CAN WE TRUST LONG-RANGE WEATHER FORECASTS?

PASCAL J. MAILIER[*]
Swiss Re, United Kingdom

Abstract. Long-range weather forecasts are widely used in the energy industry, but too often their properties and limitations are not understood well enough. This chapter reviews the characteristics, methods and reliability of long-range weather prediction, and makes recommendations regarding its use. Despite their limited skill, long-range forecasts can still be a valuable tool for managing weather risk provided the necessary caution is exercised.

Keywords: Long-range weather prediction; seasonal forecast; weather risk management; energy weather; weather forecast quality

1. Introduction

What does 'long range' actually mean in weather prediction? In the energy sector, the appellation 'long range' commonly refers to time horizons of one to several years. In weather prediction however, the definition of 'long range' is based on the notion of atmospheric predictability. Figure 1 shows the skill of eight competing forecasts of daily average surface temperatures at one location (London Heathrow, United Kingdom) produced by purely atmospheric (i.e. the evolution of oceans is not predicted) numerical prediction models. In this case, skill is measured in % through comparing the accuracy (errors) achieved by the forecasts with that obtained using seasonal normal temperatures obtained from climatology through the period considered (6 months). Positive (negative) skill means that the forecasts are more (less) accurate than simple seasonal normal temperatures.

It can be seen that in this case none of the forecasts on Figure 1 has positive skill beyond day 9. With other weather variables like surface wind speed or daily accumulated precipitation, forecast skill usually drops faster than with daily average surface temperature. In all cases though, information

[*] To whom correspondence should be addressed: Dr Pascal Mailier, Royal Meteorological Institute of Belgium, Avenue Circulaire 3, 1180 Brussels, Belgium. E-mail: pascal.mailier@meteo.be

A. Troccoli (ed.), *Management of Weather and Climate Risk in the Energy Industry*,

on the initial state of the atmosphere at the start of the forecast fades away rapidly as a result of inexorable error growth and contamination. Because of this ‘memory loss’, exact knowledge of the atmosphere’s initial state becomes irrelevant after about 2 weeks or less depending on the degree of predictability of the situation, and as a result predictions of daily weather fluctuations produced by atmospheric models are not more accurate than long-term climatology.

However, persistent forcing from the Earth’s surface can have long-term effects on the *average* state of the atmosphere, like e.g.:

- The patterns of sea-surface temperature anomalies in the North Atlantic (North Atlantic Oscillation tripole), in the tropical Pacific (El Niño/La Niña), or in the North Pacific (Pacific Decadal Oscillation)
- The extent of snow cover over Eurasia

Some predictability of average weather conditions can therefore be gained by including these non-atmospheric factors in the forecast. The *long range* refers to time horizons from one to several months (e.g. seasonal) where average weekly or monthly weather conditions still enjoy some predictability. The *short* and *medium ranges* refer to time horizons of 1–2 days and from 2 days to 2 weeks, respectively. In these ranges, transient weather systems such as storms, fronts and anticyclones can be predicted by models, but uncertainty as to their intensity and timing increases rapidly. Because of this, it is more suitable to communicate forecasts using confidence intervals or probabilities, more particularly so in the medium range and beyond. The transition time window between 2 weeks and 1 month is often designated as *extended medium range*. Climate forecasts attempt to predict the response of the earth climate system to long-lasting environmental changes such as global increases in greenhouse gas concentration in the atmosphere and the depletion of the tropical rain forest. They look at time horizons from one to several decades. Although these forecasts are becoming increasingly relevant for long-term decisions in the energy sector (see e.g. EP2, 2008), climate prediction will not be considered in this chapter.

In order to correctly interpret and use the information provided by long-range forecasts, users of these products should be well aware of what makes them inherently different from the more common short- and medium-range forecasts. These differences are highlighted in Table 1.

