Draft to the 2004 *User Guide to ECMWF forecast products (Oct 2003)*

by Anders Persson

Preface

This User Guide to the ECMWF Products is not like most other "user guides", which provide clear and straightforward instructions how to "plug in", "get started", "execute" and "switch off". Nor is this Guide a handbook in NWP, dynamic meteorology or weather forecasting; the aim of the Guide is to facilitate the use of traditional ECMWF medium range forecast products, and encourage the use of newer, more advanced products such as the wave forecasts, seasonal forecasts and forecasts from the Ensemble Prediction System (EPS).

After a short overview of the history of NWP and a background to the creation of the ECMWF, there follows a non-technical description of the forecast system, including the data assimilation. For the same reason that it is possible to drive a car without knowing exactly how the automatic gearing system works, so is it possible to make use of the output from a NWP forecast system without knowing the crucial mechanical details. Strong and weak sides of the forecast system will be addressed in a special chapter.

After presenting the most common forecast fields the guide will address issues related to the interpolation and plotting of meteorological fields. A chapter on forecast verifications discusses the problem of forecast quality. The interpretation of statistics is never trivial: what at first sight appears to be "good" might be "bad", and what looks "bad" might turn out to be "good".

There are some basic principles of interpreting deterministic NWP products, in particular medium-range forecasts. Most important of these is the relation between the atmospheric scale and its predictability. Another important aspect is forecast "jumpiness", which is normally seen as a nuisance. It can, however, be used in a productive way to indicate possible alternative forecast development and thus serve as an introduction to the EPS.

The EPS is sometimes portrayed as a new, revolutionary way of making weather forecasts. Rather it is a logical development of traditional weather forecasting. The aim of weather forecasting is not to predict meteorological parameters for their own sake, but to provide input in decision making processes. Forecasters have always been trying to tell what is *most likely* to happen, what *might* happen and what will *probably not* happen. The optimum way to convey this information is in terms of probabilities.

The EPS provides an overwhelming amount of probability information and offers an almost unlimited combinations of products. It is a challenge to the forecaster to convey the relevant parts of the EPS information to the end-customers or the public, taking the system's shortcomings into account. For this purpose automatic use of the forecasts products becomes increasingly necessary. One chapter is therefore devoted on discussing some typical problems of statistical modification or interpretation of the NWP output.

A fairly new field of great potential value is forecasts beyond ten days. Although detailed day-to-day weather forecasts might not be possible beyond a week, experiments have shown that there at present is some skill in forecasting the large scale air mass patterns up to two weeks. The ECMWF monthly and seasonal forecasts try to go even further by taking advantage of the mutual interaction between the atmosphere and the oceans.

Acknowledgement: This Guide is the fruit of several years of discussions with scientists at EMWF and weather forecasters, both from Europe and elsewhere. It has been the interaction with these two specialized groups, trying to draw the scientific consequences of the forecasters' experiences, trying to see the practical implications of the scientific exploration, which has been the main driving force and inspiration for this publication.

1. Presentation of ECMWF

The European Centre for Medium Range Weather Forecasts (ECMWF) is the result of more than 100 years of development in dynamic and synoptic meteorology and more than fifty years of development in numerical weather prediction (NWP). This long period has seen times of optimism vary with pessimism, during which a slow but steady improvement of the quality of weather forecasts has been made.

1.1. The history of NWP

In 1904, the Norwegian hydrodynamist, Vilhelm Bjerknes, then professor in Leipzig, suggested that the weather could be quantitatively predicted by applying the complete set of hydrodynamic and thermodynamic equations to carefully analysed initial atmospheric states. But lacking both the theoretical and practical means to make any quantitative predictions he initiated instead the qualitative approach that later became known as the “Bergen school”. The apparent failure at about the same time by the British genius Lewis F. Richardson’s hand-calculations seemed to confirm that NWP was a practical impossibility[1].

After the second world war two technological developments made mathematical forecasts along the lines suggested by Bjerknes possible: the establishment of a hemispheric network of upper-air stations and the development of the first electronic computers. However, it soon turned out that the nature of the problem was much more complicated than envisaged. That is why the first useful NWP forecast systems during the 1950’s were very simple: in principle based on conservation of absolute vorticity in the mid-troposphere.

Only in 1962 could the US launch the first operational quasi-geostrophic baroclinic model, followed by Britain in 1965. By that time, work was already under way, to introduce more realistic models, based on the primitive equations (PE). Theses models could more easily incorporate important physical processes like convection. Since they were not restricted by any geostrophic constraints, they could also cover the tropical latitudes. The first global PE model began operating in 1966 at NMC Washington, with a 300-km grid and six-layer vertical resolution. This model turned out to have great similarities with Richardson’s model 45 years earlier.

1. Richardson had based his forecast on a situation 20 May 1910. On that day the earth passed through the tail of Halley’s comet and a lot of balloon based measurements of wind, pressure and temperature were made all over Europe to detect any extra-terrestial influence on the atmospheric flow.

1.2.The creation of ECMWF

There were in the late 1960's moves in Europe to build up a similar system as the Americans. From the experience gathered with short-range and climatological simulations there was enough know-how to motivate an attack on the medium-range forecast problem, defined as the interval from 3 to 10 days ahead. The scientific and technological problems were still formidable, which made the subject ideal for multi-national co-operation.

In October 1967 the Council of Ministers of the European Communities adopted a resolution to implement a programme to promote joint scientific and technical research. A proposal for a "European Meteorological Computer Centre for Research and Operations" occupied the first place on a list of meteorological projects submitted by an expert group in April 1969. The proposal was accepted and other European nations were invited to participate. In April 1970 an expanded expert group initiated two study groups to look into the economic and scientific motivation for the project.

A decision was made late in 1971 create the European Centre for Medium-Range Weather Forecasts with the aim to produce weather forecasts ten days ahead with the five-day forecasts having the same accuracy as subjective two-day forecasts in the 1950's. The ECMWF convention was signed in October 1973 by nineteen European States: Austria, Denmark, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, United Kingdom and Yugoslavia. The objectives of the ECMWF were laid down as follows:

To develop dynamic models of the atmosphere with a view to preparing medium-range weather forecasts by means of numerical methods;

To prepare, on a regular basis, the data necessary for the production of medium-range weather forecasts;

To carry out scientific and technical research directed towards the improvement of these forecasts;

To collect and store appropriate meteorological data;

To make available to the meteorological offices of the Member States, in the most appropriate form, the results of the studies and research provided for in the first and third objectives above and the data referred to in the second and fourth objectives;

To make available a sufficient proportion of its computing capacity to the meteorological offices of the member States for their research, priority being given to the field of numerical forecasting. The allocation of the proportions would be determined by Council;

To assist in implementing the programmes of the World Meteorological Organization;

To assist in advanced training for the scientific staff of the meteorological officers of the Member States in the field of numerical weather forecasting.

Since 1979 co-operation agreements have been concluded with Iceland, Hungary, Croatia, Slovenia, the Czech Republic, Romania, Malta, WMO, EUMETSAT and ACMAD.

The first operational forecast was produced on 1 August 1979. Since then ECMWF has made at least one ten-day forecast per day, and distributed it from its computer system to the systems of the national meteorological services of its Member States via a dedicated telecommunication network.

1.3. The ECMWF forecasting system since 1979 – an overview

The ECMWF forecasting system consists of five components: a general circulation model, an ocean wave model, a data assimilation system and since 1992 an ensemble forecast system. In 1998 a seasonal forecasting system started to operate and in 2002 a monthly forecasting system.

1.3.1.The general circulation model

The first ECMWF numerical model was a grid-point model with 15 levels in the vertical up to 10 hPa, and horizontal resolution of 1.875 degrees of latitude and longitude, corresponding to a grid length of 200 km on a great circle. In April 1983 this grid-point model was replaced by a model with a spectral representation in the horizontal with a triangular truncation at wave-number 63. In 1985 the resolution was increased to 106 spectral components, in 1991 to 213, in 1998 to 319 and in 2000 to 511. In 1985 the number of vertical levels increased to 19, in 1991 to 31, in 1999 to 50 (up to 0.1 hPa) and later in the same year to 60 levels.

At that time of the introduction of the spectral technique it was more accurate than the grid point model for the same computational cost. With increased resolution and the introduction of the semi-lagrangian technique, there is no longer any significant difference in accuracy between the two representations.

In 1995 an explicit cloud scheme was introduced with clouds as prognostic parameters. It not only improved the cloud and precipitation forecasts, but it had also a significant impact on the model dynamics, not only in the 10-day integration, but also on the preliminary fields for the analysis. Ozone was added as a predicted variable in 1999.

1.3.2.The ocean wave model

A global wave model plus a limited area model for the North Atlantic and the European waters became operational in 1992. In 1998 the wave model was integrated into the atmospheric model allowing two-way interaction of wind and waves. It is now also incorporated in the monthly, seasonal and ensemble systems.

1.3.3. The data assimilation and analysis system

The "optimum interpolation" analysis method remained within the ECMWF system up to 1996. By then the increasing availability of asynoptic data, particular over the oceans, had stimulated research into more advanced analysis procedures, like the variational data assimilation where the concept of a continuous feed back between observations and model was put on a mathematical foundation.

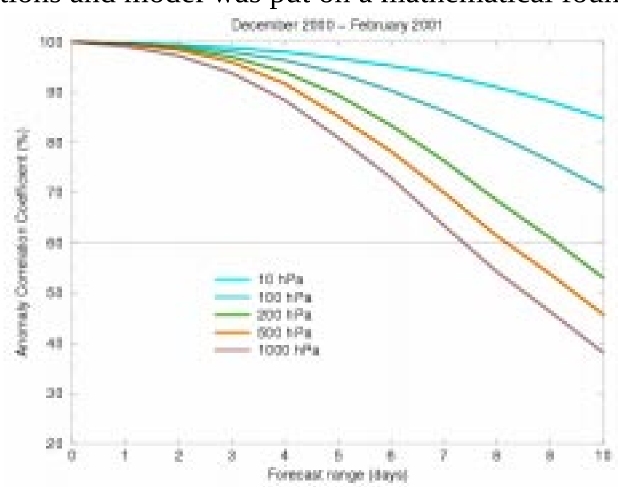


Figure 1 : Anomaly Correlation Coefficient for ECMWF forecasts for different levels over the Northern Hemisphere winter 2000-2001. The skill of the forecasts increase with height. Skilful ten day forecasts has been a reality for the stratosphere for long times.

In 1991 a one-dimensional variational scheme (1DVAR) was introduced for satellite radiance assimilation, in 1996 a three dimensional (3DVAR) for all types of observations. In 1997 this scheme was upgraded with the implementation of a four-dimensional system (4DVAR). The variational technique has become even more important due to the last decades' gradual reduction of the radio sounding network and the explosive increase in satellite data. Today the impact of satellite radiances together with other satellite data is much larger than the impact of radio sondes in the Northern Hemisphere.

The development of variational techniques has progressively allowed for a direct assimilation of infrared and microwave sounder radiances, which impact on analysed temperature and humidity fields.

1.3.4. The Ensemble Prediction System (EPS)

The EPS simulates possible initial uncertainties by adding, to the original analysis, small perturbations within the limits of uncertainty of the analysis. From these alternative analyses, a number of alternative forecasts are produced. At its start in December 1992 the ensemble system was run with 32 members using a T63 model with 31 vertical levels. In autumn 1996 the number of members was extended to 50 and the model was upgraded to TL159, to 40 levels in 1999. In autumn 2000 the resolution was increased to TL255. In 1998 a wave model was included together with a crude allowance for the uncertainty of physical processes. In connection with tropical cyclones specially designed perturbations are created in the tropics.

1.3.5. Monthly and seasonal forecasts

In 1997 a seasonal forecasting system started and is currently run at a TL95L40 resolution. In March 2002 a programme for experimental monthly forecasts in a TL159L40 resolution was started. It is currently running twice a month.

2. The ECMWF global atmospheric model

The ECMWF general circulation model, T_L511L60, consists of a dynamical component, a physical component and a coupled ocean wave component. The model formulation can be summarised by six basic physical equations, the way the numerical computations are carried out and the resolution in time and space.

2.1. The model equations

Of the six equations governing the ECMWF model, two are diagnostic and tell us about the static relation between different parameters:

-The GAS LAW gives the relation between pressure, density and temperature

-The HYDROSTATIC EQUATION gives the relation between the air density and the decrease of pressure with height. This is only an approximation of the real atmosphere, valid for horizontal scales larger than about 20km. It eliminates fast propagating sound waves which would otherwise be part of the solution.

The remaining four equations are prognostic and describe the dynamic changes over a short time interval of the horizontal and vertical wind components, temperature and water vapour contents of an air parcel, and the surface pressure

-The EQUATION OF CONTINUITY is an expression for the conservation of mass and is needed to determine the vertical wind speed and the change in surface pressure.

-The EQUATION OF MOTION describes the acceleration and deceleration of the speed an air parcel due to the pressure gradient force, and how the Coriolis force contributes to affect the direction. Included are also the effects of turbulent drag, gravity wave breaking and momentum transport due to moist convection.

-The THERMODYNAMIC EQUATION expresses how an air parcel's temperature is changing by adiabatic cooling or warming during vertical displacements. Other physical processes like condensation, evaporation, turbulent transport and radiative effects are also included.

-The CONSERVATION OF MOISTURE AND OZONE is an expression for the conservation of moisture content, except for losses due to condensation and precipitation, or gains by evaporation from clouds and rain, or from oceans and continents. There are also specific prognostic equations for the cloud fraction, water, ice content and ozone.

Physical processes like radiation, turbulence, friction and formation of clouds are also governed by the basic equations, but are, due to their small scales, described in a statistical way as Parametrization processes (see 3.3).

2.2. The numerical formulation

Despite the advances in computer technology there is a pressing need to solve the model equations as efficiently as possible. This requires elaborate numerical schemes to ensure stability and accuracy while using large time-steps to progress the forecast[1].

2.2.1. Introduction

The traditional numerical scheme since the start of NWP has been a so-called Eulerian scheme. To illustrate its main principle, consider a simple one-dimensional case with a parameter Q varying along the x-axis. The local change due to advection by the wind U (blowing along the x-axis) is

$$\frac{\partial Q}{\partial t} = -U\left(\frac{\partial Q}{\partial x}\right)$$

Eulerian schemes required rather small time steps to avoid numerical instability: the quantity Q must not be advected more than one grid length per time step. The maximum time step is therefore defined by the strongest winds.

This problem is overcome by a Lagrangian numerical scheme where the quantity Q is assumed to be conserved for an individual particle in the advection process along its trajectory.

$$\frac{dQ}{dt} = 0$$

1. In the early 1980's NWP products would sometimes contain systematic anomalies which partly resulted from the numerical schemes used at the time. Some knowledge of the underlying principles of the numerical scheme would allow users of NWP products to distinguish meteorological features from numerical artefacts. However, due to the increased complexity of all components of the forecasting system it is usually not possible any more to easily make this distinction 'by-eye'

The drawback is that with a pure Lagrangian framework it would be impossible to maintain uniform resolution over the forecast region. A set of “ear marked” particles, would ultimately result in dense congestion at some geographical locations, complete absence in other. To overcome this difficulty a *semi-Lagrangian* scheme is has been developed

2.2.2. The semi-Lagrangian numerical scheme

In this numerical scheme at every time step the grid-points of the numerical mesh are representing the arrival points of backward trajectories at the future time. The point reached during this back-tracking defines where an air parcel was at the beginning of the time-step. During the transport the particle is subjected to various physical and dynamical forcing. Essentially, all prognostic variables are then found through interpolation (using values at the previous time-step for the interpolation grid) to this departure point.

Figure 2 : A schematic picture of the semi-lagrangian scheme

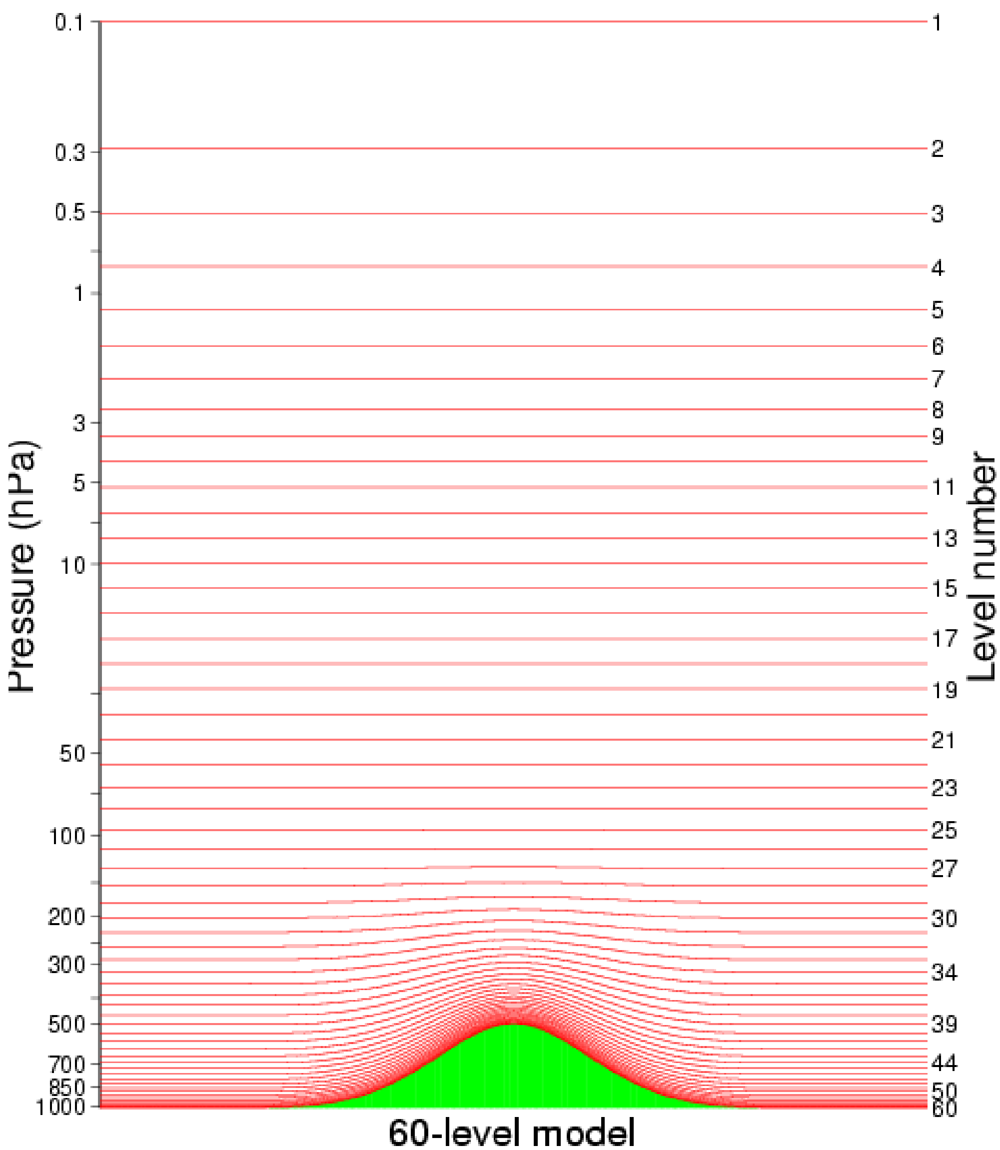
In contrast to the Eulerian framework the semi-Lagrangian scheme allows the use of large time-steps without limiting the stability. The limitations for stability are that trajectories should not cross and particles should not "overtake" another. Therefore, the choice of time-step in the semi-Lagrangian scheme is only limited by numerical accuracy. However, despite its stability properties severe truncation errors may cause misleading results.

2.2.3. The horizontal resolution in the free atmosphere

A spectral method is used for the representation of upper-air fields and the computation of the horizontal derivatives. It is based on a spherical harmonic representation, triangularly truncated at total wave number 511. This roughly corresponds to a grid length of about 40 km.

2.2.4. The vertical resolution

The atmosphere is divided into 60 vertical layers up to 0.1 hPa (about 64 km) just above the stratopause. The vertical resolution (measured in terms of geometrical height) is finest in the planetary boundary layer and coarsest in the stratosphere and lower mesosphere. There are as many levels in the lowest 1.5 km of the model atmosphere as in the highest 25 km. There are also four layers in the soil down to 1.9 meters.



The layers are not necessarily horizontal or isobaric. These, so called η-levels, are actually a hybrid of the ways to define horizontal levels. In the *lower most troposphere*, where the Earth's orography accounts for large variations, the η-levels follow the earth's surface. In the *upper troposphere, stratosphere and lower mesosphere* the η-levels are identical to surfaces of constant pressure. Between these types there is a smooth transition.

2.2.4. Resolution in time

The present T_L511 system uses a temporal resolution of 15 minutes, i.e. the dynamic equations describe the change of state of the atmospheric variable over 15 minutes period. This 15-minute forecast defines a new state from which another 15-minute forecast is made. The choice of 15 minutes has been made to on one hand ensure enough accuracy, on the other hand avoid numerical instabilities. It takes 960 such quarter-of-an-hour time steps to complete a ten-day forecast.

2.2.5. Different model versions

The different versions of the ECMWF forecast models are named after their horizontal and vertical resolutions. The current operational model with 511 linear spectral components and 60 levels is called T_L511L60, the slightly simplified version for the ensemble forecasts T_L255L40, the version used for the 4DVAR T_L159L60 and monthly forecasts T_L159L40 and the model for the seasonal forecast T_L95L40.

2.2.6. The resolution at the earth's surface

For the representation at the surface and for the model physics a grid point system is used instead of a spectral formulation. However, due to the convergence of the meridians toward the poles, a longitude-latitude grid is unsuitable. The rapidly decreasing east-west distance between the grid points would easily favour numerical instabilities near the poles, apart from redundancy of data due to over-representation of information.

The problem is alleviated in a so-called *reduced Gaussian grid,* which is almost regular in latitude. It keeps the east-west separation between points on different latitudes almost constant by decreasing the number of grid points towards the poles at every latitude. A regular Gaussian grid is only applied in a band between 24N and 24 S.The average distance between the reduced Gaussian grid points is about 40 km.

Figure 3 : An example of an Gaussian grid which shows that it is not exactly the same as a lat-long grid.

2.3. Climatological and geographical fields

A large part of the driving energy of the atmospheric motion comes from energy supply from the earth's surface, both land and sea. Its characteristics is partly derived prognostically, partly prescribed by climatological and geographical fields.

2.3.1. The model orography

The orographic information stems from a data set with a resolution of about 1 km which contains values of the mean elevation above the mean sea level, the fraction of land and the fractional cover of different vegetation types. This detailed data is aggregated ("upscaled") to the coarser model resolution.

