University of Reading
Department of Meteorology

Maximising the Value of Medium Range Weather Forecasts for Predicting Gas Demand

Markus L. Rasswallner

Submitted: August 15th 2005

A dissertation submitted in partial fulfilment of the requirements of the degree of Master of Science in Applied Meteorology.

Acknowledgements

First and foremost I would like to thank my beloved parents for supporting me in many ways, not least financially, during my year at Reading.

A special thank you goes to my dissertation supervisors, Dr. Warwick Norton and Prof. Alan O'Neill, for their efforts, guidance and trust at all stages of this project. I am particularly grateful to Dr. Norton for taking the time to teach me IDL from scratch, for the many useful ideas and comments on my work, and for putting up with me. In this context, I wish to express my sincere gratitude to Prof. O'Neill's and Dr. Norton's company, Weather Informatics Ltd., for allowing me to use their extensive archive of past forecast data, and for providing me with technical insights into the production of user-specific weather forecast products.

I am grateful to RWE Transgas, a.s., especially Dr. Frank Starrmann, for investing considerable time and giving me the opportunity to research a concrete and highly relevant application of meteorology.

Last but by no means least I would like to thank my friend and mentor, Mr. John Thompson, for his invaluable advice and encouragement all the way.

Without the contribution of all the above persons and organisations, this project would not have been possible.

Abstract

The skill of medium range (3-16 days) weather forecasts has increased dramatically in the past decade or so. However, the focus amongst the meteorological community has primarily been on improving the forecasts of certain meteorological variables in their own right. Whilst this approach is perfectly valid from a scientific point of view, end users mostly do not judge a weather forecast in terms of meteorological criteria, but in terms of how it will aid their decision-making process, and ultimately, how much money the forecast will make or save them. The value of a forecast may differ radically between users, owing to their individual needs, response-variable models and decision-making processes. Hence, an integrated end-to-end approach of forecast development, as well as skill and value assessment from the viewpoint of the end user is essential. Motivated by these issues, this study applies medium range temperature forecasts produced by the European Centre for Medium Range Weather Forecasts (ECMWF) and the National Centers for Environmental Predictions (NCEP) ensemble systems in end-to-end forecasting and decision-making processes of a specific user in the highly weather-dependent natural gas industry in Prague, the Czech Republic. Both user-specific skill and value of the forecasts are assessed. This study also innovatively contributes to the general issue of forecast calibration research. It is the first study to analyse forecast data from two different numerical weather prediction models, both comparatively as well as jointly, by extending the multi-model concept used in other areas of meteorology to location-specific medium-range temperature and user-response-variable forecasts. It applies commonly used post-processing methods, which have mostly been developed and tested exclusively on London Heathrow temperatures, to a different location. The following key findings were made: Ensemble forecasts add considerable skill in deterministic predictions of gas demand. The ECMWF's Gain over climatology in the 2004-2005 heating season is 3.8mil m^3 (~14.1% of mean daily weather-dependent demand) at a lead time of 1 day and 1.8mil m^3 at a lead time of days 10. Employing multi-model methods further enhances deterministic skill by up to 0.2mil m^3 up to a lead time of 7 days. Commonly used post-processing methods applied at other locations do not enhance the deterministic skill of Prague temperature and gas demand forecasts. Deterministic forecasts produced from the ECMWF ensemble add value to the decision-making process of gas demand nominations, especially at longer lead times, generating profits of almost 1 mil. currency units at lead 1, and around 62 mil. currency units at lead 10 over the 2004-2005 heating season. Probabilistic information contained in the raw ensemble spread adds only limited additional value to this specific application (approximately 100000 currency units at lead 1, and 340000 currency units at lead 10). Future research into forecast calibration is essential to determine whether this can be further increased.

Table of Contents

Chapter 1: Introduction

The skill of medium range (3-16 days) weather forecasts has increased dramatically in the past decade or so (Rodwell and Doblas-Reyes, 2004), especially with the improvement of dynamical models, greater computing power, and the advent of ensemble prediction techniques. As Palmer (2002) notes, the focus amongst the meteorological community has primarily been on improving the forecasts of certain meteorological features (e.g. 500mb geopotential height) in their own right, and measuring increases in forecast quality purely with regard to these variables.

Whilst this approach is perfectly valid from a scientific point of view, it does not necessarily aim to optimise, nor quantify the value a specific forecast has to an end user. In addition, Jewson (2004b) points out that forecast quality has mainly been measured with metrics devised and used by meteorologists, though not broadly known to or accepted by forecast users who pertain to other disciplines (e.g. economics). Thus, this does not enable a decision-maker in a weather dependent industry to determine if or how a specific forecast product can improve a commercial decision, which could make or lose the company millions. End users mostly do not judge a weather forecast in terms of meteorological criteria, but in terms of how it will aid their decision-making process, and ultimately, how much money the forecast will make or save them. Hence, what is required is an integrated end-to-end approach of forecast development and value assessment from the viewpoint of the end user.

Fulfilling this aim requires identifying the needs of the specific user, developing a tailored forecast product, and demonstrating how the product can be applied to provide value to the user. This necessitates close quantitative cooperation between forecast providers and forecast users. Forecast providers must understand the weather-sensitivity and decision-making process of the user in order to produce and format the forecast in a manner that will yield the maximum benefit to the user at an acceptable cost, whilst users must understand how to apply and interpret the forecast. Though seemingly simple, this can prove to be difficult. In some cases users do not even know their precise quantitative exposure to weather, let alone how to relate forecast information to their decision-making process. This is especially the case with probabilistic forecasts (Palmer, 2002). However, the greatest obstacle is that owing to the pivotal importance of weather to some industries, their quantitative weather exposure and decision-making processes are confidential and thus not accessible to forecast providers.

Whilst much effort and funding in the meteorological community is allocated to improving numerical weather prediction models, there is a widely held view that large amounts of untapped value lie in the output of currently available forecast systems (O'Neill pers. com., Palmer, 2002). Hence, the initial onus should be on analysing, post-processing and tailoring output of existing models with regard to user applications. Only then should conclusions be drawn on how to improve numerical weather prediction systems to benefit the user.

Motivated by the points made above, this study applies medium range temperature forecasts produced by the European Centre for Medium Range Weather Forecasts (ECMWF) and the National Centres for Environmental Predictions (NCEP) ensemble systems in end-to-end forecasting and decision-making processes of a specific user in the highly weather-dependent natural gas industry. The end-to-end approach is crucial for two main reasons:

1. The forecast user wants to know the goodness of a weather forecast in terms of predicting a response-variable - the actual variable of interest.
2. The user's models of temperature response-variables as well as his or her decision-making models may incorporate lags, non-linearities and threshold. Therefore, the value of a temperature forecast cannot simply be inferred form a standard skill assessment or linear scaling of the meteorological forecast.

In the case of gas demand, well-developed statistical models relating temperature, amongst other meteorological and non-meteorological variables, to gas consumption already exist (e.g. van den Berg, 1994). Hence, the weather-dependency of gas demand is already known. In addition, well-defined decision-making processes are also established. Hence, the key question at this point is how to best integrate weather forecasts into the user's demand forecasting and decision-making processes.

In this study, raw and post-processed temperature forecasts for Prague Ruzyne Station are used as an input to the company's gas demand function to produce deterministic and probabilistic end-to-end ensemble gas demand forecasts for the Czech Republic. This was made possible by the unique situation of both a sufficiently long archive of past temperature forecasts, provided by Weather Informatics Ltd., as well as a realistic gas demand function being made available by RWE Transgas.a.s., for the purpose of this study. The two principal aims are to determine and enhance forecast skill in predicting gas demand as well as forecast value when used in the company's economic decision-making process, including the probabilistic information contained in the ensemble spread. The latter aspect is explored by an approach based on von Neumann and Morgenstern's (1944) economic utility theory. This was made possible by having access to a realistic economic utility function used by the gas company. The study therefore sets out to answer the following questions:

1. What is the deterministic skill of the ensemble mean of an end-to-end gas demand forecast using raw ECMWF and NCEP temperature forecasts?
2. Can the deterministic skill be improved by post-processing methods?
3. Can the deterministic skill be improved by combining NCEP and ECMWF forecasts?
4. What is the economic value of using raw and post-processed deterministic and probabilistic forecasts in a decision-making process?

Apart from integrating temperature forecasts into a realistic user-specific and user-defined application, rather than using imaginary problems and scenarios devised by meteorologists, this study also innovatively contributes to the general issue of forecast calibration research. As far as the author is aware, it is the first study to analyse forecast data from two different numerical weather prediction (NWP) models, both comparatively as well as jointly, by extending the multi-model concept used in other areas of meteorology to location-specific medium-range temperature and user-response-variable forecasts. This study also applies commonly used post-processing methods, which, in the literature, have mostly been developed and tested exclusively on London Heathrow temperatures, to a different location.

Chapter 2 reviews the underlying concepts of ensemble forecasting and multi-model forecasting.

Chapter 3 summarizes the climate of Prague with regard to temperatures.

Chapter 4 describes the modelling of gas demand and the methods employed in this study to produce an end-to-end demand forecast.

Chapter 5 reviews the post-processing methods that are employed to calibrate and combine the temperature forecasts to produce a deterministic best estimate, as well as presenting the metrics used to assess the deterministic skill of end-to-end gas demand forecasts. Thereafter, results of the deterministic skill assessment of end-to-end demand forecasts for the Czech Republic using raw temperature forecasts, as well as three bias correction and two multi-model methods are discussed.

Chapter 6 reviews methods for generating probabilistic temperature and end-to-end demand forecasts, as well as for calibrating the ensemble spread.

Chapter 7 discusses the concept of economic value of weather forecasts and proposes the use of a method based on von Neumann and Morgenstern’s economic utility theory to distil and assess the value of probabilistic forecasts in a decision-making process. Thereafter, the results of a comparative value assessment of deterministic and probabilistic forecasts in the context of a decision-making process of the gas company are discussed.

Chapter 8 draws conclusions from the study and suggests avenues to be explored in future research.

Chapter 2: Principles of ensemble and multi-model forecasting

In this study, temperature output from the European Centre for Medium Range Weather Forecasting's (ECMWF) Ensemble Prediction System (EPS) (Persson, 2001) and the National Centers for Environmental Prediction (NCEP) Medium Range Forecast (MRF) (Kalnay and Toth, 1996) are used. They were chosen since they are the pre-eminent operational ensemble systems currently in use[1]. The EPS is widely believed to be the world's leading medium range forecast system at present, owing to its superior modelling and data assimilation capabilities, as well as its large ensemble size (Buizza et al., 2005). Furthermore, the fact that an extensive archive of past forecasts was made available for this study represented an opportunity for a project investigating the socio-economic application of ensemble forecasts.

Both the ECMWF's EPS and NCEP's MRF temperature forecasts are produced by Atmospheric General Circulation Models (AGCM), which use discrete numerical methods to solve the governing equations of atmospheric flow. The atmosphere is divided into a three-dimensional grid. For each gridpoint some variables, such as temperature, pressure, wind velocity and humidity are directly calculated, whilst other sub-gridscale variables and processes, such as clouds and rainfall, are parametrised (McGuffie and Henderson-Sellers, 1999). A comparative summary of the main specifications of the NCEP and ECMWF models is given it table 1.1. To obtain a point-specific forecast of temperature, model output must be downscaled to the location of interest. Several methods ranging from simply using raw model output to computationally intensive optimal weighting of output at surrounding gridpoints related to synoptic conditions and the location's climate (e.g. kriging), have been proposed (Hervada-Sala et al., 2000, Gutierrez et al., 2004). Owing to the limited scope of this study, only the most commonly used method, linear interpolation of the 2m temperature model output of the four nearest gridpoints to the location of interest, is employed in the case of Prague Ruzyne.

[1] For a summary of the most up-to-date developments in ensemble forecasting, the reader is referred to the WMO website. Currently, the Japan Meteorological Agency and the Canadian Meteorological Centre also produce ensemble forecasts. The latter uses different model versions and assimilation processes to generate an ensemble, rather than simply perturbing initial conditions (Lefaivre et al.,1997). If archived forecast data were to become available, the potential benefits of using these models should also be investigated in future.

	ECMWF	NCEP
Horizontal Resolution	T255 ~80km	T126 ~160km
Vertical Levels	40	28
Temporal Resolution	15 mins	20 mins
Lead Time	10 days	16 days
Ensemble Size	51	12
Method for generating Initial Conditions	Singular Vectors	Bred Vectors

Table 1.1: Summary of some key characteristics of the ECWMF and NCEP ensemble systems (adapted after Buizza et al., 2005 and Toth et al., 2004).

Owing to the chaotic nature of the atmosphere (Lorenz, 1969), the two main sources of error in medium-range forecasting are model error and uncertainty in initial conditions. Initial conditions represent the best-estimate of the actual state of the atmosphere at the time the forecast is initialized and are the product of observations and other shorter model runs conducted in the data assimilation process. Despite the great effort and expense devoted to the assimilation process (Lyster et al., 2004), uncertainty remains in the initial conditions, because of the paucity of and errors in observations. Due to the non-linear nature of atmospheric processes, small initial errors can grow rapidly with time and affect large scale flow, if they occur in sensitive areas of the atmosphere (Palmer, 2000). Hence, even a forecast using a perfect model of the atmosphere would, in the course of a few days, be spoiled by errors in the initial conditions.

Ensemble forecasting addresses the problem of uncertainty in initial conditions by using samples within the range of uncertainty to initialize several model runs. For the EPS (Persson, 2001), a set of 50 initial conditions is generated for the northern hemisphere by adding small perturbations to the analysis within the limits of uncertainty. These, as well as the unperturbed analysis are used to initialise the 51 members of the ensemble, which are then integrated forward in time to produce a set of possible future states of the atmosphere. These perturbations are calculated by the singular vector technique, which identifies the regions of greatest dynamical instability, in which errors in the initial conditions would lead to maximal forecast divergence (based on a 48-hour model integration) (Buizza and Palmer, 1995). Thus, initial conditions are not chosen at random in a statistical simulation, but in a manner that will result in the maximum growth in ensemble spread (Palmer, 2002). Hemispheric structures which could generate significant forecast divergence are produced from the leading 25 singular vectors. These perturbations are then mirrored by reversing their signs to give the total of 50 perturbations (Persson, 2001). In addition, the ECMWF employs stochastic perturbations of the model physics (Palmer, 2002). According to Persson (2001), the skill of the EPS over Europe is largely determined by the ability of the system to identify uncertainties in upstream baroclinic development and model alternative development scenarios. For predicting European weather on timescales of a few days, for example, initial conditions over the Western Atlantic and US East Coast are important, whilst for timescales of a week or so, initial conditions over the Western US and North-Eastern Pacific are crucial.

Initial conditions for the NCEP model are generated using the so-called breeding method. For a more detailed discussion the reader is referred to Toth and Kalnay (1997).

2.1 Benefits of ensemble forecasting for temperature forecasts

One aim of ensemble forecasting is to produce a better estimate of the most likely outcome by using the ensemble mean as a deterministic forecast rather than a single integration (Leith, 1974). When considering the allocation of limited computing time, a compromise has to be struck between augmenting model resolution and augmenting ensemble size. Taylor and Buizza (2004) observed that the best estimate derived from the EPS 51 member (T255 L40) ensemble mean showed higher skill than the ECMWF's single high resolution (T511 L60, i.e. with a resolution of around 40km and 60 vertical levels) integration at all lead times. This was especially the case at longer leads. It must be noted, though, that other studies (e.g. Mailier, pers. com.) found the skill of the single T511 integration to be higher than that of the T255 ensemble mean at short leads (up to around 2 days). Hence, it appears that a better representation of the atmosphere (given by a higher resolution model) is comparatively more important at short lead times than sampling errors in initial conditions.

Figure 1.1, produced in this study, shows the plot of EPS Prague temperature Root Mean Squared Error (RMSE) of the ensemble mean, the unperturbed control member and perturbed members 1-4 against lead time (days). The ensemble mean exhibits the lowest RMSE at all lead times. At leads 1-3 the control member is as skilful as the ensemble mean. However, with increasing lead time, the skill of the former decreases more rapidly.

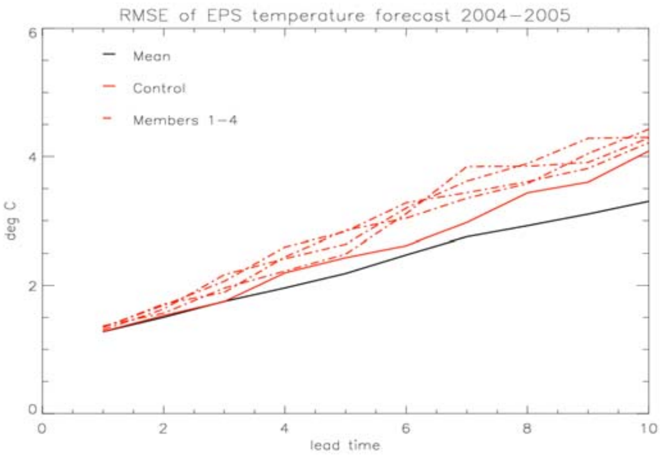


Figure 1.1: Root Mean Squared Errors of the 2m temperature forecasts for Prague of the EPS ensemble mean, control and perturbed members 1-4 against lead time (days) for the period October 2004 – March 2005.

Although the control integration was run at the same resolution as all other members, (data from the T511 L60 run were not available to this study), this shows that using the ensemble mean increases skill vis-à-vis a single integration. The skill of the individual perturbed members is lower than that of the control run at all lead times, which brings into question the assumption (Persson, 2001) that all ensemble scenarios are equally likely. Mailier (2001) argues that giving the control integration a higher weighting could be justified, since it is initialised with the best estimate of initial conditions.

A further benefit of ensemble forecasts, which a single deterministic forecast is by definition not able to give, is a qualitative as well as quantitative estimate of the flow-dependent uncertainty in the forecast (NCEP Ensemble Homepage). Whilst probabilistic forecasts can be generated using a deterministic forecast and adding a historical error distribution, this does not enable forecast uncertainty estimates to vary with atmospheric state. However, this information may be vital for a forecast user who is severely affected by extreme temperatures that occur during rapid transitions. The divergence of the ensemble members is dependent on the sensitivity of the atmosphere to slightly varied initial conditions at the point in time in question. For this purpose, information contained in the ensemble spread (e.g. frequency distribution of the ensemble members) can be used to estimate a probability density function of temperature (Taylor and Buizza, 2004). Figures 1.2a and 1.2b illustrate a qualitative example of high and low certainty for Prague 2m temperatures, respectively.