The reader should always keep in mind that the purpose of long-range forecasts is definitely not to predict the weather that will be observed at some distant time in future (e.g. on some day next month) in future. Its goal is rather to enlighten the user on a range of plausible weather scenarios which

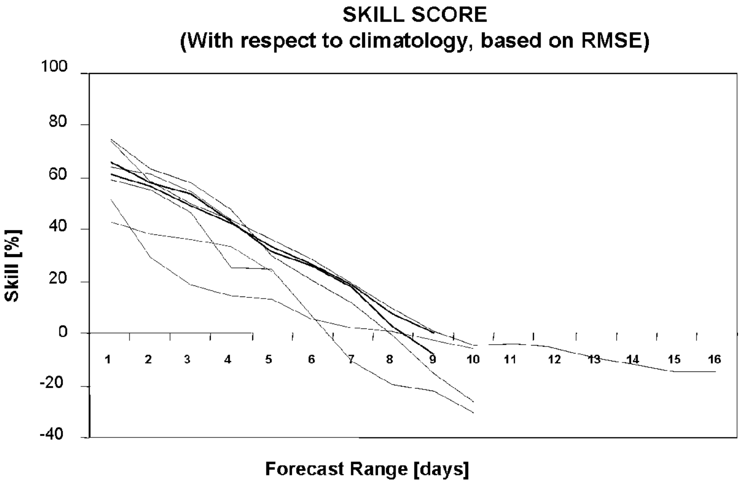

Figure 1. Skill scores of a set of eight competing forecasts of the daily average temperature at London Heathrow. Positive (negative) skill scores indicate that the forecasts are more (less) accurate than long-term climatology. Accuracy was measured by calculating root mean squared errors (RMSE).

are consistent with observed or projected patterns of temperature anomalies above the earth surface. In order to better appreciate the limitations of long-range forecasts, is also important to have a smattering of how such forecasts can be produced. The main techniques used in long-range weather prediction are discussed and exemplified in Section 2. Issues with long-range forecast communication and skill will be examined in Section 3 and Section 4 is devoted to conclusions and recommendations.

2. Outline of Methods for Long-Range Weather Prediction

There is no space in this chapter to discuss the many methods available in detail, but these can be classified in three basic categories: the method of analogues, statistical models and dynamical models.

2.1. THE METHOD OF ANALOGUES

2.1.1. *Method*

This method is the cheapest and quickest to realise, which explains why it is also the most popular approach to produce a view on the weather in the long-range. Basically, it consists of selecting past situations that were similar initially to what is currently observed and see what sort of scenarios unfolded in the weeks/months that followed. It can be seen as a 'naïve' form of ensemble forecasting using past observed scenarios as members. The method of analogues can somehow be paralleled with an experienced forecaster making inferences on future weather based solely on cases from the past and not on dynamical or physical thinking. Mechanisms like ocean-atmosphere interactions are not described statistically or explicitly by means of a model, but are believed to be included implicitly in the past scenarios themselves. This method is therefore essentially empirical. Of course, statistical techniques may be used to detect/enhance any interesting pattern(s) and/or summarise the results, e.g. cluster analysis (i.e. group scenarios into possible families) or extract mean, percentiles or anomalies from the distribution of scenarios.

TABLE 1. Comparison of short- and medium-range forecasts vs. long-range forecasts.

Short and medium range	**Long range**
Transient weather systems (e.g. storms, fronts, anticyclones) have some predictability	Transient weather systems are no longer predictable
Forecasts are able to pick up the day-to-day variability of weather variables (temperature, pressure, wind speed and direction, precipitation, etc.)	Forecasts predict overall/average conditions or a range of possible outcomes over an extended period of time (month, season)
Deterministic (one single scenario only) and probabilistic (distribution of possible scenarios, confidence intervals for predicted values of weather variables)	Should be probabilistic
The *surface* of the ocean is important, but its state evolves very slowly and does not change significantly over the forecast period (only the atmosphere does)	The state of the ocean changes and must be predicted *over a significant depth*. In order to achieve this, coupled ocean-atmosphere models are used instead of purely atmospheric models
Intensive quality control is made possible by the availability of frequent forecasts and observations, and by the existence of many established verification methods	Quality assessment is more problematic due to reduced forecast/observation frequency and less suitable verification methods
'Mature' operational models	'Young' models with limited track record, often experimental

The main criticism that can be made against the method of analogues is that usually it does not contain a scientific understanding on the mechanisms involved in the forecast. This absence of model constitutes a severe limitation to forecast improvement. Another weakness of the method is that it is heavily dependent on what is meant by ‘similar’. Furthermore, when selecting analogues, some balance must be found between two antagonistic constraints: the sample of analogues must be sufficiently large and at the same time the analogues must be close enough. The criteria used to choose the analogues as well as the sample size should always be mentioned when communicating the forecast.

