Integrating Information, Misinformation and Desire: Improved Weather-Risk Management for the Energy Sector

Leonard A. Smith

Abstract Weather-risk management has many facets. One particularly costly challenge comes from "chasing the forecast". The Forecast Direction Error (FDE) approach was deployed to address the dilemma facing decision-makers who face this challenge: today's probabilistic weather forecasts contain too much information to be ignored, but not enough information to be safely acted on as probability forecasts. Success was obtained by focusing on the information content of forecasts, and restricting their use to tasks in which the forecasts are informative in practice.

1 Introduction

The success reported in this chapter reduced the pain from "chasing the forecast", a phenomena encountered when managing a commodity whose ultimate value depends on the weather and whose current price fluctuates with the weather forecast. Modern weather forecasts contain vital information; that said, it is straightforward to show that, as probabilities, today's forecasts are not reliable[1]. And the competition can, of course, exploit the fact that one is using systematically flawed information. Rather than attempt to "fix" the probabilities, tools were developed to allow effective use of the information contained in the existing, imperfect forecast systems. The Forecast Direction Error (FDE) approach ingests a forecast-outcome archive of predictive distributions which are systematically mis-informative (unreliable) if interpreted as probabilities; it provides access to the relevant, reliable information (if any) they contain. By focusing on the relevant rather than the optimal, the FDE identified what was useful in practice given the raw materials in hand; it did **not** attempt to interpret (or construct) a probability forecast allowing optimal "trader behaviour." A tool is more likely to be embraced when it provides incentives for the

Leonard A. Smith
London School of Economics, Houghton St, London WC2A 2AE and Pembroke College, Oxford, e-mail: L.Smith@lse.as.uk

user ("it will increase my bonus") and not disincentives ("if this works, I will lose my job and a machine will replace me"). The FDE carried three incentives. First, the traders could quickly evaluate whether or not they could ignore it, today. Second, it targeted an acknowledged shortcoming in trading practice, one which both traders and their managers were aware of. Third, it could be fine-tuned (within limits) to better inform the interests of the particular trader; here the aims of traders and risk-managers in the same firm might diverge, again emphasizing the importance of incentives if one is to achieve a success story via improved mathematics.

2 Actors and Aims: A Broad Schematic of the Energy Sector

Different decision-makers have distinct goals and incentives; the inadequacies of available forecast systems will impact them differently. We personify several of these players in this Section. Traditionally, Charlie represents those who have to make binary decisions based on a probability forecast - say, whether or not to take down a coal-fired power plant for scheduled maintenance today or to keep it on-line another week given a forecast of heavy demand due to unseasonably hot weather. While the better choice for the company may be clear, the implications for Charlie himself, should something go wrong, introduce mixed incentives.

Charlize is trading energy in the forward market for a generating company, her aim is to have neither too much natural gas, nor too little when the day to burn it arrives. Charles is a day trader, looking for today's best bet. He has never considered taking delivery of whatever commodity he is trading today. Charlotte is the weather impacts officer of a national grid, and must consider mixed generation: wind and solar power, Combined Cycle Gas Turbine (CCGT) generation[1], and the security of long-lines transmission. Finally, Charlemagne is an international energy policy czar, aiming to improve the generation mix over 50 years, taking into account carbon dioxide emissions and the fact that future improvements in our ability to predict the weather alter his target. Charlie, Charlize, Charles and Charlotte are concerned with weather-like tasks: they can generate a forecast-outcome archive. Charlemagne's climate-like tasks involve more complex varieties of uncertainty[8].