The resulting *mean orography* gives quite a realistic description over most of the land areas, but is insufficient in high mountain areas where the sub-grid orographic variability becomes important. This is for example the case when cold air drainage in valleys makes it difficult for air from outside to penetrate the mountain at its true geographical height; the cold air effectively acts to "lift" the orography.

When stable stratified airflow crosses a mountain ridge gravity waves are excited into the flow. They play an important role for making the large scale flow slightly less zonal and increasing the frequency of blocking highs and cut-off lows. Depending on the local dynamic and thermal conditions the orography can, rather than make it pass over, block the low level flow and give rise to local winds.

Figure 4 : The model mean height in dekameters for northern Europe

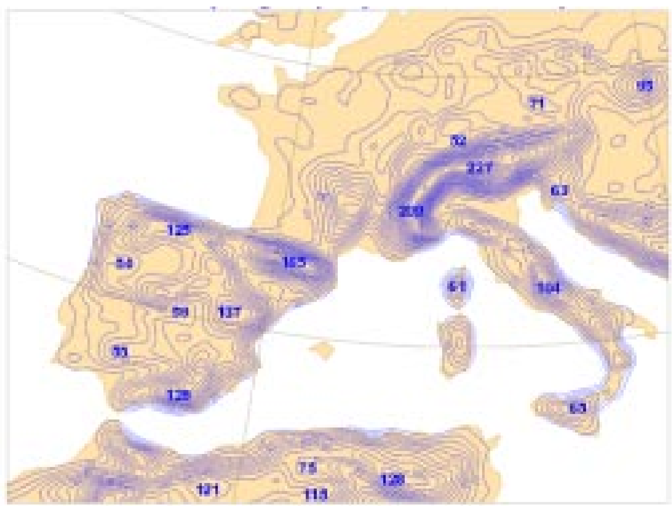


Figure 5 : The same for southwestern Europe.

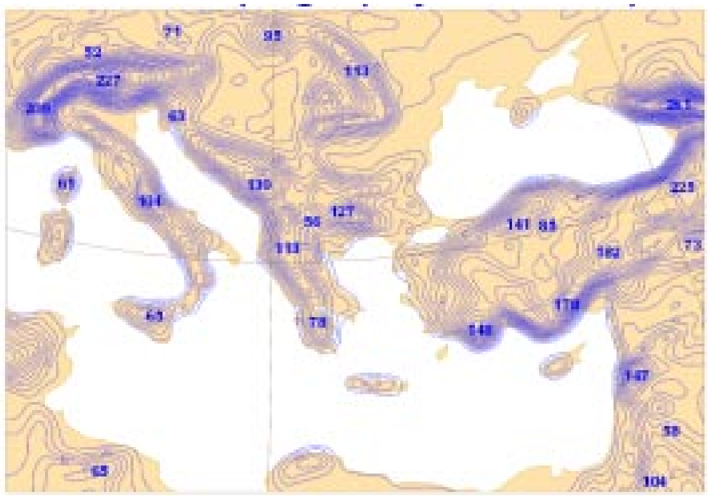


Figure 6 : The same for southeastern Europe and the Middle East.

To represent these mountain effects the mean orography is supplemented by four additional sub-grid fields: the standard deviation, anisotropy (aspect ratio of the obstacle), slope and geographical orientation of the sub-grid orography. They are added to the mean orography to provide flow dependent blocking of the air flow and to generate gravity waves.

2.3.2. The land-sea mask

The model surface is divided into sea and land points, defined by a land-sea mask taking values between 0 (100% sea) to 1 (100% land). A grid point is defined as a land point if its value > 0.5 indicating that more than 50% of the actual area within the grid-box is covered by land. Small islands and narrow peninsulas can take values <0.5.

2.3.3. The sea surface temperature (SST)

The *sea surface temperature* (SST) is based on analyses received daily from NCEP, Washington in a 0.5 x 0.5˚ grid. It is based on ship, buoy and satellite observations. In small waters where rapid changes due to upwelling can take place close to land the observed SST can sometimes differ as much as 5 deg from the NCEP analysis. The SST is kept constant over the integration.

2.3.4. The albedo

The *albedo* is determined as a combination of background monthly climate fields and forecast surface fields (e.g. snow depth). Over sea-ice the albedo is set to 0.5 and 0.7 for two spectral bands. Open water has a diffuse albedo of 0.06, and a direct albedo function of the cosine of the solar zenith angle. Over land the albedo varies between 0.07 to 0.20 for snow in forest, but can go up to 0.80 for exposed snow.

2.3.5. Aerosols

Continental, maritime, urban and desert aerosols are provided as monthly means from data bases derived from transport models covering both the troposphere and the stratosphere.

2.3.6. The surface vegetation

Within the grid area the surface physical conditions can vary significantly. To account for this subgrid-variability it is quite insufficient to describe the surface in the normal grid. Instead a sub-grid “mosaic” has recently been introduced, called TESSEL (Tiled ECMWF Scheme for Surface Exchange over Land)

At the interface between the surface and the atmosphere each grid box is divided into fractions or “tiles” with up to 6 tiles over land (bare ground, low and high vegetation, intercepted water, shaded and exposed snow). Each tile has its own properties defined by its typical heat and water fluxes. Special attention is devoted to the evaporation of bare ground and vegetated surfaces.

The vegetation is divided into low and high vegetation. The former covers 80% of the land points (out of which 22% is crops and mixed farming, 14% semi-desert, 13% tall grass, 10% short grass and the remaining 21% other types of vegetation or arid areas), the latter 63% of the land points (out of which 30% is interrupted forest, 13% broad leaf trees and the remaining 20% other types of trees).

2.3.7. Snow

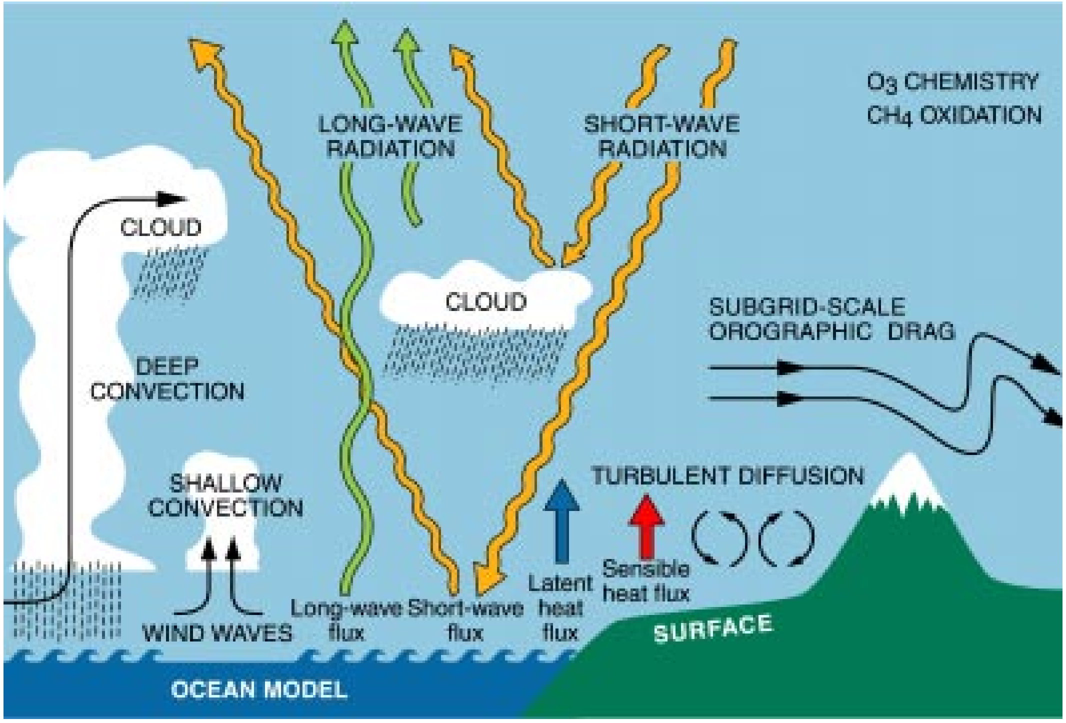
Snow cover is also included in the TESSEL scheme. It is analysed through a combination of observations and forecasts. The depth and extension evolves through the combined effect of snowfall, evaporation and melting. The thermal properties depend only on the snow mass per unit area. As the snow ages the albedo decreases and the density increases.

2.3.8. Sea surface conditions

Over sea the TESSEL contains only one or two tiles (open and frozen water). As mentioned above, open water points have a fixed SST throughout the forecast. The *sea-ice fraction* is based on satellite observations. The temperature at the ice surface is variable and calculated according to a simple energy balance/heat budget scheme. The horizontal distribution of ice is kept constant during the forecast. In case of sea ice the surface temperature is regulated by a 4-layer ice model. The SST of the underlaying ocean is assumed to be -1.7˚C.

2.4. The formulation of physical processes

The smallest scales of motion that can be resolved are those which have a wave length of two grid lengths or more. So even with increased resolution there will always be weather systems and physical processes, from convective clouds down to molecular processes, which develop on scales that are shorter than can be explicitly resolved by the model. These are described by *parametrization.*



2.4.1. Parametrization

Parametrization is the process in which the effect of a subgrid process is expressed in terms of resolved model variables. So for example, radiation is parametrized by computing the temperature tendency as a function of temperature and moisture profiles.

Parametrization can be both statistical and physical, or a combination. Statistics is used not only to calculate typical values of, for example, the surface characteristics, but also to link sub-grid scale motions to large scale variables such as wind, temperature, specific humidity or cloud fractions.

Sub-grid scale processes can also be treated physically as adjustment processes. The air closest to the earth's surface adjusts towards surface conditions, radiation adjust temperature differences and convection adjusts unstable air to neutral stability. Radiative-convective adjustment is a dominant process controlling the vertical structure of the troposphere.

Although the physics computations are performed only in the vertical, the complexity of processes and feedback mechanisms between the various processes, makes the computations complex and expensive. While the dynamics as such only occupy 23% of the computational time, the physical processes (including radiation) account for 36% and the ocean wave model alone 10%[1].

2.4.2. The importance of the Planetary Boundary Layer

The Planetary Boundary layer (PBL) plays a fundamental role for the whole atmosphere-earth system through exchange of momentum, heat and moisture. Below 1.5 km there are 13 levels at around 10, 30, 60, 100, 160, 240, 340, 460, 600, 760, 950, 1170 and 1400 meters above the model surface. Soil temperatures and moisture in the ground are calculated with levels at 7, 28, 100 and 289 centimetres depth.

Even with this fairly high resolution, the vertical gradients of temperature, wind, moisture etc. cannot be described very accurately, let alone their turbulent transports. For this the model uses the large scale variables such as wind, temperature and specific humidity, with the assumption that the transports are proportional to the vertical gradients. At the earth's surface, the turbulent transports of momentum, heat and moisture are computed as a function of air-surface differences and surface characteristics.

1. The remaining 31% are spent on communications, numerical transformations and spectral computations.

2.4.3. Radiation

The radiation spectrum is divided into a long wave part (thermal) and a short wave part (solar radiation) with 6 bands in the short wave spectrum, and 16 in the long wave spectrum. The forecast parameters influencing the emission, absorption of long wave radiation, and the absorption and scattering of short wave radiation are pressure, temperature, moisture, cloud cover and cloud water content. The pre-defined parameters, apart from the solar constant, affecting the gaseous absorption are the concentration of CO_2, O_3, methane, nitrous oxide, CFC-11 and CFC-12. The CO_2 for example has a constant mass mixing ratio over the whole globe corresponding to a volume concentration of 353 ppmv. Various types of aerosols are also taken into account in the long wave absorption and short wave absorption and scattering.

The radiation scheme takes the cloud-radiation interaction into account in considerable detail. For computational efficiency the radiation scheme is called only every 3 hours (*every* hour during the 12-hour first guess forecast used in the analysis). Overall, it covers less than 5% of the overall computational time. For cloudy grid points, computations are made both for clear and overcast conditions. The total amount is weighted together according to the forecast cloud amount at every model level using a maximum-random cloud-overlap algorithm (see below).

2.4.4. Cloud formation and dissipation.

Clouds are handled by a cloud scheme with prognostic explicit equations for cloud water/ice and cloud cover. The cloud processes are strongly coupled to other parameterised processes, in particular the convective scheme. Tendencies for condensate and cloud cover are generated by large-scale ascent/descent, cumulus convection, boundary layer turbulence and radiative cooling. The scheme also takes into account several important cloud processes like cloud top entrainment, precipitation of water and ice and evaporation of water. Fog is represented in the scheme as clouds that forms in the lowest model level.

An important role is played by the cloud-overlap algorithm which tries to calculate the relative placement of clouds in upper and lower levels. This is important for the radiation scheme and for the “life history” of falling precipitation: from a level with cloud to a level with clear sky and vice versa.

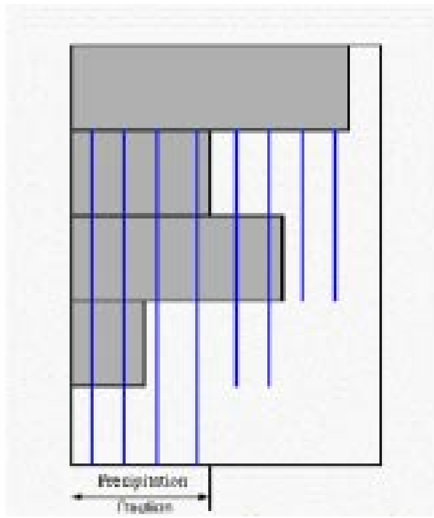


Figure 7 : Precipitation is treated separately in clouds and clear skies. The figure shows how model clouds cover areas which vary with height. Only a fraction of the grid box is covered by precipitation at all levels, including the surface.

The stratiform cloud processes have their own prognostic equations to forecast both cloud fraction and cloud/ice content. The physical processes, such as radiative cooling, are in principle part of the equations that govern the cloud scheme.

2.4.5. The convective cloud scheme

The convection scheme fulfils several objectives. Apart from computing the cloud production it also computes the convective precipitation, the vertical transport of moisture and momentum, and the temperature changes due to release of latent heat (heating) or evaporation (cooling). It distinguishes between deep, shallow and mid-level convection.

Deep convection is represented by a single pair of entraining/detraining plumes which describe updraught and downdraught processes. Downdraughts become negatively buoyant through evaporative cooling of precipitation.

Mid-level convective cells have their roots not in the boundary layer but higher up at rain bands at warm fronts or in warm sectors of extra-tropical cyclones.

Shallow convection has clouds that are less than 200 hPa deep with no or little precipitation e.g. trade wind cumuli or day time cumulus over land, or when cold air flows out over a warmer ocean.

2.4.6. Stratospheric processes

Methane oxidation is a source of stratospheric humidity. A sink representing photolysis in the mesosphere is also included.

Ozone is fully integrated into the ECMWF forecast and analysis system in a way similar to humidity, but is not interactive with radiation. Radiative heating is computed from a monthly mean zonal climatology.

2.4.7. Precipitation and the hydrological cycle

Two different sources of precipitation mechanisms are included in the ECMWF model: the convective and stratiform precipitation.

Convective precipitation is formed when the amount of the condensate formed in the updraft of the convective Parametrization exceeds the value that can be sustained by the vertical velocity. The precipitation is formed as water or snow. All condensate with a temperature above 0° C is defined as water, below -23°C as snow, and a mixture of water and snow in between.

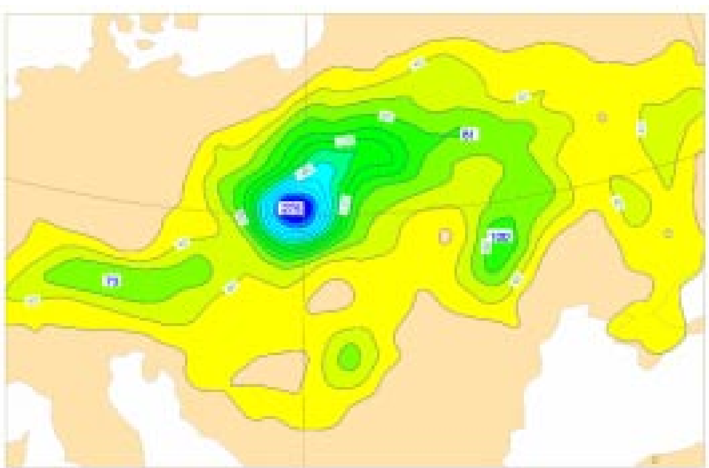


Figure 8 : The flooding in eastern Europe summer 1997 was well forecasts by the ECMWF model. The figure shows the accumulated rainfall during the first 96 hours of the T213 operational forecast on 5 July 1997. The maximum of 400 mm in southern Poland was slightly underestimated.

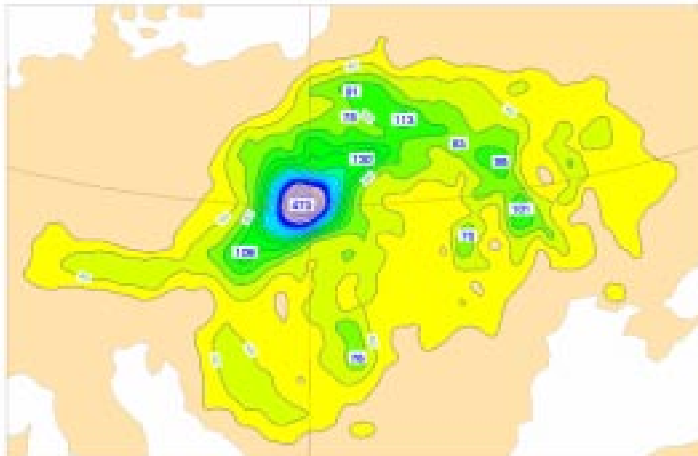


Figure 9 : An experimental T639 forecast of the same situation with the maximum amount well captured and the effects of the orography described in more detail.

Stratiform precipitation is formed depending on the water/ice content. Precipitation processes, such as collection of cloud water by precipitation and the Bergeron process are also taken into account.

Evaporation of the precipitation, before it reaches the ground, is not assumed to take place within the cloud, only in the cloud free, non-saturated air beside or below the model clouds.

Melting of falling snow occurs in a thin layer of a few hundreds of meters below the freezing level. It is assumed that snow can melt in each layer whenever the temperature exceeds 0°C. The melting is limited not only by the snow amount, but also by keeping the induced cooling of the layer such that the temperature of the layer after melting is not below 0°C.

2.4.8. The increasing skill of the weather parameters

For many years after the ECMWF model became operational there was a certain hesitancy to recommend the direct use of the produced forecasts of weather parameters like 2 m temperature, 10 m wind, cloud and precipitation. The impact of a change of a parametrization process was not primary judged from its effect on the forecasts of weather parameters, but on the improved simulation of the large scale flow. An important objective with the cloud scheme is to provide input to the radiation computations and to calculate precipitation.

With increasing computer power, finer resolution and much improved parametrization these contradictions became much more rare. Although the impact of the large scale flow is still seen as being of paramount importance, ECMWF provides forecasts of weather parameters of high skill and usefulness.

2.4.9. Remaining problems

Even with increased resolution there will always be a scales that border between full-grid and sub-grid scale. At present one such feature is sea-breeze circulations which have a typical extension of 20 km. They are too large and organized to be parameterized, but too small to be described by the current resolution (40 km). This has some negative consequences. During the forecast integration an over-sized “sea-breeze” on a scale of about two grid lengths (80 km) develops. Similar effects are observed with wind systems near heated mountains and in large-scale convective systems.

2.5. The ocean wave model

During the 1980’s it became evident that including the frictional effect of the ocean waves on the atmospheric flow would increase the realism of the atmospheric modelling, which in its turn affects the ocean circulation and the storm surge.

Although the positive impact on the atmosphere was the main reason for the introduction of the coupling, forecasts of ocean waves are also potentially very valuable products by themselves. There are a number of activities at sea where ocean waves are important and where the risk of high waves must be considered, such as the towing and maintenance work on oil rigs or the construction of under water pipe lines.

2.5.1. The model dynamics

The wave model used at ECMWF is the so called WAM (WAve Model) which describes the rate of change of the wave spectrum due to advection, wind input, dissipation due to white capping and non-linear wave interactions. The model gives the distribution of wave energy over frequency and direction, and gives a complete specification of the sea state.

2.5.2. The wave models in the ECMWF forecast system

Two versions of the WAM model are running at ECMWF: the global model and a limited area model.

The global model has an irregular latitude-longitude (0.5° × 0.5°) grid with an average resolution of 55 km. The advection time step is 15 minutes, the same as for the source term integration (the wind input, non-linear effects and dissipation). The wave spectrum has 30 frequency bins and 24 directions (15° intervals).

The limited area models (0.25° × 0.25°) cover the North Atlantic, Norwegian Sea, North Sea, Baltic Sea, Mediterranean and the Black Sea. They have a resolution of 28 km. Shallow water effects are included and the advection and the source time steps are 10 minutes. Like the global model they have 30 frequency bins and 24 directions.

2.5.3. Wave data assimilation

Since 1993 assimilation of altimeter data has been made for the global wave model. Buoy data are not assimilated, instead, they serve as an independent check of the quality of modelled wave heights.

2.5.4. Performance of the wave model

Since the introduction of the T511 model, verification of significant wave height and peak period against Northern Hemisphere buoy data has shown a good performance of wave analysis and forecasts, in particular near the coasts and during extreme events. There may, however, be underestimation of the wave forecast near the coasts and in enclosed basins such as the Baltic and the Mediterranean. Furthermore, in rapidly varying circumstances such as occur near fronts or at the peak of the storms, the limited resolution of the atmospheric and wave model may prevent a realistic representation of the sea state.

Table 1: Wave forecast products

2D-spectra
Peak period of 1D-spectra
2D-spectra for total sea, wind:
Significant wave height, mean wave direction, mean wave period
Global model:
0.5˚ x 0.5˚ latitude/longitude T+0 to T+240 h forecasts every 6 hours
N. Atlantic, Mediterranean and Baltic model:
0.25˚ X 0.25 ˚ latitude/longitude T+0 to T+240 h every 6 hours

2.5.5. The wave ensemble forecasts

In June 1998, the EPS was coupled to the ocean wave model. From then on, daily ensemble wave forecasts have been available. The present version of the EPS wave model runs on a 110 km grid resolution equivalent to approximately 80 km, with shallow water physics, 12 directional bins and 254 frequencies bins. For the waves, all ensemble members use the unperturbed analysis as the initial condition. The divergence between the wave ensemble members is therefore due only to different wind forcing when the coupled atmospheric ensemble members are subject to different evolutions.