2.1.2. *Example*

An example of forecast obtained through the method of analogues and its verification are presented in Figure 2. The divisional temperature dataset from the US National Climatic Data Centre (NCDC, 1994) was used in this case. The map on the left shows the distribution of mean surface temperature anomalies (in °F) predicted over the United States for the winter of 2008–2009 (from December to February, or DJF). This forecast was made at the end of August 2008 using 11 past cases with similar neutral to weak El Niño conditions between 1950 and 2007. The plotted mean anomalies suggest that, on average, such conditions are consistent with a cooler regime across the north and eastern half of the United States and possibly a warmer regime over central areas.

The map on the right shows the mean surface temperature anomalies that were actually observed during the winter of 2008–2009.

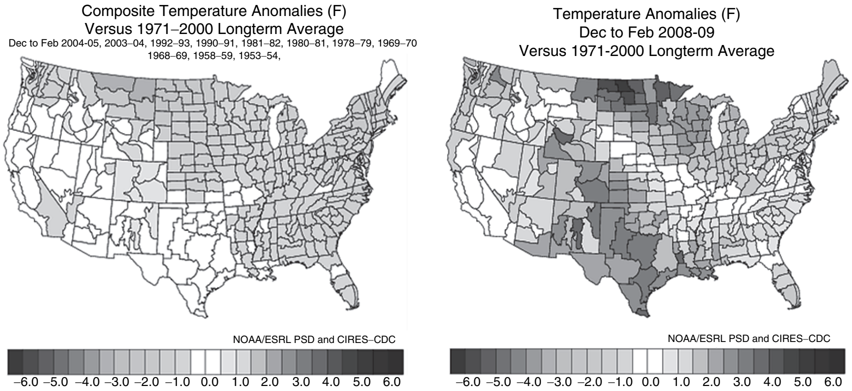
Composite Temperature Anomalies (F)
Versus 1971–2000 Longterm Average
Dec to Feb 2004-05, 2003–04, 1992–93, 1990–91, 1981–82, 1980–81, 1978–79, 1969–70
1968–69, 1958–59, 1953–54,
NOAA/ESRL PSD and CIRES–CDC
–6.0 –5.0 –4.0 –3.0 –2.0 –1.0 0.0 1.0 2.0 3.0 4.0 5.0 6.0
Temperature Anomalies (F)
Dec to Feb 2008-09
Versus 1971-2000 Longterm Average
NOAA/ESRL PSD and CIRES–CDC
–6.0 –5.0 –4.0 –3.0 –2.0 –1.0 0.0 1.0 2.0 3.0 4.0 5.0 6.0

Figure 2. Mean surface temperature anomalies over the continental United States as predicted by the method of analogues (*left*) and observed (*right*) for the winter of 2008/09. Source: NOAA/ESRL Physical Sciences Division.

The sign of observed anomalies was predicted correctly over the north, much of the northeast, and over some of the central and southern sections of the US. However, the sign of the anomalies was not well predicted over the south and much of the southeastern quadrant (expected negative, observed positive). Note that the predicted mean anomalies are often quite small compared to the observed. This is mainly due to the averaging process over all 11 cases, which smoothes out extremes. Users should be aware that forecasts obtained through averaging a number of scenarios (e.g. ensemble mean) typically underpredict significant anomalies. Because climate is not stationary, it is also important to specify which reference climatology has been used. In this case, the mean temperature anomalies were calculated relative to 1971–2000 long-term averages.

As mentioned earlier, a significant drawback of the method of analogues is that it is a kind of 'black box' that does not 'explain' the forecast. The forecasting methods discussed below use models to overcome this problem.

2.2. STATISTICAL MODELS

2.2.1. *Method*

Variations in average weather conditions can be forecasted quantitatively using statistical relationships between one or a set of several chosen explanatory variables (predictors) and a dependent variable to be predicted (predictand), e.g.:

- Use sea-surface temperature (SST) anomalies in the North Atlantic and/or the extent of the snow cover over the North-American and Eurasian continents to predict the state of the North Atlantic Oscillation (NAO) during the following winter (Rodwell and Folland, 2002; Saunders and Qian, 2002; Saunders et al., 2003).
- Use the states of the Atlantic Multidecadal Oscillation (AMO), the Quasi-biennial Oscillation (QBO) and El Niño Southern Oscillation (ENSO) to forecast the frequency and intensity of Atlantic hurricanes (Klotzbach, 2007).