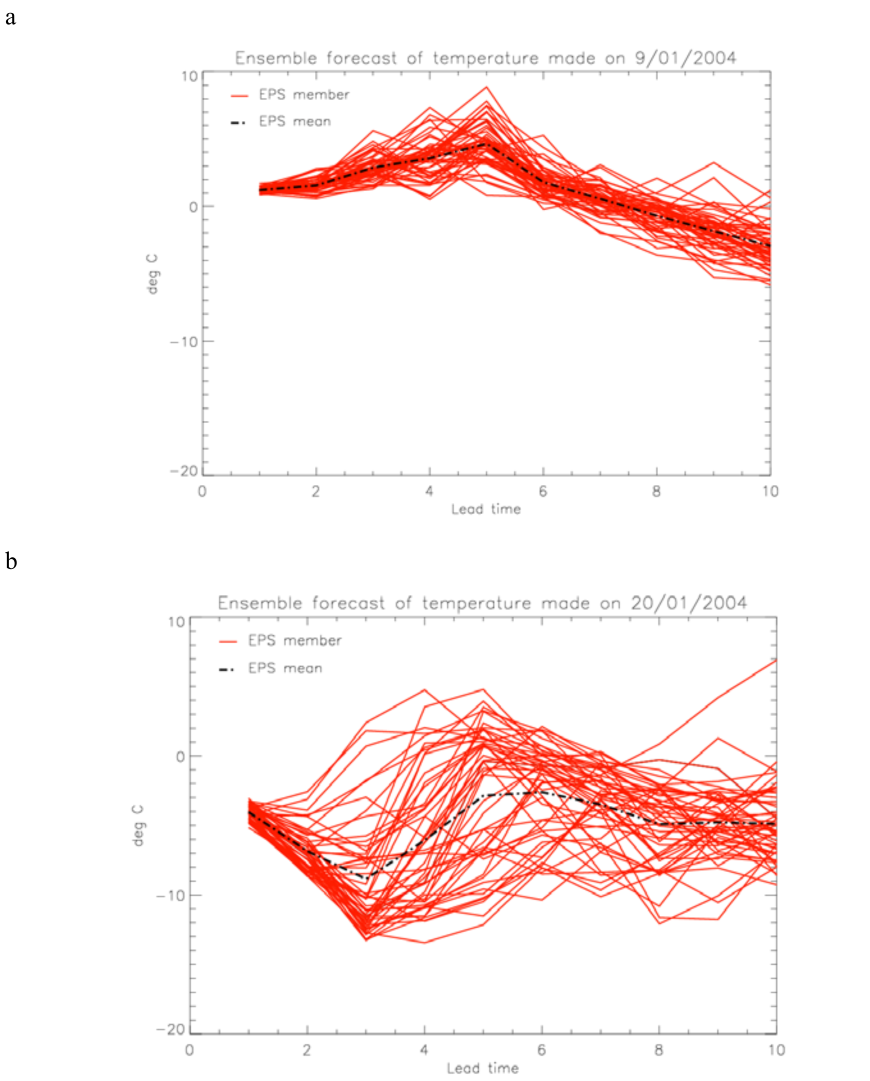


Figure 1.2: ECMWF Ensemble forecasts of Prague daily mean temperatures. a: Produced on 09.01.2004 - the narrow spread of the ensemble indicates high certainty in the forecast. b: Produced on 20.01.2004 - the wide spread of the ensemble indicates low certainty in the forecast.

However, caution must be taken when interpreting the ensemble spread in a quantitative fashion. Several studies (e.g. Toth et al., 2003) have noted that recalibration of the spread is necessary in order to gain reliable probabilistic information. This is mostly due to the ensemble spread of both systems tending to be too narrow (observations at times fall outside the range of possible outcomes predicted by the ensemble), indicating that not all of the uncertainties in the forecast, especially with regard to model error, are incorporated (Buizza et al, 2005). Toth et al. (2003) found that this underestimation of uncertainty increased with lead time.

2.2 Multi-model forecasting

The aim of multi-model forecasting is to reduce the second major source of error in numerical weather prediction - model error - by combining two or more models, which show similar levels skill on their own, but have different dynamics and physics. These technical differences tend to result in different strengths, weaknesses and biases in different synoptic conditions and geographical areas. Just like ensembles sample uncertainty in initial conditions, multi-model techniques sample model uncertainty.

Mylne et al. (2002) showed that combining United Kingdom Meteorological Office's (UKMO) Unified Model (UM) and ECMWF EPS model increased both deterministic skill of the ensemble mean (assessed by RMSE) and probabilistic skill (assessed by Brier score; Brier, 1950) at times when one of the two models performed poorly. This was attributed to the fact that the combination of the models was producing solutions synoptically more different than each individual system. The multi model ensemble almost always performed as well as the best individual ensemble and on occasions better than either of them. At times when the EPS performed well, adding ensemble members from an ensemble performing less well did not degrade overall skill. The relative performance of the two ensemble systems varied from day to day. Apart from increasing ensemble size, the ensemble spread was also more representative – more so than it was from increasing the number of ensemble members of the individual models.

The benefits of multi model ensembles were found to be flow-dependent and to vary in time and geographically. The multi model technique was found to be especially beneficial in the northern hemisphere during the December to February period. This is encouraging for the potential usefulness of this approach in forecasting winter gas demand in Prague. To maximise the benefit of multi model forecasting, Mylne et al. (2002) recommend that the models used should exhibit similarly high skill individually and, in terms of their physics and dynamics, be as different to each other as possible, in order to increase the likelihood of one model doing well when the other does poorly. Additionally, Mylne et al. (2002) state that using different operational analyses widens the sampling of analysis errors. This hypothesis cannot be tested in this study, since the necessary data are not available. In any case, perturbations generated for one model but input into a different model would produce sub-optimal results, since perturbations are generated to specifically maximise ensemble spread in the model they are generated for (e.g. singular vectors for the EPS, and breeding method for the MRF).

Although the results from Mylne et al. (2002) are encouraging, their study only verified model fields with ECMWF analyses for MSLP, 500hPa geopotential, 850hPa temperature, and 24 h precipitation accumulation, not with point-specific observations or a user response variable. Hence, this study will examine the multi-model technique in the end-to-end demand forecasting process. Apart from the user-specificity aspect, the need for verification with station data is compounded by the fact that analysis fields are widely regarded as being less accurate than ground observations (Jewson and Ziehmann, 2004). In addition, Mylne et al. (2002) only used a one year period of data and did not conduct any calibration of the forecasts.

A hybrid of ensemble and multi model concepts is what Ziehmann (2000) terms a ‘poor man’s ensemble’. This consists of single unperturbed forecasts from different NWP centres. Applying this concept, Ziehmann (2000) found that despite the small size of the ensemble used and the lack of perturbations, this approach proved to be an effective way of generating ensembles compared with the EPS. Since only EPS and NCEP data sets are available for the current study, this approach cannot be explored.

Chapter 3: Temperatures at Prague

Before attempting to calibrate and interpret temperature forecasts for a specific site, it is essential to consider the underlying climatic features of the location. In this study, temperature model output was downscaled to and verified with observations at Prague Ruzyne station (latitude: 50:06:03°N, longitude: 14:15:28°E, elevation: 364m asl, WMO number: 11518; see fig. 3.1).

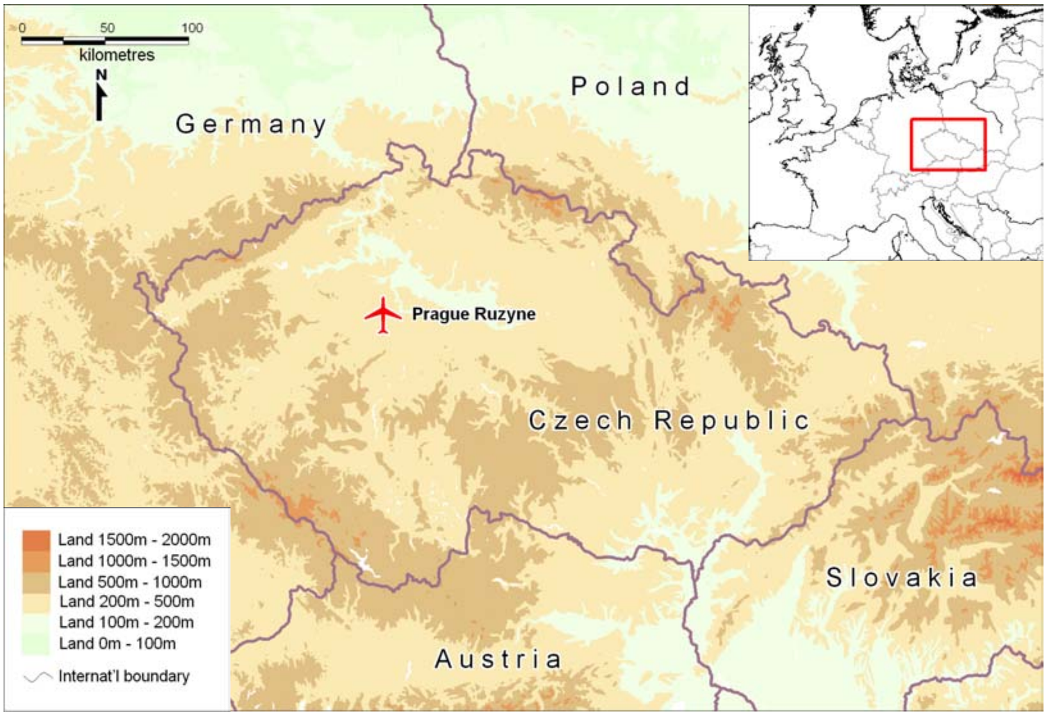


Figure 3.1: Map segment of Central Europe, showing orography surrounding Prague Ruzyne. (Produced for this study using MapInfo GIS. Data Source: Bartholomew Digital Data).

Since the station is located near the runway of Prague airport (around 10 km away from major conurbations; source: Czech Airport Authority) the effects of urban heat islands and increasing urbanisation should not be of great concern.

This area of Central Europe is influenced by both oceanic and continental air masses. Years with warm winters and cold summers coincide with decreased continentality of the European climate, whilst years with cold winters coincide with increased continentality (Kysely, 2002). Due to the influence of the Atlantic, extremely low temperatures are not as common in Central Europe as they are in North

America or Asia. However, during blocking conditions[2], cold polar and continental air masses can penetrate into central Europe resulting in temperatures plummeting as low as 20°C below climatological average (Domonkos and Piotrowicz, 1998). In addition, the Czech Republic's location near the source regions of strongly contrasting air masses, makes transient eddies an important feature for the advection of air masses with very different thermodynamic properties (Wallen, 1977). This is especially the case during winter, when mean horizontal temperature gradients are stronger, leading to rapid transitions in temperature (Domonkos et al., 2003). Figures 3.2.a and b show the mean December to February and June to August potential temperature at the 1000hPa level. Potential temperature, rather than temperature was chosen, since it is independent of topography and thus gives a clearer indication of the thermal property of air masses. Apart from showing stronger gradients in potential temperature in winter, the plots indicate a more zonal gradient in winter and a more meridional gradient in summer.

[2] See section 3.2 for a discussion.

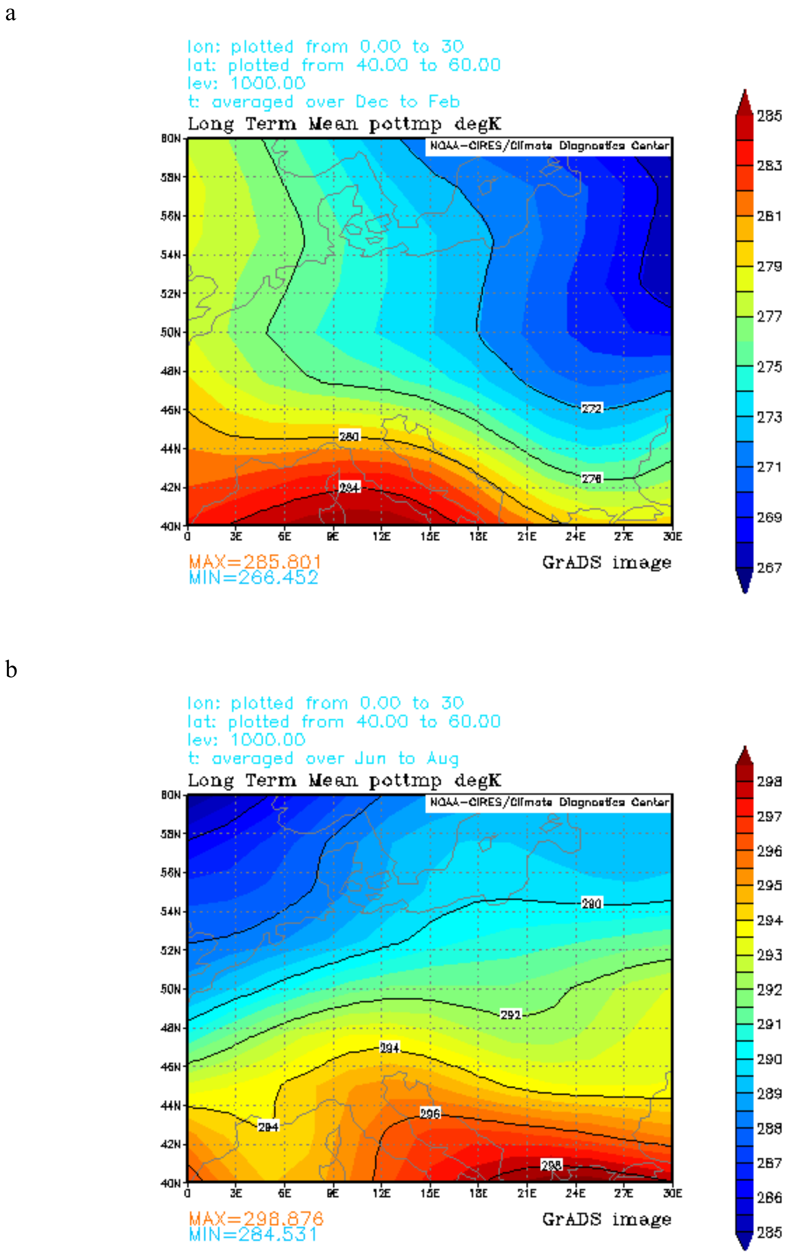

a
lon: plotted from 0.00 to 30
lat: plotted from 40.00 to 60.00
lev: 1000.00
t: averaged over Dec to Feb
Long Term Mean pottmp degK
NOAA-CIRES/Climate Diagnostics Center
60N
58N
56N
54N
52N
50N
48N
46N
44N
42N
40N
0
3E
6E
9E
12E
15E
18E
21E
24E
27E
30E
272
280
276
284
MAX=285.801
MIN=266.452
GrADS image
285
283
281
279
277
275
273
271
269
267
b
lon: plotted from 0.00 to 30
lat: plotted from 40.00 to 60.00
lev: 1000.00
t: averaged over Jun to Aug
Long Term Mean pottmp degK
NOAA-CIRES/Climate Diagnostics Center
290
292
294
284
296
298
MAX=298.876
MIN=284.531
GrADS image
298
297
296
295
294
293
292
291
290
289
288
287
286
285

Figure 3.2: 1000hPa long term mean potential temperature fields over Central Europe. a: December to February (major contour intervals: 6 K). b: June to August (major contour intervals: 2 K). Source: NCEP Reanalysis.

The intrusion of continental and maritime air masses is to some extent impeded by the mountain ranges (mostly with heights above 1000m) which almost completely surround the Western and Central part of the Czech Republic (see fig. 3.1).

The complex orography of the Czech Republic can pose a problem for NWP models to simulate near-surface temperatures, as well as for point predictions made by interpolation of gridpoint values (Kysely, 2002). In a study related to Sardinia, Italy, Boi (2004) notes that differences in elevation between the station of interest and the adjacent gridpoints, as well as unrepresentative boundary layer and land surface parametrisations can have a particularly pronounced effect on forecasts for locations surrounded by complex orography. In conjunction with this, radiative and thermo-dynamic processes (e.g. strong horizontal variations in radiative fluxes and Foehn effects leeward of mountains) can become an issue for temperature forecasting at Prague.

Although some maritime influence is evident, the climate at Prague is more continental than at London Heathrow. This can be observed by comparing the seasonal temperature cycles of the two stations (fig. 3.3).

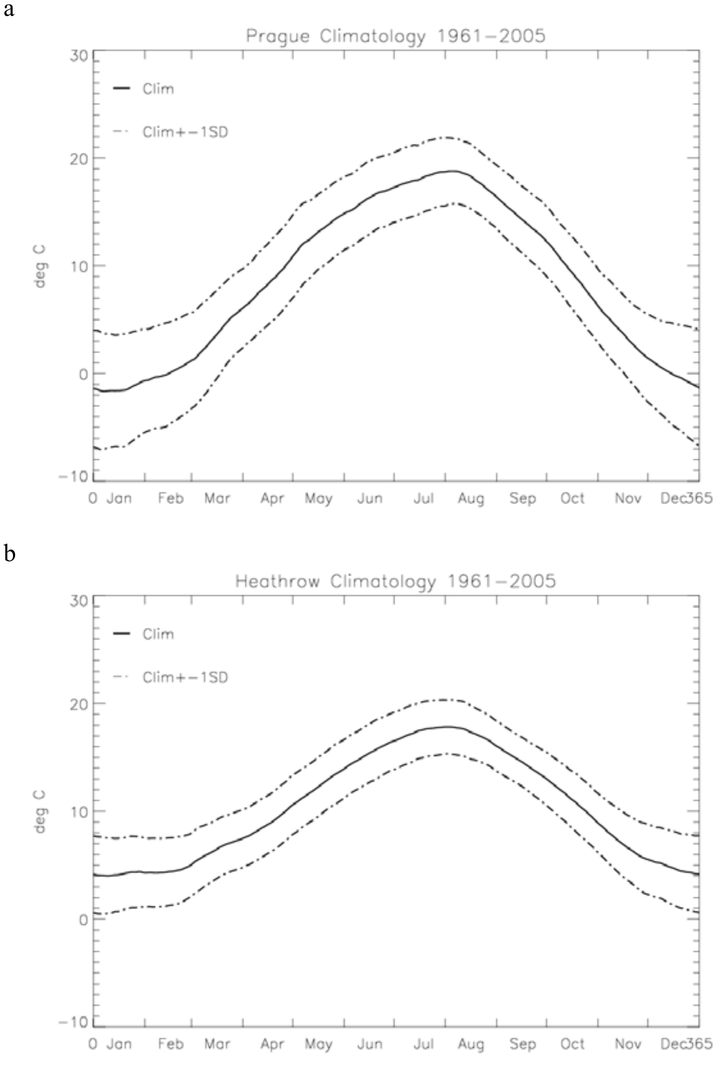


Figure 3.3: Temperature climatology (a: at Prague Ruzyne, b: at London Heathrow) +- 1 standard deviation. Derived from 1961-2005 measurements.

Prague temperatures exhibit a higher standard deviation (5.5°C versus 3.5°C), and a stronger seasonal variation of the mean (19°C versus 12°C) and standard deviation (2.5°C versus 1.5°C). London Heathrow is used as a comparison since most assessments of skill and commercial value as well as associated post-processing methods of ECMWF temperature forecasts mentioned in the literature use data from London Heathrow. The issues raised in this chapter must be borne in mind, since the climate of a location may affect forecast errors and the performance of specific bias correction methods.

3.1 Synoptic types

Many attempts have been made to summarize synoptic situations on particular days in the form of weather types. For the British Isles some subjective (e.g. Lamb, 1972), as well as objective (e.g. Jenkinson and Collison, 1977) categorisations exist. Although all approaches have severe weaknesses, some form of categorisation may be useful in relation to forecast calibration, since forecast errors may be dependent on atmospheric state. A commonly used classification of synoptic patterns over Western and Central Continental Europe are the Hess-Brezowsky circulation types (Gerstengarbe et al., 1999). Though originally devised for Germany, they were found to be useful for central Europe, since they refer to the large-scale circulation (Kysely, 2002). The classification allocates the synoptic circulation on a particular day to one of 9 major and 29 minor circulation classes (HBCs) according to the degree of zonality, direction of the prevailing flow and cyclonicity/anticyclonicity (Domonkos et al., 2003). Domonkos et al. (2003) found a strong connection of HBCs to temperatures, especially in winter. For example, extreme winter cold events at Prague (defined as days with mean temperature < -5°C) were shown to have a strong correlation with meridional and anticyclonic situations. Most extreme cold events (1902 to present) occurred under anticyclonic northerly flow and were rare under westerly and southerly flow. However, in blocking situations, the location of the anticyclone relative to Prague is crucial in determining the direction of large-scale airflow. On its eastern side, for example, the flow is northerly, advecting cool air southwards. The mean residence time of HBCs is 4 to 7 days (Domonkos and Piotrowicz, 1998). Hence, this appears to be the typical timescale of major changes in temperature. It is worth noting, though, that the persistence of types has increased in the 1990s (Kysely, 2002).