2.5.6. The performance of the EPS wave forecasts

The potential benefits of using the EPS have been demonstrated using buoy and platform data as well as altimeter data for the period between September 1999 to March 2002 (Saetra and Bidlot 2002, Saetra and Bidlot 2003). The ensemble spread turned out to be a good measure of the uncertainties in the deterministic forecasts. The EPS forecasting system for decision making indicate that the model displays skill compared with those of traditional deterministic forecasts including a “a poor-man’s ensemble” of deterministic forecasts.

2.5.7. Remaining problems

There are still improvements to be made to the wave model. For example, the propagation of swell is handled by a simple scheme which gives rise to a smoothing of the wave field; errors due to this are in the order of 10-20 cm in significant wave height.

Furthermore, many islands in the Pacific are so small that they are not resolved by the model. Nevertheless, their presence blocks the propagation of wave energy. Inclusion of the effect of unresolved islands will remove some long standing systematic errors in the wave height fields.

2.6 Plans for the future

The ECMWF plans for the coming four years (2004-2007) is to provide increasingly skilful deterministic and probabilistic medium-range forecasts, in particular of extreme weather events. Among other things, the horizontal and vertical resolutions will be increased to T_L799L91 for the operational model, T_L399L91 for the EPS and T_L255L65 and T_L65L65 for the 15 day and monthly forecasts.

3. The data assimilation system

3.1. General overview.

A very large amount of observed data is available for use by the assimilation and forecast system. In a typical 12-hour period there is a total of 75 million pieces of data available, around 98% from satellites. Most of the available data are considered for use but the quality control, redundancy checks and thinning of locally dense data will reduce the numbers.

The observations can roughly be divided into conventional, in situ observations, and non-conventional, remote-sensing observations. The earth-atmosphere system can be measured directly by conventional in situ instruments, and indirectly by remote sensing instruments. The latter can be done in two different ways: *passively* and *actively*. The various data types have different characteristics in terms of geographical coverage, vertical structure and temporal distribution, which determine their ability to affect the analysis.

With increased availability of non-conventional observations the analysis system has developed into higher sophistication to be able to cope with off-time data and, in particular, indirect measurements such as radiances from satellites instead of direct observations of temperature, humidity, pressure, ozone and wind.

Many conventional observing platforms report at 12, 6 or 3 hourly intervals, but some report every hour. Aircraft, buoy and satellite data tend to be continuous in time. The ECMWF operates with long data collection window (12 hours). The analysis starts around 7 hours after 00 and 12 UTC to allow late observations 5 extra hours to enter in the system. Still 5-10% of mainly satellite data arrive too late, due to delay in the processing of the data.

3.2. Conventional observations

Since the conventional data report the values in the same units as the model variables, and on pressure or height levels, they can be used more or less directly by the analysis system after vertical interpolation.

Reports of pressure and humidity are used SYNOP (conventional surface weather station reports). 10 meter wind observations are not used, not even from marine locations such as coastal stations or minor islands. Pressure and winds are used from SHIP (conventional weather reports from moving ships) and DRIBU (drifting buoys).

Temperature, wind and humidity from TEMP (upper air observations from radio sonde stations) are used and their position defined according to the pressure at all reported levels. Temperatures in the stratosphere are corrected for estimated mean errors (bias correction). Humidity observations from drop sondes are not used.

Winds from PILOT (wind measurements in the free atmosphere from stations launching balloons) are used, except when there is duplicates with radio sonde data. PROFILERS (measuring winds with remote sensing) provide wind speed and direction.

Temperature and wind reports are used from AIREP (manual air craft reports), AMDAR (Aircraft Meteorological Data Relay) and ACARS (automatic air craft reports). The traditional AIREP observations now only account for ~4% of all used aircraft data, as most commercial aircrafts operate the AMDAR or ACARS systems. During landing and take-off the latter provide data in quantity, quality and location comparable to radio sondes.

PAOB is a hybrid between conventional and satellite data. They are manually derived pseudo-observations of MSLP on the Southern Hemisphere, made by the Australian Bureau of Meteorology from satellite images.

Humidity observations reported as relative humidity or dew point, are transformed into specific humidity using the reported temperature. SYNOP dew points are used, together with the temperature to calculate the 2 m specific humidity.

Other data types are used to analyse snow, ice, SST, soil wetness and ocean waves. These are at present analyzed separately, but might in the future be incorporated in the full data assimilation system.

3.3. Satellite observations

During the last 5-10 years there has been a significant increase in the quantity, quality and diversity of satellite observations. At the time of writing (autumn 2003) ECMWF routinely receives data from more than 15 satellites, some of which are equipped with several instruments that provides in total 27 satellite data sources.

ECMWF has during the last 10-15 years developed new assimilation techniques to make use of this new information. Although satellite data is slightly less accurate than conventional observations such as radio sonde observations, their great advantage is their broad geographical coverage. While the data assimilation system has to spread out the information in space of radio sonde observations, this is less of an undertaking with satellite observations.

Another advantage is that the use of satellite data ensures that the elusive small amplitude-large scale errors over the oceans are corrected for, something which isolated measurements would have difficulties to do. Although the amplitude of the analysis increments are weak, their large-scale nature becomes important after some days integration when they have “cascaded” into smaller scales, which might develop and affect synoptic scale weather systems.

Consequently there is now a strong benefit from satellite data in the ECMWF system and the influence of other conventional data types are becoming less critical. In particular over the Southern Hemisphere, where there is a lack of conventional data, the satellite data has had a large impact the scores which are now almost as good as in the Northern Hemisphere.

3.3.1. Different agencies

Satellites are operated by two types of agencies: research institutions (such as NASA, NASDA, ESA and others) and operational centres (such as EUMETSAT, NOAA, JAXA and agencies in Russia and China). Research agencies provide new technology and pave the way for future operational missions, but their satellites might also have a short life span of only 3-5 years. The investment to make use of the data operationally has then to be carefully managed.

Operational agencies can on the other hand provide real time data and maintain their satellite series for decades. Their products might be more conventional, but will gradually improve thanks to the knowledge obtained from research satellite programs.

3.3.2. Different satellites

There are two types of a satellites, geostationary and low earth orbiting satellites. Geostationary satellites are positioned in the earth’s equatorial plane at about 36 000 km height. They have a wide spatial coverage and high temporal resolution (observations up to every few minutes). Due to their position they cannot cover the whole planet and in particular not the polar regions. Due to the high altitude they do not measure in the micro wave spectrum at present.

Low earth orbiting satellites circle the earth at heights between 400 and 800 km. They are able to cover the whole spectrum, including the micro wave band. But since it takes several days before the satellite comes back to the same point it needs many orbits to provide a full coverage of the globe. A constellation of 2-4 platforms is therefore necessary to ensure reasonable temporal sampling.

3.3.3. Satellite passive measurements

Atmospheric parameters can be measured by *passive technologies* through sensing natural radiation emitted by the earth and atmosphere, or solar radiation reflected by the earth and atmosphere. Most current satellite data originate from this approach.

Radiances are provided by several satellites (the orbiting NOAA 15, 16 and 17, and the geostationary Meteosat 5 and 7, and GOES 9, 10 and 12). For regions with clear sky the radiances provide information on temperature and moisture content in broad layers. However, when clouds are present, no attempt is made to determine temperature and moisture. Instead the radiances are used to derive atmospheric motion vectors (previously known as cloud winds or SATOB).

Two MODIS (Moderate Resolution Imaging Spectroradiometer) instruments, the first launched on 18 December 1999 on board the Terra Platform, the second on 4 May 2002 on board the Aqua platform, are uniquely designed. They have a wide spectral range, high spatial resolution, and near daily global coverage to observe and monitor these and other changes of the earth's surface.

Ozone total column and profiles are measured by the ENVISAT and NOAA satellites by agencies such as KNMI and DRL.

3.3.4. Satellite active measurements

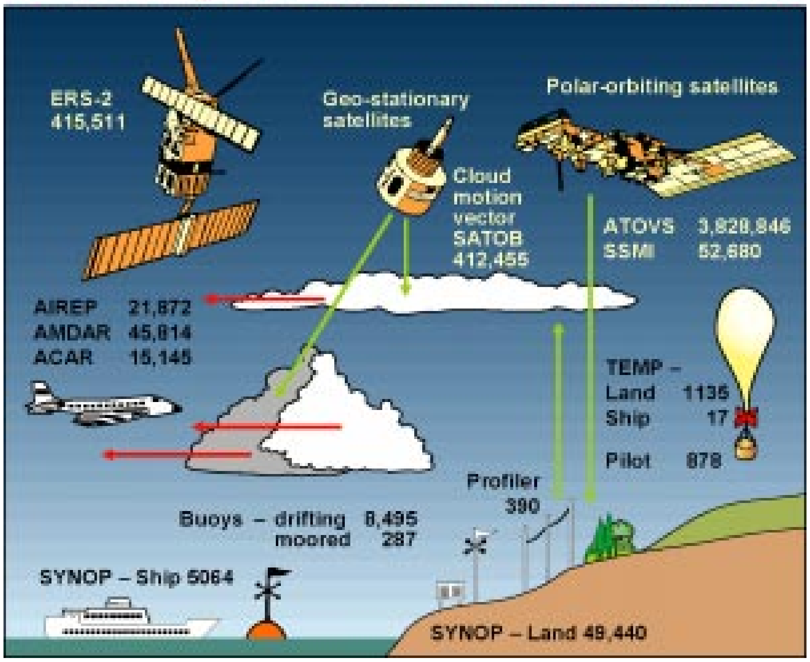
Atmospheric variables can also be measured by *active technologies* which transmit radiation towards the earth and then sense how much is reflected or scattered back. This is done for the so called Seawinds (on board the QuickScat satellite) on the ERS-2 scatteometer providing wind vector information derived from the influence of the ocean capillary waves on the back scattered signal. Depending on the frequency of the radar, the data may be affected by the presence of rain (seawinds) or not (ERS-2).

3.3.5. Bias corrections

Unfortunately, use of satellite data introduces the problem of biases. The spectrometers are calibrated to measure certain wave lengths. Due to technical circumstances (vibrations at the launch of the satellite, sun storms etc.) this calibration can be slightly distorted. A second source of systematic errors lies in the radiation model used to convert model temperature and moisture into radiances. A bias correction scheme is designed to take into account all these factors.

Bias corrections are also applied to ERS winds, but not to QUICKSCAT data. Winds from the GOES platforms are corrected before dissemination. The negative wind speed bias at jet cores are thereby corrected. Cloud winds from METEOSAT are not corrected, but are flagged with a quality index. Moreover, an asymmetric first-guess check is applied to the winds to penalize too weak wind observations.

24 hour summary of observations received at ECMWF, 18 March 2000



3.4. Quality control of observations

To ensure that only good quality data are used for the analysis an intricate quality control is applied. There are several ways data can be prevented from affecting the analysis.

3.4.1. Thinning

Many stations, platforms or satellites report with high density in time or space. To avoid flooding the system with superfluous data a thinning procedure is applied which removes data that is either redundant or has highly correlated errors. This applies in particular to satellite data and some aircraft and buoy data. Multiple reports from the same (geographically fixed) station are used at up to half-hourly frequency. All reported radio sonde levels are used.

3.4.2. Blacklist procedure

The ECMWF puts great emphasis on extensive daily and monthly monitoring of all the platforms to establish any systematic mean errors or erratic variations. Whereas the daily monitoring mainly is based on in-depth investigations of bad or very inconsistent forecasts, the monthly monitoring is mainly statistical. Platforms which are found to report biased or erratic observations are put on a so called blacklist. There are two types of blacklists:

Permanently blacklisted platforms have reported data with large systematic differences compared to the model state. This is most often because the station is badly calibrated or equipped, but can also affect a station for which the model orography is very different to the actual station height. All data of experimental type is blacklisted until a careful monitoring has shown that it can be allowed into the system. Observations, which cannot be decoded satisfactorily, for instance because they do not follow the WMO conventions, are also blacklisted.

Temporarily blacklisted platforms have been detected by the daily or monthly monitoring to suffer from a sudden deterioration in quality. This blacklist is updated manually once a month and whenever a sudden deterioration occurs. It involves comparison of the data with model fields, neighbouring observations of the same or other data types. The blacklist is lifted when the careful monitoring has shown that the quality is back to an accepted standard.

3.4.3. The automatic quality control

The data assimilation system also acts as an automatic quality control. It can still reject non-thinned or non-blacklisted data if they are climatologically unrealistic, appear as duplicates (or triplicates), or are very different from the first-guess field of the model or disagrees significantly with its neighbours.

3.4.4. The number of used observations

During the quality control the number of observations are reduced significantly, in particular radiance observations from satellite radiances where only around 5% are actually used. Due to the enormous amount of satellite information, this small fraction in absolute terms is still ten times the total amount of all other types of observations.

Table 2: Number of received and proportion of used conventional observations per day

Type	Received	Used (%)
SYNOP pressure	55 00	80
SYNOP humidity	55 000	45
SHIP pressure	6 500	75
SHIP wind	6 200	55
DRIBU pressure	14 000	40
DRIBU wind	13 000	80
TEMP temp.	45 000	90
TEMP wind	37 000	95
TEMP humidity	22 000	90
PILOT wind	17 000	85
PROFILERS	30 000	75
AIREP wind	15 000	20
AIREP temp.	14 000	20
AMDAR wind	35 000	75
AMDAR temp.	35 000	75
ACARS wind	70 000	70
ACARS temp.	70 000	65
PAOB	560	70
GOES WV	45 000	20
GOES IR	50 000	20
METEO7 WV	30 000	20
METEO7 IR	30 000	15
METEO7 VIS	20 000	30
METEO5 WV	80 000	10
METEO5 IR	30 000	15
METEO5 VIS	15 000	25
QUICSCAT		50

3.5. The analysis

In the early years of NWP initial conditions were obtained from manually analysed meteorological charts, laboriously interpolated to pre-defined grid points. During the 1950's the current concept of fitting a prognostic first-guess field to observations was suggested and successfully tried.[1]

3.5.1. The optimum analysis

During its first fifteen years the ECMWF performed the analysis according to the principles of Optimum Interpolation (OI). A short (6 hour) forecast, serving as a "first guess", was matched against observations. The final analysis was a weighted compromise between the observations and the "first guess". The weights depended on the typical errors of different types of observations and the first guess.

The OI had, however, certain weaknesses. So for example is the method local; only a limited area could be considered at each time to influence a given grid point. It could not cope well with non-conventional data. It relied only on statistically derived "structure functions" to describe how the information from the observations would be spread out into the intermediate areas.

In the 1990's the increased availability of data at off-synoptic times, in particular over the oceans and, most importantly, the increasing amount of non-conventional data from satellites and remote sensing instruments, contributed to make OI obsolete[2]. Most of the observations from satellites arrive in form of radiances which had to be converted into temperatures and humidity observations, as was done with the SATEM observations.

3.5.2. From three to four dimensional variational analysis

To avoid dependence on external agencies the ECMWF in 1992 put into operation a conversion system specially suited for satellite data. It operated only

1. The idea of a first-guess was introduced into automatic NWP analysis in 1955 by Bo Döös and Pal Bergthorsson. Bergthorsson had worked as a forecaster in Reykjavik, where he had to analyze North Atlantic maps with few observations. To improve the analysis he had been taught to put the most recent chart on a light-table with the previous analysis underneath as a "first guess". Bergthorsson represented Iceland at the ECMWF Council well into the 1990's.

2. In the late 1980's satellite retrievals were removed from the operational analysis system in the Northern Hemisphere due to their negative impact. In the early 1990's the use of 1DVAR radiance retrievals were introduced to the Northern Hemisphere but the impact was very small. In the mid 1990's operations changed to 3DVAR and direct use of radiances which improved the Northern Hemispheric forecasts and their impact was similar to that of radio sondes.

vertically and for individual grid points. There were several possible methods; the reason why the ECMWF used a so called one-dimensional variational scheme (1DVAR) was because of future plans to expand the variational schemes to higher dimensions.

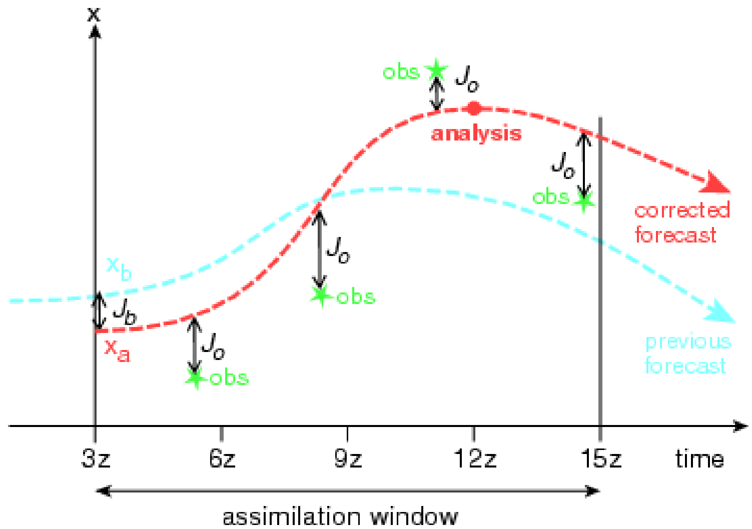
Apart from that the *variational* analysis methods can assimilate non-conventional data, they are global. Instead of analysing small regions of the atmosphere, one by one as in OI, the variational methods can analyse the whole globe simultatneously. This can be achieved by having all observations enter as element in a gigantic vector.

Instead of converting radiance observations into temperature and moisture, the opposite is done: model temperature and moisture variables are converted into model radiances provided by the first guess. The variational analysis system compares what the satellite observed in terms of radiances with what the “first guess” would expect it to observe and modify the radiance first guess to bring it closer to the observed radiance.

A first version of this variational system (3DVAR) was made operational in 1995 which solved the estimation problem globally. But only with the introduction of the four-dimensional variational analysis (4DVAR) in November 1997 was the influence of an observation in space and time controlled by the model dynamics.

3.5.3. The four-dimensional data assimilation (4DVAR)

By having the background errors modified by the model dynamics over the assimilation period in a flow dependent way increased the realism of the spreading out of the information. Observations inside baroclinically unstable jet streams or deepening cyclones, where the short range forecast is uncertain, will have larger weights and impact than it would have in a stable system such as a stationary barotropic cut-off low for example. This enables the 4DVAR to correct the phase, tilt and deepening rate of a developing storm.



A correction of one model variable will generate corrections of other variables wherever this is consistent with the dynamics. For instance, a sequence of observations of humidity from an infra-red instrument in a satellite, that shows that a displacement of atmospheric structures, will entail a correction, not only of the moisture field but also of the wind and temperature fields that is consistent with the motion of these structures[1].

Since 4DVAR is also an integrated part of the quality control, to provide initial values for the short range background fields, against which the observations are compared, erroneous observations have less possibilities to cause errors in the analysis which cause major forecast failures.

3.5.4. Land-surface analysis

Soil temperature and soil water content are prognostic variables of the forecasting system and, as a consequence, they need to be initialized at each analysis cycle. Currently the land surface analysis is performed every 6 hours, but is decoupled from the 4DVAR atmospheric analysis. The absence of routine observations

1. The 4DVAR is close to the principles that guide traditional manual weather map analysis over data sparse areas, for example the extra-tropical oceans. By going backward and forward between the last 3-4 weather maps with his pen and eraser. The forecaster could construct a set of mutually fairly consistent analyses, in particular for the last map.

of soil moisture and soil temperature requires to use proxy data such as SYNOP 2 metre temperature and relative humidity.

In a first step, the background field (6 h or 12 h forecast) is interpolated horizontally to the observation locations and background increments are estimated at each observation location. The increments are then analyzed using a bilinear interpolation scheme and added to the background. From the surface analysis the soil temperature in three layers are inferred. No analysis is performed in a gridpoint if the rainfall the last 6 hours has exceed 0.6 mm, the wind speed > 10 m/s, the 2 m temperature <0˚ or there is snow on the ground.

3.6. The ECMWF analysis cycle

Several types of analyses are run at ECMWF.

3.6.1. Operational analyses

At ECMWF four global analyses per day are produced at 00, 06, 12 and 18 UTC. These are obtained by two 4DVAR minimization cycles running from 03 and 15 UTC, and from 15 UTC to 03 UTC. From these analyses forecasts that have been constrained by the data of the optimal state of art model are produced, in particular as 3 and 9 hour forecasts valid at 00, 06, 12 and 18 UTC.

3.6.2. Preliminary analyses

To allow limited area forecasts to be run in the Member States the ECMWF provides boundary conditions four times a day (the BC-project). These forecasts are based on preliminary 3DVAR analyses made for all four analysis times, based on slightly reduced data (6-hour data collection window and only 1-hour cut-off time).

3.6.3. Cut-off time

As a forecasting centre with the emphasis on the medium–range, ECMWF operates with long data collection times, varying between 15 hours for the 00 UTC analysis and 8 hours for the 12 UTC final analysis.Real–time operational constraints mean that the data can only be used if it is received in time for each analysis. An example of the data collection windows and the cut-off times are given below for each of the daily operational analyses.

Table 3: Data collection window and cut-off times for different analysis cycles

Nominal analysis time	00 UTC short cut-off	00 UTC	12 UTC short cut-off	12 UTC
Earliest data time	21:00 UTC	15:00 UTC	09:00 UTC	03:00 UTC
Latest data time	03:00 UTC	03:00 UTC	15:00 UTC	15:00 UTC
Times the analysis start to be computed	04:00 UTC	07:15 UTC	16:00 UTC	19:15 UTC

This schedule ensures the most comprehensive global data coverage, including the Southern Hemisphere surface data and global satellite sounding data.

3.7 Future developments of the data assimilation cycle

A main future direction for the evolution of the analysis is to resolve smaller-scale structures by using more data (particularly from satellites) and by increasing the model resolution.

Even more use will be made of satellite data with a large research effort being directed towards use of satellite radiances affected by cloud and rain. A research effort is also directed to improving the use of ‘clear air’ satellite data over land. New variables like carbon dioxide will also be introduced.

4. Why do forecasts go wrong?

4.1. Introduction

A quarter of a century ago the rapid advance in computing technology, remote sensing from satellites and an ever increased sophistication of the primitive equation models, fostered a sense of great optimism. But progress in predictive skill did not advance as much as had been hoped for, and gradually the question arose if there was an ultimate limit to atmospheric predictability.

The question was not new. It had been discussed in mainly philosophical terms for decades. Already before W.W.II certain simple analytical models had quantified the effects of “infinitesimal perturbations” affecting the cyclone development. In the 1930’s there was saying among American forecasters “a sneeze in China may set people to shovel snow in New York”. Among European forecasters there was a saying that “flips of a sea gull wings” would cause unexpected storms.