The main advantage of statistical models is that they can offer a scientifically sound methodology to produce long-range forecasts that is still relatively cheap to develop, maintain and run. Another significant advantage is that many of these models are documented and discussed in the scientific literature. Their focus is mainly regional (e.g. Western Europe, North America).

The significance and physical interpretation of statistical relationships must be treated with particular care. For example, measures of association like correlation do not necessarily imply causality. A good statistical model should contain statistical relationships that reflect connections or links believed to take place between key physical processes.

A point worth noting about the statistical modelling approach is that it is essentially based on linear thinking whereas weather and climate processes are subject to non-linear interactions. Dynamical models, which are more suitable to deal with non-linearity, will be dealt with in the next subsection.

2.2.2. *Example*

Winter climate over the North Atlantic and European sector is modulated by a phenomenon known as the North Atlantic Oscillation (the NAO, see http://www.ldeo.columbia.edu/NAO/, for more details). In the United Kingdom, the Met Office has used a statistical model that uses the SST anomaly pattern over the North Atlantic in May to predict the average state of the NAO for the next DJF winter (Rodwell et al., 1999; Rodwell and Folland, 2002). In November 2005, the onset of a cold spell in Europe triggered a considerable rise in UK wholesale gas prices. The main factor which had made energy markets particularly sensitive was the expectation by the Met Office that a negative phase of the North Atlantic Oscillation (NAO) would favour colder-than-usual conditions in northwest Europe over the winter. The Met Office had stated that their system was able to correctly predict the sign of the NAO two times out of three, which is a slight advantage over random guesses. For instance, one might reasonably expect that a prediction based on tossing a coin will be correct roughly 50% of the time. The dashed line in Figure 3 shows all Met Office (UKMO) hindcasts/forecasts of the winter NAO index from 1948/49 until 2008/09 (61 consecutive DJF winters). The solid line shows the observed indices. It can be seen that the sign of the NAO index was correctly predicted for the winter of 2005/06 (larger circles), though the expected amplitude (–0.86) was twice as large as the observed (–0.42). Despite this apparent success, there are also some winters where the UKMO forecast fails badly. The predictions appear to follow the same trends (low-frequency signal) as the observations, but there is no convincing evidence that they manage to capture the year-to-year variability (high-frequency signal) of the observed indices. Therefore, the claim that the Met Office statistical model provides a useful forecasting advantage is questionable.

Alternative forecasts of the winter NAO index are shown in Figure 4. These forecasts were not produced using a statistical model, but much more simply by predicting the moving average of the observed NAO indices of the two most recent winters (MA-2). Mathematically, MA-2 is good for picking up the trends in observed NAO indices while being unable to realistically reproduce their variation from 1 year to the next. For the winter of 2005/06, the negative sign of the NAO index is also predicted correctly by MA-2, but more accurately (–0.24) than by UKMO.

The statistics presented in Table 2 compare the performances of the two forecasting systems over the period 1950/51–2008/09. The reader is referred to Jolliffe and Stephenson (2003) for an exhaustive discussion of the verification metrics used. The results suggest that the Met Office statistical model does not really provide a clear advantage because the forecasts it produces do not perform better than those obtained through a simple moving average.

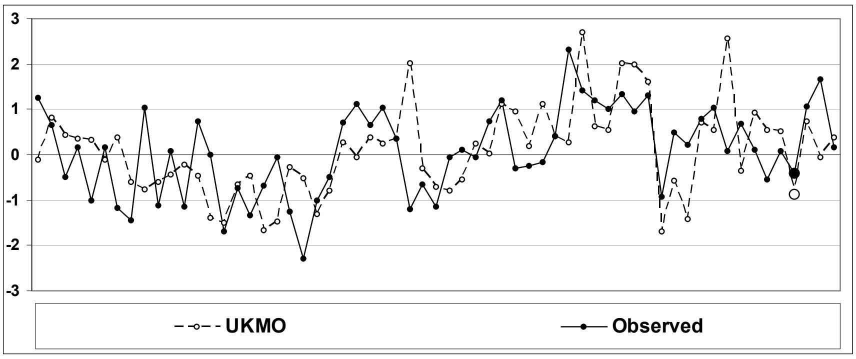

Figure 3. Time series of observed (solid line with black circles) and UK Met Office predicted (dashed line with white circles) winter NAO indices from 1948/49 (first on the *left*) until 2008/09 (last on the *right*). The larger circles to the right highlight the winter of 2005/06. Source: UK Met Office.

