It is not of paramount importance how the probability is communicated (numerically or verbally) as long as the right action is taken. Using odds, like '4 to 1' instead of 20% will be understood by people who are regular gamblers. Intervals implicitly suggesting probabilities of 50-80% that the truth will lie within the interval are also suitable.

b) Temporal and/or spatial intervals? On a more basic level it must be known if the probability refers to the chance that a certain point will be affected during a time interval, or at a specific time the chance that some point in an area will be affected? Or a combination of both? Studies have shown, although not conclusively, that laymen grasp a 20% probability better if it is expressed in words like 'in ten cases similar to the one we have now, rain will follow in two cases'.

c) The 50% probability problem: The 50% probability is often misinterpreted as complete ignorance, but is only so in the case of tossing a coin where the probabilities of either outcome is 50%. A 50%

probability for snowfall in Barcelona or for 26°C in Reykjavik would not be a trivial forecast.

d) Total uncertainty: In case of total uncertainty the probabilities should coincide with the climatological averages. Normally this only applies for longer lead times when the predictability has decreased to zero, but it might of course in rare cases also signify shorter forecasts. It can be shown both mathematically and practically that it is optimal for any user to be told about total uncertainty rather than to be given a deterministic forecast which would be no better than a pure guess, for example from the last deterministic forecast from the state-of-art NWP model.

e) The popularity or optimism bias: In a quest to remain popular with the audience the forecasters cannot always resist the temptation to "look on the bright side", i.e. exaggerate the probabilities for favourable or popular weather.

f) The framing effect: When a major city council once was warned about a 20% probability for severe thunderstorms not much was done, in contrast to when they heard that there was a 70% probability for the same extreme weather somewhere in the region. If the climatological probability is 2% then the forecast probability 20% can be presented as something '10 times more probable than normal'.

When the uncertainty information has been estimated and presented in a useful way one would assume that there is nothing more the forecaster could do. But according to the current literature there is a global problem with decision makers having difficulties to make rational decisions in light of uncertainty. Weather forecasters with their unique expertise in meteorology and probability theory can therefore still make a contribution.

Decision making from probability information

We now enter into a field that is well covered by an extensive literature outside meteorology. Decisions are not only based on economic considerations, but equally often on political or highly subjective criteria where prestige, pride and status play an important role. The popularity of deterministic, categorical forecasts has, as mentioned earlier, its objective root in that it can relieve the decisions maker from at least some of the burden of responsibility for their decisions, and then perhaps also the blame if something goes wrong.

But even if we restrict ourselves to monetary values it is important to realise that the commonly taught 'cost-loss' model for decision making is just the first approximation. Its says that if a loss L is likely with a p% probability then a cost of c ‹ pL is well spent. But does it work in the following case?

-Would the reader participate in a (free) lottery where there is a 80% chance of winning 1000 euros or be given 700 euros straight in the hands?

According to the 'cost-loss' reasoning the expected outcome of the first choice is 100 euros larger than the second. But even professors in mathematical statistics would have taken the second choice – if it had been an isolated opportunity. If it is repeated, as a way to pay your salary, the first choice is optimal.

According to the literature people asked to choose between a 50% chance of winning 2000 euros or getting 1000 euros straight in the hand, prefer the second 'safe' option. However and interestingly, faced with a choice between a 50% chance of losing 2000 euros or a certain loss of 1000 euro people prefer to gamble, offering a 50% chance of losing nothing. The conclusion is that people tend to be more motivated to avoid a loss than a chance to win.

In weather forecast applications this would perhaps mean that a customer with a potential loss of 100 euros and a protection cost of 30 euros would not, as suggested by the cost-loss model, take action when the probability is just › 30%, but when it is › 30%.

Studies of decision making in weather forecasting are important in order that we avoid dismissing as 'stupidity' decisions by our customers that at a closer inspection turn out to be rational.

Summary

When NWP was introduced around half a century ago it was said that the future role of the forecasters would be to communicate the NWP information to the customers and, knowing their needs, also help them to make optimal decisions in their own interest.

As with other aspects of operational weather forecasting there is not, and has never been, much

documentation about what this would involve. An attempt to define the role of the forecaster in the 'computer age' has been made in this article. However, the conclusion would have been equally valid before now; good forecasters have always acknowledged the intrinsic uncertainty in weather forecasting and cleverly use it to demonstrate their skill.

In spite of the prophesies that 'in 5-10 years there will be no need for weather forecasters', first heard by the author in 1966, it seems there have never in the history of meteorology been so many forecasters around. This is particularly true for the private sector which cannot be suspected of employing meteorologists on humanitarian grounds.

If the aim had been to create an automatic weather service of the 1966 standard, this could technically have been possible in the 1980s with the emergence of the high-resolution primitive equation models. The reason why this did not happen must among other factors involve the fact that with increasing forecast skill the demands from the public and paying customers have increased – and will continue to increase!

The scenarios oulaid in this article are therefore just as valid now as they were in when Fitzroy started operational weather forecasting in the 1860s, and will continue to be just as valid in the futre.