3.2 Predictability of blocking

Predicting transitions in weather regimes on timescales longer than 1 or 2 days is a major problem in the extratropics (Pelly and Hoskins, 2003). As noted above, these transitions can lead to very abrupt changes in temperature. Perhaps the most crucial transition is that into and out of blocking conditions. Blocking conditions are quasi-stationary synoptic-scale high pressure systems in the mid-latitudes with sufficiently large amplitude to disrupt the prevailing Westerly flow. Both model uncertainty and uncertainty in initial conditions have been found to put an upper bound of around 3 to 4 days on the predictability of blocking by single integration deterministic forecasts if blocking conditions are not already included in the analysis (Tibaldi and Molteni, 1990). However, since the ECMWF's EPS is designed to sample uncertainty in initial conditions and thereby generate several possible scenarios of atmospheric development, it can provide a probabilistic estimate of the onset or decay of blocking.

In an assessment of one year of operational probabilistic blocking forecasts in the Euro-Atlantic sector using an objective blocking index based on potential vorticity and potential temperature, Pelly and Hoskins (2003) found that the EPS was skilful at predicting the onset and decay of blocking conditions out to 10 days. The EPS was more skilful than the control integration at all lead times. With increasing lead time, though, the EPS forecasts tended towards the model's climatology, which has a westerly bias and hence underestimates the frequency of blocking observed in reality. The EPS was more skilful at predicting the decay than the onset of blocking. This was related to the fact that the onset is usually very rapid and primarily controlled by synoptic and planetary scale dynamics, whereas the decay is usually less rapid and related to a spin-down or diabatic decay of the system.

Chapter 4: End-to-end gas demand forecasting

A meaningful assessment of a forecast with regard to its skill in aiding the prediction of a non-meteorological response variable and, more importantly, with regard to the economic value the forecast can offer to its user, is only possible if a specific forecast application is considered in an integrated manner from the perspective of the user. Hence, this study analyses the specific application of ECMWF and NCEP ensemble temperature forecasts in end-to-end gas demand predictions in the Czech Republic and decisions based upon these predictions. This is achieved by integrating the temperature forecasts into the demand and decision-making models of the gas company.

Accurate energy demand forecasts have become increasingly important to network operators and distributors alike due to increasing uncertainty in supply and demand, fiercer competition and thus smaller profit margins, as well as the need to wisely plan investment decisions (McSharry et al., 2005). Improved demand forecasts can help decision-makers in the energy industry reduce risks and increase efficiency. In addition, accurate demand predictions play a pivotal role in energy commodity trading.

Natural gas consumption is highly weather dependent, since it is predominantly used for space heating. Temperature is the key parameter, though other meteorological factors such as wind speed and luminosity have also been found to affect demand (van den Berg, 1994). In addition, non-meteorological variables, such as economic cycles and trends, days of the week and fluctuating customer numbers, to name but a few, play an important role (McSharry et al., 2005). Since these variables are non-meteorological, they will not be considered in this study. Taylor and Buizza (2003) investigated the use of ECMWF model output of several meteorological parameters for electricity demand forecasting in England and Wales. However, this present study on gas demand will solely focus on temperature, since the version of the demand model available for this project only contains temperature as a meteorological variable. In any case, it is important to analyse the use of each meteorological forecast parameter individually at first, in order to determine its skill and value to the end user’s application.

Initially, gas demand must be related to temperature. Several approaches to demand forecasting exist, including time-varying splines, judgemental forecasts, artificial neural networks and multiple regression models (McSharry et al 2005). In practice, most gas companies use regression models based on historical consumption data to predict future gas demand. However, Palmer (2002) notes that a major difficulty faced by academic researchers is the fact that most demand models are proprietary. Fortunately, this obstacle was overcome in the case of this study through close cooperation with the gas company. A simplified version of the gas company's daily consumption model was made available and is given in eqn. 4.1:

$$\begin{aligned}\hat{G}_i = 0.13 + 1.10 * HDD_i + 0.14 * HDD_{i-1} + 0.45 * HDD_{i-2} \\ +1.71 * ln(d) - 4.37 * day\end{aligned} \quad , \qquad (4.1)$$

where
G_i is the estimated gas demand on day i,
HDD_i is the number of heating degree days on day i,
HDD_{i-1} is the number of heating degree days on day i-1,
HDD_{i-2} is the number of heating degree days on day i-2,
ln(d) is the natural log of the day, continuously numbered from 01.01.1996, and
day is a dummy variable set to 0 on working and 1 on non-working days.

Unlike electricity demand, which tends to be quadraticly related to temperature (McSharry et al., 2005), gas demand generally exhibits a linear increase with decreasing temperature below a threshold temperature (Quayle and Diaz, 1980). Gas is only used for heating when temperatures are low, and not for cooling when temperatures are high. The energy industry generally uses a threshold of 18°C for trading and derivative contracts (Banks, 2001) and defines the Heating Degree Day (HDD) as

$$HDD_i = Max(0, 18 - \overline{T}_i) \quad , \qquad (4.2)$$

where
$\overline{T}_i$ is the mean temperature on day i.

However, thresholds for consumption may vary from one geographical location to another (Heerdegen, 1988, Boehm, 1989), and the threshold used in the Czech Republic is 18.3°C. Hence, the model uses HDDs instead of temperatures to exclude days with temperatures above 18.3° C. In addition, gas demand is not only affected by temperature on the specific day of interest, but over several days preceding the day of interest. Therefore, lagged HDDs are included in the model. The temperature-independent base-load is represented by the intercept term as well as the day variable, accounting for reduced consumption on non-working days, and the logarithmic trend, representing increases in demand over time due to economic factors.

Similar to the study by McSharry et al. (2005), the demand model was developed with historical consumption data, using meteorological and non-meteorological predictor variables. Subsequently, all variables unaffected by the weather (*day* and *ln(d)*) were taken out of the model, whilst leaving regression coefficients of weather-dependent variables unchanged. In this manner, the weather dependent, and thus weather forecast dependent, portion of demand could be analysed separately.

Taylor and Buizza (2003) note, that "the expected value of a non-linear model of random variables is not the same as the non-linear function of the expected values of the random variables." Although the above gas model is linear, it uses lagged HDDs with different weightings. This may affect the shape of the PDF of gas demand in ways that are different to merely a linear scaling of a temperature PDF derived from the temperature forecast for the specific day of interest. In the case of probabilistic forecasts, using an end-to-end approach in demand forecasting translates the uncertainty in the weather into future uncertainty in gas demand. Therefore, an end-to-end approach was applied in this study to assess the skill and value of the temperature forecasts for predicting gas demand. It was designed as follows:

The 2m temperature values of each ensemble member for the four nearest gridpoints of the ECMWF and NCEP forecasts were linearly interpolated to Prague Ruzyne station (Norton, pers. com). Individual ensemble members were then converted into HDDs and used as input in the Gas Demand Model. At lead times of two days or shorter, a combination of forecasts and available observations were used. Gas demand calculated by running the demand model with observed temperatures was taken as actual demand against which the forecast could be verified, as done by Taylor and Buizza (2003) in their study on end-to-end electricity demand forecasting in the UK.

Since the temperature dependency of gas demand is highest during the heating season, this study investigated forecast performance for the period October to March (generally regarded as the heating season in the gas industry). The seasons of 2003-2004 and 2004-2005 were used in the analysis.

Chapter 5: Deterministic skill assessment

As set out in chapter 2, one of the aims of ensemble forecasting is to provide a superior estimate of the most likely outcome of a variable to be predicted. Hence, in this respect the mean of the ensemble gas forecast members is treated as a deterministic forecast. Thornes and Stephenson (2001) remind us that when assessing the goodness of a deterministic forecast, we must distinguish between forecast accuracy - the correspondence between forecasts and observations – and the economic value of the forecast. This chapter deals with the former. After a brief review of the skill metrics and post-processing methods used, the accuracy of deterministic gas demand forecasts produced with the end-to-end method and various post-processing techniques is assessed.

5.1 Metrics

The choice of skill metric depends on what information and quality of the forecast is desired by the user (Jolliffe and Stephenson, 2003). It is often the case that several skill scores need to be used to deduce a more comprehensive picture of forecast performance. For a more extensive treatment of skill measures for forecasts of a continuous variable the reader is referred to Deque (2003). A commonly used measure of accuracy to determine the quality of a forecast in meteorology and business applications is the Root Mean Squared Error (RMSE) (Wilks, 1995):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(F_i - O_i\right)^2} \quad , \tag{5.1}$$

where

F_i is the forecast for day i,

O_i is the observation for day i, and

n is the number of days.

One of its advantages is that it gives an indication of the average forecast error in units of the forecast quantity. In addition, it penalizes large errors more heavily than small errors, which can be beneficial if a user is more sensitive to large deviations. However, it is unrelated to the degree of difficulty in predicting the variable of interest (Mason, no date), and is strongly affected by the background variability of temperature at a location (Deque, 2003). Goeber et al. (2004) note that the accuracy of weather forecasts is also strongly dependent on the actual weather at a location, not just the forecasting system used. Taking an example from this current study, the background variability of Prague temperatures is higher than at London Heathrow (compare figs. 3.3 a and b), suggesting that temperatures may be more difficult to predict at Prague in an absolute sense.

A further measure of accuracy is Gain over climatology (Gain):

$$Gain = \frac{1}{n}\sum_{i=1}^{n} | C_i - O_i | - | F_i - O_i | \quad , \qquad (5.2)$$

where

F_i is the forecast for day i,
O_i is the observation for day i,
C_i is climatology for day i, and
n is the number of samples.

Unlike the RMSE, this measure expresses the skill of the forecast as an improvement of accuracy relative to a reference forecast, e.g. climatology, or any other forecast. Gain gives an indication of skill directly in the forecast quantity and is related to the difficulty in predicting the variable of interest. In addition, the Gain score is more resistant to outliers than the RMSE, and the statistical significance of the skill gained in estimating first order moments relative to a reference forecast can be tested with Student's t-test (Deque, 2003).

In addition to measures of *accuracy*, it is useful to consider a measure of *association*, i.e. the strength of the relationship between forecasts and observations. The Anomaly Correlation Coefficient (ACC) is a measure of the strength of the linear relationship between observed and predicted anomalies from climatology:

$$ACC = \frac{\mathrm{cov}(F', O')}{s_{F'} s_{O'}} \quad , \qquad (5.3)$$

where

F' are the forecast anomalies: $F' \equiv F - C$,
O' are the observed anomalies: $O' \equiv O - C$,
$S_{F'}$ is the standard deviation of forecast anomalies, and
$S_{O'}$ is the standard deviation of observed anomalies.

Using correlation of full forecast data and observations is not a sensible skill measure due to the dominant effect of seasonality (even poor forecasts can predict lower temperatures in winter than in summer). Therefore, anomalies must be used. An advantage as well as a disadvantage of the ACC is that it is insensitive to shifts in the mean and rescaling of the forecast or observations. Hence, high correlation is not sufficient to provide a good forecast. Transposing or rescaling of the forecast may be necessary to achieve this. However, it also follows that correlation can therefore be taken as a measure of the potential skill of a forecast if its (linear) bias could be removed (Deque, 2003). A further point to consider, is that no correlation does not mean no association, since there could be a non-linear relationship between forecasts and observations.

5.2 Stochastic temperature model as a baseline

Wilks (1997) notes that calibrated persistence is a better baseline forecast to measure the skill of a forecast system against than climatology, since it exploits the high degree of memory inherent in the temperature time series over a few days. The time series of October-March daily mean temperature anomalies at Prague (1961-2003) exhibits a lag 1 auto-correlation coefficient of 0.84. An example of a calibrated persistence forecast is a first order autoregressive model (AR1).

$$T_{i+1} = \alpha\,(T_i - C_i) + C_{i+1} \qquad , \qquad (5.4)$$

where

T_{i+1} is the predicted temperature on day i+1,

T_i is the temperature on day i,

C_i is climatology on day i,

C_{i+1} is climatology on day i+1, and

α is the lag 1 auto-correlation coefficient of anomalies.

Since AR1 models are inexpensive and simple to produce, a numerical weather prediction system must show considerably higher skill or added value to justify its expense. This is especially the case in a commercial setting where the bottom line is the ultimate target.

5.3 Post-processing and calibration

Whilst Jewson and Caballero (2003) acknowledge that end-to-end forecasting may indeed be useful for some applications, Jewson (2004a) stresses that any necessary post-processing of the meteorological forecast (e.g. removal of bias) should be performed before the data are input into a response-function. In order to investigate the effects of post-processing mid-way, this is tested empirically.

5.3.1 Forecast bias

A forecast may exhibit a high ACC with observations, i.e. possess the ability to resolve the predicted variable, but still exhibit large RMSE and low Gain. This indicates bias in the forecast. Bias can consist of a conditional and/or an unconditional element which can be defined as follows (Mason, no date):

Unconditional bias is the mean difference between forecast and observation. The unconditional bias can be removed by subtracting it from the forecast.

Conditional bias (Type I) is the degree to which the correspondence between forecast and observation varies with the forecast. An example of this is a forecast variance that is smaller than the observed variance. In this case, the conditional bias can be removed by scaling the forecast variance. Calibration can thus improve the accuracy of the forecast by removing biases. Potts (2003) points out that the worst forecast is a forecast that is statistically independent from the observations. In that case, no calibration can extract valuable information.

5.3.2 Methods of calibration

Various methods of forecast calibration are proposed in the literature, ranging from simply subtracting unconditional bias to forecast assimilation using Bayesian multivariate normal models (Stephenson et al., 2005). Jewson (2004b) suggests that bias correction should follow the principle of parsimony. Research should always begin with the simplest model against which progressively more sophisticated models can be evaluated. Any method devised should be tested empirically – in the case of regression models, out-of-sample tests are necessary.

Running mean

The simplest method of bias correction consists of subtracting the mean forecast bias over the most recent preceding time period (typically 60 days) from the current forecast. This is computed separately for each forecast lead time.

$$B = \frac{1}{nd} \sum_{i=-1}^{-nd} T_{E(i)} - T_{O(i)} \qquad , \qquad (5.5)$$

where
nd is the number of preceding days over which mean bias is calculated,
$T_{E(i)}$ is the ensemble mean forecast temperature for day i, and
$T_{O(i)}$ is the observed temperature on day i.

The greatest advantages of this method are its simplicity as well as the fact that the estimated bias relates to the most recent performance of the model and thus implicitly takes updates in the numerical model into account. The method was found to significantly improve the skill of temperature forecasts at London Heathrow, as can be seen from fig. 5.1, produced in this study. A two-tailed paired sample t-test showed that the mean absolute errors of corrected and uncorrected forecasts were significantly different at the 0.05 level at all lead times.

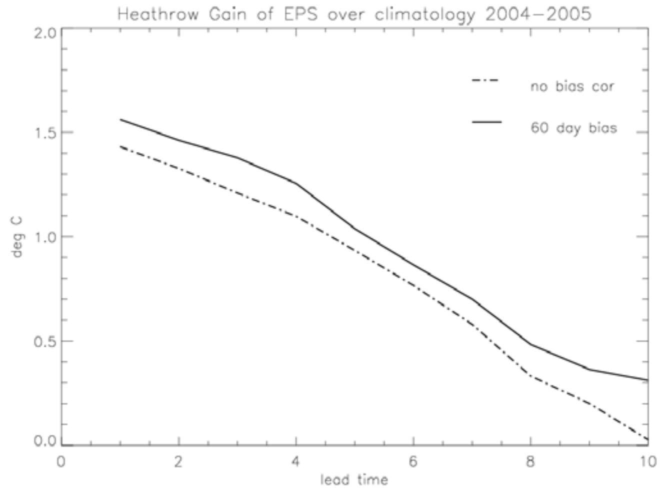


Figure 5.1: Comparison of Gain over climatology against lead time (days) achieved by the raw EPS ensemble mean and the ensemble mean corrected with a 60 day moving average bias correction for October 2004 to March 2005.

When selecting the length of the bias calculation period, a balance between several factors has to be established: On the one hand, any seasonal variations in forecast bias and potential numerical model updates require the period to be short (Jewson et al. (2005) suggest a running mean correction using no more than the last 90 days), whereas on the other hand the presence of noise in the bias requires a large sample size. With regard to seasonal adjustment, the problem can only partially be solved by taking shorter averaging periods, since the correction is lagged in any case. Apart from seasonality, errors may depend on synoptic conditions at the time at which the forecast is produced or at the time at which it validates. Since large scale flow patterns can change on timescales of several days (the mean persistence of a Hess-Brezowsky circulation type over central Europe is around 4 to 7 days; Domonkos et al., 2003) the bias correction relating to past forecasts, and thus past synoptic flow patterns, may not be representative of current forecast biases. A further problem is that this method only corrects the mean (unconditional) bias, not the magnitude-dependent (conditional) bias.

Bias correction by regression:

As suggested by Jewson (2004b), a temperature forecast can be bias corrected by linear regression between observed temperatures and the corresponding ensemble mean forecasts for each lead time. The regression and subsequent correction should be performed on temperature anomalies.

$$T_{O'} = \alpha + \beta T_{E'} + \varepsilon \qquad , \qquad (5.6)$$

$T_{O'}$ are the observed temperature anomalies,
$T_{E'}$ are the corresponding ensemble mean forecast temperatures,
α, β are the regression coefficients, and
ε are the residuals.

This model corrects mean bias (unconditional bias) using α, as well as optimally scaling the variance of the ensemble mean (type I conditional bias) using β. If α and β are significantly different from 0 and 1 respectively, then the correction improves the forecast in sample. In addition, *ε* can provide an estimate of the flow-independent uncertainty of the forecast, if *ε* is assumed to be normally distributed (generally a valid assumption for temperature; Jewson and Caballero, 2003). Jewson (2004b) found a regression model to improve ECMWF forecasts for London Heathrow. However, this method also has disadvantages. In order to ensure that the model is not over-fitted, it is essential that the model is tested out of sample. Due to the limited availability of past forecast data, this can be difficult. In addition, potential numerical model updates are not taken into account, since past data must be used to derive coefficients. Furthermore, not all skill measures are improved to the same extent by regression correction. Primarily, the RMSE is reduced, since least squares regression minimizes squared residuals.

Seasonally varying parameters

As noted before, forecast biases tend to vary with season (usually being negative in summer, and positive in winter). Standardised anomalies of Heathrow 60-day running mean forecast biases of the MRF and EPS at a lead time of one day are shown in fig. 5.2 as an example. Taking standardised anomalies transforms two time series with different means and variances into the same dimensionless scale in order to enable comparison of their correlated fluctuations (see Wilks, 1995). Standardised anomalies, *z,* are calculated as follows:

$$z = \frac{(\overline{F-O}) - (F-O)}{s_{(F-O)}} \quad , \qquad (5.7)$$

where

F are the forecasts

O are the observations, and

$S_{(F\text{-}O)}$ is the standard deviation of forecast errors.

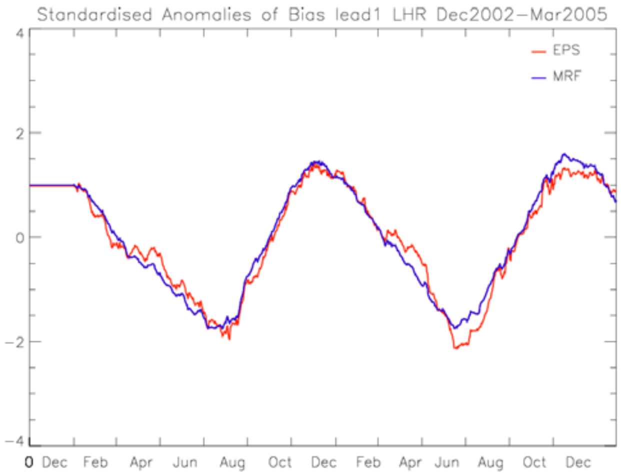


Figure 5.2: Standardised anomalies of the 60-day running mean bias of MRF and EPS forecasts at lead 1 for London Heathrow between December 2002 and March 2005, showing a distinct seasonal cycle.