Our focus will be on Charles and Charlize. Charles is often the competition for Charlize: given his aims it is quite reasonable to argue that a good forecast of "the next forecast for next Friday" is more valuable than a good forecast of the actual temperature next Friday. For concreteness, and with not too much loss of generality, suppose Charlize is trading natural gas to burn in the next three to twenty days. She is required (by her boss, or by law) to hold sufficient reserves to meet the demand of a large region; that bound is set by the most recent reference weather (point) forecast of her national Meteorological Office. Thus she can be required to buy, especially when the forecast shifts to indicate an increase in demand and the price rises. (Charles knows this, of course: in winter the price rises as the forecast tem-

[1] CCGT efficiency is a nonlinear function of temperature, humidity and pressure; the challenge of weather risk is significantly more complex than a one dimensional time-series problem.

perature falls. Ideally he does not know how much reserve she is already holding.) When a new day enters her time-window of responsibility, Charlize starts with some initial reserve of natural gas secured long ago. The size of this reserve was based on the climatological distributions of that calendar week, the risk tolerance or her company, the capital reserves of her company, perhaps a seasonal weather forecast, and without doubt a long range economic forecast. Where is her pain?

One source of pain amounting to annual losses in the hundreds of thousands of pounds, was from "chasing the forecast.". On Monday morning. Charlize arrives to learn that the reference forecast proclaims a week from Friday will be unseasonably cold. While many factors are at work in the energy market, the price will almost certainly go up. Tuesday morning, she sees the reference forecast jump back to near normal conditions. She knows the price will drop. Should she sell early or hold? (Or buy late!) Traditionally, she would sell. If in the next few days the reference forecast for next week drops back to low temperatures, then she would buy this same gas back, at a higher price. Even today's best singleton forecasts often move significantly from day to day when the target date is more than a few days ahead; arguably that might happen with a Perfect Model (if such a thing exists) given the uncertainty in today's observation systems. This loss-making "sell low, buy high, sell low, buy high" scenario is explicitly what the FDE was designed to relieve.

From Ensemble Simulation to Probabilistic Forecast Major weather centeres make not only one *singleton* simulation, but also an *ensemble* of lower-resolution (computationally cheaper) simulations. Combining the information from these simulations allows one to form a probabilistic forecast[2]. Consider forecasts for temperature, T, observed by a specific station x, at a particular time of day, at a lead-time τ days from now. Every probability forecast is conditioned on some set of information, I. The climatological distribution $P_\mu = P(T(x,t)|I_{historical})$ is a probability distribution for the temperature on a particular day of the year based on historical observations, independent of current meteorological conditions.

There are many ways to relate an ensemble of model-space simulations to a real-world observable. Kernel dressing and blending[2] obtains a predictive distribution via the sum of two terms, one from the simulations, P_{ens_1}, and the other from the climatology, P_μ. The forecast is then $P_{M_1}(T(x,\tau)|I^*) = \alpha P_{ens_1} + (1-\alpha)P_\mu$ where I^* is the union of information from simulations and from history. We expect $\alpha \approx 1$ for $\tau \approx 0$, and then to decrease towards zero. While many weather forecast systems prove informative out-of-sample, in my experience not one of them is reliable. This is due, in largest part, to structural model error[7, 3, 8]. If $P_{M_1}(T(x,\tau)|I^*)$ is not reliable, then the fraction of a collection of events with predicted probability near p which actually occur is statistically far from the target fraction p for most p . This is observed easily in binary forecasts (Bröcker and Smith [1]); full PDF forecasts have corresponding shortcomings quantified in their information deficit[3].

In actionable probabilistic systems the outcome is arguably a random draw from distribution. "Actionable" implies we can expect to apply all the tools of Decision Theory 301 profitably; forecasts are of course "subjective" but only in I.J. Good's [4] sense that they are conditioned on the information currently in hand. To be clear:

today's best weather forecasts are not actionable; probability-odds based on them are not sustainable in the sense that a cooperative market offering odds based on a probability distribution for the temperature at London Heathrow (LHR) a week from Friday could be driven bankrupt. There is enough information in the forecasts to out-perform the climatological probabilities, but not enough to offer alternative probability odds. Happily, building a tool to improve the lot of traders in Charlize's position does not require a complete solution of this wicked problem.