So the suggestion that “small effects might have large consequences” was not new when the American meteorologist Edward Lorenz made his now famous coffee break[1]. When he returned to his computer he found that his computational simulations had deviated after he had unintentionally made some very small initial changes. The “butterfly effect” was born[2].

There are, at least, three reasons why forecasts do not agree 100% with the observed weather. Apart from “flaps of butterfly wings” also know deficiencies in the model formation and clearly visible analysis errors make sure that the forecasts do not agree 100% with the observed weather

4.2. Monitoring the ECMWF forecast system

The ECMWF puts a lot of effort into n monitoring the forecast system, both the model characteristics and the quality of the observational network

1. In the late 1950’s Lorenz was regarded mainly a skilful statistician. It was as such he was recruited to a project to explore alternative computational methods to NWP, for example advanced statistics. His part was to find out if it was possible to reproduce non-linear time evolutions by statistical methods. For this purpose he designed a set of non-linear equations with a similar structure as the atmospheric equations.

2. The name “The Butterfly effect” was probably inspired by a short story “A sound of thunder” from 1952 by the American sc-fi author Ray Bradbury. A group of time travellers return from a pre-historic excursion only to find that the political regime in the US has changed drastically. It turns out that one of the travellers accidentally had stepped on a butterfly in the pre-historic era.

4.2.1. Monitoring the model

The main tool to investigate the properties of the model is statistical analyses, both with respect to mean deviations between model and reality ("model biases") and its ability to simulate the atmospheric variance ("model activity").

4.2.2. Monitoring the data assimilation

Platforms with any systematic biases or erratic variations can be detected through statistical monitoring. Deficient platforms are blacklisted, with the exception of stratospheric radio sonde data and radiances from satellites, where stable mean errors are corrected. Plans are under way to also correct surface pressure data from SYMOP, SHIP and DRIBU. Daily inspections of the analysis material and case studies of bad or very inconsistent ("jumpy") forecasts provide an important source of information about erroneous observations come.

4.2.3. Summarizing the monitoring

The results of the daily monitoring is published daily on ECMWF's internal website to inform the research staff about possible problems within their areas of expertise. In case of more serious forecast failures special investigations are launched which often involve several of scientists. In addition to this, four time a year, the whole scientific staff gathers for a one-day informal "quarterly seminar" where the performance over the last 3-4 months is summarized and discussed.

4.3. Model errors (systematic forecast error)

Remaining limitations in the model properties, in particular the model physics, is a source of forecast errors. This is in particular the case where release of latent heat is crucial, for example in connection to cyclonic developments where warm and moist air is involved.

Other sources of model errors are when the radiative fluxes are underestimated which for example might have the result that the deep mid-winter temperatures are difficult to obtain. Frictional processes in particular gravity wave drag are not yet well described. Cloud processes under thermally stable conditions has also been a long lasting problem.

At the time of writing (autumn 2003) the ECMWF monitoring points to the following systematic model deficiencies related to the flow patterns. They are extracted from investigations of *extended* forecast ranges, but do also to some degree apply for the medium range.

4.3.1. Global model errors, including the stratosphere

There is a general tendency to have an underestimation of the dynamic activity, reflected in a decrease of 15-30% of the kinetic energy of transient eddies (moving low and high pressure systems) on all scales. This is particular the case at the start of the storm tracks and might be coupled to a tendency to parts of the westerly flow to be “too zonal”, underestimating blocked events. There is also a weak tendency to shift the circulation slightly poleward. The upper-stratosphere is slightly too warm.

4.3.2. Model errors over the Northern Hemisphere

The zonal flow, storm tracks and jet streams seem to be slightly displaced poleward and he synoptic activity is underestimated in the Polar regions. There is a certain lack of stratocumulus in the subtropics and too little clouds in the mid-latitudes.

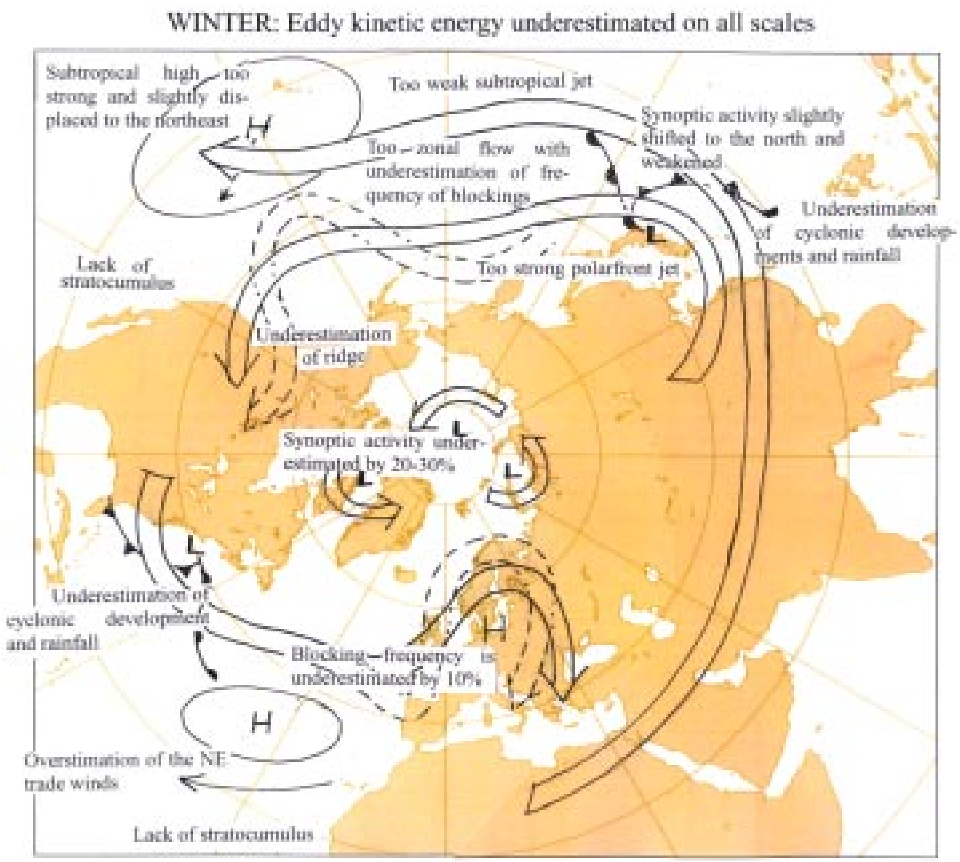


Figure 10 : The extended range model errors in Northern Hemisphere winters

4.3.3. Model errors over the Europe-Atlantic area

There is a certain underestimation of cyclonic developments over western most Atlantic, at the entrance of the storm track. The subtropical high over the Azores is slightly displaced toward NE in summer. Over Europe the blocking frequency is underestimated mainly in summer. The duration of blocks is well caught in general.

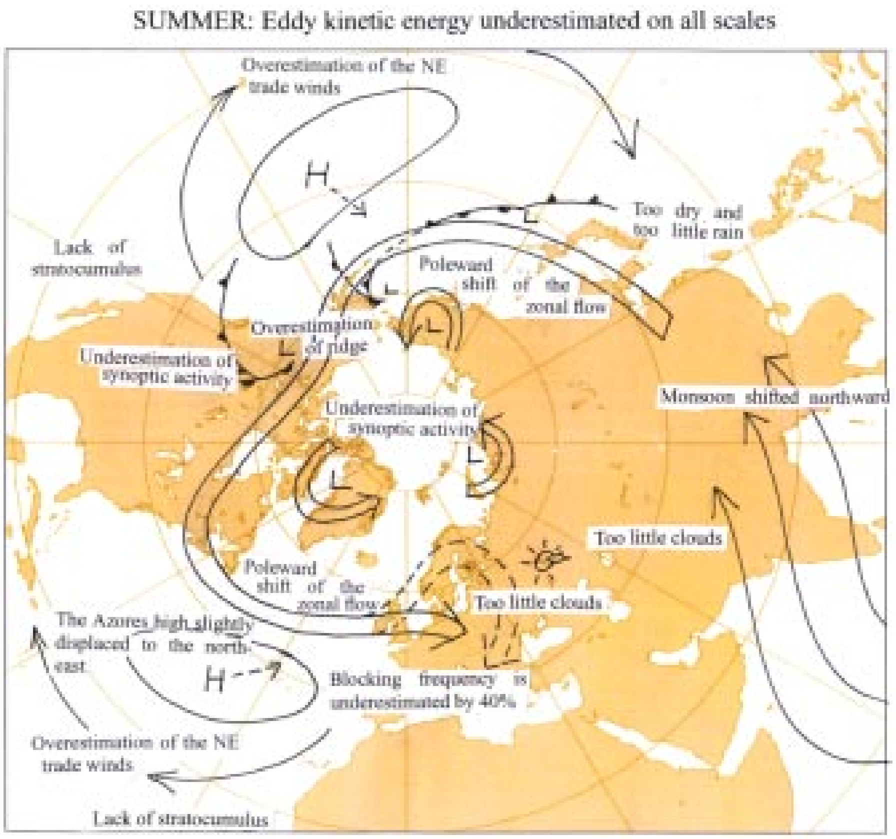


Figure 11 : The extended range model errors in Northern Hemisphere summer

4.3.4. Model errors over the North Pacific area

There is an underestimation of cyclonic developments at the start of the storm track east of Japan. The flow is too zonal with underestimation of blockings. The subtropical jet stream is slightly weakened, the polar front jetstream slightly strengthened. The subtropical high is strengthened on the northern side and displaced toward NE.

4.3.5. Model errors in the Tropics

The summer ITCZ is slightly displaced northward over Africa. There is a loss of humidity in the troposphere. The model has also difficulties to simulate the slow-propagating velocity potentials related to the so called “Madden-Julien Oscillation” in the upper-tropospheric tropical latitudes

4.3.6. Model errors over the Southern Hemisphere

(to be explored and included)

4.4. Analysis errors (non-systematic forecast errors)

The major cause for forecast failures in the medium range are errors in the initial conditions. They are mostly related to the assimilation of data.

4.4.1. What causes an analysis error?

All analyses contain errors, but it does not necessarily mean that the forecast will fail. To have a serious or more wide spread impact an error must occur in a dynamically sensitive region, in particular where baroclinic systems develop. During a baroclinic cyclogenesis huge amounts of kinetic energy is fed into the upper mid-latitude westerlies, in particular into the jet-stream. It will transport the kinetic energy out of the area and downstream.

In a short period of time, the dynamical and thermal conditions will be affected, which will affect the timing as well as the intensity of a downstream cyclogensis. If the forecast of this development is in error, more or less kinetic energy will feed into the upper tropospheric flow. This will affect the conditions downstream. A new cyclone in this location might start to develop erroneously. This may affect the interaction with other dynamically unstable systems in particular downstream. As a consequence the error will have more wide spread ramifications.

The investigation of a particular bad forecast is conducted to find out *when* the error entered into the analysis, *where* it happened and *what* caused the error. It is then assumed, as a working hypothesis, that the forecast error or inconsistency is due to an analysis failure.

The effect of an erroneously forecast cyclogenesis over the western part of the Atlantic may be felt in the forecast for Europe within 48-60 hours, at a time when the cyclone itself has yet be well west of Ireland. The effects of an erroneously forecast cyclone over the North Pacific might be felt in the forecast over

Europe within 5-6 days, at a time when the cyclone itself might already be filling somewhere over Canada. In a few days time the errors from this system will spread from one system to the next in a process, reminiscent of a “domino effect”.

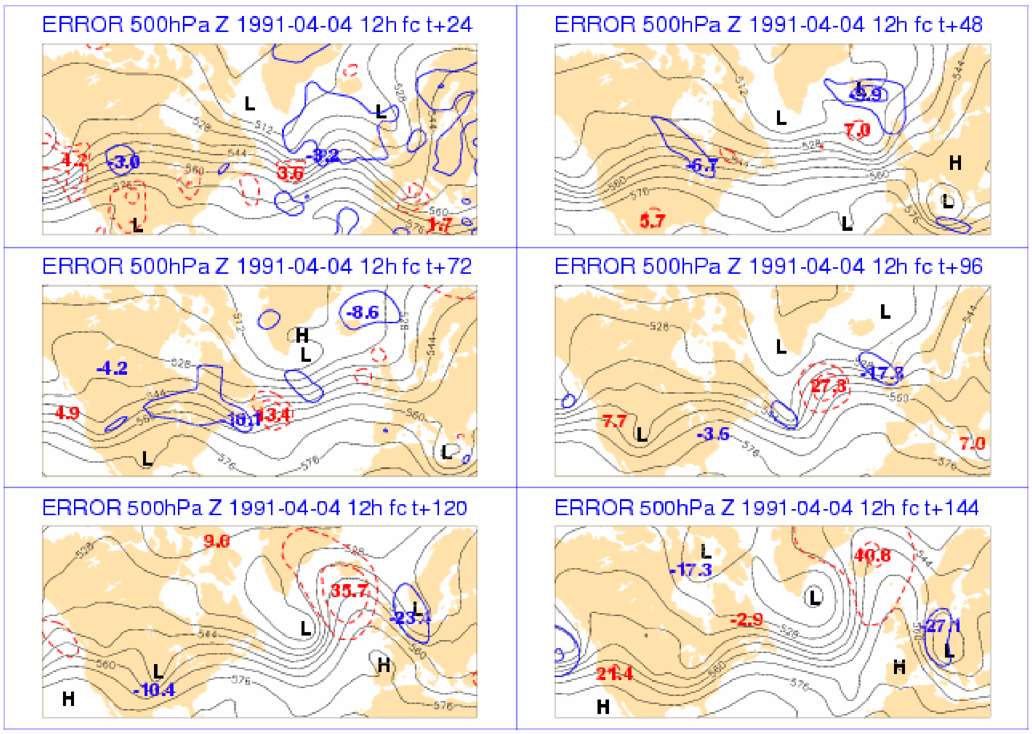


Figure 12 : The 500 hPa forecast and errors (red dashed lines positive, blue full lines negative errors) from 4 April 1991 12 UTC. A TEMP SHIP on the northeast Pacific, west of Vancouver, had been wrongly decoded causing a 2-3 dgpm positive analysis error. Early in the forecast a new negative error)-3.0 dgpm) is created downstream over W USA. While it moves eastward under amplification (from -6.7 to -18.1 dgpm), another error of opposite sign (amplifying from +13.4 to 27.8 and +35.7 dgpm) is followed by a fourth error (amplifying from -17.3 to 23.4 and -27.2 dgpm). The latter error was associated with a spurious cut-off over NW Europe.

The speed of this transport is roughly determined by the upper tropospheric flow. For a velocity of 30 m/s this corresponds to 30°/day at 45° latitude[1]. In the Southern Hemisphere the typical speed of influence is 40°/day due to the predominant zonal flow and the low frequency of blocked patterns.

4.4.2.Determining the timing of an analysis error

By looking at the verification scores or the forecasts themselves the time when the error entered into the analysis is mostly determined. The timing of the

1. This agrees well with the theoretical calculations of the “group velocity” of dispersive Rossby waves.

errors is easy when a good forecast is followed by a bad, more difficult when a sequence of forecasts have not been good, and perhaps in different ways. Then the right signal has never been in the system and no specific error can be found.

4.4.3. Tracing the geographical location of an analysis error

To trace the geographical origin of the error, different methods, empirical as well as objective, have been developed: forecast error maps, EPS perturbations and sensitivity analysis.

On an *error map*, displaying the forecast minus analysis fields, this process appears as an initial error, which will move slowly downstream (with the typical phase velocity) while it generates a wave train with increasing longitudinal extension.

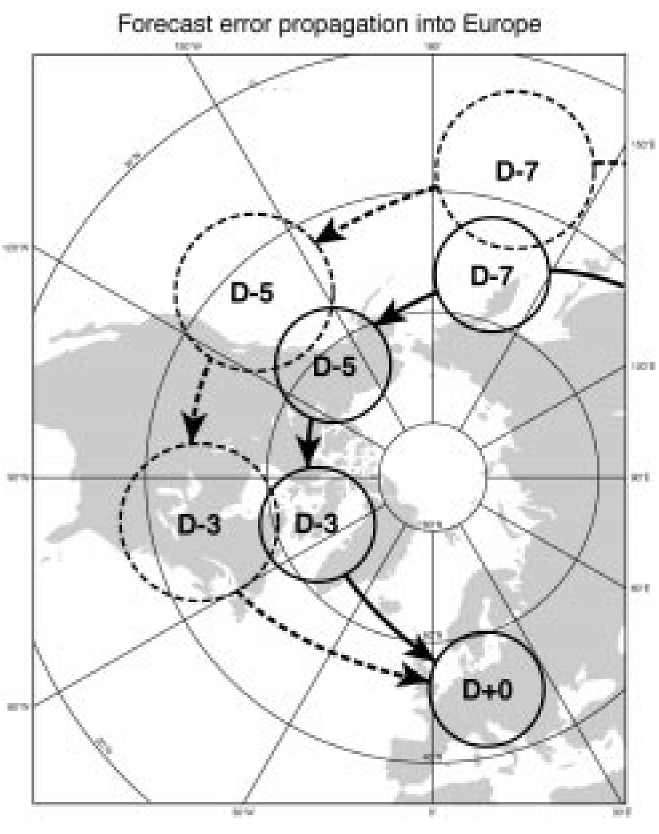


Figure 13 : The areas in the NH where analysis errors D-N days back will a most effect on the forecasts over Europe at final time. During winter the zonal flow is stronger and passes over the dense network over the US. In summertime when the US is covered by a subtropical ridge, the main flow is slower and passes over Alaska, northern Canada and Greenland where the density of observing station is low.

The *EPS perturbations* point out regions which are potentially dynamically unstable and where minor errors might amplify. The mathematics for this a priori analysis (independent of occurred errors) can also be used for posteriori analyses, so called *sensitivity analyses*.

By using the adjoint model it traces the observed 48 hour forecast errors backward in time. The system points out those regions where a small error would have effects which are very similar to the ones that occurred. This is often close to, but not identical to where the real error occurred which might be in a less unstable region.

4.4.4. The cause of the error

When the time and geographical location of the origin of the error has been established, the real ”trouble-shooting” starts when the observations in the area are carefully investigated to find possible causes of the error. The observation might be biased, a duplicate, been subjected to coding errors etc. The experiences are that observations cause errors when:

-An area where cyclones can form has not been covered with not enough relevant observations. This may happen preferably in non-inhabited areas, like the Polar Regions or the subtropical regions.

-Erroneous observations have been accepted by the analysis system. This happens for example when the first guess is in error and happens to agree with erroneous observations, or when lack of other observations in the area makes any internal check impossible.

-Correct, but unrepresentative data have been accepted. Observations reflecting extreme or small-scale (sub-grid scale) weather conditions, might have been wrongly interpreted as reflecting large-scale dynamical systems.

-Correct and representative data have been rejected. This may happen if the first guess deviates so much from the truth that correct observations appear to be wrong. It can also be that other observations in the region are in error and have made internal checks difficult or misleading.

-Correct observations can give rise to analysis error if the analysis system spreads out the influence in an erroneous way. This was often the case with older analysis systems, but is less common with the 4DVAR.

Depending on the error the platform is blacklisted or a short-coming in the quality control corrected. If possible the operator, responsible for the platform, is notified.

4.4.6. Re-running forecasts

Re-running the analysis and forecast with the suspected observations left out or corrected to establish the cause of the forecast failure, is not 100% conclusive. A change in one of a few observations might change the assimilation of other observations complicate the search for the cause of the bad forecast.

Re-running bad forecasts often leads to increasing understanding of the forecast failure. However, they are not always reliable due to the statistical “regression to the mean” effect. A random change of observations, analysis method or model properties is more likely to make the forecast better than worse.

4.5 What about the butterflies?

Some cases of bad forecasts remain unsolved even after extensive investigations. Then, but only then, could it be conjectured that a “butterfly” could have entered the system, a very small error almost impossible to detect.

A grid point value does not really represent a “point value”, but rather a “grid box average”. It has been suggested (Palmer, 2001, QJ 127, pp. 279-304) that the methodology used to approximate the equations of motion (i.e. neglecting the variability of the unresolved scales) is itself a source of large scale systematic errors.

Further, for purely mathematical-numerical reasons, features with a spatial extension of less than three grid lengths cannot be resolved by the model. Consequently, features of an spatial extent smaller than around 100 km can not be described by the T511 model, features smaller than about 200 km by the T255.

Experiences of assimilation of satellite data has shown that there are not only “butterfly” errors, but also “silk blanket” errors, small in amplitude but large in scale. Such esoteric errors might not appear to change the analysis in any noticeable way. But we know from theory and praxis that they become important during the forecast, when energy cascades from large to small scale atmospheric features.

5. The Ensemble Prediction System (EPS)

5.1. Introduction

The advance of weather forecasting is progressing along two fronts: on one hand improving the observational network, the data assimilation system and the NWP models; on the other hand, accepting that the forecasts will always be imperfect, trying to provide a measure of the degree of uncertainty.

5.1.1. The idea of “forecast forecast-skill”

Adding to the problem of non-perfect forecasts is their highly varying degree of quality. Periods of good or rather good forecasts are followed by shorter or longer spells with repeated failures. The value of the NWP forecasts would be highly enhanced if the quality of the forecasts could be assessed a priori. -How good or bad is today’s forecast likely to be? -How certain can we be that a forecast extreme weather event will actually verify?

Questions like these led in the late 1960’s to the idea of including a stochastic element in the NWP. But it had to wait until the late 1980’s until sufficient computer power made experiments possible[1].

5.1.2. The principles behind the ensemble prediction system

The ECMWF ensemble prediction system is based upon the notion that small analysis errors in sensitive parts of the atmosphere may affect the large scale flow during the course of the ten day forecast period. In other words, a slightly different, but equally accurate analysis, might yield a different forecast. Since the alternative analysis is equally possible as the original (Control) analysis, the alternative (Control) forecast will be equally likely to verify. This leads us to the profound insight that the deterministic forecast, run from the Control analysis, is just one possible development of a number of alternative evolutions, and not necessarily the most likely.

1. In the late 1980’s, awaiting sufficiently powerful computers, attempts were made to assess the skill a priori (“forecast forecast-skill”) using statistical methods. The results were indifferent, partly because of a misinterpreted correlation between forecast consistency and skill (see 8.4.1)

Depending on the particular hemispheric flow pattern, forecasts originating from perturbed analyses develop more or less differently during the course of the ten day forecast. A small spread among the EPS members should be an indication of a predictable situation. In other words, whatever small errors there might be in the initial conditions, they should not seriously affect the deterministic forecast. By contrast, a large spread should indicate a large uncertainty of the deterministic forecast. In these cases the main information is to be found from *probabilistic* conclusions.