Hence, this seasonal variation should be incorporated into bias correction regression models. Jewson (2004b) found a considerable improvement for London Heathrow forecasts if the regression parameters in eqn. 5.6 are represented by a set of optimally tuned sinusoids as follows:

$$\alpha = \alpha_0 + \alpha_s \sin\phi_i + \alpha_c \cos\phi_i \qquad (5.8a)$$

$$\beta = \beta_0 + \beta_s \sin\phi_i + \beta_c \cos\phi_i \qquad (5.8b)$$

where
ϕ_i is the day of the year.

Again, the noise in the error time series can lead to over-fitting and forecasts being degraded by bias removal, rather than enhanced. As Deque (2003) suggests in general, if only small samples of forecasts are available, it may be better not to bias correct at all. An issue with all the above methods is that using past forecasts to assess skill and correct errors, implies that past performance and error patterns will reflect current performance and error patterns. However, Gilmour (2004) notes that the frequent updates in the dynamical models pose a challenge and mean that forecasts need to be tested and recalibrated frequently, often with insufficiently large sets of past forecasts produced by the model currently in operational use. The latter issue can only be resolved if modelling centres produce and make available re-forecasts using the most recent model version.

Atmospheric state may be a key controlling factor on model bias, and thus other variables may need to be included in calibration. Therefore, determining which predictable or measurable atmospheric variables, if any, show a significant relationship to temperature forecast errors is a key task. Huth (1999 and 2002) investigated various methods of statistical downscaling of temperature forecasts in Central Europe. He tested two temperature variables (850hPa temperature and 1000-500hPa thickness), as well as two circulation variables (surface pressure and 500hPa geopotential height) as potential predictors - both in the form of gridpoint values and principal components of their fields. Pressure was used as an indicator of large-scale flow, and upper air data since they are generally considered to be simulated more reliably by numerical models than surface variables. The method of multiple linear regression of gridpoint values (as opposed to full fields) with stepwise screening yielded the best results. In terms of predictors, Huth (2002) found that a combination of one temperature and one circulation variable gave the most accurate results.

5.3.3 Multi-model

The simplest way of combining two forecasts is to take their mean. Deque (2003) notes that this may in many cases lead to a cancelling out of biases. However, models may have different magnitude errors or have systematic biases of the same sign. In addition, one forecast may perform better than the other and vice versa depending on lead time. Hence, a regression model can be fitted for each lead time to optimally weight both forecasts and correct their joint unconditional bias.

$$T_{O'} = \alpha + \beta_{EPS} T_{EPS'} + \beta_{MRF} T_{MRF'} + \varepsilon \qquad , \qquad (5.9)$$

where

$T_{EPS'}$	are the forecast temperature anomaly of EPS mean,
$T_{MRF'}$	are the forecast temperature anomaly of MRF mean, and
$\alpha, \beta_{EPS}, \beta_{MRF}$	are coefficients.

However, as is the case with individual model performance, the relative performance of each model may also depend on atmospheric state. Incorporating these factors could involve Bayesian methods and neural networks, where algorithms could be tuned to weight forecasts differently according to synoptic situation. Methods such as those proposed by Coelho et al. (2004) for combining and calibrating numerical and empirical forecasts of El Nino Southern Oscillation could be used. Whilst being a future possibility, the latter method is beyond the scope of this project, since it would require longer records of forecasts to enable a statistically stable relationship between forecasts and synoptic variables to be established. In this regard, Mylne et al (2002) note that differences between models are often subtle and impossible to identify synoptically.

5.4 Results

As far as the author is aware, this study is the first to assess the quality of end-to-end forecasts of a user specific variable using two different NWP systems, both comparatively and jointly. In addition, this study addresses two recommendations for calibration research proposed by Jewson (2004b) – testing calibration methods on locations other than London Heathrow and using longer forecast records. Whilst Jewson (2004a) only used a single year of daily ECMWF temperature forecasts for London Heathrow, this study examines over two years of forecast data for Prague Ruzyne, a continental station with a distinctly different climate.

Initially, end-to-end forecasts using the raw ensemble members of the EPS and MRF were created, and the respective ensemble mean gas demand taken to be the best estimate of consumption. The heating seasons 2003-2004 and 2004-2005 were analysed. Figures 5.3 to 5.5 show RMSE, Gain and ACC of the two systems plotted against lead time (days) respectively. Figures (a) are for the first, figures (b) for the second season. The ensemble mean of both systems provides a skilful deterministic estimate of gas demand at all lead times (days 1 to 10) and both seasons investigated. Results for the 2004-2005 heating season show the EPS to have a Gain over climatology of 3.8 mil m^3 at a lead time of 1 day (around 14.1% of mean daily weather-dependent demand, estimated at 27 mil m^3) and 1.8 mil m^3 at a lead time of 10 days. The RMSE was 1.5 mil m^3 at a lead time of 1 day and 4.7 mil m^3 at a lead time of 10 days.

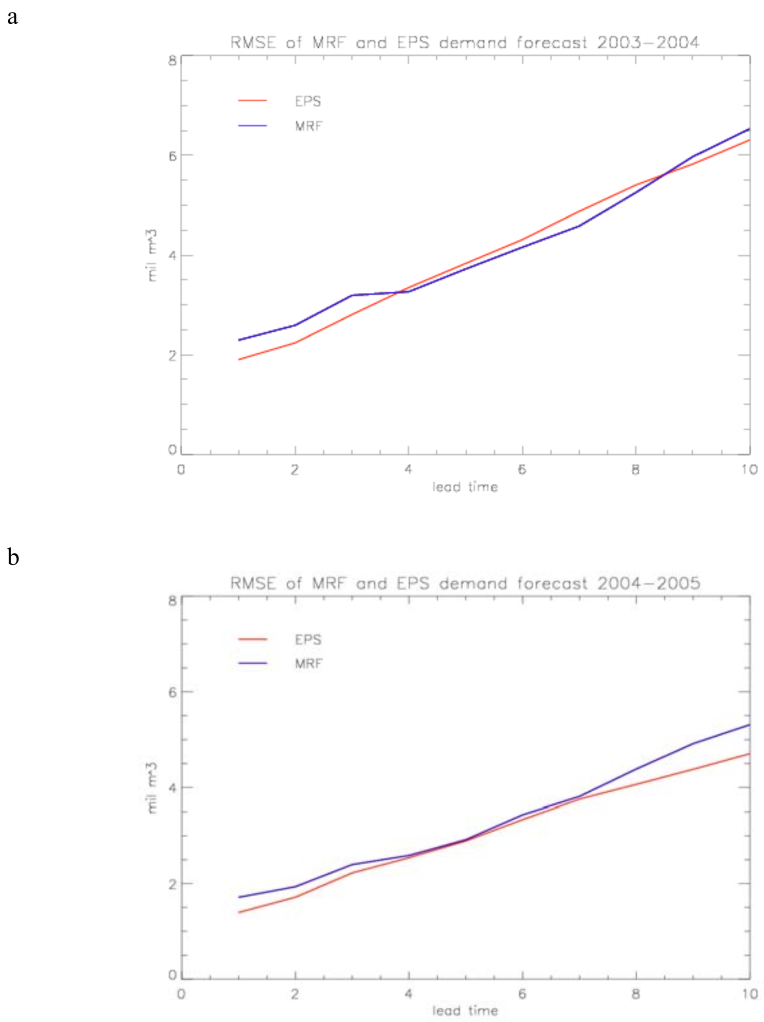


Figure 5.3: RMSE of raw end-to-end forecasts against lead time (days) for the heating seasons 2003-2004 (a) and 2004-2005 (b). RMSE increases with lead time and differs between MRF and EPS, as well as between the two seasons.

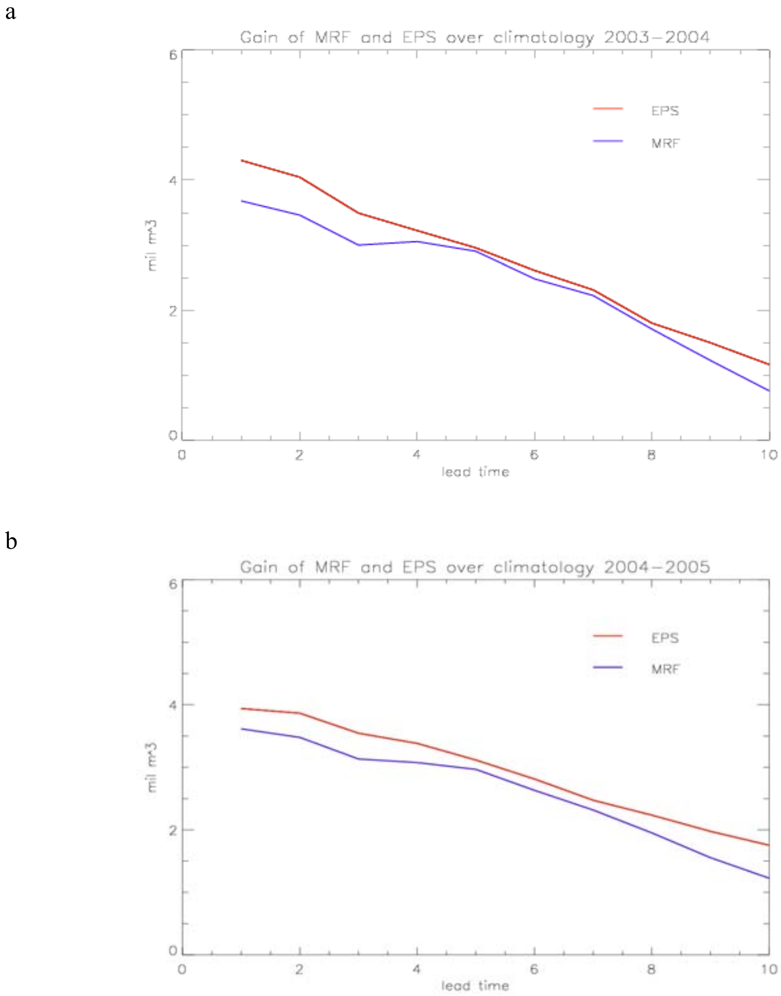


Figure 5.4: Gain of raw end-to-end forecasts against days lead time (days) for the heating season 2003-2004 (a) and 2004-2005 (b). Gain decreases with lead time and differs between MRF and EPS, as well as between the two seasons.

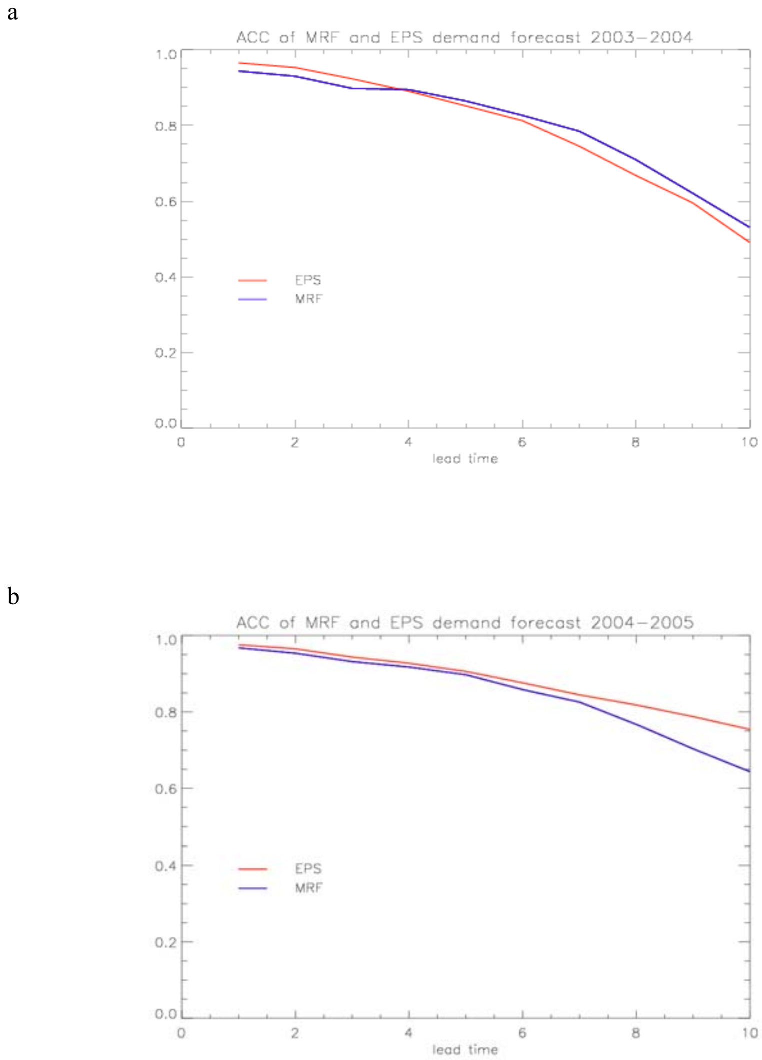


Figure 5.5: ACC of raw end-to-end forecasts against days lead time (days) for the heating season 2003-2004 (a) and 2004-2005 (b). ACC decreases with lead time and differs between MRF and EPS, as well as between the two seasons.

5.4.1 Comparison of two seasons

The RMSE of both NWP systems was lower in 2004-2005 than in 2003-2004 at all lead times. At a first glance, this could either be due to improved model performance or the temperatures being more difficult to predict in the first season. To obtain a greater insight, the Gain score was considered. In 2003-2004, Gain over climatology was higher at short lead times compared to 2004-2005, but dropped off more rapidly with increasing lead time. Hence, the fact that Gain was higher in 2003-2004 at short lead times, whilst RMSE was also higher suggests that temperature anomalies from climatology were greater and more difficult to predict in the first season. Figures 5.6 a and b show the temperature time series of the two seasons.

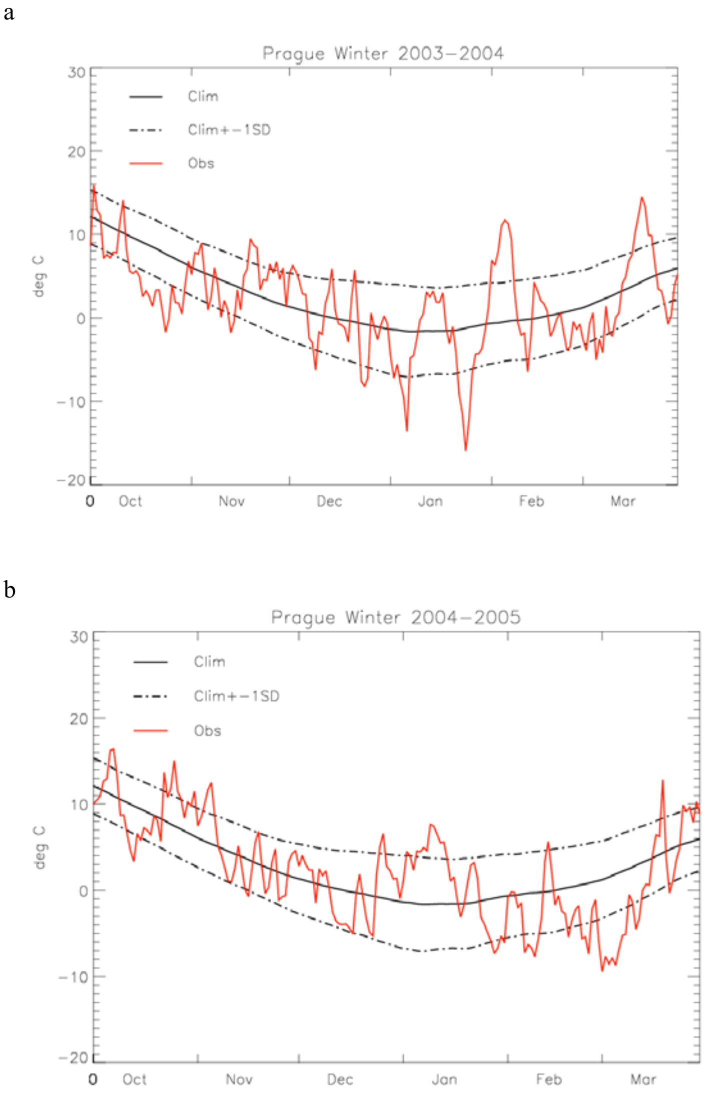


Figure 5.6: Recorded temperatures at Prague Ruzyne during the heating seasons 2003-2004 (a) and 2004-2005 (b), with climatological mean +- 1 standard deviation. 2003-2004 exhibits more rapid and extreme transitions.

Although the overall variance of temperature was slightly higher in 2004-2005 (σ = 5.74°C) than in 2003-2004 (σ = 5.65°C), 2003-2004 was characterised by several rapid transitions, which resulted in high magnitude forecast errors. Whilst the ensemble mean exhibits greater skill at resolving rapid transitions at short lead times compared to climatology, the ensemble mean loses much of its skill over climatology at longer lead times. This is due to the fact that the ensemble spread grows more rapidly and merges sooner with the climatological spread in situations of large forecast uncertainty (e.g. large and rapid transitions), than in more predictable situations. This more rapid decline in skill of the ensemble mean in the first season is also diagnosed by the ACC (figs. 5.7a and b), which decreases more rapidly with lead time in 2003-2004.

This raises questions as to the usefulness of the ensemble mean at different lead times. Figures 5.7a,b,c show the climatological and observed gas demand for the 2003-2004 season, as well as the ensemble mean best estimate gas demand using the raw EPS forecasts at lead 1 (fig. 5.7a), lead 5 (fig. 5.7b) and lead 9 (fig. 5.7c).