3 The Challenge: Chasing the Forecast

Charlize wants to avoid selling (or buying) today only to wish tomorrow that she had maintained her current position. Is there anything she really really wants which we might give her? She is not concerned about temperature fluctuations within a narrow range, ϕ, about the target reference forecast, but she would like to know if future reference point-forecasts for each day in the next few weeks (i) are "likely" to move outside the ϕ-range of today's point-forecast, (ii) are "likely" to fall above (or below) the ϕ-range if they are likely to fall outside it, (iii) are particularly "likely" to change drastically (unusually uncertain, given τ) or (iv) if something somewhere in the entire system appears not to be internally consistent (as in "broken"). The Forecast Direction Error tool (FDE) supplies just that information.

Every morning, she wants a quick idea of whether she has to worry about weather more than usual today. Glance at the Figure: green circles above and below indicate temperatures are expected to fall within their ϕ-ranges: no reason to worry has been detected, red circles indicate an alert that that day may be significantly warmer (the temperature will be above the ϕ-range), blue circles alert for colder. Charlize requests the target-probability thresholds θ which define the "likely's" in the previous sentences. The request is rejected if her forecast system cannot reliably supply this information: the FDE does not issue alerts by blindly interpreting $P(event|I^*)$ as a probability, but rather by determining the information in the forecast, if any . In practice, the FDE may detect days where the inputs are badly inconsistent with historical data and physical insight, suggesting that the forecast system is broken and the FDE cannot be relied upon. A purple light[2] flags this internal **in**consistency.

"The good Bayesian beats the nonBayesian but the bad Bayesian gets clobbered!" [4] Given an actionable probability forecast one could compute the forecast probability mass within the ϕ-range, and turn on green, red and blue lights for any θ. Meeting Charlize aims is straightforward given a forecast system[3] M^{IRO} that produced actionable probability distributions. Let $P_{in}(T_\tau|M^{IRO})$ be the probability mass within the ϕ-range at lead time τ, define $P_{above}(T_\tau|M^{IRO})$ and $P_{below}(T_\tau|M^{IRO})$

[2] The purple light was activated once in training almost surely due to an upstream keystroke error (12 degrees being entered as 21). In practice, traders requested "deep red" and "deep blue" lights.

[3] I.J. Good's Infinite Rational Org (IRO) has a PDF subjective only in the sense that it is conditioned on (all) the available information. Laplace's demon would as well, given quantum mechanics.

as the mass above and below. If $P_{in}(T_\tau|M^{IRO}) \geq \theta_{in}$ then we have a green day. If not, then turn on a red light if $P_{above}(T_\tau|M^{IRO}) \geq \theta_{above}$ and a blue light if $P_{below}(T_\tau|M^{IRO}) \geq \theta_{below}$ (in practice, both the red and blue lights do come on simultaneously now and then). Lastly, if the central ρ-range of the forecast distribution is unusually large, show it as a yellow band[4] In addition to an expected random flicker of yellow bars, The FDE discovered cascades of yellow bars: "uncertainty storms" of several days duration appearing in the long range and persisting as those target dates approach (note the yellow bars in the Figures lower panel). For one seeking risk, the volatility such storms suggest presents an interesting opportunity.

Given M^{IRO} one could supply Charlize with probability odds of: $O_{in} = 1/P_{in}$, $O_{above} = 1/P_{above}$ and $O_{below} = 1/P_{below}$ with $1/O_{in} + 1/O_{above} + 1/O_{below} = 1$. Model inadequacy suggests shortening the odds on each outcome. The implied probabilities would no longer add to one, but a cooperative market could then be sustained. It would use the information in today's weather forecasts to offer better odds than climatology without being driven bankrupt as a result of misinterpreting the forecast probabilities as actionable.