5.2. The creation of perturbed analyses

The ECMWF ensemble system applies a perturbation technique which is based on a mathematical method called singular vector decomposition, related to the techniques used in the variational data assimilation.

5.2.1. Singular vectors

While the aim in 4DVAR is to minimize the distance between a model evolution and observations, the singular vector approach seeks perturbations that will maximize the impact on a 48 hour forecast, measured by the total energy over the hemisphere (poleward of 30˚ latitude).The impact on individual weather systems can be either a strengthening or a weakening; in addition the system can be displaced.

The computations start by calculating 25 independent (orthogonal) singular vectors. Since these calculations are quite costly they have to be run at a T42 resolution. These singular vectors tend to be rather small scale and underestimate the kinetic energy compared to analysis error estimates. To compensate for this a set of evolved singular vectors are calculated by linearly bringing forward the perturbations from 48 hours earlier.

5.2.2. EPS perturbations

The 25 singular vectors are then linearly combined in a linear way to create hemispheric perturbation structures, poleward of 30˚ latitude. These perturbation fields are then scaled so that their local maxima or minima are comparable to typical local analysis errors, and to have a realistic ensemble spread after 48 hours. Like the observation and atmospheric state vectors mentioned in the context of 4DVAR, the singular vectors cover the whole global atmospheric state.

By reversing the signs of these 25 hemispheric perturbations, 25 new “mirrored” hemispheric perturbations are produced, yielding a total of 50 perturbed fields, each added to the Control analysis to provide 50 alternative analyses, all assumed *a priori* to be equally likely. Another set of 50 alternative analyses are computed separately, but with the same technique, over the Southern Hemisphere. The final result is 50 alternative *global* analyses, which define the initial conditions for 50 alternative forecasts. To save time the 50 forecasts are run on a version of the operational model which has half the horizontal resolution T_L255 with only 40 levels.

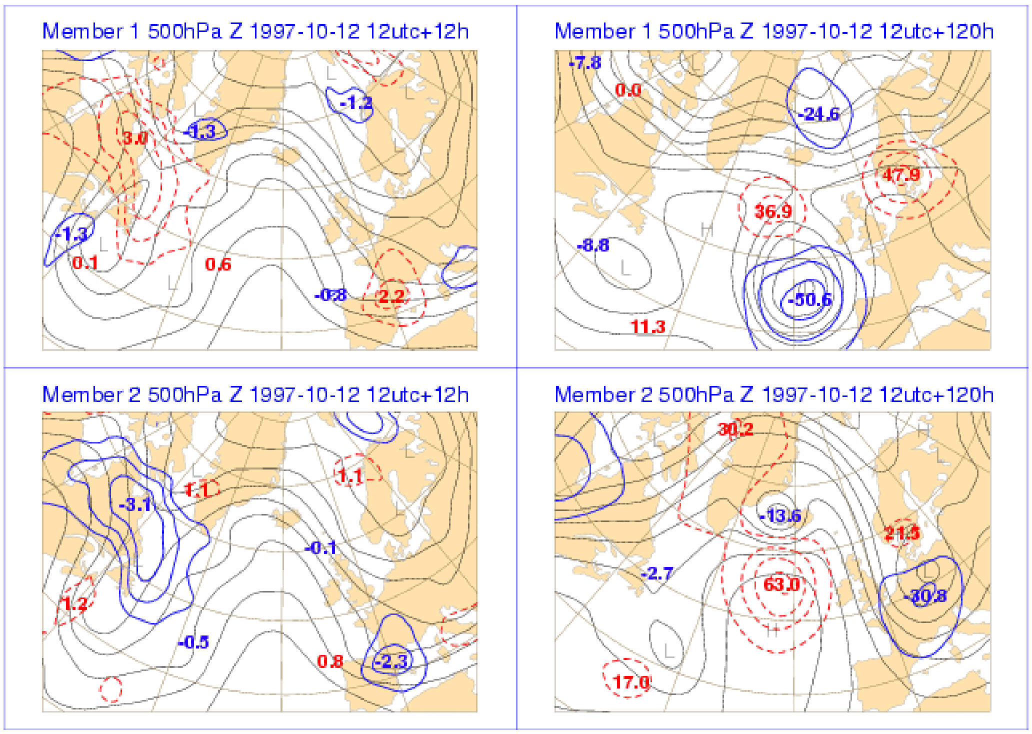


Figure 14 : The EPS perturbations at 500 hPa 12 Oct 1997 12 UTC at +12 hand +120 h for members 1 and 2. For these pairwise members the perturbations initially have opposite signs and mirror each other. This correlation gradually weakens during the forecasts and at +120 h the differences become random.

The quality of the perturbations are primarily judged by the degree to which they can correctly account for the uncertainties and alternative developments. In addition to that, it has been found that they make “synoptic sense”. Most of the EPS perturbations, which are of importance for the medium range forecast over Europe on a week’s range, are inserted in the analysis of baroclinic systems over North America and North Pacific. Error tracking of bad forecasts (see 4.4) often lead to regions which have been identified by the EPS perturbations as sensitive to potential analysis errors.

5.2.3. Stochastic physics

The source of model errors, such as the finite resolution of the model grid or simplified physical parametrization, has been represented by the introduction of stochastic physics. For each ensemble member, the stochastic physics perturbs tendencies of parametrisized physical processes by up to 50%, with a spatial correlation of 10 longitude degrees and time correlation of 6 hours. The whole globe is perturbed including the Tropics. The Control forecast is run without stochastic physics.

5.2.4. Tropical singular vectors

To further improve the ensemble forecast of developments, typical of low latitudes, in particular tropical cyclones, a scheme for creating perturbations specially designed for the Tropics has been developed, which based on diabatic singular vectors. They also seem to improve forecasts of extra-tropical developments, for example when tropical cyclones after some days into the forecast enter the mid-latitudes and interact with the baroclinic developments in the westerlies.

5.3. The operational EPS clustering

The EPS provides a large amount of information, in principle fifty times the amount from the operational forecast system. ways have to be found to condense the information. One such is “clustering”, the grouping together of synoptically similar ensemble members.

5.3.1. Basic clustering principles

To highlight the predictable and thus relevant part of the atmospheric flow, “similar” EPS forecasts are collected and averaged to constitute new forecast fields, so called *clusters*. At present this is done on the 500 hPa forecast. The norm for judging what is “similar” is their RMS differences over different domains from +120 h to +168 h to take the synoptic continuity into account. It is always the same members which make up the contents of each cluster. For two EPS members to join the same cluster they must therefore display similar synoptic 500 hPa development from +120 to +168 hours.

There are occasions when two members in the same cluster can be rather different at the beginning or end of the period, but sufficiently similar during the rest of the time interval to be placed in the same cluster. On the other hand, two members, being similar during a part of the period, may be placed in different clusters if they are sufficiently different during most of the period.

5.3.2. The number of clusters

The number of clusters depends on four factors:

-The spread of the day, i.e. the EPS standard deviation. It is varying from day to day, but follows a seasonal trend similar to the forecast errors, with higher values in winter than in summer

-The clustering threshold used to limit the clusters standard deviation. It follows the same seasonal trend as the spread and errors

-The degree of “multi modality”, the tendency of the forecasts to form discrete alternatives is taken into account. For the same spread and threshold a multi modal distribution might lead to a smaller number of clusters than a mono modal distribution. A large spread in the ensemble does therefore not necessarily lead to more clusters, nor does a small spread necessarily lead to fewer.

-The number of clusters cannot exceed six.

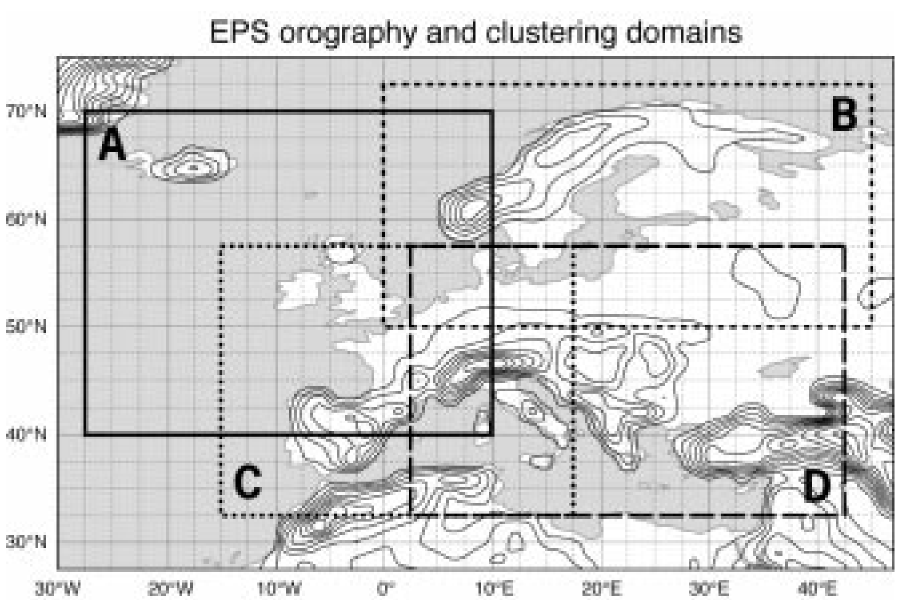


Figure 15 : The five main clustering areas, the European and the four sub-areas.

5.3.3. Cluster products

The clustering is performed separately for the whole of Europe plus four European sub–domains: Europe, NW Europe, NE Europe, SW Europe and SE Europe. The clustering over one sub-domain with respect to the position and intensity of a dominating feature will then be made without considerations of the uncertainties in the forecast of a blocking over another sub-domain.

There is no separate clustering for 1000 and 850 hPa. The clustering of the 1000 hPa geopotential height and the 850 hPa temperature depends on the clustering of 500 hPa.

5.3.4. No ideal clustering

Every possible clustering is a compromise; the advantage of condensing information has to be paid by the disadvantage of losing information by averaging fields which on some occasions, in hindsight, might have been important. There is really no superior or objective measure of which type of clustering is “best”. Clustering can be performed over larger or smaller geographical areas, on different parameters, it can be done for each forecast time or for a longer period, each with their own drawbacks and advantage.

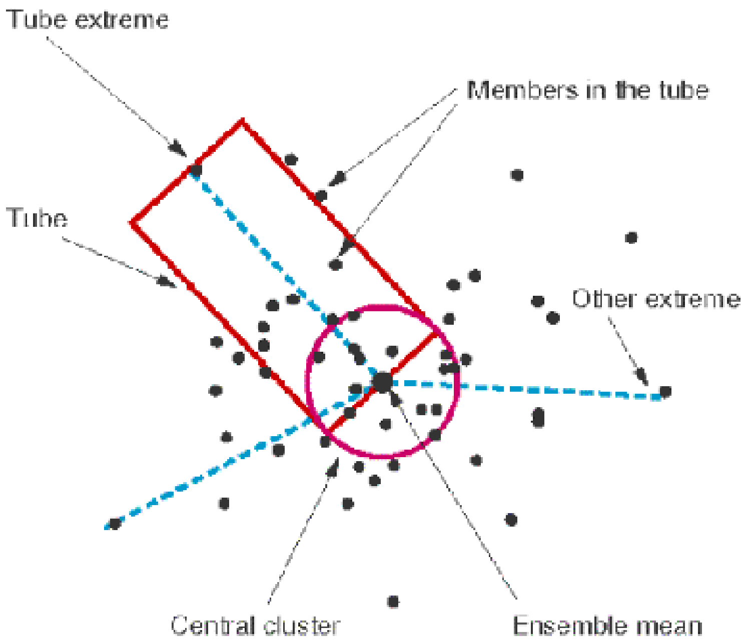
Clustering on individual forecast days will have the advantage that each day can be judged by its own merits; the disadvantage is that the temporal continuity and synoptic consistency will be lost. Clustering based on correlation measures would highlight similarities in the patterns but may group together forecasts which differ in the overall level of temperature and geopotential heights.

In cases of weak gradients rather different flow patterns might end up in the same cluster, which indeed can be the only cluster (the ensemble mean). This is an unavoidable consequence of the RMS method applied over 48 hours forecast time. Using (anomaly) correlation would avoid this problem, but introduce other. The correction would for example not be able to distinguish between two zonal flows which had a large geopotential difference due to different air masses being involved.

Ideally the forecaster should have access to more than one clustering method, since what is the "best" clustering will vary according to the weather situation. One such alternative clustering is the “tubing”.

5.4. The "tubing" clustering

Another clustering method, called *tubing*, averages all ensemble members which are close to the ensemble mean and excludes members which are significantly different. Again, the "closeness" is measured in RMS terms. The average of all the "similar" members provides a more refined ensemble mean, *the central cluster mean*. The excluded members are grouped together in a number of *tubes* (maximum 9).



The central cluster mean and the tubes are computed for the whole forecast range. For each tubing reference step (+96h, +144h, +168h, +192h and +240h), tubing products are generated over a 48-hour sequence finishing on the reference step. For example +48/+72/+96h are used for the +96h tubing, allowing a sequential view of the different tendencies. For the +168 h tubing, the +72 to +168 hour forecasts are used. The tubing results are then applied to the 1000 hPa geopotential and 850 hPa and 500 hPa temperature. Tubes are computed for the same five geographical domains as the clusters: Europe, NW Europe, NE Europe, SW Europe and SE Europe.

Each tube is represented, not by an average of members in the tube, but by its most extreme member. This allows a better visualization of the different scenarios in the ensemble. The tubes are not intended to serve as probability alternatives, only to give an indication of what is not included in the central cluster.

6. The forecast products

6.1. The operational schedule

At ECMWF two main suites are running. One produces global analyses for the four main synoptic hours 00, 06, 12 and 18 UTC, and global 10–day forecasts based on 00 and 12 UTC analyses. The second one produces analyses for the same time as above, but with a short cut-off (only one hour instead of four, from the closing time of the data collection window) as part of the “Boundary Conditions project”. As an optional project global 96 h forecasts are run four times a day from these analyses. These forecasts provide Member States with boundary conditions (frames) up to 78 hours to their limited area models.

6.2. Direct model output

The model variables for the computation of the forecasts are pressure, temperature, wind and specific humidity. From these primary parameters most meteorological parameters can be derived. Land and sea surface conditions are also described by a series of parameters such as surface roughness, albedo, etc. Tables 1 and 2 summarize the main output of the forecast model. These parameters are computed at 3–hourly intervals from 3 to 72 hours and every 6 hours from 72 to 240 hours, based on 00 and 12 UTC high resolution forecast model. Surface parameters from the EPS are available every 6 hours, upper air parameters every 6 hours.

Table 4: Upper air parameters

Geopotential height (not on model levels) Potential vorticity(*) (not on model levels) Temperature Vorticity and Divergence (*) Wind (U and V components)(*) Vertical Velocity Specific Humidity Cloud ice/water content on model levels
Upper-air parameters are produced on the original model levels and on standard pressure levels (1000, 925, 850, 700, 500, 400, 300, 250, 200, 150, 100, 70, 50, 30, 20, 10, 7, 5, 3, 2 and 1 hPa). and some of them (*) on isentropic levels (ϑ=300, 315, 330, 350, 370, 475, 600 and 850 K)

Table 5: Surface and single level parameters

Mean sea level pressure
ϑ, Z, P, u, v, q and ozone mass mixing ratio on PV+2PVU surface
10 metre wind
2 metre temperature
2 metre dew point
Maximum and minimum 2m temperature since previous post-processing
Maximum wind gust since previous post-processing
Large scale and convective precipitation
Snowfall
Surface temperature and soil wetness
Snow depth
Total cloud cover
Low, medium, high and convective cloud cover
Surface fluxes, surface stress, surface roughness, albedo
Solar and thermal radiation

The 2 metre temperature, dew point, 10 metre wind and wind gust are computed from both values at the lowest model level (approx. 10 metres above ground) and at the surface, taking into account a prescribed state of the surface (albedo, roughness etc.). Analysis fields for 00, 06, 12 and 18 UTC including additional fields such as model orography, land sea mask, percentage of vegetation, and some additional isentropic level is also available.

It should be borne in mind that except 2 metre temperature and 2 metre specific humidity, surface parameters, cloud and radiation parameters are not analyzed in the present system. The analysis and forecast output is archived into MARS (the ECMWF archiving system of meteorological data, cf. Meteorological Bulletin M1.9/2).

6.3. Dissemination products

A full set of parameters is available to ECMWF Member States through the operational dissemination system (table 3; cf. Meteorological Bulletin M 3.1 (2) for a description of the system). As a matter of fact, more parameters are produced and disseminated, than are archived. They are available in Gaussian regular and reduced grid, regular and rotated lat-lon grid forms. Upper air parameters (except humidity) are also available in spectral form.

Table 6: ECMWF dissemination EPS products

Control and Perturbed forecast products: Levels and validity: 1000, 850, 700,500,200 hPa for +0 hour to +240 hour forecasts at 12 h interval Geopotential, temperature, u- and v-velocities, specific humidity, vertical velocity, vorticity, MSL pressure, divergence Surface products: Large scale precipitation, convective precipitation, snow fall, total cloud cover, 10 m u- and v-components, 10 metre wind gusts, 2m temperature and dew point temperature, 2 metre maximum and minimum temperature, all for +0 to 240 h at 6 h interval Products on PV surface (2 PVU only): Geopotential, ozone mass mixing ratio, potential temperature, pressure, specific humidity, U velocity, V velocity
Cluster and ensemble means and standard deviations: Geopotential 1000 and 500 hPa Temperature 850 and 500 hPa Validity: for means +72 hour to +168 hour forecasts at 12 h interval, for ensemble standard deviations +0 hour to +240 hour forecasts at 12 h interval Tubes Geopotential 1000 and 500 hPa Temperature 850 and 500 hPa +48 to +96h for the 96 h tubing, +96 to +144 h for the 144 h tubing, +72 to +168h for the 168 h tubing, +144 to +192 h for the 192h tubing and +192h to +240h for the 240h tubing.

Table 6: ECMWF dissemination EPS products

Forecast probability products:
850 hPa anomaly probabilities
cold anomaly of at least -8K
cold anomaly of at least -4K
warm anomaly of least +4K
warm anomaly of at least +8K
850 hPa anomaly probabilities from day 6 to10, day 6 to 7 and day 8 to 10
average temperature at 12 UTC more than 2K below climate
average temperature at 12 UTC more than 2 K above climate
Precipitation probabilities over 24 hours
at least 1 mm
at least 5 mm
at least 10 mm
at least 20 mm
Precipitation probabilities from day 6 to10, day 6 to 7 and day 8 to 10
less than 0.1 mm over the period
mean precipitation rate less than 1 mm/day
mean precipitation rate greater than 3 mm/day
mean precipitation rate greater than 5 mm/day
Wind probabilities
at least 10 m/s
at least 15 m/s

6.4. Products on the GTS

A limited quantity of ECMWF analysis and forecast products is disseminated via the GTS. The product range is summarized in table 4. At the time of writing (autumn 2003) there is an agreement to extend the dissemination with products useful for tropical cyclone forecasting. Also added are seasonal forecast products such sea surface temperature forecasts in GRIB and 850 hPa temperature and 500 hPa geopotential anomalies in graphical format.

LIST OF PRODUCTS AGREED BY ECMWF'S COUNCIL 53 FOR DISSEMINATION ON THE GTS

Parameter	Level	Domain	Steps
Based on deterministic model			
Z	500	G	H+00,24,48,***<u>72,96,120,144,168</u>***
T	850	G	H+00,24,48,72,96,120,144,168
u,v	850	G	H+00,24,48,***<u>72,96,120,144,168</u>***
u,v	700	G	H+00,24,48,72,96,120,144,168
u,v	500	G	H+00,24,48,72,96,120,144,168
u,v	200	G	H+00,24,48,72,96,120,144,168
Rel Humidity	850	G	H+00,24,48,72,96,120,144,168
Rel Humidity	700	G	H+00,24,48,72,96,120,144,168
MSL pressure	surface	G	H+00,24,48,***<u>72,96,120,144,168</u>***
Divergence	700	T	H+00,24,48,72,96,120,144
Vorticity	700	T	H+00,24,48,72,96,120,144
Based on EPS			
Prob Precip	>10mm	NH,SH	H+72,96,120,144
Prob Precip	>20mm	NH,SH	H+72,96,120,144
Wind gusts	>15m/s	NH,SH	H+72,96,120,144
Wind gusts	>25m/s	NH,SH	H+72,96,120,144

G = Global

T = Tropics between 35S and 35N

NH = Northern Hemisphere north of 20 deg

SH = Southern Hemisphere south of 20 deg

<u><<Essential>></u> products are underlined above

6.5. Web service

ECMWF web site (http://www.ecmwf.int) makes available a wide range of documentation and an essential selection of forecast products to the general public.

It also provides enhanced products access to Member States' national meteorological services and other registered users. Updated versions of this User Guide can be accessed and downloaded through the web together with more specific informations about subjects mentioned in the current document.

6.6. Data archives

Weather forecasting makes use and generates very large volumes of data that need to be stored for long periods. ECMWF operates a comprehensive data service from its archives. In particular, it maintains an archive of level III-A atmospheric data in support of projects associated with the WMO World Climate Research Programme.

This includes observations, analysis, forecast and also research experiments. ECMWF has accumulated 150 Tbytes (150,000,000,000,000 characters). These data represent a valuable asset, providing a detailed record of worldwide weather and weather forecasts over the past 45 years. To accommodate these data, ECMWF has a dedicated Data Handling System. In order to manage and access this large archive, ECMWF has developed a dedicated software: the Meteorological Archive and Retrieval System (MARS). Data is stored in standard formats agreed with the World Meteorological Organisation, namely GRIB format for meteorological fields and BUFR format for meteorological observations. Retrievals can be easily prepared making use of a pseudo-meteorological language.

6.7. Access to archived data

6.7.1 Operational data

All authorised users within Member States and Co-operating States can access ECMWF's archive and retrieve data. This can be done either through the lines between ECMWF and Member States or through the Internet. For research or education purposes from states which are not Member States or Co-operating States. This service is provided at handling cost. Access for commercial purposes is also possible, but then the request must go via one of the Member States.

6.7.2. Re-analysis data

The reanalysis project ERA-40 covers the period from mid-1957 to 2002 including the earlier ECMWF reanalysis ERA-15, 1979-1993. The main objective is to promote the use of global analyses of the state of the atmosphere, land and surface conditions over the period.

The 3DVAR technique is applied using the T159L60 version of the Integrated Forecasting System to produce the analyses every six hours. Analysis involves comprehensive use of satellite data, starting from the early Vertical Temperature Profile Radiometer data in 1972, then later including TOVS, SSM/I, ERS and ATOVS data. Cloud Motion Winds are used from 1979 onwards.