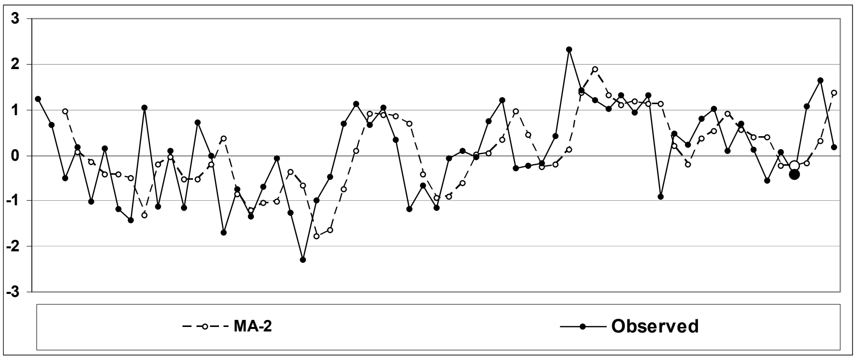

Figure 4. Time series of observed (solid line with black circles) and MA-2 predicted (dashed line with white circles) winter NAO indices from 1950/51 (first on the left) until 2008/09 (last on the right). The larger circles to the right highlight the winter of 2005/06.

TABLE 2. Performance statistics of the winter NAO index forecasts produced by the Met Office statistical model (UKMO) and by moving averages over two winters (MA-2). The scores that perfect forecasts should achieve are also indicated (Perfect) to facilitate interpretation.

Predicted attribute	**Verification metric**	**Perfect**	**UKMO**	**MA-2**
Sign	Proportion of correct forecasts	100%	68%	69%
Sign	Odds ratio skill score	1.00	0.63	0.68
Sign and amplitude	Mean squared error	0.00	1.07	0.89

2.3. DYNAMICAL MODELS

2.3.1. *Method*

This ‘number-crunching’ approach, which has been made possible thanks to the availability of ever more powerful supercomputers, consists of running numerical simulations of global coupled ocean-atmosphere models. These very complex models attempt to mimic the behaviour of the atmosphere-ocean system in a way that is consistent with the laws of physics. Because of all the technology and research efforts involved, this method is by far the most expensive. However, it also offers the greatest scope for improvements as models get more sophisticated. Much work has been done recently to obtain better simulations of key patterns such as El Niño and the Madden-Julian Oscillation. Long-range forecast models are mainly developed, run and maintained by national or international weather agencies in collaboration with academic institutions.

Given the considerable levels of forecast uncertainty present in the long range, producing one single forecast from one model does not make much sense. Instead, ensembles of forecasts are run, each individual member starting from slightly different initial conditions (different dates). Ensembles run from different models (a.k.a. multi-model ensembles or super-ensembles) like EUROSIP attempt to gauge the additional uncertainty due to model imperfection. The resulting forecast distribution provides quantitative information on forecast uncertainty that can be translated e.g. in probabilities or confidence intervals. The horizontal resolution of long-range forecast models is coarser (>100 km) than that of short- and medium-range forecast models (<100 km), so downscaling techniques are required for regional applications and extremes (e.g. weather generators).

Dynamical models and associated methods are extensively documented in the peer-reviewed scientific literature.