Interpreting actual simulation model-based forecasts as providing probability-odds can be a costly error. Some subjective Bayesians happily commit themselves to making this error; others (like Jim Berger and Susie Bayarani) acknowledge the challenges model inadequacy raises. While someone with better information can always expect to outperform you; the point here is that treating an inferior forecast as actionable places you **at risk from those with no additional information**.

Given an archive of past forecast-outcome pairs for the ϕ-range and a requested value for θ, it is straightforward to construct a reliability diagram (complete with internal consistency bars) which contrasts the relative frequency with which past probability forecasts saying the target temperature would fall in the ϕ-range actually fell in that range; allowances are made for the finite size of the sample and the fact that exactly the same P_{in} is unlikely to ever occur twice[1]. The forecast probabilities of today's best probability weather forecast systems are not consistent with the relative frequency of corresponding outcomes. Without knowing the nature of this inconsistency one can create two books, one betting a particular type of forecast is too high, the other betting that it is too low. One book will go to zero; the other will put the competitor out of business. It is for this reason that the use of model-based probabilities as probabilities is not recommended. There is more information in the forecasts than in the (reliable) climatological distribution, yet interpreting these forecasts via model-based probabilities can lead to ruin.

If weather forecasts cannot be interpreted as probability forecasts, then what good are they? It is straightforward to show that European Centre for Medium-range Weather Forecasts' (ECMWF) ensemble-based probability forecasts for next Friday have more skill [2, 5] than the climatology. This poses the fundamental dilemma facing anyone using probability forecasts of a physical system in the real world: one has two distributions, the first is a reliable probability forecast, the second is demonstrably not reliable yet contains more information than the first. If

[4] The threshold ρ-range at τ was typically set to the 97^{th} percentile of that season's distribution.

Charlize uses the climatology, then Charles will make money from her by using the imperfect ECMWF forecast. If Charlize trades according to the ECMWF forecast, Charles will use the shortcomings of the forecast to trade against her successfully. Resolving the Bayesian's dilemma is beyond the scope of this article; it requires giving up the equivalence of odds and probabilities, and moving to sustainable odds. As a first step we might identify FDE parameters for which even todays operational forecasts provide(d) alerts Charlieze found useful. In practice, of course, there need be no such set. Whereas if the probability forecasts were actionable, then all questions can be answered via Bayes Theorem and the Probability Calculus.

Define the capture rate κ as the fraction of events for which alerts were issued. Define the achieved rate γ as the fraction of alerts for which the event occurred. Consider a value of predictive distribution mass corresponding to an event merely as an informative threshold, π, **not** as a probability. Lower values of π correspond to higher capture rates κ, and perhaps lower achieved-rates γ. Given any "probability forecast" and target rate θ, one can determine whether or not there is a forecast threshold π of interest to a given user which has achieved θ historically.

To avoid chasing the forecast Charlize may be happy to hold onto a reserve (rather than sell it now and risk buying it back in the near future at a higher price) given even odds that today's reference point forecast is high. Charles, on the other hand, is looking to take on (the smallest possible) risk for a good shot at reward. He may want a significantly higher θ_{event}, generating many fewer signals, as long as it comes with a high γ. He is less concerned with a high κ if he prefers winning the bets he makes over making all possible winning bets. Charlize is deciding how to best manage her position (under constraints). The fact she wants to avoid taking a loss is reflected in the parameters she requests: Charles is looking for a signal to jump in (taking a good bet), Charlize is looking for a warning not too move (avoiding a bad bet). Charlize is in the game every day; Charles enters only when he wishes too. As both their boundary conditions and their risk appetite differ, so do their FDE parameters. Declaring forecast probabilities to be actionable allows any set of thresholds; using them is ill-advised if the forecast is not, in fact, actionable. No existing weather forecast with significant skill is reliable.