ERA-40 products will also revitalize the use of data from past field experiments such as the 1974 Atlantic Tropical Experiment of the Global Atmospheric Research Program, GATE, 1979 FGGE, 1982 Alpine Experiment, ALPEX and more recent 1992-1993 TOGA-COARE. ERA-40 will provide a new potential for studying longer term trends and fluctuations such as ENSO and QBO.

6.8. Retrieving data from the ECMWF archives

Analysis and forecast values are available for every 6 hours, for surface parameters every 3 hours up to the 72 hour forecast range. The exact value of these parameters might be affected by the way data is selected, interpolated and presented. A short description of these effects follow.

6.8.1. Temporal resolution

The range of the daily variation of the 2 metre temperature and wind gust is best estimated by retrieving the forecast daily maximum and minimum values. Care must be exercised since the maximum and minimum values refer to the last post-processing (3 and 6 hours). Note that the exact time for the extremes can not be deduced.

Precipitation forecasts are time integrated values from the start of the forecast. Values are available every 3 hours, after 72 hours for every 6 hours. For the EPS the interval is 6 hours throughout. No information about the occurrence of precipitation at the specific UTC times can therefore be deduced. The same applies to the other parameters which are accumulated, such as evaporation.

6.8.2. Horizontal and vertical resolution

The ECMWF forecast products can be retrieved at a wide range of resolutions, from coarse lat-lon grids to the original reduced Gaussian grid of about 40 km, 80 km for the ensemble forecasts. The data can be retrieved both from model, pressure, isentropic levels or iso-PV levels.

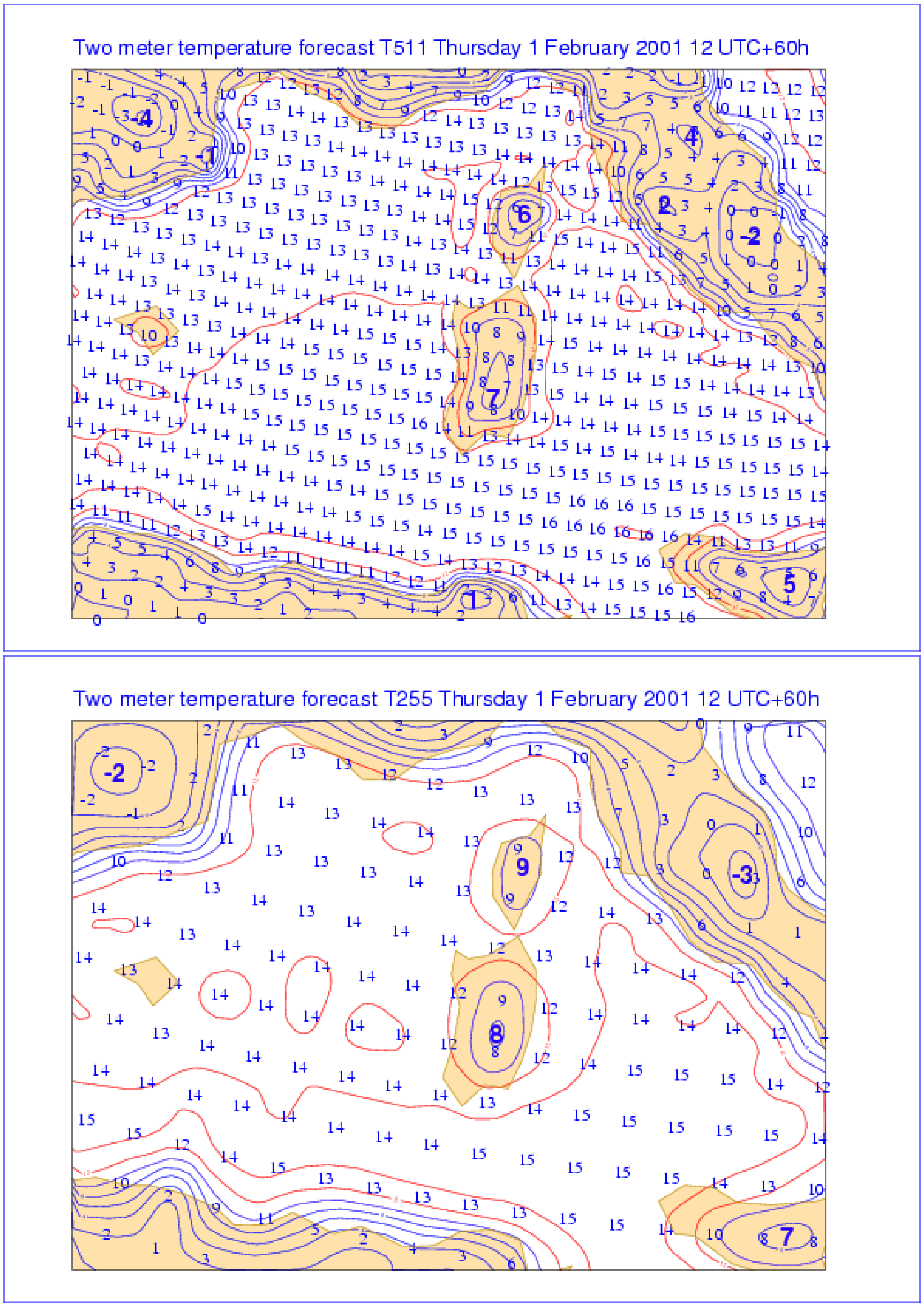


Figure 16 : The a) T319 and b) T159 Gaussian grid data. There are 4 grid points over Corsica in the T319, none in the T159. The same for Sardinia is 8 and 2 respectively.

The reduced Gaussian grid values, like all other grid values, should not be considered as representing the weather conditions exactly at the location of the grid point, but as a mean within a two– or three–dimensional grid box. Note that the 1.5˚ x 1.5˚ lat-lon grid value are point values interpolated from the reduced Gaussian grid and do not represent a mean over the 1.5˚ x 1.5˚ lat-lon area.

Often the variance of the observations within the grid area can be as large as the area average. This is particularly the case for precipitation. Any comparison or verification should then be against some spatial average around the grid point.

6.8.3. Orography

As mentioned in ch. 2.3.1, valleys and mountain peaks are smoothed out by the model orography. Due to this difference the direct model output of 2 m temperature may represent an altitude significantly different from the real one. A more representative height might be found in one of the nearby grid points. Any remaining discrepancy can be overcome by a correction using the Standard Atmosphere lapse rate, or statistical adaptation (see ch. 10)

6.8.4. Islands and peninsulas

For near surface parameters the distinction between land and sea points may be crucial, for example for 2 m temperature, precipitation or 10 m wind. As mentioned above (2.3.2.), the model land-sea mask values range from 0 to 1, allowing for mixed land and sea surface conditions near coastlines, small islands and narrow peninsulas. The marine influence if often over-estimated. Also here statistical interpretation schemes might be beneficial, in particular for the temperature forecasts.

6.8.5. Interpolation

Repeated interpolations, horizontally or vertically, will smooth the fields and dampen the extreme values. Graphical systems also introduce a slight smoothing. This smoothing might, in some applications, such as upper air fields, have a positive effect on the forecast quality, but for surface fields it might give unrealistic values.

If, due to lack of archive storage or limitation on the telecommunications links, compromises have to be made with the retrieval of fields, it is suggested that upper air fields, in particular from the EPS are retrieved with a coarse resolution, for example 5˚ x 5 ˚, while for the near surface weather parameters the use of the model's own reduced Gaussian grid is recommended.

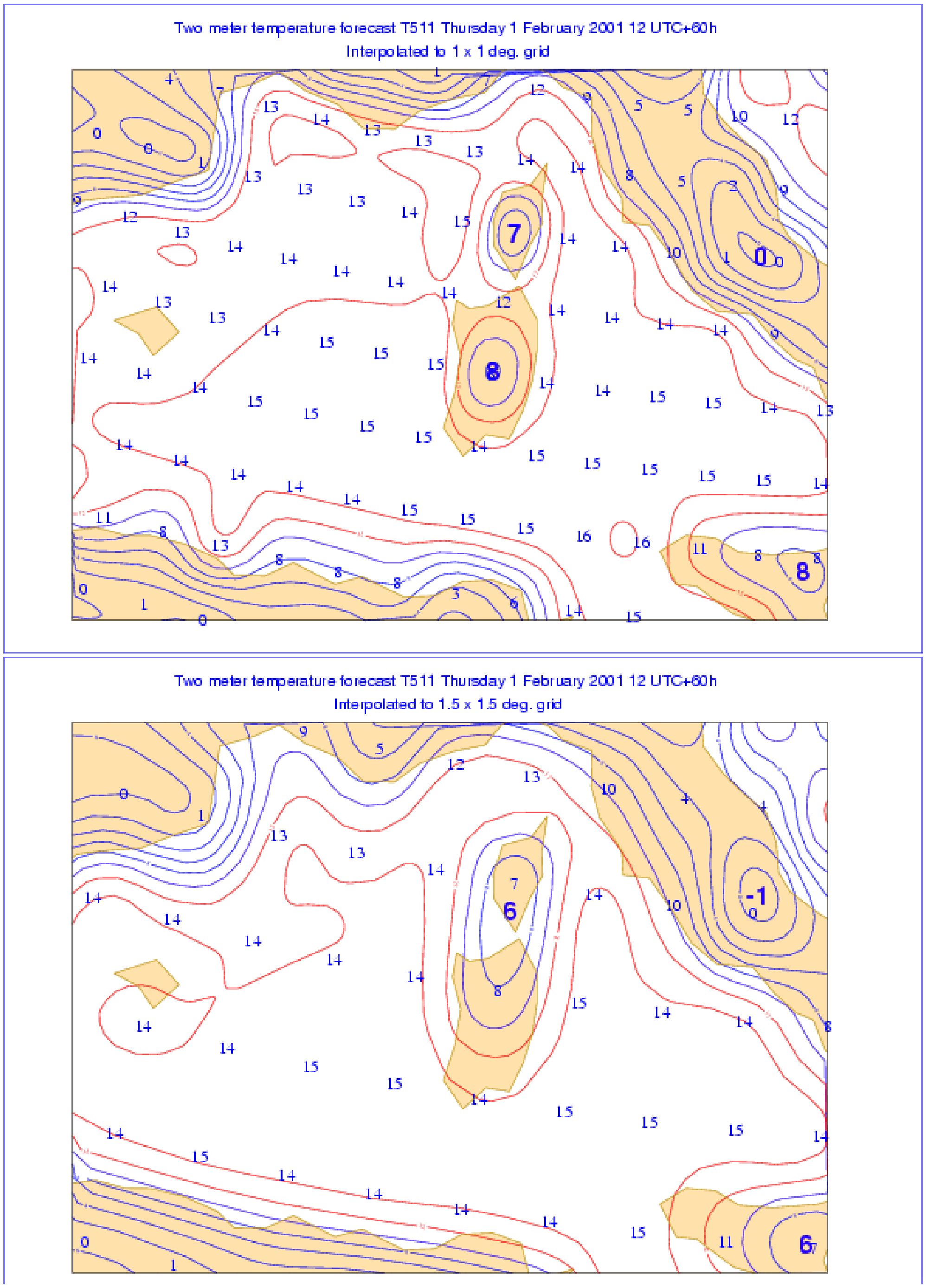


Figure 17 : The impact of interpolation to a) 1.5 deg. and b) 2.5 deg. latitude-longitude grids. Note than in contrast to the previous figure b, the 1.5 deg. grid makes the values from Sardinia "spill over" into Corsica due to the interpolation.

7. The verification of ECMWF forecasts

At ECMWF several types of statistical verification scores are computed for a number of areas and parameters, and stored in a historical data base. Most of them have been defined in agreement with other NWP centres and are regularly exchanged following WMO/CBS recommendations.

The mathematics of the must common verifications are quite simple, while the *interpretation* of them is a constant source of confusion and debate. The reason is not only that different verification methods often give conflicting indications, sometimes it is not obvious how to interpret the results from just one verification system: -*What looks good might be bad, what looks bad might be good*

Firstly, it is important to distinguish between the different types of forecasts we want to verify: Direct Model Output (DMO) from NWP, Post-Processed Products (PPP) and End-Products (EP) delivered to the public or paying customers. Here we will mainly deal with the first category although many of the conclusions are also valid for PPP and EP. Secondly, for all three categories it is important to distinguish between statistics which evaluate the *accuracy*, the *skill* and the *utility* of the forecasts.

The most common verifications measure the *accuracy* of the forecasts in relation to the analyses or observations. While the accuracy is absolute, the *skill* measure is relative, comparing with some reference like persistence, climate or some alternative forecast system. Finally, the *utility* aspect determines the success of a forecast system in terms of the monetary or political use that can be made from its predictions.

7.1. The standard verifications of deterministic forecasts

Several different types of verification scores are used at ECMWF.

7.1.1. The mean error

The mean error (ME) of forecasts (f) relative to analyses (a) can be defined

$$ME = \overline{(f-a)}$$

where the overbar denotes an average over a large sample in time and space.

The ME is not an accuracy measure as it does not provide information of the forecast errors. A perfect score, ME=0, does not exclude vary large and compensating errors of opposite signs. It is also important to remember that a non-zero ME does not necessarily imply a “flat bias”, i.e. a mean error independent of forecast value.

7.1.2. The RMSE

The most common accuracy measure is the Root Mean Square Error (RMSE)

$$RMSE = \sqrt{(\overline{(f-a)^2})}$$

The RMSE is sensitive to interpolation, phase error and the general variability or anomaly.

7.1.3. Interpreting the RMSE

Interpolation of fields, tends to smooth the features which acts to dampen the RMSE. Although higher resolution provides more detailed small scale structures and stronger gradients in the forecasts, any errors in time and space will be larger than if the same field had been interpolated into a coarser grid.

A *phase error* of half a wave length or more will score worse than if the system had not been forecast at all. The forecast is punished twice: first for *not* having a low where there is one, and second for *having* a low where there is none (“The double penalty problem”)[1].

Finally, the *flow dependence* of the RMSE males it core small and “good” values when the flow is zonal and close to the climatological mean, large and “bad” values during meridional and highly anomalous situations.

All these three effect, can be understood by a decomposing the RMSE into terms which measure different aspects of the score (see Appendix B). It shows that a large contribution to the RMSE comes from the variability of the fields themselves, both in the forecasts and in the analyses, not only from their agreement.

1. A commonly expressed reservation about the RMSE is that because of the high penalties of large forecast errors will easily lead a forecaster to a conservative forecasting practice. As will be discussed in ch. 8 deterministic forecasts *have to be* “conservative”, but complemented with probability statements for possible extremes.

The accuracy of a forecast is determined not only by the synoptic agreement between forecast and observed values or patterns, also the general “activity” of the forecast model relative to the atmosphere. A forecast system which, due to excessive diffusion, gradually decreases the frequency or amplitude of such synoptic features, will produce lower RMSE

.

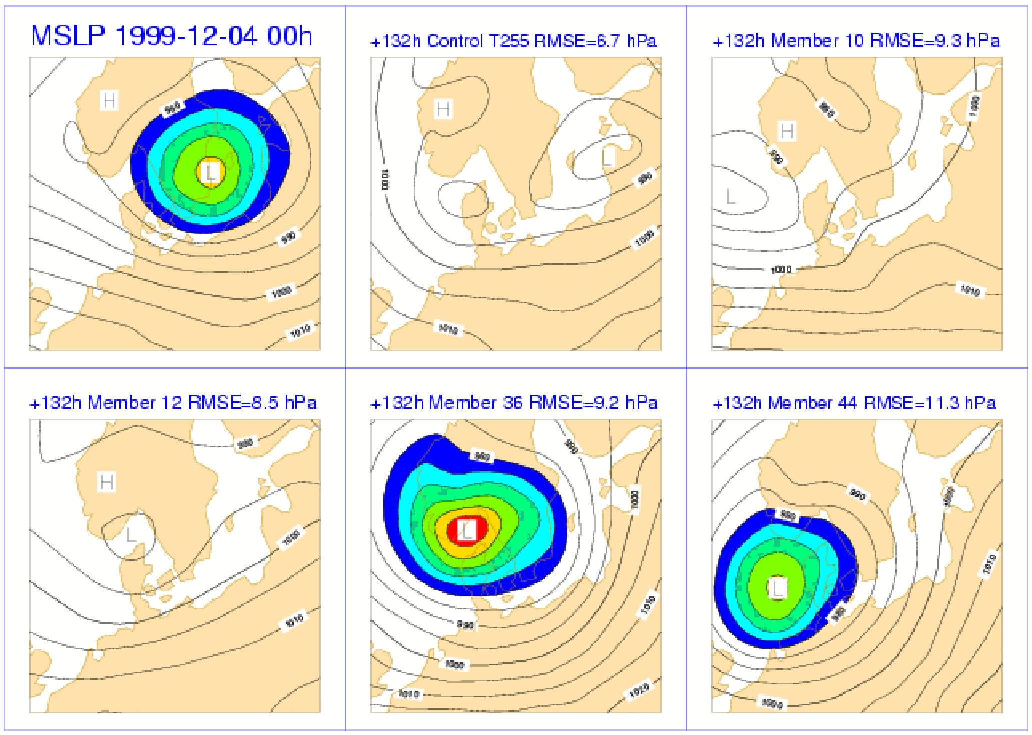


Figure 18 : On 4 December 1999 Denmark was severely hit by an extremely forceful storm. The figure shows the pressure patterns for the verifying analysis at 00 UTC, and five 132 h forecasts (EPS Control and four members, all run at T255). Those two members which actually had the storm, but with a certain phase error, scored worse than those three which had no storm at all.

7.1.4 The Mean Absolute Error (MAE)

A measure which is similar to RMSE is the Mean Absolute Error.

$$MAE = \overline{|f-a|}$$

The MAE is very similar to the RMSE but is less sensitive to large forecast errors. For small or limited data sets the use of MAE is preferred. Like the RMSE it is practical from the point of view of the duty forecaster’s intuition as it shows the errors in the same unit and scale as the parameter itself.

7.1.5. The Anomaly Correlation Coefficient (ACC)

The ACC is the correlation between the forecast and analysed deviations from climate (c):

$$ACC = \frac{\overline{(f-c)(a-c)}}{A_f A_a}$$

where $A_f^2 = \overline{(f-c)^2}$ and $A_a^2 = \overline{(a-c)^2}$. The ACC can be regarded as a skill score with reference to the climate.

7.1.6. Interpretation of the ACC

The ACC is sensitive to similarities in forecasts and analysed *patterns*, rather than their absolute values. In contrast to RMSE the ACC has a tendency to score large and “good” values during meridional flow situations, small and “bad” values during periods of predominantly zonal flow. This is particular the case in zonal situations when the forecast and observed positions of shallow waves are out of phase. ACC displays a weaker seasonal and annual variability than RMSE.

It has been found empirically that the level ACC=60% corresponds to the limit where the forecast does not exhibit any significant synoptic skill. It can be shown mathematically that ACC=50% corresponds to a categorical forecast for which the RMSE score is equally to a climatological statement.

7.1.7. What is “analysis” and “climate”?

The climate reference can be both the real climate or the model climate over some period. In both cases there is a problem of sample size and representativeness.

The analyses used in NWP verifications are to some minor extent dependent on the model forecast, in particular in data sparse areas. The forecast error at initial time is therefore rarely zero, since the verifying analysis most likely is not perfect. Replacing analysis with observations solves the accuracy problem, but introduces the problem of representativeness.

7.2. Forecast variability

To know if a certain change in the RMSE (and partly the ACC) is due to real model changes or just a consequence of variations in the atmospheric flow, ECMWF also monitors the synoptic-dynamic variability. This will also help to analyse the operational problem of “inconsistent” or “jumpy” forecasts.

Improving the model, for example but correcting the diffusion, might paradoxically result in *increasing* the RMSE. So what looks statistically "good" might be scientifically “bad”, what looks statistically "bad" might be scientifically “good”.

7.2.1. Variability measures

At ECMWF different variance measures are used to make sure that the dynamical activity in the forecasts on average is the same as the one observed.

Average spatial variance: The variance of a field over a specific geographical area (typically Europe and the northeastern parts of the Atlantic) is calculated daily. This results in a time series of analysed and forecast variabilities. If the area is sufficiently large the influence of random synoptic forecast errors are dampened.

Average temporal variance: For every grid point the monthly or quarterly variance is calculated. The result is a map of the geographical distribution and size of the average activity or variability.

Average tendency variance: For each grid point the RMS of the 12 or 24 h change in the forecasts and the analyses are computed on a monthly or quarterly basis. The result is similar to the time average maps, a map of the geographical distribution and size of the average temporal evolution in the forecast.

The three methods normally yield qualitative similar results, i.e. an over- or under-active forecast system will be detected by all three methods. There might, however, be some differences with respect to time and place due to the different methods involved.

7.2.3. Scatter plots

A further way to present verification which will highlight the variability, is the scatter diagram. Plotting forecasts vs analysis/observations provides a visual insight into the correspondence between the two data sets. Apart from depicting the variance of the two sets, it also offers the chance to identify outliers.

Additional useful ways to produce scatter plots are, as recommended by Nurmi (2003), in the form of forecasts vs forecast errors (or observations/analysis vs forecast errors). Such plots provide a visually descriptive method to see how forecast errors relate to the forecast (observation) distribution. It will reveal any potential clustering or curvature in the relationship.

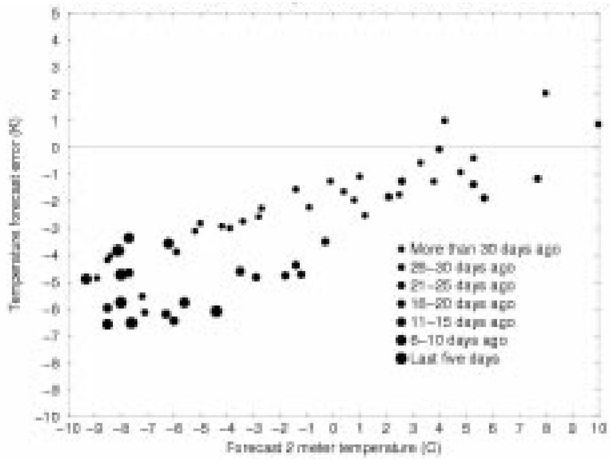


Figure 19 : A scatter diagram which depicts the relation between forecast 2 m temperature values and the forecast error. By varying the size of the circles the actuality of the values can be assessed. The mean error of about -4˚K does not correspond to any “flat bias”, but rather a systematic, almost linear relationship (error = -2.0 + 0.4 x forecast) which can be explored to correct the forecasts.