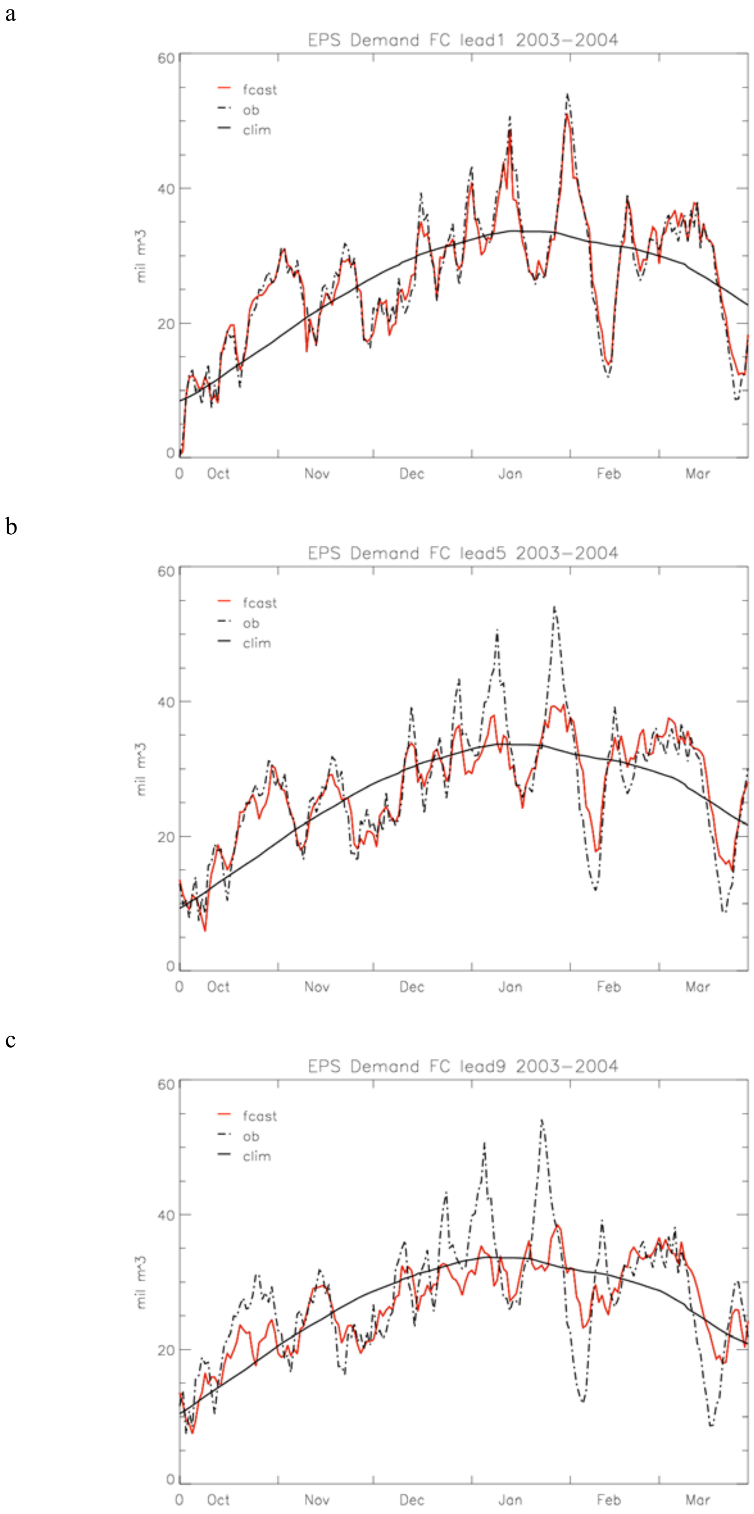


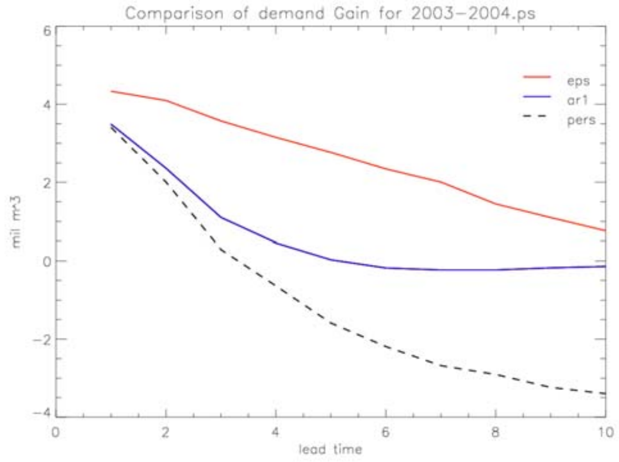
Figure 5.7: Comparison of actual and climatological gas demand during the heating season 2003-2004 with demand predicted by the ensemble mean of end-to-end forecasts using raw EPS data for leads of 1 (a), 5 (b) and 9 (c) days. The ability of the ensemble mean to resolve the magnitude of large anomalies decreases with lead time.

Figures 5.7 a, b and c reveal that at short leads, the ensemble mean resolves the timing and the amplitude of anomalies well. However, the magnitude of the most extreme peaks and troughs is not fully captured. In general, the amplitude is increasingly underestimated with increasing lead time. At longer leads the timing of anomalies is also lost. This is due to the fact that the ensemble mean tends to merge with climatology at long lead times. Hence, the ensemble mean becomes less useful with increasing lead time. In addition, even at short lead times, large extremes tend to be slightly underestimated by the mean. Thus, only a complete probability distribution, possibly inferred from the ensemble spread, may warn about potential extremes and their likelihood. Therefore it is essential to explore the potential use of probabilistic information contained in the ensemble spread. This is pursued in chapter 7.

5.4.2 Comparison with persistence and autoregressive model

To assess whether the much simpler and cheaper alternatives to NWP - persistence or a climatology reverting AR1 model (eqn. 5.4) temperature forecasts - could offer similar levels of skill in making deterministic gas demand predictions, temperature forecasts generated in this manner were used to create demand forecasts. As with NWP forecasts, a combination of forecasts and observations was used at lead times of 2 days or less. For the sake of brevity, only the Gain score is plotted. Since the raw EPS exhibited higher Gain than the MRF, the former was also plotted to provide a comparison.

a



b

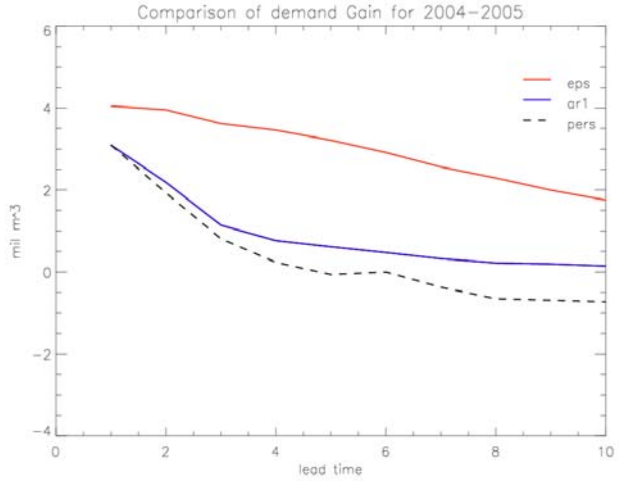


Figure 5.8: Comparison of Gain over climatology of the EPS, persistence and a climatology-reverting AR1 model for predicting gas demand for the season 2003-2004 (a) and 2004-2005 (b).

As can be seen from figs. 5.8 a and b, the EPS forecast was (as expected) superior at all lead times. However, the AR1 model showed statistically significant skill over climatology at the 0.05 level at leads 1 to 3 in 2003-2004 and at all leads in 2004-2005 (using a two-tailed paired-sample t-test for the difference in means of the absolute errors of the two forecasts). At lead 1, the difference to the EPS is just under 1mil m^3. Persistence showed statistically significant skill over climatology at the 0.05 level at leads 1 and 2 in both seasons. Whilst the skill of persistence was near identical to that of the AR1 model at lead 1, it declined rapidly thereafter. This was especially the case in 2003-2004, when persistence performed worse than climatology beyond lead 3. Interestingly, the AR1 model performs slightly worse than climatology at leads 6 to 10, but exhibits an upward trend towards climatology beyond lead 7. Once more, this is due to the large and rapid transitions in the 2003-2004 season. Since both persistence and the AR1 model are purely statistical techniques and rely on past temperatures, they perform particularly badly in these events. Hence, numerical weather prediction is especially beneficial if rapid transitions to large anomalies occur, which only a dynamical model can simulate. The greatest added skill of the EPS is at leads of around 3 to 9 days.

A similar comparison was conducted by Taylor and Buizza (2003) in the skill assessment of end-to-end electricity demand forecasts in the UK using the EPS. They, too, found the ensemble forecast to be significantly more skilful than a stochastic temperature model.

5.4.3 Comparison of EPS and MRF

In terms of Gain (figs. 5.4 a and b), the EPS is superior at all lead times, at some leads up to 0.6 mil m^3. Two-tailed paired sample t-tests of the mean absolute errors of the two systems at each lead time show the differences to be statistically significant at the 0.05 level at leads 1 to 3 and 10 in 2003-2004, and at leads 1 to 3 in 2004-2005. However, in terms of RMSE (figs. 5.3 a and b), the MRF is superior at leads 4 to 8 in 2003-2004. This suggests that whilst the MRF has greater errors overall, it has fewer large magnitude errors at leads 4 to 8, which are more heavily punished by the RMSE than the Gain score. This can be seen in fig. 5.9, showing the lead 5 forecast errors for the 2003-2004 season (note the two large positive errors of the EPS in January).

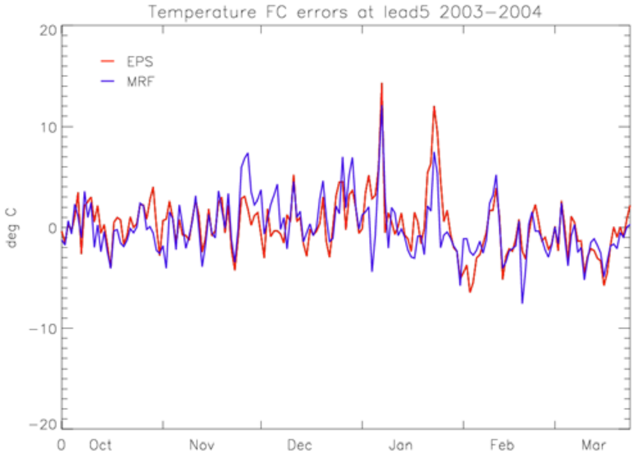


Figure 5.9: EPS and MRF temperature forecast errors at lead 5 during the 2003-2004 heating season. Note the two large positive errors of the EPS in January.

Furthermore, the ACC (fig. 5.5a) shows that the *potential* skill of the MRF is higher than that of the EPS in 2003-2004 at leads beyond day 4. Hence, a combination and calibration of forecasts may improve *actual* skill. However, care must be taken, since this is not evident in 2004-2005 (fig. 5.5b), pointing towards a statistically unstable relationship.

5.4.4 Calibration

As explained above, bias correction may improve the skill of the forecast. The standardised anomalies of the 60-day running mean bias for EPS and MRF forecasts at lead 1 are shown in fig. 5.10. Patterns at other lead times are similar.

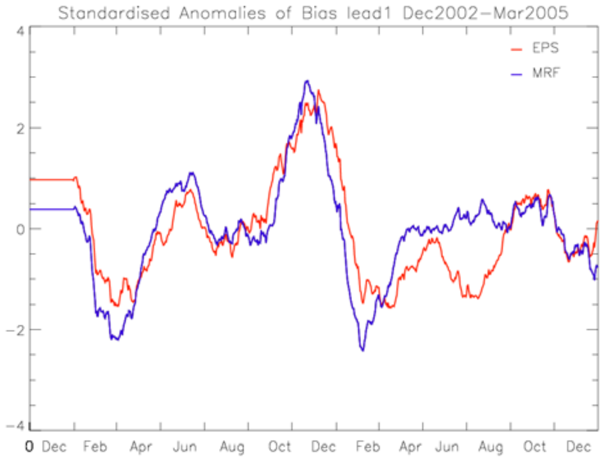


Figure 5.10 Standardised Anomalies of 60 day running mean forecast bias of MRF and EPS at a lead time of 1 day December 2002 to March 2005.

Though more erratic than the patterns at London Heathrow (fig.5.2), the mean bias of both systems appears to be affected by seasonality, with large negative biases in late winter/early spring and mid-summer, and a less negative (in 2003-2004 large positive) bias in late autumn and late spring. The methods of bias correction discussed in section 5.3 were applied.

5.4.4.1 Running mean

The running mean forecast bias at each lead time over different lengths of averaging periods (60 days, 30 days, 15 days) was subtracted from the forecast, both for the EPS and the MRF. A comparison of Gain scores for the 2004-2005 season are presented in figures 5.11a and b.

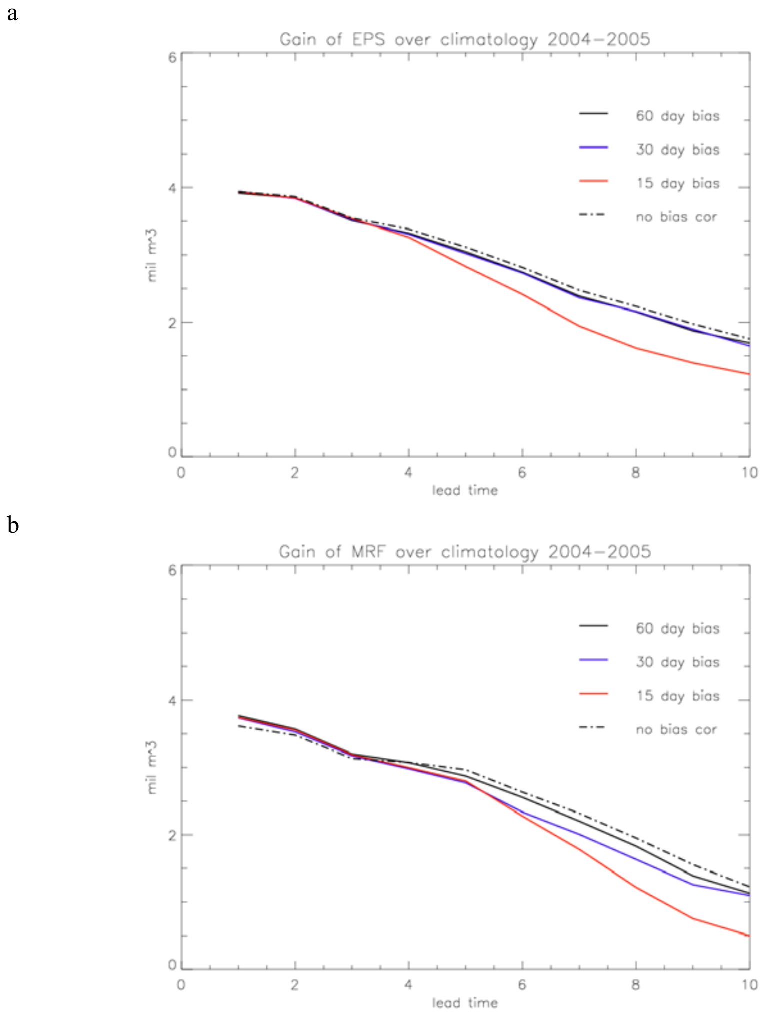


Figure 5.11: Comparison of Gain of running mean bias correction of MRF demand forecast using 15, 30 and 60 day averaging periods for the 2004-2005 (a) and 2004-2005 (b) heating seasons at leads of 1 to 10 days.

All running mean corrections at all lead times degraded the skill of the EPS. This effect increased with increasing lead time and was most pronounced when using the 15 day running mean. For the MRF, running mean bias correction slightly increased skill at leads up to 3 days, but thereafter degraded skill. Again, the 15 day mean performed worst. The general phenomenon may be due to the more erratic nature of the error patterns at Prague, compared with other stations, which could be attributed to the higher variance of temperatures. The particularly bad performance of the 15 day average, which deteriorated with increasing lead time, may partially be a reflection of the noisiness of the bias. In addition, the deterioration of the 15 day correction with lead time may be evidence of the dependence of bias on atmospheric state. Since the 15 day average samples only the most recent biases, and thus biases that occurred under the most recent past states of the atmosphere, it is itself biased towards these states. As time progresses, atmospheric state is likely to change significantly, making the bias correction estimated in this manner less and less representative of the actual bias at more distant lead times of the forecast. The improvement in skill of the MRF up to lead 3 gained with the running mean bias correction may be further evidence of this, since flow patterns will be more likely to persist at short lead times. Reducing noise may be a more important factor in bias estimation at Prague than sampling the most recent model performance and seasonality. Using a longer averaging period partially offsets this problem, but does not lead to improvement over the raw forecast.

5.4.4.2 Regression model

To account for both unconditional as well as conditional bias, the regression model set out in eqn. 5.6 was fitted between EPS forecast and observed demand anomalies at each lead time. For all regression-based calibration models the 2003-2004 season was used as the training and the 2004-2005 season as the out-of-sample test period, with the exception of the time varying parameters, which were trained on January 2003 – September 2004 data.

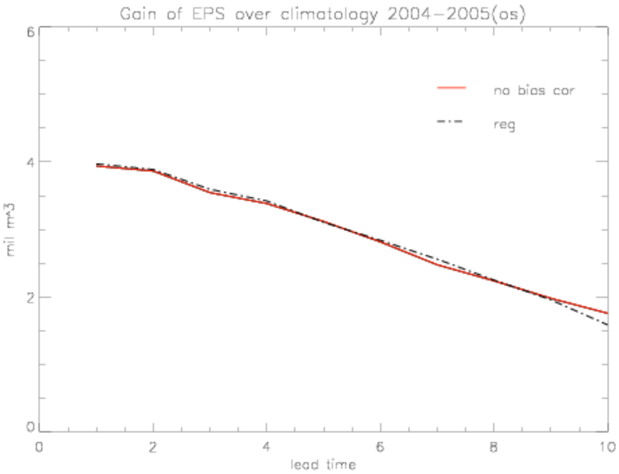


Figure 5.12: Comparison of Gain achieved by the raw forecast and by regression calibration out of sample. Training period: October 2003–March 2004. Test period: October 2004-March 2005.

Results from the regression model calibration showed minimal differences in an out of sample test. In addition, a t-test on the slope and the intercept of the regression revealed that they were not significantly different from 1 and 0 respectively, at the 0.05 level at any lead time. Hence, this form of regression model does not even offer significant correction in-sample.

5.4.4.3 Regression model with seasonally varying parameters

Suspecting a seasonal dependence of bias, the regression coefficients were represented by sinusoids (as set out in eqns. 5.7a and b) with a period of 365 days.

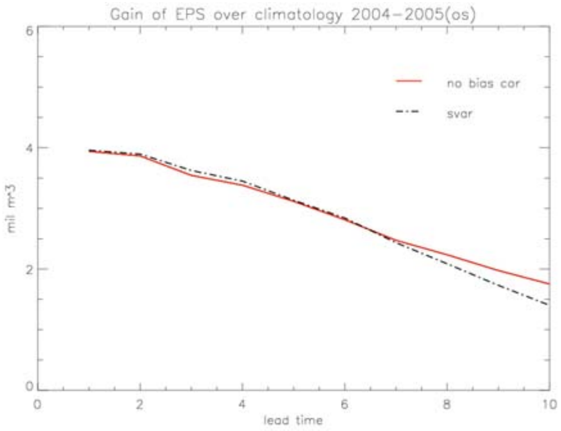


Figure 5.13: Comparison of Gain achieved by the raw forecast and by regression calibration with seasonally varying parameters out of sample. Training period: January 2003-September 2004. Test period: October 2004–March 2005.

This gives some improvement in-sample. However, nearly no improvement of skill is found out of sample at leads up to 6 days, and skill is degraded at leads 7 and beyond. Two potential explanations offer themselves: Firstly, forecast biases at Prague are more highly variable than at Heathrow, and thus any underlying seasonal cycle will most likely be obscured by noise in this relatively short time series available to train the model. Secondly, the seasonal structure also appears to be more complex than at Heathrow. A second order harmonic could be used to provide a better fit, considering the two annual peaks and the two annual troughs observed in fig. 5.10. However, the danger of over fitting the model is substantial in the absence of longer past forecast errors.

5.4.5 Multi-model deterministic forecast

The aim of the multi model approach is to exploit uncorrelated errors of two forecasts. The visual patterns of 60-day averaged standardised anomalies of forecast errors in fig. 5.10 and correlation coefficients between the errors of the two models ranging between 0.70 and 0.88 over all lead times, suggest that forecast errors at Prague are to some extent related and most probably due to atmospheric states in which both models tend to have similar problems. However, the error patterns of EPS and MRF at Prague show less similarity[3] and have lower correlation coefficients than London Heathrow (r ~ 0.80 - 0.92). Therefore, skill and value may be added by combining the forecasts. A more poignant justification for the concept of combining the two forecasts is found when analysing and comparing the timing of large error events of the EPS and MRF. In fig. 5.15 the timing of the ten largest magnitude negative and positive errors of the EPS and the MRF at lead 5 are shown.

[3] For this purpose, fig.5.2 and fig.5.10 can be compared. Standardised anomalies lend themselves useful for an initial analysis of the association of error patterns of the two forecasts, since combining two perfectly correlated forecasts would not improve skill. However, since these time series are 60-day averages, the effect of one system's error compensating for the other's on particular days is not clearly visible.