Charlize requires that the achieved rate be (at least) 50%, that (at least) half of the alerts result in events; note there is no such thing as a false alarm[5]. Given this constraint, does the forecast system offer a capture rate sufficiently high that the FDE is worth the time it takes to look at? For that last 4 years she has been chasing the forecast, and inasmuch as she has suffered during each and every event in that time, she may be happy with a relatively low capture rate. Experience suggests that relatively soon, however, she will want to start pushing up the capture rate. And she has the freedom to tune the FDE toward her desires within the constraints set by the (limited) information content of her forecast system. And that, of course, is how our work protects her from being clobbered: the limited skill of her forecast restricts the

[5] See Roulston and Smith[6]. If Charlize requested an alert given a 50% chance of an event, she cannot then reasonably complain if half of the alerts are not followed by events! It is a nonsense to balk at a system which does precisely what you asked it to do.

questions she can ask of it. The forecast-outcome archive is used to stop her from extracting more information from the forecast than it contains.

Defining Success Success in relieving a specific industrial discomfort exemplifies the goal many CATS projects have successfully targeted. The tunable knobs meant that the FDE could be customized to meet the aims of particular users where those aims were achievable given the information in hand. The FDE is informative both to traders and to risk managers: it aids those who wish to avoid chasing the forecast, and it also allows those with an appetite for risk the opportunity to find palatable instances. It would be interesting to consider how it might be generalized to inform and aid regulators as well. While there are no "false alarms" in a system that meets its design specifications[6], the fallacy in thinking that repeatedly taking action when nothing happened was "unnecessary" is difficult to overcome. Extracting just enough decisive information from simulations can prove of great value once it is overcome, not only in operations but also in the design of engineered systems given the uncertainty in the climate they will face.

The success of the FDE brought interesting questions into focus: How to better inform the diverse needs of regulators, traders, risk managers, and in-house meteorologists simultaneously? How is one to act rationally when the available probability distributions are not believed to be actionable? How to better identify and act when the purple light comes on, indicating that the probability of a big surprise is high? How to take economic cover in true uncertainty storms? These questions suggest that many opportunities for future success stories in mathematics are out there, waiting to be told.

Acknowledgements The FDE's success hinged on the engagement and enthusiasm of Dave Parker, chief meteorologist for EDF England. EPSRC and NERC grants supported the work of Jochen Bröcker, Liam Clarke, and Devin Kilminster. I am grateful for the support of Pembroke College, Oxford.

References

1. Bröcker, J. and Smith, L.A. (2007) 'Increasing the reliability of reliability diagrams', Weather and Forecasting, **22** (3): 651-661.
2. Bröcker, J. and Smith, L.A. (2008) 'From ensemble forecasts to predictive distribution functions,' Tellus A, **60** (4): 663.
3. Du, H. and Smith, L.A. (2012) 'Parameter estimation through ignorance', Phys Rev **E 86**, 016213.
4. Good, I.J. (1983) 'Good Thinking'. Dover, New York.
5. Hagedorn, R. and Smith, L.A. (2009) 'Communicating the value of probabilistic forecasts with weather roulette,' Meteorological Applications **16** (2): 143-155.
6. Roulston, M.S. and Smith, L.A. (2004) 'The boy who cried wolf revisited: the impact of false alarm intolerance on cost-loss scenarios', Weather and Forecasting, **19** (2): 391-397.
7. Smith, L.A. (2000) 'Disentangling uncertainty and error: on the predictability of nonlinear systems', in Mees, A.I. (ed.) Nonlinear Dynamics and Statistics, Boston: Birkhauser, $31-64$.
8. Smith, L.A. and Stern, N. (2011) 'Uncertainty in science and its role in climate policy', Phil. Trans. R. Soc. A, **369**, 1-24.

.