7.2.4. Forecast “jumpiness”

Closely related to the dynamic activity and accuracy is the problem if “inconsistent” or “jumpy” forecasts: the re-occurring problem that for example today’s D+5 is quite different to yesterday’s D+6 verifying at the same time. The “jumpiness” is closely monitored by the ECMWF in the same way as the forecast accuracy, measuring the RMS difference between consecutive forecasts verifying at the same time (see 8.4 and Appendix B).

7.3. Hit rate and False alarm rate

Verification measures like the RMSE and the ACC will value equally the case of an event being forecast, but not observed, as an event being observed but not forecast. But in real life the failure to forecast a storm that occurred will normally have more dramatic consequences than forecasting a storm that did not occur. To assess the forecast skill under these conditions another type of verifications must be used.

For any threshold (like frost/no frost, rain/dry or gale/no gale) the forecast is simplified to a yes/no statement (categorical forecast). The observation itself is put in one of two categories (event observed/not observed). Let H denote “hits”, i.e. all correct yes-forecasts - the event is predicted to occur and it does occur, F false alarms, i.e. all incorrect yes-forecasts, M missed forecasts (all incorrect no-forecasts that the event would not occur) and Z all correct no-forecasts. Assume alto-

Table 7: A forecast/verification table

forecast\obs	observed	not obs
forecast	H	F
not forecast	M	Z

gether 100 forecasts of this type with H+F+M+W=100. A perfect forecast sample is when F and M are zero. The ratio of observed to non-observed cases (H+M)/(F+Z) is determined by the verification sample (sample climate) and is independent of the forecast quality.

The Proportion of Perfect Forecasts is PPF=(H+Z)/100.

The Hit Rate HR=H/(H+M), the proportion of perfect forecasts when the weather occurred.

The False Alarm Rate FAR=F/(F+Z), is the proportion of forecasts of the event when it did not occur. HR and FAR can be combined into a diagram, the Relative Operating Characteristics (ROC) to compare deterministic and probabilistic forecasts (see 7.3.2. below)

The Probability Of Detection, POD=H/(H+M), is the proportion of perfect yes-forecasts.

The Frequency Bias Index, FBI=(H+F)/(H+M), measures the relative frequency of occurrence in the forecast with respect to observations that is not measured by the True Skill Score. It serves the same function as the dynamic activity in the forecast verification using the RMSE or the ACC.

A very simple measure of success of categorical forecasts is the difference POD-FAR which is known as the Hansen-Kuiper or True Skill Score. Among other properties, it can be easily generalised for the verification of probabilistic forecast (see 7.4 below).

7.3.2. The the Relative Operating Characteristics (ROC) diagram

Probabilistic forecasts can be transformed into a categorical yes/no forecasts defined by some probability threshold. For different thresholds the corresponding hit rates HR and false alarm rates FAR can be computed. A very powerful way to display and interpret this information is the two-dimensional so called Relative Operating Characteristics or ROC-diagram. A point in the ROC diagram is defined by the FAR on the x-axis and the HR value on the y-axis.

Figure 20 : An example of a ROC diagram to be inserted

The upper left corner of the ROC-diagram represents a perfect forecast system where there are no false alarms and only hits. The closer the point is to this upper left corner (low value of F and high of H) the higher the skill. The lower left corner, where both HR and FAR are zero, represents a system which never warns of an event. The upper right corner, where both HR and FAR take the value 1, represents a system where the event never occurs.

In reality a non-perfect system will have its values on a long convex curve pointing to the upper-left corner (the “ROC curve”). The area between ROC curve and the x-axis and the x=100% axis measures the skill of the forecasts

The ROC curve enables a comparison between a probabilistic and a deterministic forecast system. If the deterministic (FAR, HR) value lies over the ROC curve, the deterministic system is more skilful than the probabilistic. However, in terms of utility, greater advantages might be gained from the probabilistic information. It takes very good deterministic forecasts to be more *useful* than probabilistic ones.

7.4. Verification of probabilistic forecasts

In contrast to a deterministic forecast an individual probabilistic forecast can never be "right" or "wrong", except when 100 or 0% have been stated. Due to its nature the performance of the EPS can therefore *only* be evaluated from large samples of forecasts. However, like the deterministic forecast system the performance is determined not only by its predictive accuracy but also its ability to account for the variability of the atmosphere.

7.4.1. The Brier score

The most common verification method for probabilistic forecasts, the Brier score BS is similar to the RMSE, measuring the difference between a forecast probability of an event (p) and its occurrence (o), expressed as 0 or 1 depending on if the event has occurred or not. As with RMSE, the lower the Brier score the "better"

$$BS = \overline{(p-o)^2}$$

A Brier Skill Score (BSS) is conventionally defined as the relative probability score compared with the probability score of a reference forecast

$$BSS = (BS_{ref} - BS)/(BS_{ref})$$

In a similar way as the value of the RMSE depends on more factors than just agreement between forecast and analysis, so the value of the Brier Score is dependent on three factors: reliability, resolution and uncertainty (see Appendix C).

7.4.2. The reliability

The *reliability* measures the ability of the system to forecast accurate probabilities. Out of a large number of, for example 20% probability forecasts, the predicted event should verify for 20% of the forecasts, not more, not less. The reliability can be displayed in a reliability diagram where the x-axis is the forecast probability and the y-axis the frequency it occurs on those occasions.

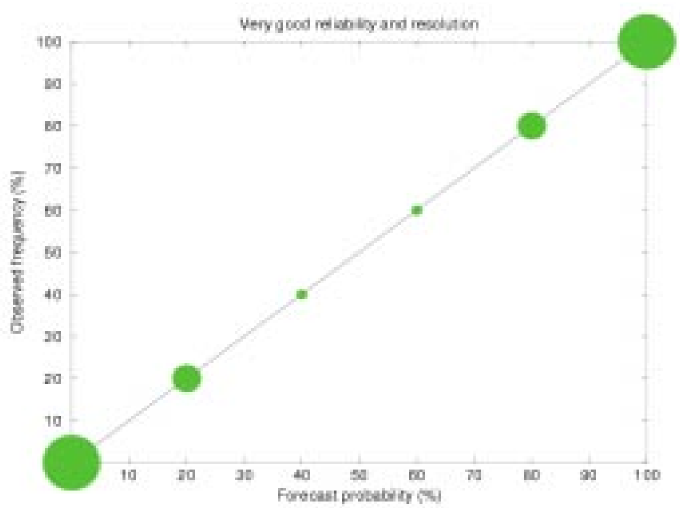


Figure 21 : A schematic example of reliability diagram. The x-axis indicates the forecast in intervals of 20%, the y-axis the observed proportions of occurrences in each probability interval. The size of the filled circles indicate the total number of forecasts for each category. This forecast system is very good with perfect reliability and a majority of forecasts in the high and low probabilities (good resolution.

Ideally the distribution should lie along the 45˚ diagonal. Most forecasts systems, both subjective and objective tend to give verifications where the distribution is flatter than 45˚. This is because most forecast system tend to be *over-confident*: low risks are underestimated and high risks overestimated. This can, as with any bias, be overcome by calibration (see 9.4.5.)

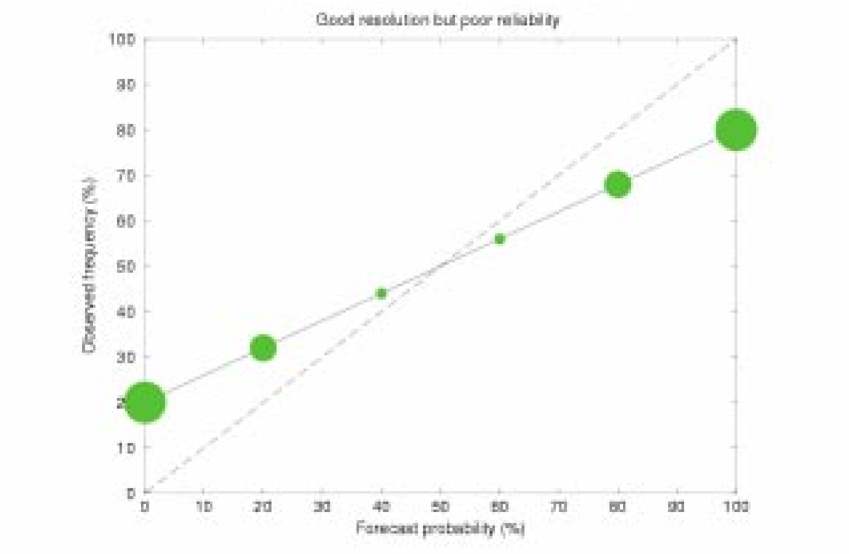


Figure 22 : An example of a typical over-confident probability forecast system. When the system is absolutely sure a certain weather will not occur (0%) or will occur (100%) it is wrong by 20%.

On the other hand, if the low risks have been overestimated and the high risk underestimated the forecasts are *under–confident* and the distribution is steeper than 45˚. Also this systematic error can be overcome by calibration.

7.4.3. The resolution

The *resolution* indicates the ability of the forecast system to correctly separate the different categories, whatever the forecast probability. For a given reliability, the resolution thus indicates the "sharpness" of the forecast. The maximum resolution corresponds to a deterministic forecast (only 0% and 100% are forecast), the minimum resolution corresponds to a climatological forecast (the same probability is always forecast).

Reliability and resolution are independent. For example, if the observed frequency is 90% in the 10% probability category, and 10% in the 0% probability category, the resolution is high but the reliability is poor. Fig. 23 shows an example of high reliability (45 deg slope) but poor resolution (most forecasts around the mean state).

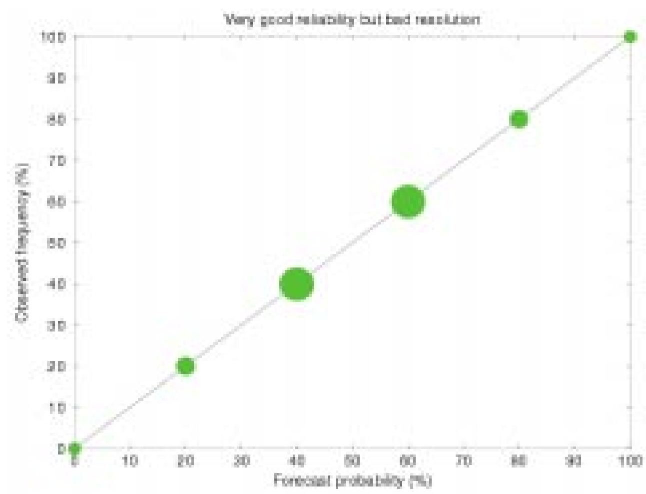


Figure 23 : An example of verification of a probabilistic forecast system with good reliability, but poor resolution. Most forecasts are centred around the climatological mean, with few cases when it provides confident forecasts (low or high probabilities).

For operational purposes, the resolution term is the most relevant, since the reliability can generally be improved by a calibration. However, this is obtained at the expense of sharpness. The resolution is not modified by the calibration if the number of categories remains the same and the EPS error characteristics remain stable from one season to another.

7.4.4. The uncertainty

The *uncertainty* indicates the intrinsic difficulty in forecasting the event during the period. It is also the probability score of the sample climatology forecast. The uncertainty is independent of the forecast system: being the same for the reference forecast and the forecast under evaluation, it plays no role in the skill score - but it can be shown to be an upper bound for the resolution.

7.4.5 Talagrand diagram

Due to the limited number of EPS members, the verifying analysis will be outside the ensemble range. For a system with 50 members this will happen 2/51 (~4%) of the time. In reality around 10% of the analyses verifying outside the ensemble. This means that the EPS at present does not spread out sufficiently.

A more detailed way of analysing the EPS spread is to construct a so called *Talagrand diagram*. It is constructed from the notion that in an ideal EPS system the verifying analysis is equally likely to lie between any two ordered adjacent members, including the cases when the analysis will be outside the ensemble range on either side of the distribution.

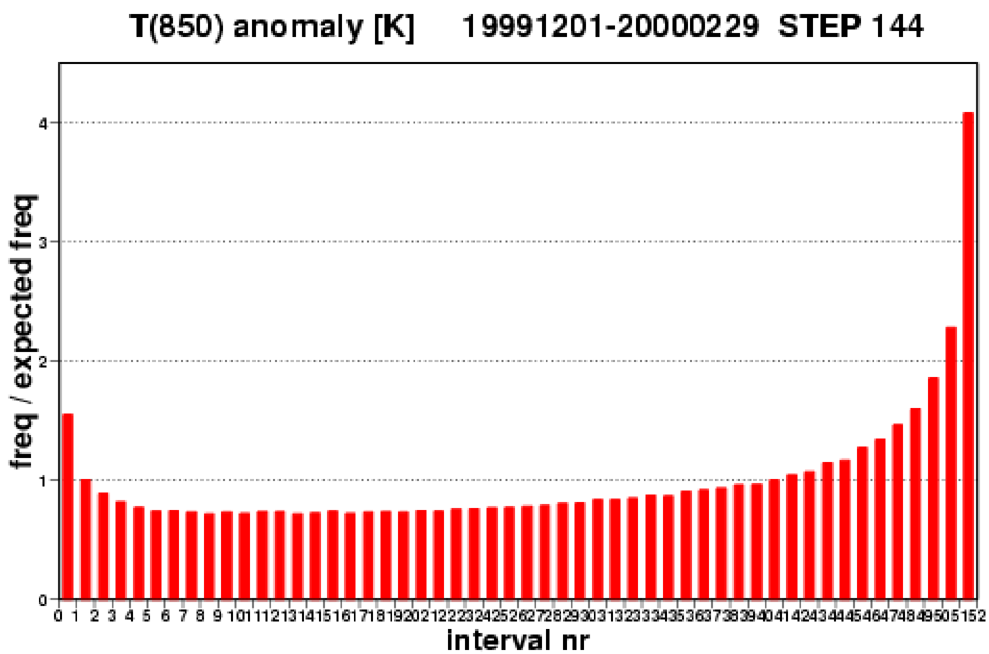


Figure 24 : The x-axis in a Talagrand diagram constitutes all the members in the ensemble, ordered according to some norm, for example the 850 hPa temperature in a certain geographical location. For an EOPS with 50 members there are 51 intervals including the unlimited intervals at the edges of the distribution. The y-axis indicates the number of cases the analysis is found between two ordered adjacent members on the x-axis. The U-shape indicates that the ensemble does not spread out sufficiently, the tendency to have a J-shape indicates that the forecast has a cold temperature bias.

In an ideal EPS system the long term “Talagrand distribution” should be flat with equally many verifications in each interval. In reality the distribution is slightly U–shaped due to over-representation of cases when the verification falls outside the ensemble and under-representation when it falls in the ensemble centre. For some parameters the U-shape degenerates into a J-shape, which indicates that the system has a bias for this parameter.

Improving the spread is a necessary, but not sufficient condition; a random sampling of weather parameters from the same season from the last 50 years would provide a flat distribution in the Talagrand diagram, but of course with no predictive skill.

7.5. Decision making from meteorological information

Ultimately the motivation for weather forecasts are the guidance they give in decision makings, their *utility.* A forecast system that provides good scores normally also provide good guidance for a wide range of needs. But there are exceptions. A forecast system that over-forecasts the occurrence of rain will score badly, but will be very useful for anyone who is sensitive for rain. Our intuitive sense of “usefulness” can to some degree be mathematically modelled.

7.5.1. The cost-loss ratio

A common situation in decision making is to weight the cost (C) of taking a protective action against the risk of making a loss (L) when no protection is made. Assume that (p) is the climatological risk for some adverse weather. The expected average of the day-to-day loss when there is no protection is pL.

If pL > C, i.e. the expected loss is greater than the cost of protection, then of course it will pay to invest in protection. On the other hand if pL< C the cost of protection is regarded as too high and should be avoided. The breaking point occurs when pL=C or p=C/L.

The "cost–loss ratio", C/L, is an important indicator of the sensitivity to weather information. If the climatological risk exceeds the “cost-loss ratio” there are reasons to invest into a permanent protection.

7.5.2. A simple cost-loss example

At a certain location it rains two days a week (climatological risk p= 2/7 ≈ 28%). Someone who regularly organizes an outdoor public event hesitates having to pay for rain protection, costing C=£200. If he doesn’t, and there is rain, he will lose L=£1000. With rain two days out of seven, his *expected* daily loss (pL) would be slightly above £280 (=2000/7). This exceeds the cost of protection £200 so he is well advised to make this investment. On average it will reduce his expenses with on average £80.

But the day-to-day probability of rain is not constant at 20%, but varies according to the synoptic weather situation. Since dry conditions seem to be more frequent than rainy, he would benefit from knowing when this will be the case. Knowing when the risk is < 20% would prevent him from investing unnecessarily in protection.

7.5.3. The importance of weather forecasts

The organizer starts to consult a weather service which issues rain probability forecasts. Assume that on 30 occasions their probability forecasts indicated risks > 20%, which, we assume, were followed by 10 occasions of rain[1]. On the remaining 70 occasions the forecast risk <20%, which consequently were followed by 10 occasions of rain. His total cost of protection £6 000 plus total loss £10 000 would yield £16 000.

In case the weather forecasts improves, thanks to an increased ability of the system to identify high and low risk weather situations, the organizer would save even more money. If the 70 occasions of <20% were followed by bad weather only on 5 occasions his total expenses would be just £11 000.

By relying on an increasing sophisticated weather information the organizer cut down his expenses from £ 280 to £ 200, by relying on climatological information, to £ 160 by consulting a weather service and to £110 by benefiting from their increased skill.

7.5.4. Probabilities as a way to compete

It can easily be shown that a weather service can improve its standing without really have to improve the forecasts in a meteorological sense. The way to improve is to accept that forecasts are imperfect, but finding out how much!

Figure 25 illustrates schematically the situation in a certain location where there is a climatological risk of 30% having bad weather (rain) on a particular day. A wide range of enterprises are operating in the region, each suffering a loss of £1000 is hit but rain. However, their cost of protective measures differ from £0 to £1000. In other words, their individual C/L ratio varies between 0 to 1.

In case nobody ever protects, they will on average suffer a loss of £300 per day (horizontal green line). Those who have C/L > 30%, i.e. their protection costs more than £300, will never protect. But the other might do every day and pay accordingly (steep brown line).

Weather forecasts will of course help to make better decisions when and when not to protect. But before we come into realistic examples it is important to know that even perfect forecasts be associated with expenses in those cases they motivate protection (sloping blue line).

1. This values do not follow from any formula, but are freely invented, but reflect typical forecast skill.

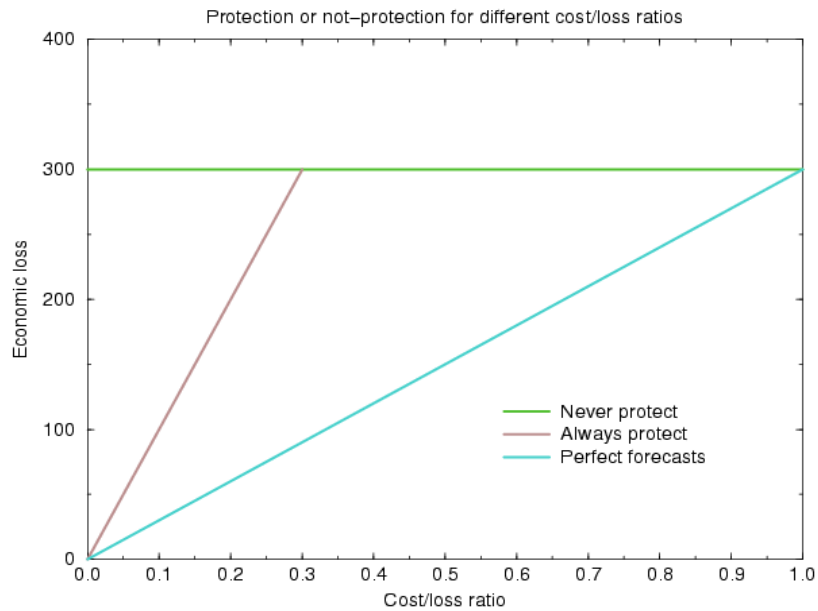
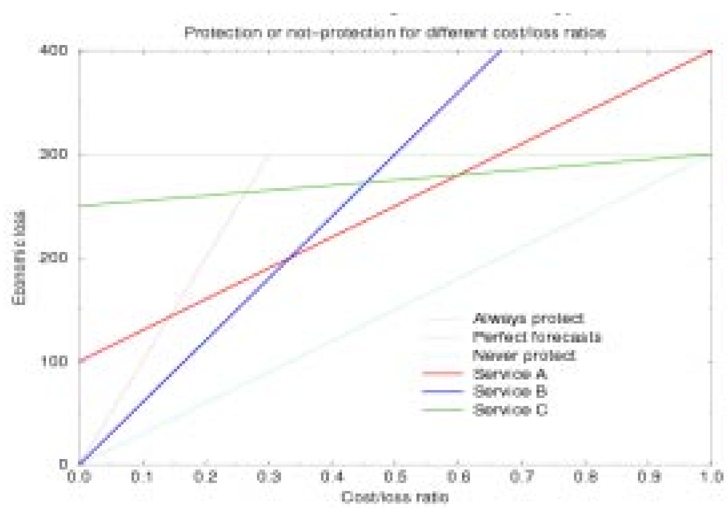


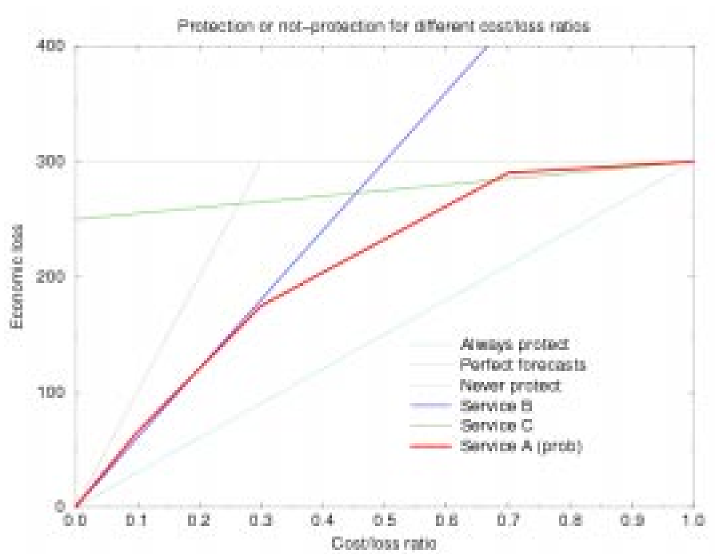
Figure 25 : A diagram which shows the average daily expenses for different cost-loss ratios and actions based on climatological information only. The curve for perfect forecast information indicates the lowest possible cost level.