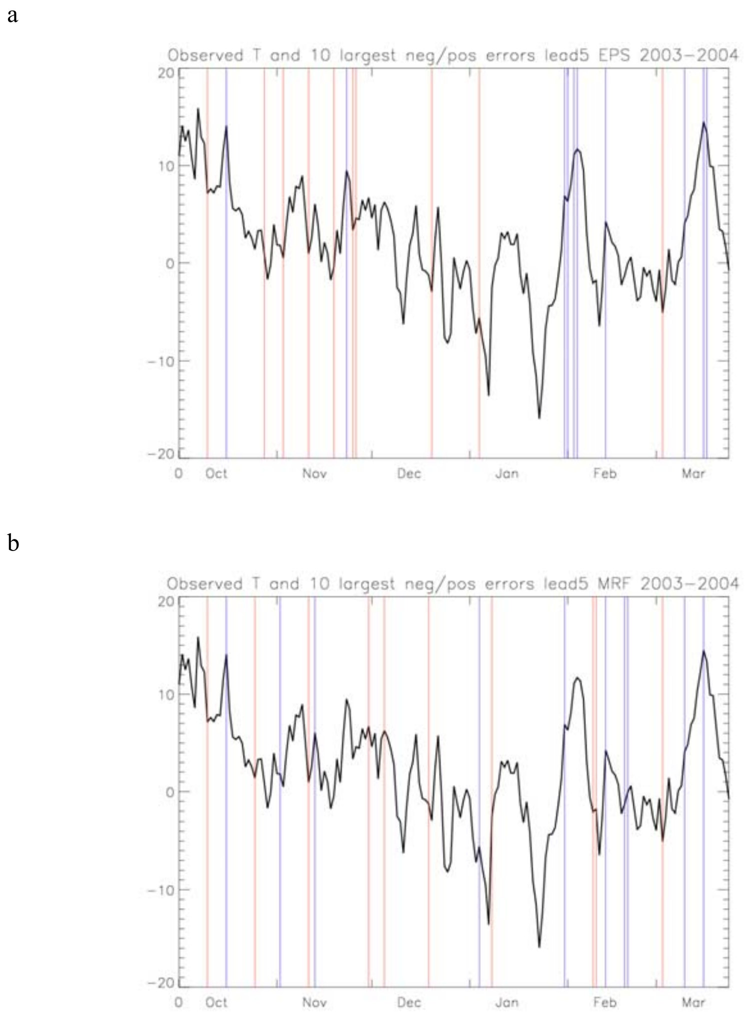


Figure 5.15: Observed temperatures and timing of the ten largest negative (blue lines) and positive (red lines) errors of the EPS (a) and MRF (b) at a lead time of 5 days for the season 2003-2004.

This reveals that most of the large magnitude errors occurred in the prediction of transitions to extreme temperature peaks and troughs – high temperature peaks were predicted too cool, and cold temperature troughs too warm. However, whilst some of the top ten errors occur at the same time in both systems (indicating common weaknesses), most of the large magnitude errors of the EPS and MRF do not coincide. High magnitude negative errors of the EPS appear to be clustered around the end of January and beginning of February, and large positive errors in October and November. In the case of the MRF, errors are spread more homogenously over the season. This indicates that the overall information content can be increased by combining the two forecasts. Both the approach of taking the mean of the two forecasts as well as regression calibration (eqn. 5.8) were tested for this purpose. Results for Gain are shown in fig. 5.17.

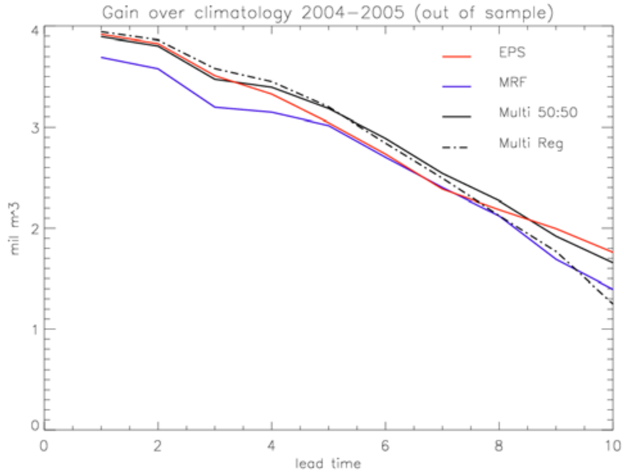


Figure 5.17: Comparison of Gain achieved by raw EPS and MRF forecasts as well as multi-model methods using the mean of both models and regression calibration out of sample. Training period: October 2003 – March 2004. Test period: October 2004-March 2005

The out-of sample test revealed that the multi-model regression calibration increases skill vis-à-vis the EPS up to lead 7, by up to 0.2 mil m^3, especially between leads 3 and 6. However, this is not statistically significant at the 0.05 level. Beyond lead 7, though, multi-model regression calibration decreases skill. This points towards an unstable statistical relationship at long lead times, which was also detected above in the comparison of ACC results of the two heating seasons (figs. 5.5a and b). At leads 4 to 8, the two-model average also increases skill. This suggests that although both forecast systems tend to exhibit similar bias patterns dependent on synoptic conditions, some skill can be extracted from combining the forecasts. Once more, though, the short archive of past forecast errors means that care must be taken when using forecasts corrected by models with coefficients estimated from a relatively short time series.

Chapter 6: Probabilistic forecast information

As McSharry et al. (2005) note in the context of electricity demand forecasting, prediction intervals or probability densities could provide crucial information about uncertainties inherent in the forecast. In the case of gas demand forecasts, extremes of temperatures could be very damaging, if the gas company was not prepared for the event. However, if forecast correctly, extreme temperatures could also represent a commercial opportunity. Extremely low temperatures would trigger a surge in demand, making it challenging to ensure an adequate supply of gas to customers, whilst trying to avoid having to pay high prices on the spot market to make up for any shortages in the gas company's own reserves, for example. Extremely warm temperatures lead to low demand and thus lower revenues. However, if predicted early, surplus gas supplies could be traded off in advance, minimizing the loss. This requires accurate and reliable demand forecasts, especially warnings of potential extremes. As shown in chapter 5, the ensemble mean does not convey forecast uncertainty, which becomes an increasingly important issue with increasing lead time. Hence, a probabilistic demand forecast should be used. In end-to-end demand forecasting, producing a probabilistic demand forecast requires a temperature forecast from which probabilistic information can be extracted.

A recent survey of weather forecast users revealed that only a small minority of them use probabilistic forecasts (Mailier, 2005). In earlier work, Jewson (2004b) seems to suggest that the paucity of applications of probabilistic forecasts is mainly the fault of forecast vendors, either because very few of them produce such forecasts or because they do not calibrate these forecasts correctly. Furthermore, he points out that the terms probabilistic forecast and ensemble forecast should not be confused. A probabilistic forecast states probabilities of occurrence, whilst an ensemble forecast consists of several members. Probabilistic forecasts can be produced both from single integrations using past error statistics or from ensembles, using statistical methods.

As described by Jewson et al. (2005), the most basic form of a probabilistic forecast is to fit a Gaussian distribution around the best estimate (e.g. the ensemble mean) with a standard deviation obtained as a by-product from the error distribution from a calibration regression of the best estimate, such as eqn.5.6. This means that the probability density function (PDF) will always have the same shape for a given lead time, with only its location altered depending on the best estimate forecast for a specific day. Hence, only the flow-independent uncertainty is considered. However, forecast uncertainty can vary greatly depending on atmospheric state, as illustrated in chapter 2.

It is widely believed that the ensemble spread contains useful quantifiable information about flow-dependent uncertainty in the forecast, since the spread is related to the atmospheric state (Jewson and Ziehmann, 2004). However, vigorous debate surrounds the issue of how information contained in the ensemble spread, should be used. Jewson (2004b) found that incorporating the raw ensemble spread as a measure of uncertainty led to worse probability interval estimates for London Heathrow, compared to using past error statistics (the goodness of fit of the distribution was assessed in terms of the log-likelihood; Fisher, 1912). This is believed to be due to the ensemble spread usually underestimating forecast uncertainty (also noted by Taylor and Buizza, 2004). In addition, the limited resolution due to the finite ensemble size can be problematic if the user is particularly interested in the tails of the distribution.

Hence, the ensemble spread needs to be recalibrated and for some applications transformed into a continuous PDF. A possible method of recalibrating the ensemble was developed by Norton (unpublished work). The absolute forecast errors of the ensemble mean at a certain lead time are related to the variance of the ensemble spread and a constant term at a certain lead time by way of linear regression. Assuming a normal distribution of forecast errors, the recalibrated variance is then used to derive the standard deviation for a continuous forecast PDF, or used to scale individual ensemble members. The latter option has the advantage of preserving ensemble members, which is necessary in the case of end-to-end forecasts if calibration is performed before inputting temperature forecasts into a response function. Ensemble members can be scaled by multiplying their departure from the ensemble mean by the ratio of corrected to uncorrected forecast standard deviation.

A similar method relating the forecast errors of the ensemble mean to the standard deviation of the ensemble members, which does not use the climatological standard deviation as a normalising factor, is employed by Jewson (2004b) and is termed ‘spread regression’ (eqn.6.1). The actual standard deviations (estimated from the absolute errors of the ensemble mean from the observations on individual days) are regressed on the standard deviations of the ensemble members on the corresponding days:

$$\hat{\sigma} = \gamma + \delta s_i \tag{6.1}$$

where

$\hat{\sigma}$ is the estimated standard deviation for the target day,

γ is the mean level of forecast uncertainty,

s_i is the standard deviation of the ensemble members on day i,

δ is a coefficient.

Such a regression model optimally blends the mean forecast uncertainty, γ, and the flow-dependent uncertainty, s_i. If s_i is not significantly different from 0, then the ensemble spread does not contain useful information for estimating σ.

In his analysis using EPS forecasts for London Heathrow for the time period of one year, Jewson (2004b) found that the mean of the uncertainty is best predicted by increasing the ensemble spread, whilst the variability of the uncertainty is best predicted by decreasing the amplitude of the variability of the ensemble spread. Results of the calibration methods were assessed by comparing their log-likelihood scores. The calibrated variability of the uncertainty was found to be relatively small (5% to 20% of the mean level), suggesting that the ensemble spread does not contain much useful information. Jewson (2004b) postulates that singular vector systems, such as ECWMF, would be expected to overestimate the amplitude of the variability in the uncertainty. Furthermore, he found that using flexible kernel density models showed no improvement over an ordinary linear regression model which assumes Gaussian error distributions. This suggests that any non-normality potentially present in the ensemble does not contain useful information. Therefore it appears to be justified to assume a normal distribution of errors of temperature forecasts.

In analogy to the bias correction of the ensemble mean (chapter 5), Jewson (2004b) accounts for the effects of a seasonal cycle on the relationship of the ensemble spread to forecast uncertainty by letting γ and δ in eqn. 6.1 vary seasonally in the form

$$\gamma = \gamma_0 + \gamma_s \sin \phi_i + \gamma_c \cos \phi_i \tag{6.2a}$$

$$\delta = \delta_0 + \delta_s \sin \phi_i + \delta_c \cos \phi_i \tag{6.2b}$$

As in the case of calibrating the ensemble mean, Jewson (2004b) found this to offer noticeable improvement of forecast uncertainty estimates for London Heathrow.

In order to assess whether the estimate of the forecast uncertainty could be improved, Jewson (2004c) also investigated different potential relationships of the ensemble spread to forecast uncertainty in addition to that based on the standard deviation (eqn.6.1):

variance-based: $$\hat{\sigma}^2 = \gamma^2 + \delta^2 s_i^2 \tag{6.3a}$$

inverse standard deviation-based: $$\frac{1}{\hat{\sigma}} = \gamma + \frac{\delta}{s_i^2} \tag{6.3b}$$

inverse variance-based: $$\frac{1}{\hat{\sigma}^2} = \gamma^2 + \frac{\delta^2}{s_i^2} \tag{6.3c}$$

Though visible differences in the calibrated spread were observed, no major differences in log-likelihood skill scores was noted, suggesting a low information content of the ensemble spread. However, Jewson (2004c) acknowledges that longer time series of forecasts may lead to progress on this matter. Furthermore, these different estimations of forecast uncertainty may prove to be beneficial at other geographic locations, e.g. Prague.

In a further study of probabilistic forecast calibration, Taylor and Buizza (2004) analysed forecasts of constituent quantiles of the temperature PDF derived from ECMWF temperature ensemble predictions for London Heathrow. They compared bias correction methods by determining how well the percentage of forecasts which fall into a specific prediction interval matched the value of the prediction interval. Quantiles of the distribution were debiased using several versions of quantile regression. To address the problem of updates in the numerical model, Time Varying Parameters (TVPs) were used, which increased regression parameters if temperatures in the current period exceeded the estimated quantiles, and conversely. Although the study involved categorical predictions, as opposed to predicting continuous variables, Taylor and Buizza's (2004) results indicate, contrary to Jewson (2004c), that the ensemble spread does contain quantifiable information related to uncertainty, and that there therefore exists a strong potential for using ensemble predictions in temperature density forecasting. Therefore it can be justified to further explore the possibility of extracting quantifiable probabilistic information form the ensemble spread. In addition, the use of time-varying parameters should be considered.

Chapter 7: Economic value of weather forecasts

As Jolliffe and Stephenson (2003) note, weather forecasts have become increasingly complex - more variables can be predicted, and more sophisticated techniques are employed (e.g. ensemble forecasting). However, users mostly require very specific information, summarised in an accessible and practical format. Whilst science aims to develop a better understanding of nature, business wants to use forecasts to increase efficiency and profitability. First and foremost, it is important to make a distinction between skill and value of a forecast. *Skill* is a measure of the increase in accuracy in predicting a variable (either in a deterministic or probabilistic sense) achieved by using the forecast, compared to a baseline. *Value*, or rather Value of Information (VOI), is the economic utility the forecast user gains, if the user acts on the forecast, as opposed to not acting on a forecast or using a different forecast (Wilks, 1997).

Not only do different users require different forecast information, but the economic value the same information can provide to one user can be strikingly dissimilar to the value it can provide to another. Richardson (2003) notes that the value of a forecast depends both on the forecast itself as well as the weather sensitivity of its user. This places a premium on knowing the end user's requirements and decision-making framework, and requires cooperation between end users and forecast providers (Palmer, 2002).

A number of studies have aimed to quantify the value of weather forecasts to business. These assessments have mostly attempted to generalize the concept of value and group users simply by Cost/Loss Ratios (e.g. Zhu et al., 2002; Palmer, 2002). Zhu et al. (2002) employ cost-loss analysis in the context of a binary decision framework – i.e. a forecast of an event occurring or not and the user's decision based on the forecast of whether to take mitigating action or not. An event is defined either as a catastrophic event or the exceedence of a certain threshold (e.g. daily mean temperature exceeding 30°C), which is of significance to the user. In the case of a deterministic forecast, the forecast would either be 'threshold will be exceeded' or 'threshold will not be exceeded' and the decision of the user, assuming he or she acts on the information, would be to mitigate or not, respectively.

In the context of a probabilistic forecast, however, the user has to define a minimum forecast probability threshold of the event occurring. If this probability is exceeded in the forecast, the user will take mitigating action. Zhu et al. (2002) postulate that this probability level is equivalent to the so-called Cost/Lost Ratio, i.e. the ratio of the cost of protection to the loss that can be protected. The higher the Cost/Loss Ratio (i.e. the higher the cost of mitigating with respect to potential losses), the higher the probability threshold required. The value of the forecast is determined by the resulting net profit or loss vis-à-vis a baseline (e.g. an existing forecast or climatology) over many forecasts, assuming the user always acts on the forecast.

This approach assumes that all forecast users can be categorised in terms of a cost-loss ratio. Whilst this may hold true in some cases, it does not in all. In many applications, it is a continuous variable or quantity which has to be forecast, rather than a binary event. As Smith et al. (2001) note, this means moving away from questions such as ‘will the event occur or not’, to questions along the lines of ‘how much is expected’. In energy demand forecasting, for example, a utility company wants to know the optimal amount of energy demand to plan for, not simply whether energy will be consumed or not. Furthermore, in this case, a predicted weather variable such as temperature is not the final variable of interest to the user. Rather, meteorological variables are used as one of several predictors in the end-user’s model, which forecasts the actual quantity of interest, e.g. the price of or demand for a commodity. Like in weather derivative pricing, gas demand forecasting involves assessing an entire continuous probability distribution, not binary events or categorical forecasts. Hence, most of the meteorological probabilistic skill scores, such as the Brier Score (Brier, 1950) which are designed for categorical events, are not relevant.

7.1 Economic utility theory

Johnson and Holt (1997) offer a more satisfactory approach to this problem. They assert that meteorological information should be viewed as a factor (in most cases one of several) which reduces uncertainty in a decision-making process. In order to include meteorological information in the decision process a user-specific decision-analytic model needs to be developed. This model should be based on subjective probabilities with the assumption that economic agents can assign a unique economic utility to each pair of decisions and possible outcomes. Wilks (1997) notes that decision-analytic models divide the problem into four basic components:

1. the possible actions available to the decision maker,
2. the possible future unknown events that may occur,
3. the probabilities associated with these events, and
4. the specific known consequences of each possible action-event pair.

Taking a Bayesian approach, the user's subjective probability distribution can be modified by information, e.g. that contained in a weather forecast. If the user is risk-neutral, he will choose the action that maximizes the *expected economic utility*. In the case of a non-linear utility function this decision may not be the same as the decision that would be taken based on the most likely outcome.

This process represents a prescriptive approach to decision-making, since it specifies the best action to be taken in the face of particular circumstances. A descriptive approach, by contrast, is an analysis in which weather information is used as a reference point for a subjective decision. A further distinction between decision frameworks is whether the decision problem is static or dynamic. A static decision problem involves an isolated decision, which is not affected by or does not affect other decisions based on a forecast, whilst a dynamic decision problem represents actions dependent on each other.

Examples of using probabilistic end-to-end response-variable forecasts in conjunction with von Neumann and Morgenstern's utility theory are given in studies by Smith et al. (2001) in relation to energy demand forecasting for several locations world wide and Roulston et al (2003) for wind energy production in the UK. PDFs of response variable forecasts were applied to the user's utility function to determine the decision that is associated with the maximum expected economic utility. In both cases, output from ECMWF's EPS was used as forecast data. Through this method, uncertainty indicated by the probabilistic forecast as well as the financial utility to the end user in relation to the response variable are incorporated into the decision-making process. Smith et al. (2001) showed the inclusion of probabilistic forecast information to be especially beneficial for users with tight profit margins or heavy penalties associated with asymmetric or non-linear utility functions. Similar findings made by Palmer (2002) in relation to the cost/loss model suggest that users with a low cost/loss ratio could face potentially ruinous losses if only considering the best-estimate, rather than the entire PDF.

According to Smith et al.’s (2001) findings, the skill of different methods of estimating the PDF (e.g. use of raw ensemble spread, calibrated ensemble spread, or adding the historical error distribution to the ensemble mean) varied with geographical location. However, the best estimate deterministic forecast on its own performed poorly at all locations. In concluding, Smith et al. (2001) noted that the actual value of a forecast to an end user depends not only on the quality of the forecast, but on the response variable model and the user’s utility function. Whilst the value of the forecast cannot be generalized for all users, the utility maximising approach can be used as a framework of assessing value for a specific user.

7.2 Application to a gas demand decision-making scenario

This current study draws on and extends Smith et al.'s (2001) conceptual framework as a basis for assessing economic value of temperature forecasts generated by EPS and MRF model output in relation to a specific decision-making process of the gas company as follows:

The Gas Company (GC) is responsible for the wholesale of gas to Regional Distribution Companies (RDCs), which it supplies through a pipeline system operated by the Transmission System Operator (TSO). Each day every RDC must nominate to the GC the amount of gas it wishes to purchase for the next days. In turn, each day the GC must nominate to the TSO how much gas it wishes to transport through the pipeline system. If the RDC's actual demand on the day nominated for deviates by more than the nomination tolerance allowance (Nt_{RDC}), the RDC must pay penalties to the GC (see fig.7.1 for a graphical example), which are calculated as follows:

$$P_{RDC} = \mathrm{Max}(0, C * (|Y - X_{RDC}| - Nt_{RDC})) \quad , \qquad (7.1)$$

where

C is the unit price of deviations,

Y is the actual demand,

X_{RDC} is the nomination of the RDC,

Nt_{RDC} is the nomination tolerance given by

$$Nt_{RDC} = (0.0151 * X_{RDC} + 0.0248 * (K - X_{RDC})) \quad ,$$

where

K is the RDC's contractual maximum daily offtake of gas.