2.3.2. *Example*

Tercile probabilities are commonly used to summarise the forecast distribution in a way that is relevant to users in the energy sector. Each temperature forecast falls in one of three climatologically equiprobable categories labelled as ‘below normal’ (the lowest third of the climate distribution), ‘above normal’ (the upper third of the climate distribution), and ‘normal’ in-between. Probabilities of terciles can then be estimated from the proportion of forecasts counted in each of these categories. The seasonal forecasting system of the European Centre for Medium-range Weather Forecasts (ECMWF) routinely produces probabilistic forecasts of terciles for the monthly mean temperature out to a horizon of 7 months. The performance of these forecasts over Southern England has been gauged for all three terciles from a set of 252 hindcasts. The statistic used in this case is the ROC skill score (ROCSS). This metric measures the overall ability of the forecasts to discriminate between event and non-event (maximise the hit rate and minimise the false alarm rate). The score is 1 for perfect forecasts (hit rate of 100% and false alarm rate of 0%), and 0 for forecasts that are not more informative than climatology (i.e. always forecasting a probability of 33% for each tercile). Once again, the reader is referred to Jolliffe and Stephenson (2003) for more details on the ROCSS. The results shown in Figure 5 indicate some modest skill in the first month, which dwindles rapidly at longer lead times.

3. Current Issues with Long-Range Weather Forecasts

The examples given in Section 2 prompt to some issues with the communication of long-range weather forecasts as well as in the interpretation of their skill.

3.1. COMMUNICATION OF FORECASTS

Weather forecasts should always be presented to users in formats that are fit for purpose. Long-range forecasts are inherently uncertain and so the level of confidence that can be placed in them constitutes crucial information to the decision maker. Forecast uncertainty can be conveyed to users by means of confidence indices, confidence intervals and probabilities. The forecast and its uncertainty must preferably be quantified so that the user can assess their quality in an objective manner. In spite of this, a large number of long-range forecast products available on the markets are based on a single consensus scenario (e.g. the case presented in Section 2.1) with no or little verifiable information on uncertainty or on alternative scenarios. Users

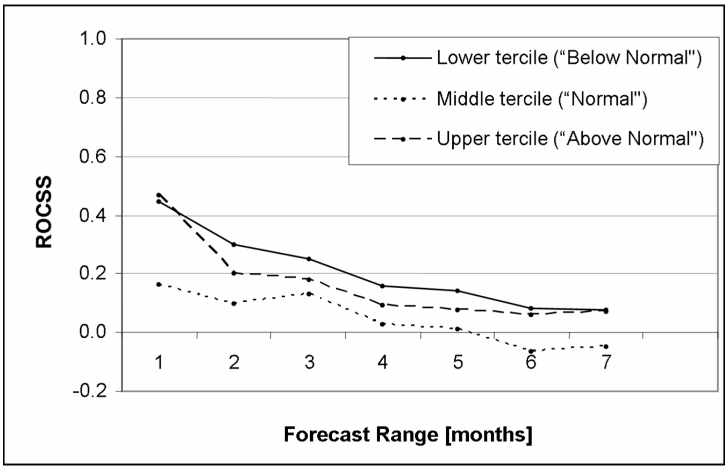

Figure 5. ROC skill scores of a set of 252 ECMWF probabilistic seasonal hindcasts of terciles for the monthly mean temperature over Southern England out to 7 months.

should also keep in mind that consensus scenarios are typically obtained through averaging, so they tend to under-forecast or even remove significant events that can be detected in the scenarios that have been averaged to produce the consensus.

In a risk management perspective, probabilistic products offer more value because they allow users to treat forecast uncertainty as information that allows economically optimal decisions (Jolliffe and Stephenson, 2003, Chapter 8). Nonetheless, many users may still be deterred by the difficulty to understand and process probabilistic information, by the negative connotation of probability implying ignorance, and by some reluctance to transfer the ‘Yes/No’ decision stage from the weather forecaster to the user (Mailier et al., 2008).

3.2. SKILL OF LONG-RANGE FORECASTS

In 2006, the hedge fund Amaranth lost $6 billion and collapsed after speculating wrongly that a very active hurricane season would disrupt the US oil production in the Gulf of Mexico with soaring natural gas prices as a result (Dealbreaker, 2006). Their bet was based on long-range forecasts published in December 2005 and April 2006 by Klotzbach and Gray (2005, 2006). The high level of confidence placed in their predictions was due to Dr Gray’s successful forecasts in 2002, 2003, 2004 and 2005. This story along with the case discussed in Section 2.2 illustrates how forecast performance results can be misinterpreted. The example of Subsection 2.2 also recalls that a forecasting system deemed skilful by meteorologists may turn out to be less attractive for practical applications. This aspect is too often neglected in the weather forecasting industry. For instance, predictions based solely on ensemble means are popular as they tend to score best in terms of accuracy because they minimise forecast errors. However, they smooth out potentially crucial features like extreme events.