Side Caption FDE Screens. **The x-axis is lead-time, the y-axis temperature. The + symbols are the reference point forecast, the mauve colored vertical bar centered on the + is Charlize's (acceptable) ϕ-range. The top panel is what Charlize would have seen on 3 December 2005, the red disks on days 7, 8, 9, and 10 alert her to expect the temperature at LHR above her ϕ-ranges. The middle panel adds the outcomes "+". The lower panel is the FDE of 28 October 2003 for LaGuardia, New York. It shows the outcomes and the onset of an uncertainty storm at lead-time 7 to 10 days.**

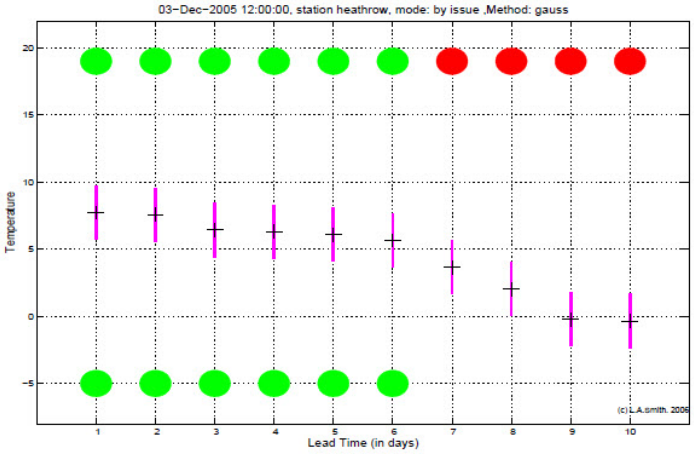

Figure 125: Daily FDE plots for ECMWF hi resolution forecast at LHR for two degrees error for likely success level 3 in 4 .

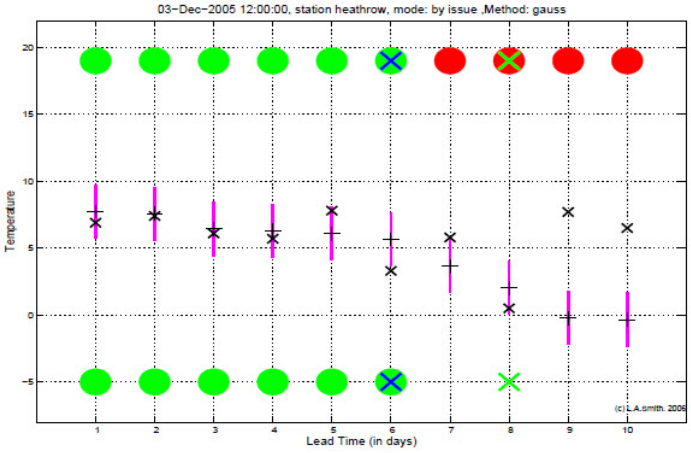

Figure 126: Daily FDE plots for ECMWF hi resolution forecast at LHR for two degrees error for likely success level 3 in 4 .

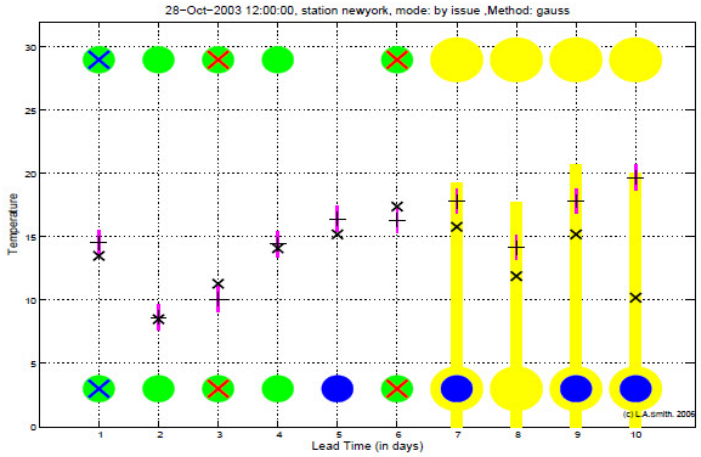

Figure 52: Daily FDE plots for NCEP hi resolution forecast at LAG for one degree error for likely success level 3 in 4 .