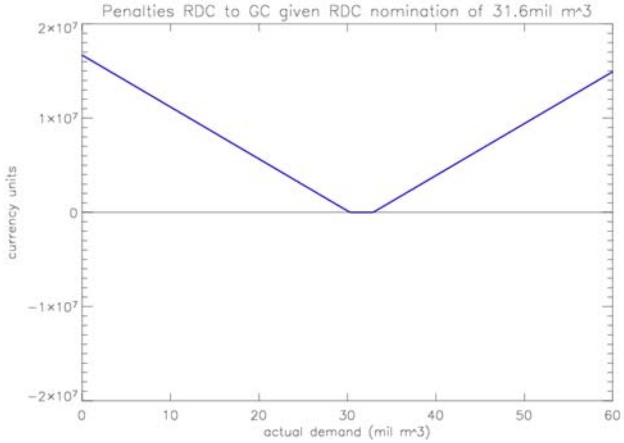


Figure 7.1: Penalties paid by the RDC to the GC as a function of actual demand, given a specific nomination of the RDC (in this case 31.6mil m^3). Amounts are positive since they represent positive cashflow from the viewpoint of the GC.

Equally, if the GC's actual amount of gas transported on the day nominated for deviates by more than the nomination tolerance allowance (Nt_{GC}), the GC must pay penalties to the TSO (see fig.7.2 for a graphical example), which are calculated as follows:

$$P_{GC} = -[\mathrm{Max}(0, C * (|Y - X_{GC}| - Nt_{GC}))] \quad , \qquad (7.2)$$

where

C is the unit price of deviations,

Y is the actual demand,

X_{GC} is the nomination of the GC,

Nt_{GC} is the nomination tolerance given by

$$Nt_{GC} = (0.0151 * X_{GC} + 0.0248 * (K - X_{GC})) \quad ,$$

where

K is the GC's contractual maximum daily delivery allowance of gas to the RDC via the Transmission System to the RDC.

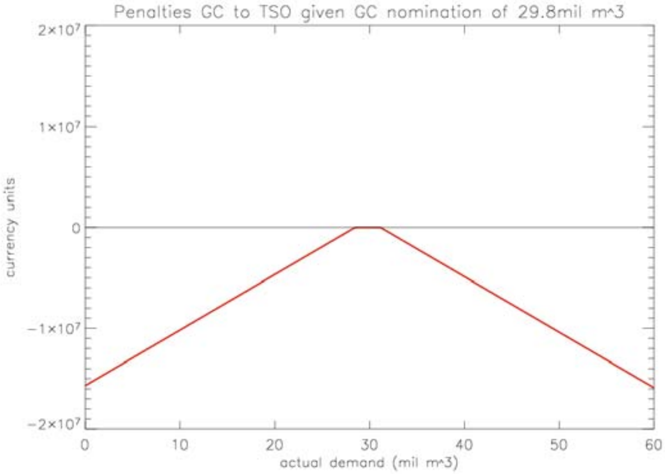


Figure 7.2: Penalties paid by the GC to the TSO as a function of actual demand, given a specific nomination of the GC (in this case 29.8mil m^3). Amounts are negative since they represent negative cashflow from the viewpoint of the GC.

Hence, the penalty payments of both RDC and GC, given a certain actual demand, vary depending on their respective nominations. Economic utility to the GC is represented by the net profit or loss to the GC. Summing the two penalty functions yields the economic utility function U(x_{gc},x_{rdc},y) to the GC for nominating an amount X_{GC}, given a specific nomination of the RDC (X_{RDC}) and an actual demand of Y (eqn. 7.3 and figure 7.3). (This assumes that the GC is risk-neutral).

$$U(x_{gc}, x_{rdc}, y) = P_{RDC}(x_{rdc}, y) - P_{GC}(x_{gc}, y) \qquad (7.3)$$

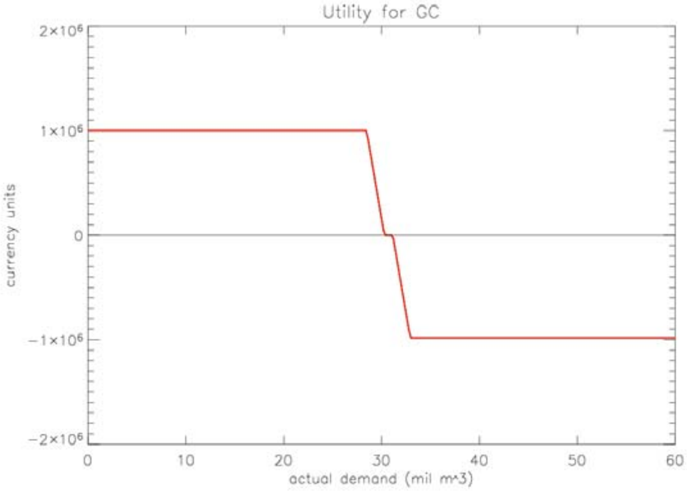


Figure 7.3: Economic Utility to the GC of nominating 29.8mil m^3, given the RDC nominates 31.6mil m^3, as a function of actual demand.

The example of figure 7.3 reveals that the economic utility function is asymmetric. This is difficult to visualize due to the overall magnitude of penalties. However, the magnitude of the cap (upper limit) of the depicted function is around 20000 currency units greater than the magnitude of the floor (lower limit). This asymmetry is due to the variability of Nt_{GC} and Nt_{RDC}, which are functions of X_{GC} and X_{RDC} respectively. Thus, if the probabilities of all possible outcomes of actual demand were symmetrically distributed around the RDC's nomination (X_{RDC}), it would make economic sense for the GC to nominate lower than the RDC, even if the RDC's nomination were the best estimate of actual consumption. However, this does not suggest exactly how much lower the GC should nominate. For this purpose it may be beneficial to use information about the uncertainty of demand estimates. Hence, the attention should turn to utilizing probabilistic information contained in demand forecasts in order to optimise the GC's nominations.

PDFs of gas demand, P(y), produced from end-to-end gas demand forecasts (as described in chapter 6) can be applied to the utility function to estimate the expected economic utility for a specific nomination X_{GC}, thereby incorporating the uncertainty in the forecast (see fig.7.4 for a graphical example). Due to the limited scope of this study, forecast errors are assumed to be normally distributed (this has generally been found to be a valid assumption for temperature forecasts at London Heathrow, e.g. Jewson, 2004c) and the PDFs modelled as Gaussians. However, preliminary analysis of the demand error distributions in this study suggests that this may not necessarily be the case. Hence, further investigation is needed into exploring the use of other probability models.

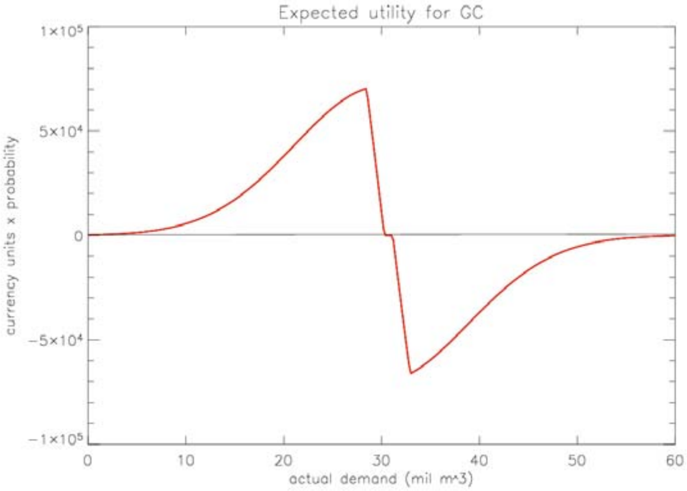


Figure 7.4: The utility function to the GC of nominating 29.8mil m^3, given the RDC nominates 31.6mil m^3, multiplied by a probabilistic demand forecast in the form of a Gaussian with μ = 30mil m^3 and σ = 6.27mil m^3. Expected utility is given by integrating over all possible values of actual demand.

The resulting function (utility function x forecast PDF) is then integrated over all possible demand values to obtain the expected utility for the GC's decision to nominate a specific gas amount X_{GC} (eqn. 7.4).

$$E[U](x_{gc}) = \int_y U(x_{gc}, x_{rdc}, y)p(y)dy \tag{7.4}$$

The GC nominates X_{GC} so as to maximise the integral and thus the expected utility:

$$E[U](X_{gc}) = \max_{xgc}\{E[U](x_{gc})\} \tag{7.5}$$

In fig. 7.4, the positive area under the curve is slightly larger than the negative area, if the GC nominates 29.8mil m^3. This implies higher economic utility than nominating according to the best estimate given by the ensemble mean of the same demand forecast probability distribution, 30mil m^3. The maximal expected utility based the probabilistic forecast in this example is given if the GC nominates 29.8mil m^3. If the GC nominated the identical amount which the RDC nominated, both areas would be zero, implying a guarantee of zero losses but also no possible profit. Hence, if the PDF of a probabilistic demand forecast is to a sufficient degree sharper than the climatological PDF, whilst still being consistent with observed probabilities, its use would lead to improved nominations and potential profit for the GC.

7.2.1 Empirical testing

Since the economic value of a forecast is determined by how much money it will make or save a user, a comparison of the GC's total profits/losses achieved over the entire 2004-2005 heating season by nominating according to different deterministic and probabilistic MRF and EPS multi-model forecasts was conducted.

However, to enable this comparison the nominations of the RDC must be defined first. Since it is assumed that the RDC uses some sort of temperature forecast, setting the RDC's nominations to climatology would not be a stringent enough test. For the purpose of this study, it was therefore assumed that the RDC nominates according to an end-to-end demand forecast based on a single integration deterministic temperature forecast. This was taken to be the control integration of the EPS. The GC would then nominate according to the best estimate, in the case of deterministic forecasts, or the maximal expected utility, in the case of probabilistic forecasts, derived using one of the following end-to-end demand forecasts:

EPS_Det	EPS deterministic best estimate (ensemble mean),
MRF_Det	MRF deterministic best estimate (ensemble mean),
Multi_Det	Multi-model deterministic best estimate (ensemble mean),
EPS_Hist	EPS probabilistic forecast using historical errors,
MRF_Hist	MRF probabilistic forecast using historical errors,
Multi_Hist	Multi-model probabilistic forecast using historical errors,
EPS_Ens	EPS probabilistic forecast using the raw ensemble spread,
MRF_Ens	MRF probabilistic forecast using the raw ensemble spread,
Multi_Ens	Multi-model probabilistic forecast using the raw ensemble spreads.

The probabilistic forecasts for this study were generated by fitting a Gaussian distribution around the ensemble mean. In the case of historical errors, the standard deviation was taken to be the standard deviation of the 2003-2004 error distribution of the ensemble mean demand forecast around demand observations. For PDFs based on the raw ensemble spread, the standard deviation of the demand forecast ensemble members for a particular target day was used. For the multi-model forecasts, the mean was taken as the mean of the EPS and MRF means, and the standard deviation as the mean of the standard deviations of the EPS and MRF.

7.2.2 Results

The first objective was to assess the value added by all forecasts derived from the two ensemble system and their combination in the multi-model vis-à-vis using the EPS single control integration. Total net payouts at each lead time for the 2004-2005 heating season are shown in fig. 7.5.

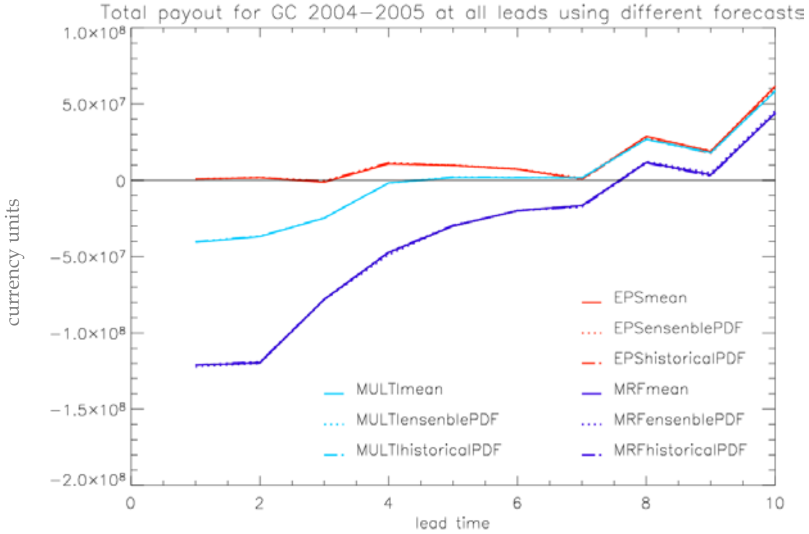


Figure 7.5: Total net payouts for the GC over the 2004-2005 season at specific lead times using different forecasts to nominate. Red lines represent EPS forecasts EPS_Det (solid), EPS_Ens (dashed) and EPS_Hist (dashed-dotted). The blue lines represent the equivalent forecast types using MRF data, and the turquoise lines those of the multi-model.

All forecasts based on the MRF ensemble (MRF_Det, MRF_Ens and MRF_Hist) exhibit large losses at short lead times, compared to the EPS control integration. However, losses decrease rapidly with lead time and profits are made beyond day 7. All forecasts based on the EPS ensemble (EPS_Det, EPS_Ens and EPS_Hist) generate profits at all lead times compared to the single integration, except on day 2. The amount of profit tends to increase with lead time, especially beyond day 7 (total profits for the 2004-2005 are just under 1mil currency units at lead 1 and around 62mil currency units at lead 10). This shows that whilst the MRF ensemble forecasts tend to cause losses, the value of all of the EPS ensemble forecasts to the GC in the nomination process is superior to that of the single integration at almost all lead times.

The multi-model profits/losses lie in between those of the two systems on their own (except at lead 7) with profits/losses being closer to those of the more profitable forecast, the EPS. However, relative to the MRF and EPS, the profits/losses are not as one may expect based on either the assumption that they may be the mean of the MRF and EPS, or that they would follow a similar relative pattern to the MRF and EPS as indicated by the Gain score in figure 5.17 (i.e. being superior to both forecasts at leads 4 to 8). Losses are incurred up to lead 4. Thereafter, profits are made and are similar to those of the EPS beyond lead 6. At lead 7, multi-model profits even exceed those of the EPS. These results confirm the hypothesis that the economic value of a forecast, or the relative value of several forecasting techniques, can only be determined when they are integrated into a specific-end-to-end application and decision-making process. Although profits are for the most part below those achieved by the EPS, further investigation into appropriate calibration methods is necessary to determine whether the multi-model technique could potentially add value in this context.

Furthermore, fig. 7.5 shows that revenues generated by the different forecasts (probabilistic and deterministic) produced from EPS ensembles are very similar at all lead times. This is also the case for all forecasts produced from MRF ensembles and the multi-model. This indicates that including probabilistic forecast information does not affect the net revenues as much as the choice of forecasting system or, in the case of the EPS, whether to use the ensemble mean or the control integration. A probable reason for the small differences between revenues achieved by probabilistic and deterministic forecasts is that the asymmetry of the utility function (fig. 7.3) is only slight, meaning that even in the event of a large spread of the probabilistic forecast distribution, the nomination giving the maximal expected utility based on the probabilistic forecast will be close to the nomination according to the ensemble mean best estimate. This once more highlights the user- and application-specific nature of forecast value. Other applications and decision-models with more asymmetrical utility functions may benefit more form probabilistic forecast information (e.g. as found by Smith et al., 2001).

Having found that using the EPS ensemble mean adds considerable value in the nominations process, but that the probabilistic forecasts appear to add similar amounts of value on the whole, the question remains whether probabilistic information in the form of past errors or that contained in the ensemble spread adds any additional value at all. To investigate this, the value of probabilistic EPS forecasts vis-à-vis the EPS ensemble mean was assessed by subtracting the profits achieved by the latter from those achieved by the former (fig. 7.6).

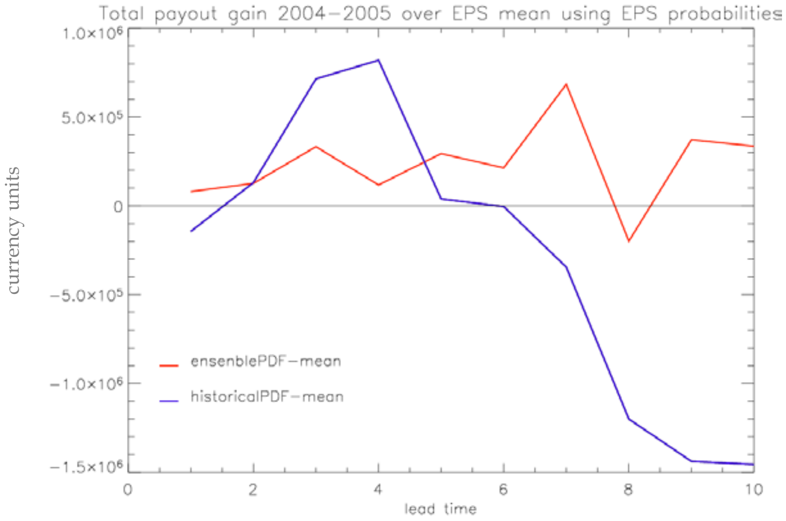


Figure 7.6: Total net payouts gained by the GC relative to using the EPS control integration during the 2004-2005 season at specific lead times using EPS_Ens (red line) and EPS_Hist (blue line).

Additional revenues achieved by the probabilistic forecast using the historical error distribution are very erratic over lead times. Profits are only made at leads 2-5 and comparatively heavy losses are suffered at longer leads. Hence, this probabilistic forecast produced from historical errors should not be used in its current form in the nomination process. On the other hand, the probabilistic information contained in the ensemble spread of the EPS does appear to add some limited value to optimizing nominations. Figure 7.6 shows a gain in profits at all lead times (except lead 8), with profits generally increasing with lead time. This gain in value (even if comparatively small) is encouraging, since it suggests that the ensemble spread may contain useful probabilistic information. Calibration of the spread (e.g. starting with methods outlined in chapter 6) or blending past error statistics and ensemble spread (e.g. Jewson's, 2004b, spread regression) may improve results. In addition, non-Gaussian probability models should also be investigated. Naturally, longer test periods would be necessary to establish robust results. As mentioned in chapter 6, the issue of how to calibrate the ensemble spread is vigorously debated, and any assertive statements about the value of the spread should at this stage be made with care.

Chapter 8: Conclusions and future work

The following answers can now be given to the four questions set in the introduction:

Question 1: What is the deterministic skill of the ensemble mean of an end-to-end gas demand forecast using raw ECMWF and NCEP temperature forecasts?

The ensemble mean provides a skilful deterministic estimate of gas demand. Results for the 2004-2005 heating season show the EPS to have a Gain over climatology of 3.8 mil m^3 at a lead time of 1 day (~14.1% of mean daily weather-dependent demand) and 1.8 mil m^3 at a lead time of 10 days. The RMSE was 1.5 mil m^3 at a lead time of 1 day and 4.7 mil m^3 at a lead time of 10 days.

Using the raw means of both ECMWF EPS and NCEP MRF ensemble forecasts individually to produce deterministic end-to-end demand forecasts for 1 to 10 days ahead considerably increases skill over climatology, persistence and a stochastic AR1 model. The difference in absolute errors is statistically significant at all lead times at the 0.05 level. The gain over persistence and the AR1 model is especially great at lead times of 3 to 9 days. In terms of Gain, the EPS is superior at all leads. The difference of absolute errors is significant at the 0.05 level at leads 1 to 3 and 10 in 2003-2004, and at leads 1 to 3 in 2004-2005. In terms of RMSE, the MRF is superior at leads 4 to 8 in 2003-2004. This suggests that whilst the MRF has greater absolute errors, it has fewer large magnitude errors at leads 4 to 8.