Finally, the example of Subsection 2.3 demonstrates that the typical skill of long-range weather forecasts is not particularly high. All the methods tend to perform best in situations with strong persistence in sea-surface temperature anomalies (e.g. El Niño/La Niña), which is not always the case.

4. Conclusion

Because of the chaotic nature of the atmosphere, long-range weather forecasts will never achieve the same level of detail and confidence as short- and medium-range forecasts. Users in the energy sector must take this fact into account so that they can adjust their trust and base their strategies on realistic expectations. Long-range forecast products based on obscure (unpublished or undisclosed) methods should always be treated with suspicion.

Any claim of skill should not be taken at face value. The meaning of "skill" is strongly user-dependent and metrics used to assess skill should be chosen carefully so that they are appropriate for the user application. Even when the methods of forecasting appear to be scientifically sound, useful skill is often modest in the mid-latitudes, particularly so in Europe. However, the limited reliability of long-range forecasts should not discourage their users in the energy sector. Indeed, careful usage of these forecasts can still make them a valuable tool for managing weather risk, and meteorologists need feedback from energy users in order to make these forecasts more useful for industrial applications.

ACKNOWLEDGEMENTS

The author wishes to thank the NATO Science for Peace and Security Committee for funding this Advanced Research Workshop, and Swiss Re for allowing him to participate in this event as co-organiser and contributor. Dr Adam Scaife of the UK Met Office, is acknowledged for kindly providing the winter NAO index dataset.

References

Dealbreaker, 2006, *Who Killed Amaranth? Academic Weather Forecasters Admit They Called 2006 Wrong*, http://dealbreaker.com/2006/10/who-killed-amaranth-academic-w.php

EP2, 2008, http://www.metoffice.gov.uk/climatechange/businesses/casestudies/energy.html

Jolliffe, I.T., and Stephenson, D.B. (Eds.), 2003, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Wiley, Chichester, 254 pp.

Klotzbach, P.J., 2007, Recent developments in statistical prediction of seasonal Atlantic basin tropical cyclone activity, *Tellus A*, **59**:511–518.

Klotzbach, P.J., and Gray, W.M., 2005, *Extended Range Forecast of Atlantic Seasonal Hurricane Activity and U.S. Landfall Strike Probability for 2006*, http:// hurricane.atmos.colostate.edu/Forecasts/2005/dec2005/

Klotzbach, P.J., and Gray, W.M., 2006, *Extended Range Forecast of Atlantic Seasonal Hurricane Activity and U.S. Landfall Strike Probability for 2006*, http://hurricane.atmos.colostate.edu/Forecasts/2005/april2006/

Mailier, P.J., Jolliffe, I.T., and Stephenson, D.B., 2008, Assessing and reporting the quality of commercial weather forecasts, *Meteorol. Appl.*, **15**:423–429.

NCDC, 1994, Time Bias Corrected Divisional Temperature-Precipitation-Drought Index, Documentation for dataset TD-9640, available from DBMB, NCDC, NOAA, Federal Building, 37 Battery Park Ave. Asheville, NC 28801-2733, 12 pp.

Rodwell, M.J., and Folland, D.P., 2002, Atlantic air-sea interaction and seasonal predictability, *Q. J. R. Meteorol. Soc.*, **128**:1413–1443.

Rodwell, M.J., Rowell, D.P., and Folland, C.K., 1999, Oceanic forcing of the wintertime North Atlantic Oscillation and European climate, *Nature*, **398**:320–323.

Saunders, M.A., and Qian, B., 2002, Seasonal predictability of the winter NAO from north Atlantic sea surface temperatures, *Geophys. Res. Lett.*, **29**(22):2049, doi:10.1029/2002GL014952.

Saunders, M.A., Qian, B., and Lloyd-Hughes, B., 2003, Summer snow extent heralding of the winter North Atlantic Oscillation, *Geophys. Res. Lett.*, **30**(7):1378, doi:10.1029/2002GL016832.