Forecast performance varies between the two seasons. RMSE of both systems is higher in 2003-2004 at all lead times. Gain over climatology is higher at short leads (around 1 to 4 days), but decreases more rapidly with increasing lead time. This is due to the rapid transitions between extremely warm and cold anomalies occurring during the first season, making temperature anomalies from climatology more difficult to predict than in the second season. Although rapid transitions lead to larger absolute forecast errors, using the MRF and EPS yields even higher skill in these situations than achieved by climatology or statistical models, since dynamical atmospheric processes can only be forecast by numerical models.

Despite providing a skilful estimate of the most likely outcome, the ensemble mean mostly underestimates extreme anomalies. This is observed even at leads as short as 1 day, and becomes increasingly pronounced with increasing lead time. It is especially evident in situations of high forecast uncertainty, when the ensemble spread grows rapidly and thus rapidly merges with the climatological spread. By consequence, the mean also rapidly merges with the climatological mean. Hence it is essential to consider the use of probabilistic information, in addition to the ensemble mean, especially if the user is sensitive to extremes.

Question 2: Can the deterministic skill be improved by post-processing methods?

Conventional bias correction methods of the ensemble mean add no further skill for predicting temperatures and gas demand at Prague, and in some cases degrade skill. Bias correction methods developed on one location cannot simply be transferred to another location without prior research, even if the same numerical weather prediction system is used.

Conventional bias correction methods of the ensemble mean add no further skill for predicting temperatures and gas demand at Prague, and in some cases degrade skill. This is due to the more erratic nature of forecast bias at Prague, which is probably attributable to climatic differences between Prague and London Heathrow, the station for which most existing methods have been developed. Prague's climate exhibits a stronger continental influence and a higher variance in temperatures, as well as being influenced by complex surrounding orography which modifies large scale flow modelled by the NWP systems. London Heathrow, on the other hand, has a lower variance in temperatures, is more maritime and not surrounded by major orography, and exhibits and a distinct annual cycle in its forecast bias pattern.

All running mean bias corrections degrade the skill of the EPS at all lead times. This effect increases with increasing lead time and is most pronounced when using the shortest (15 day) averaging period for calculating the unconditional bias. In the case of the MRF, running mean bias correction increases skill at leads up to 3 days, but thereafter substantially degrades skill. Again, the 15 day mean performs worst. This is probably due to the bias being estimated form preceding days. However, actual bias may vary depending on the state of the atmosphere when the forecast is produced and when the forecast verifies. This seems to explain why the degrading of skill increases with lead time, since the state of the atmosphere is more likely to change with increasing lead time.

Owing to the erratic nature of the forecast biases the simple linear regression model attempting to correct unconditional as well as flow-dependent bias does not lead to improvement in skill. Fitting first order harmonics to regression parameters in order to account for seasonal variations in biases results in some improvement in-sample, but in reduced skill at lead times beyond 6 days out of sample. This again points towards the high variability of errors at Prague, as well as to a more complex seasonal structure of bias patterns.

Question 3: Can the deterministic skill be improved by combining NCEP and ECMWF forecasts?

Both simple averaging as well as optimal weighting of the two ensemble means by linear regression enhances deterministic skill up to a lead time of around 7 days.

Though similar (probably due to underlying physical causes, such as synoptic conditions in which both forecasts have similar problems), error patterns of the EPS and MRF at Prague show greater dissimilarities than at Heathrow (exemplified by lower correlation coefficients across all lead times ranging for 0.70 to 0.88, compared with 0.80 to 0.92 at Heathrow). Qualitative analysis of large error events revealed the ten highest magnitude positive and negative errors of the EPS mostly do not coincide temporally with those of the MRF. This proves the potential for exploiting uncorrelated forecast errors in order to enhance overall skill. Additional skill is gained by combining the two ensemble means by linear regression, which optimally weights the two forecasts, up to a lead time of around 7 days. Beyond lead 7, the model is over-fitted and leads to a decrease in skill. Simply taking the mean of the two ensemble means adds skill at leads 4 to 8.

Question 4: What is the economic value of using raw and post-processed deterministic and probabilistic forecasts in a decision-making process?

Substantial value can be gained at almost all lead times when using the ensemble mean of the EPS as a deterministic forecast instead of the EPS single integration control for the user-specific application of nominating gas demand. This is especially the case at longer lead times, with total profits over the 2004-2005 season ranging from just under 1 mil currency units at a lead time of 1 day to around 62 mil currency units at lead 10. Additional value added by probabilistic forecasts derived from the ensemble spread is only limited (approximately 100000 currency units at lead 1, and 340000 currency units at lead 10).

Using the MRF ensemble mean generates losses up to lead 7 and profits thereafter are smaller than those achieved by the EPS ensemble system. Incorporating probabilistic information from a Gaussian distribution of historical errors does not add value, though it is necessary for this to be further investigated, e.g. by using different probability models. Although the added value of probabilistic information from the raw ensemble spread is comparatively small, it is encouraging that it was shown to add *some* value. The limited size of this added value may be due to the decision model in this specific application only having a small asymmetry and hence maximal utility often coinciding with the deterministic best estimate. In addition, the spread had not even been calibrated and only a Gaussian probability model was tested. Hence, it is essential that further research is conducted into forecast calibration, since this is the only means by which maximum benefit can be extracted from the information inherent in the forecast.

Using the simplest version of the multi-model forecast, the mean of the two demand forecasts, produced no added value vis-à-vis using the EPS ensemble on its own. However, judging by the added deterministic skill of multi-model forecasts shown in chapter 5, research into appropriate calibration may in future make it possible to gain some additional value with regard to gas demand nominations.

8.1 Future work

Calibration of probabilistic forecasts

The most important continuation of this project is the calibration of probabilities derived from the ensemble spreads of the MRF, EPS and the multi-model method for use in the nomination decision-making framework. Only then could the full value residing in the EPS and MRF forecasts be fully exploited. This may include Bayesian methods relating bias to atmospheric state, as well as applying non-Gaussian probability models. Relaxing the normality assumption is an equally important avenue to be explored in the context of probabilistic inferences derived from historical errors of the ensemble means, since analyses conducted in relation to this study have shown that forecast errors are not always normally distributed.

Identifying other variables that may affect errors

As this study suggests, bias correction is no trivial matter. Forecast error structures can be complex, location-specific and are to some extent related to atmospheric state. Hence, it is essential that synoptic flow and pressure patterns (e.g. the North Atlantic Oscillation), as well as upper air data are considered as potential predictors. These may provide important and perhaps more reliable information on bias structures. The greatest obstacle currently posed to this research remains the lack of long records of past forecasts from the model in current operational use. At the moment, over-fitting of statistical models due to small sample size presents a great danger. Nonetheless, in the absence of these data, a partial solution to dealing with the noise in the NCEP error pattern is to make use of re-forecasts produced by an older version of the NCEP model, which have just become available (Norton, pers. com). Key variables showing a relationship to model bias could be identified by fitting statistical models to this data set. Thereafter, regression coefficients could be tuned using the comparatively short archived data of the current NCEP model.

Interpolating from numerical model grid

In addition, more sophisticated methods of interpolation should be investigated. For the present study, 2m temperatures values of the four nearest numerical model gridpoints were linearly interpolated to Prague Ruzyne. Variable weighting of the gridpoints, related to flow patterns or stability conditions on a particular day, i.e. by drawing on statistical relationships between a location's climatological data and gridded model output using methods such as kriging and spatial clustering (Gutierrez et al., 2004) may provide a more representative interpolation.

Extending research to other stations and fields

The application considered in this study requires point-specific temperature forecasts, since the gas model uses HDDs at a specific station as an input. However, this could be expanded in future research to include a group of stations or even larger spatial fields. This may not only prove to be highly valuable for the gas industry (gas consumers are usually spread out over a larger geographical area), but also for developing more comprehensive theories and solutions for forecast calibration in general.

Extending research to other decision models

Having confirmed that the value of a temperature forecast not only depends on the forecast itself, but also on its user-specific application, it should be the priority of applied meteorologists and forecast users alike to assess the value of forecasts such as those used in this study in the context of different end-to-end forecasting and decision-making processes. This would be especially important in the context of highly non-linear demand and utility functions.

References

Banks, E. (2001). Weather Risk Management: Markets, products and applications. Palgrave.

Boehm, R. (1989). Klimatisch bedingte Heizkosten im Raum Wien in Abhaengigkeit von Bebauungsdichte und Orographie. *Wetter und Leben 41, 259-267.*

Boi, P. (2004). Probabilistic temperature forecast by using ground measurements and ECMWF ensemble prediction system. *Meteorological Applications, 11 (4), 301-309.*

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78, 1-3.*

Buizza, R., Palmer, T.N. (1995). The singular vector structure of the atmospheric general circulation. *Journal of Atmospheric Science, 52 (9), 1434-1456.*

Buizza, R., Houtekamer, P.L., Toth, Z., Pellerin, G., Wei, M., Zhu, Y. (2005). Assessment of the Status of Global Ensemble Prediction. *Monthly Weather Review, 133 (5), 1076-1097.*

Coelho, C.A.S., Pezzulli, S. (2004). Forecast Calibration and Combination: A Simple Bayesian Approach for ENSO. *Journal of Climate, 17 (7), 1504-1516.*

Czech Airport Authority. http://www.csl.cz/en/letiste/pr_technickeudaje.htm (accessed 04.05.2005).

Deque, M. (2003). Continuous Variables. Pages 97-113 in Jolliffe, I.T., Stephenson, D.B. (eds) Forecast Verification: A Practitioner's Guide in Atmospheric Science, Wiley.

Domonkos, P., Piotrowicz, K. (1998). Winter temperature characteristics in central Europe. *International Journal of Climatology, 18 (13), 1405-1417.*

Domonkos, P., Kysely, J., Piotrowicz, K., Likso, T. (2003). Variability of extreme temperature events in south-central Europe during the 20th century and its relationship with large-scale circulation. *International Journal of Climatology, 23 (9), 987-1010.*

Fisher, R. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics, 41, 155-160.*

Gerstengarbe, F.-W., Werner, P.C., Ruege, U. (1999). Katalog der Grosswetterlagen Europas nach Paul Hess und Helmuth Brezowski 1881-1991. Deutscher Wetterdienst, Offenbach.

Gilmour, I. (2004). Using weather forecasts in energy trading. ECMWF Formal Seminar.

Goeber, M., Wilson, C.A., Milton, S.F., Stephenson, D.B. (2004). Fairplay in the verification of operational quantitative precipitation forecasts. *Journal of Hydrology, 288 (1-2), 225-236.*

Gutierrez, J.M., Cofino, A.S., Cano, R., Rodriguez, M.A. (2004). Clustering methods for statistical downscaling in short-range weather forecasts. *Monthly Weather Review, 132 (9), 2169–218.*

Heerdegen. R.G. (1988). Evaluation of the heating degree-day index. *Weather and Climate, 8 (2), 69-75.*

Hervada-Sala, C., Pawlowsky-Glahn, V., Jarauta-Bragulat, E. (2000). A statistical method to downscale temperature forecasts. A case study in Catalonia. *Meteorological Applications, 7 (1), 75-82.*

Huth, R. (1999). Statistical Downscaling in Central Europe: evaluation of methods and potential predictors. *Climate Research, 13 (2), 19-101.*

Huth, R. (2002). Statistical Downscaling of Daily Temperature in Central Europe. *Journal of Climate, 15 (13), 1731-1742.*

Jenkinson, A.F., Collison, F.P. (1977). An initial climatology of gales over the North Sea. *Synoptic Climatology Branch Memorandum No. 62, Meteorological Office, Bracknell.*

Jewson, S. and Caballero, R. (2003). The use of weather forecasts in pricing weather derivatives. *Meteorological Applications, 10 (4), 377-389.*

Jewson, S. (2004a). Making use of the information in ensemble weather forecasts: comparing end-to-end and full statistical modelling approaches. www.stephenjewson.com (accessed 15.05.2005).

Jewson, S. (2004b). Probabilistic temperature forecasting: a summary of our recent research results. www.stephenjewson.com (accessed 15.05.2005).

Jewson, S. (2004c). Probabilistic temperature forecasting: a comparison of four spread-regression models. www.stephenjewson.com (accessed 20.05.2005).

Jewson, S. and Ziehmann, C. (2004). Five guidelines for the evaluation of site-specific medium range probabilistic temperature forecasts. www.stephenjewson.com (accessed 20.05.2005).

Jewson, S., Brix, A., Ziehmann,C. (2005). Weather Derivative Valuation: The Meteorological, Statistical, Financial and Mathematical Foundations. Cambridge University Press.

Johnson, S.R., Holt, M.T. (1997). The Value of Weather Information. Pages 75-108 in Katz, R.W., Murphy, A.H. (eds) Economic Value of Weather and Climate Forecasts. Cambridge University Press.

Jolliffe, I.T., Stephenson, D.B. (2003) Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley.

Kalnay, E., Toth, Z. (1996). Ensemble Prediction at NCEP. Preprints, 11th AMS Conference on Numerical Weather Prediction.

Kysely, J. (2002). Temporal fluctuations in heat waves at Prague-Clementinum, the Czech Republic, from 1901-97, and their relationships to atmospheric circulation. *International Journal of Climatology, 22 (1), 30-50.*

Lamb, H.H. (1972): British Isles weather types and a register of daily sequence of circulation patterns, 1861-1971. *Geophysical Memoir 116, HMSO, London, 85pp.*

Lefaivre, L., Houtekamer, P.L., Bergeron, A., Verret, R. (1997). The CMC Ensemble Prediction System. *Proc. ECMWF 6th Workshop on Meteorological Operational Systems, Reading, UK, ECMWF, 31-44.*

Leith (1974). Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review, 102 (6), 409-418.*

Lorenz, E. (1969). The predictability of a flow which possesses many scales of motion. *Tellus, 21, 298-307.*

Lyster, P.M., Guo, J., Clune, T., Larson, JW (2004). The computational complexity and parallel scalability of atmospheric data assimilation algorithms. *Journal of Atmospheric and Oceanic Technology, 21 (11), 1689-1700.*

Mailier, P.J.A. (2001) Ensemble prediction of extreme mid-latitude cyclones. MSc dissertation, University of Reading.

Mailier, P.J.A. (2005). Forecast Verification. Climate Analysis Group Seminar, University of Reading.

Mason, S.J. (no date). Definition of technical terms in forecast verification. http://iri.columbia.edu/outreach/education/ForecastTerm/ (accessed 0.6.06.2005).

McGuffie, K., Henderson-Sellers, A. (1999). A Climate Modelling Primer. Wiley.

McSharry, P.E., Bouwman, S., Bloemhof, G. (2005). Probabilistic forecasts of the magnitude and timing of peak electricity demand. *IEEE Transactions of Power Systems, 20 (2), 1166-1172.*

Mylne, K.R., Evans, R.E., Clark, R.T. (2002). Mulit-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Quarterly Journal of the Royal Meteorological Society, 128 (579), 361-384.*

NCEP Ensemble Homepage http://wwwt.emc.ncep.noaa.gov/gmb/ens/info/ens_detbak.html (accessed 02.05.2005).

NCEP Reanalysis http://www.cdc.noaa.gov/cdc/reanalysis/ (accessed 05.07.2005).

Palmer, T.N. (2000). Predicting uncertainty in forecasts of weather and climate. *Report on Progress in Physics, 63 (2), 71-116.*

Palmer, T.N. (2002) The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society, 128, (581), 747-773.*

Pelly, J.L., Hoskins,B.J. (2003). How well does the ECMWF Ensemble Prediction System predict blocking? *Quarterly journal of the Royal Meteorological Society, 129 (590), 1683-1702.*

Persson, A. (2001). User Guide to ECMWF forecast products. www.ecmwf.int (accessed 02.07.2005).

Potts, J.M. (2003). Basic Concepts. Pages 13-36 in Jolliffe, I.T., Stephenson, D.B. (eds) Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley.

Quayle, R.G., Diaz, H.F. (1980). Heating degree day data applied to residential heating energy consumption. *Journal of Applied Meteorology 19 (3), 241-246.*

Richardson, D.S. (2003). Economic Value and Skill. Pages 165-186 in Jolliffe, I.T., Stephenson, D.B. (eds) Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley.

Rodwell, M.J., Doblas-Reyes, F. (2004) Predictability and Prediction of European Climate. ECMWF, Reading.

Roulston, M.S., Kaplan, D.T., von Hardenberg, J., Smith, L.A. (2003). Using medium-range weather forecasts to improve the value of wind energy production. *Renewable Energy, 28 (4), 585-602.*

Smith, L.A., Roulston, M.S., von Hardenberg, J. (2001). End to end ensemble forecasting: Towards evaluating the economic value of the Ensemble Prediction System. *ECMWF Technical Memorandum No.336.*

Stephenson, D.B., Coelho, C.A.S., Doblas-Reyes, F.J., Balmaseda, M. (2005). Forecast Assimilation: A Unified Framework for the Combination of Multi-Model Weather and Climate Predictions. *Tellus A, 57 (3), 253-264.*

Taylor, J.W., Buizza, R. (2003). Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting, 19 (1), 57-70.*

Taylor, J.W., Buizza, R. (2004) A comparison of temperature density forecasts from GARCH and atmospheric models. *Journal of Forecasting, 23 (5), 337-355.*

Thornes, J.E., Stephenson, D.B. (2001). How to judge the quality and value of weather forecast products. *Meteorological Applications, 8 (3), 307-314.*

Tibaldi, S., Molteni, F. (1990). On the operational predictability of blocking. *Tellus 42 A, 343-365.*

Toth, Z., Kalany, E. (1997). Ensemble Forecasting at NCEP and the Breeding Method. *Monthly Weather Review, 125 (12), 3297-3319.*

Toth, Z., Talagrand, O., Candille, G., Zhu, Y. (2003). Probability and Ensemble Forecasts. Pages137-162 in Jolliffe, I.T., Stephenson, D.B. (eds) Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley.

Toth, Z., Zhu, Y., Wobus, R. (2004) March 2004 upgrade of the NCEP global ensemble forecasting system. http://wwwt.emc.ncep.noaa.gov/gmb/ens/NAEFS-pdf/256,1,MARCH 2004 UPGRADE OF THE NCEP GLOBAL ENSEMBLE FORECAST SYSTEM (accessed 23.07.2005).

van den Berg, W.D. (1994). The role of various weather parameters and the use of worst case forecasts in prediction of gas sales. *Meteorological Applications 1, 33-37.*

von Neumann, J., Morgenstern, O. (1944). Theory of Games and Economic Behavior. 1953 edition, Princeton, NJ: Princeton University Press.

Wallen, C.C. (1977). Climates of Central and Southern Europe. In World Survey of Climatology, vol. 6, Landsberg, H.E. (ed.). Elsevier.

Wilks, D.S. (1995). Statistical Methods in the Atmospheric Sciences. Academic Press.

Wilks, D.S. (1997). Forecast value: prescriptive decision studies. Pages 109-146 in Katz, R.W., Murphy, A.H. (eds). Economic Value of Weather and Climate Forecasts. Cambridge University Press.

WMO http://www.wmo.ch/web/www/DPS/EPS-HOME/eps-home.htm (accessed 19.07.2005).

Ziehmann, C. (2000). Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus, 52, 280-299.*

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., Mylne, K. (2002). The economic value of ensemble-based weather forecasting. *Bulletin of the American Meteorological Society, 83 (1), 73-83.*

Map:

The map segment of central Europe was created using the MapInfo GIS package and Bartholomew Digital Data.

Personal communications:

Mailier, P.J.A.: Weather consultant, University of Reading.

Norton, W.A.: Senior research scientist, University of Reading. Director of Technology at Weather Informatics Ltd.

O'Neill, A.: Director of the Data Assimilation Research Centre, University of Reading. Chairman of Weather Informatics Ltd.