In the region three weather services are operating. They all make categorical *yes/no* forecasts of rain, but have different characteristics. Service A tries to neither over- or under forecast rain, whereas Service B tends to over-forecast and Service C under-forecast. The average daily expenses for different C/L ratios are shown in figure 26.



Over- or under-forecasting rain would not yield favourable statistical verification results, but might be beneficial for certain users. Over-forecasting rain is of course interesting for those who have low protection cost (dark blue line), under-forecasting interesting for those with high protection costs (slightly sloping green line). Service A is only interesting for those who not have extreme protection costs.

The way Service A can compete with B and C is by supplementing their yes/no forecasts with probabilities. Instead of issuing a no-rain forecast, it might be expressed as 0, 10, 20, 30 or 40% probability. An enterprise with a C/L ratio < 50%, that would not have taken action if just a no-rain forecast would have been issued, might do otherwise when confronted with a 30% probability if their C/L ratio is 20%.



On the other end of the scale, an enterprise with a C/L ratio of 70% might chose not to protect if the risk is only 60%. Service A can thus provide a better foundation for decisions just by telling how certain the rain and no-rain forecasts are (broken red line in figure).

7.5.5. When weather forecasts do not matter

If the losses are small and the protection high, (high C/L ratio), there would be no need to protect. If the protection cost had cost more than £280 he would not have made the investment and rather taken the risk. Even if the protection cost had remained at £200 it would have been uneconomic to protect if the loss had been less than L=£700.

Nor would there be any interest in weather forecast if the protection was cheep and the potential losses high (low C/L ratio). The organizer would then always protect. Generally, the greatest interest in weather information are for users with a C/L ratio close to the climatological mean risk.

7.5.6. Limitation of the cost-loss model

The fact that the users do not follow the simple cost-loss model is not necessarily because they “do not understand probabilities”. Even a professor in statistics prefer to have £5 000 in his hands rather than having a chance of 70% of winning £10 000 although the expected gain is £7000. However, had the professor been a millionaire he would perhaps have appreciated the thrill of having a 70% chance of winning a large sum of money.

The simple cost-loss model must be extended to take into account the dependence of such factors as the decision maker’s total economic conditions.If the organizer has come into deep financial trouble and has only £300 left, he will of course be more careful to take the risk of losing £1000 and might choose to protect even for lower risks than 20%.

7.5.7. The human “irrationality”

Adding to the complexity of evaluating the social and economic benefits of a probabilistic weather forecast system is the fact that human judgment may take heuristic short-cuts that systematically depart from basic principles of probability.

Research during the last decades has increasingly focused on how human decisions may systematically depart from those predicted by standard economic theory. The (shared) 2002 Nobel Prize in Economic Sciences went to Daniel Kahneman, Princeton University, USA "for having integrated insights from psychological research into economic science, especially concerning human judgment and decision-making under uncertainty”[1].

The road to introduce a rational understanding of the best way to make use of weather forecasts in general and probability forecast in particular will be long, but full of interesting challenges.

1. Not all research is of Nobel prize class. The BBC recently showed on their web-site an academic explanation of probabilities which appeared to show how probability weather forecasts helped an out-door restaurant owner to decide when and when not to hire an extra waiter. However, a close examination of the figures on the web site showed that the owner made the best choice by employing the waiter!

8. The deterministic use of ECMWF forecasts (to be more worked on)

8.1. Introduction

The numerous meteorological parameters which are produced by the medium range forecast system do not seem to need any "interpretation". Time series of the forecast temperature, cloud cover, wind, rain etc. presented graphically for a specific location (“meteogram”), can be read off by meteorologists as well as laymen. However, care must be observed in dealing with direct model output. The figure below shows two consecutive forecasts of the 2 meter temperature for a location in the Netherlands.

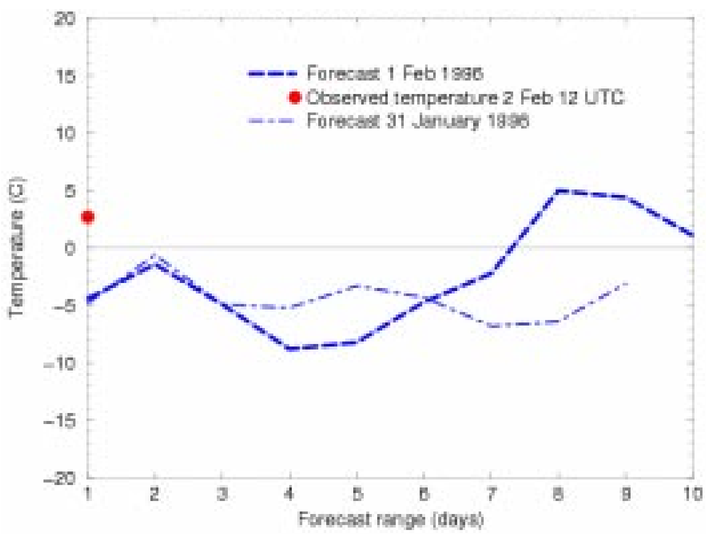


Figure 26 : The forecast 2-meter temperature forecast for Volkel in the Netherlands in winter 1995 according to two consecutive ECMWF forecasts. They are 8° K wrong 24 hours into the forecast and inconsistent during the second half of the period.

–The last forecast indicates a return to milder conditions a week into the forecast whereas the forecast from the day before indicated continued cold conditions. *Does this indicate that today’s forecast is less reliable?* A further potential problem might be that the temperature on the first day does not correspond with the forecast. *Does this mean that the rest of the forecast can be trusted?*

8.2. What can the forecaster do?

In contradiction to what is often said the forecaster can indeed improve on the NWP forecast in many ways, both in the short range and medium-range. But the methods are different depending on the forecast range and the geographical location.

8.2.1. Synoptic quasi-linear update

Whereas in short range forecasting it is a useful technique to modify the NWP products quasi-linearly in light of later information, this is not possible in the medium range. The impact of an analysis change remains approximately linear only up to 48 hours. Beyond this range it is normally impossible, without computer based calculations, to deduce how later information ought to modify the forecast.

8.2.2. Correction for systematic errors

The forecasters have possibilities to make positive correction for *systematic* model deficiencies due to poor representativeness, limitations in the horizontal resolution or in connection to some physical processes. This is in particular true for mountainous regions where the model orography differs from the real.

Although the forecasts of the large scale flow do exhibit systematic mean errors (see 4.3.) they are mostly of a much smaller magnitude than the non-systematic errors.

8.2.3. Correction of non-systematic errors

The forecasters best opportunity to add value to the forecast rests with addressing the *non-systematic* errors, in particular of the movements, positions and intensifies of synoptic features. Paradoxically, they can add substantial value to the NWP not by adding information, but by *removing* information.

8.3 Scale and predictability

Both operational verification and theoretical studies have shown that the larger the scale an atmospheric system, the more predictable it normally is.

8.3.1 Large scales are more predictable

For a realistic NWP model the range of meteorological scales is the same throughout the forecast; a D+10 forecasts *looks* like an analysis of the atmosphere: all scales are represented in a realistic way. But most of these realistic looking features in the D+10 forecast will of course not verify. They will be in the wrong place, with the wrong intensity and mostly not exist at all.

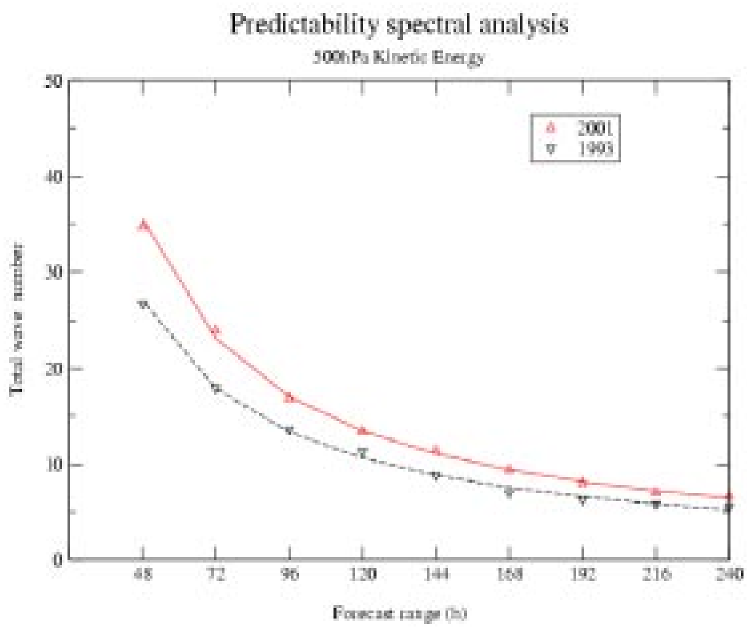


Figure 27 : The relation between the scale of an atmospheric feature at 500 hPa, measured by the number of spectral components needed to describe it, and its predictability. Forecasts beyond six days normally contain useful information only in the first 10-15 spectral components. Since 1993 the predictability measured in this way has increased by two days.

The predictability of atmospheric motion scales decreases rapidly throughout the forecast, starting with the smallest scales. Small baroclinic systems or fronts are well forecasted up to around D+3, cyclonic systems around D+5 and the long planetary waves around D+7 (see table 1).

Table 8: The current skill in NWP

Feature	<D+3	D+3 to D+5	D+5 to D+7	D+7 to D+10
Hemispheric flow transitions	Excellent	Excellent	Good	Some skill
Blocking creation and breakdown	Perfect	Good	Fair	Low skill
Cyclones’ life cycle	Perfect	Fair	Low skill	—
Fronts and 2nd developments	Good	Fair	—	—

Table 8: The current skill in NWP

Feature	<D+3	D+3 to D+5	D+5 to D+7	D+7 to D+10
Temperature/ wind	Very good	Skill in daily extremes	Skill in 5–10 day mean	
Acc.precip./ mean clouds	Good	Some skill	Some skill in precipitation 5–10 day acc. values	

By relying on his experience of what is normally predictable at a certain range, the forecaster using the ECMWF deterministic forecasts can disregard the small and unpredictable scales, and concentrate on the large and predictable scales Doing so he will most of the time be able to make useful forecasts on average up to a week ahead.

8.3.2. Methods to highlight the predictable scales

There are different techniques to highlight the larger, more predictable scales. The most consistent way is to use the Ensemble Prediction System (see ch. 9). But for non-EPS users there are plenty of useful alternatives through different types of smoothing or averaging.

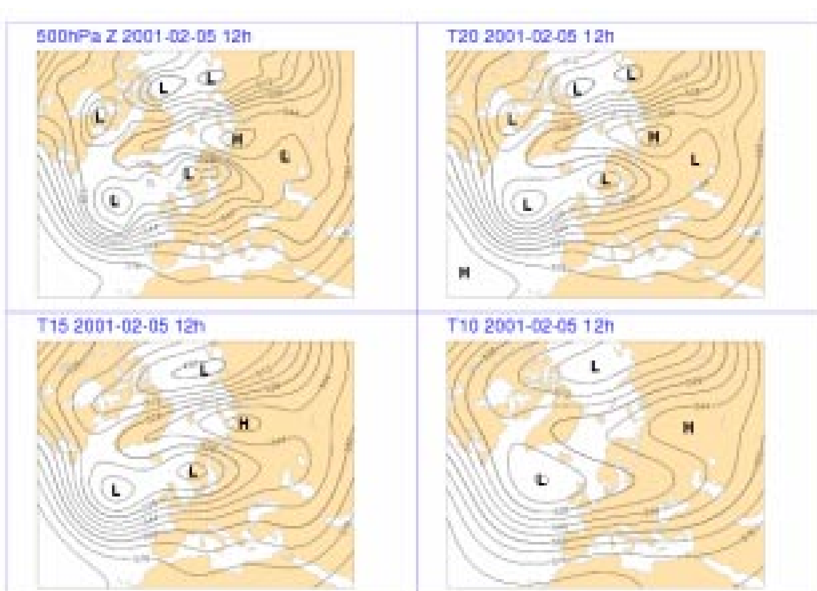


Figure 28 : The 500 hPa analysis from 5 February 2001 at different levels of spectral truncation: T511 (upper left), T20 (upper right), T15 (lower left) and T10 (lower right). The last truncation is suitable for highlighting the predictable scales beyond D+6.

Spatial smoothing: The average of the D+4, D+5 and D+6 forecasts from the same run might serve as a useful complement to the proper D+5 forecast. Similarly might the average of the current D+5, yesterday's D+6 and the D+7 from before yesterday highlight the large scale flow patterns. A third, slightly more technical technique, is to retrieve the spectrally archived forecasts with a reduced resolution, typically at T15 for D+6, T10 for D+8.

For medium range purposes the geographical area on display can be enlarged. This will automatically suppress the impression of the smaller scales. This is also in line with the speed by which atmospheric systems influence each other downstream. Forecast beyond three days are best understood when also the western part of the Atlantic and the easternmost part of North America are included. For forecasts beyond five days the whole of the North American continent and easternmost Pacific ought to be included.

Temporal smoothing: Mean temperatures are always more predictable than instantaneous values. The same is true for maximum and minimum values. The reason is that we have compromised with the time of the event to acquire knowledge about the value.

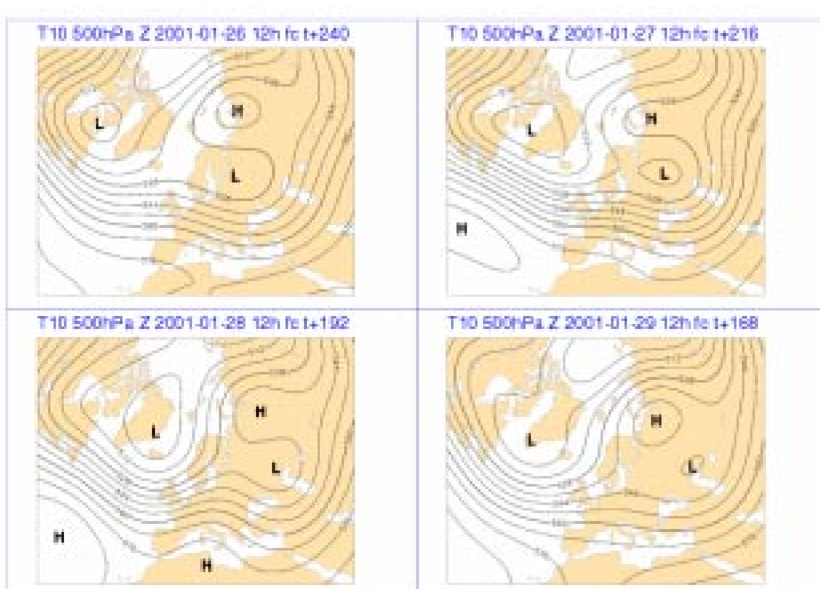


Figure 29 : Four 500 hPa forecasts from D+10 on 26 January to D+7 from 29 January 2001, all verifying on 5 February 2001 (see figure 28 plotted in truncation T10.

The predictability will increase with increasing time window, Rainfall can for example be more skilfully forecast if accumulated over two days rather than 12 or 24 hours. The fact that the deterministic forecast has indicated an event of wind speed >20 m/s becomes more significant if it is attributed to a three day period than to the exact time provided by that particular forecast. The advantage of condensing information by spatial or temporal filtering has, of course, to be paid by the occasional risk of losing information which, in hindsight, might have been important.

8.4 The day–to–day inconsistency

Closely related to the dynamic activity and accuracy is the problem if “inconsistent” (“jumpy”) forecasts: the re-occurring problem that today’s medium-range forecast is quite different to yesterday’s. There is a wide spread (mis)conception that “jumpy” forecasts are of lower reliability than “consistent” forecasts.

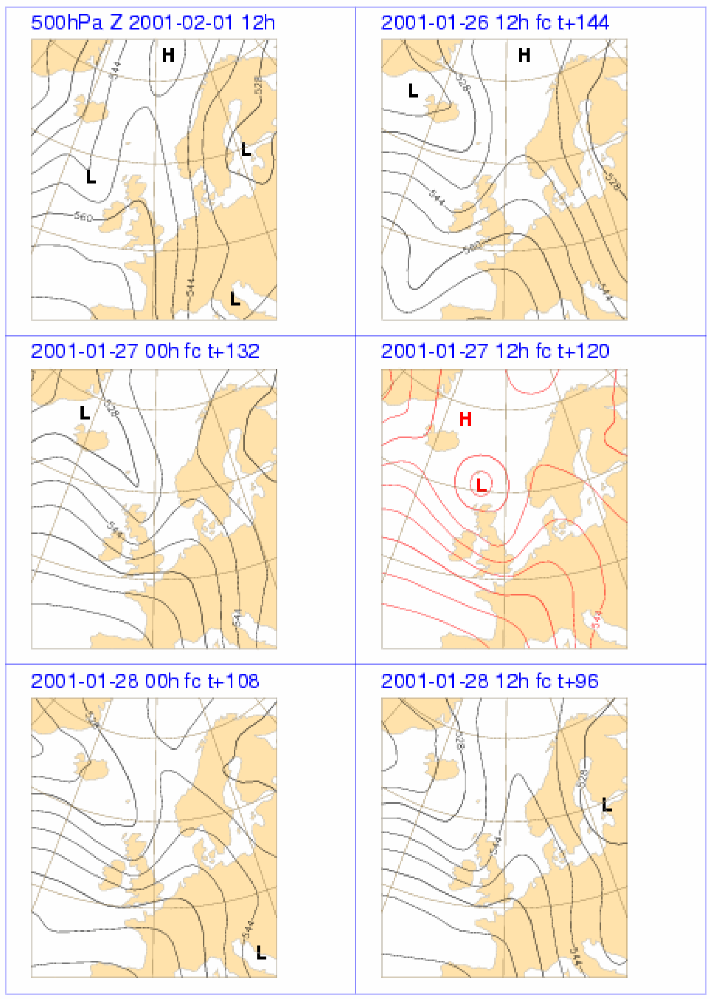


Figure 30 : An example of forecast “jumpiness” from the end of January 2001. Over a period of two days, the D+4 to D+6 are quite consistent, only with the exception of the D+5 from 27 January.

8.4.1. Forecasts have to change

The predictability can vary considerably from situation to situation. Sometimes a D+4 forecast can be wrong even in the large scale, occasionally the D+7 can be perfect even in details.

Changes in the forecast from one day to the other are necessary to enable a forecast system that take full benefit of new observations and modify previous analyses of the atmospheric state. Since the latest forecast is based on more recent data than the previous forecast, it is on average better. Although in most cases the changes in successive NWP forecasts are quite small, at least for the first three or four days, occasionally there might be large differences.

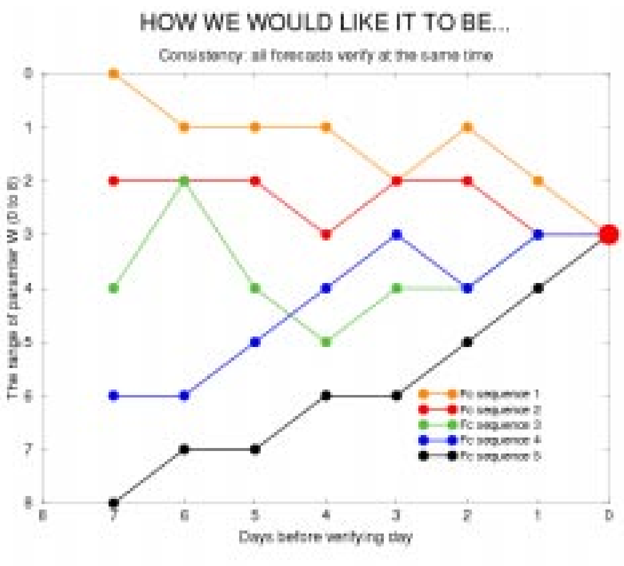


Figure 31 : A schematic representation of forecast evolutions from D+7 to the verification on five different occasions. All forecasts along the same time trajectory verify at the same time. Seven days

before the valid date most of the forecasts are in error. But as the verifying time approaches the successive forecasts consistently approach the analysis.

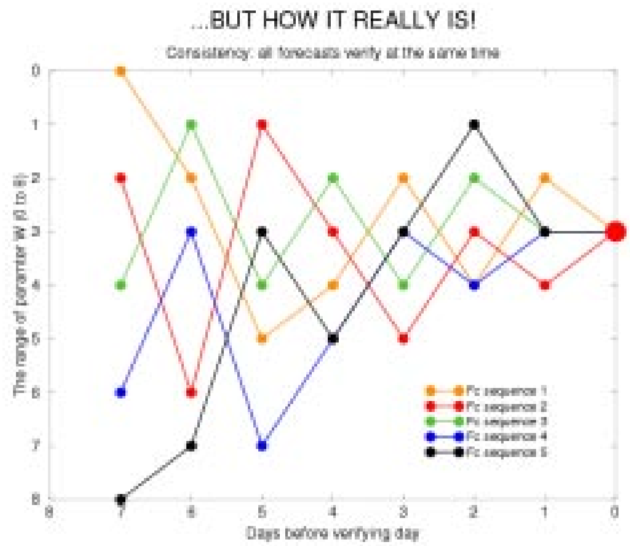


Figure 32 : In reality the successive forecast do not approach the analysis in an orderly, but rather chaotic way, with sometimes large changes from one day to the other.

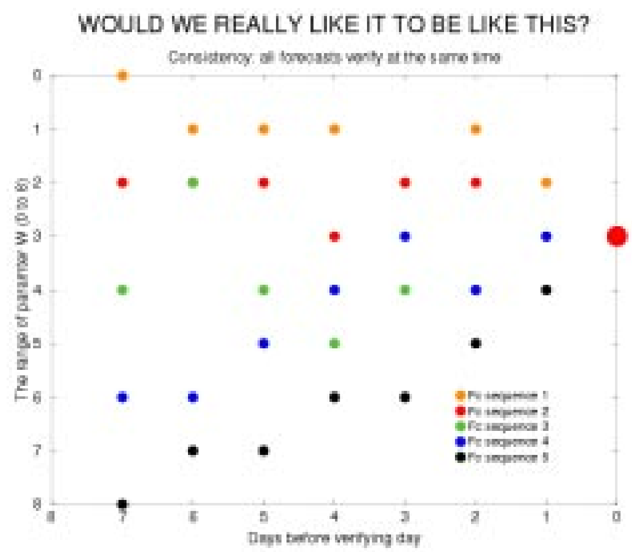


Figure 33 : The same as figure 31 but with the connecting lines removed.

If we remove the lines connection the individual forecasts and look at the scatter of forecasts individually, it turns out that the “orderly” system is less accurate than the “chaotic”. The forecaster would in these cases have been wise to ignore the jumpiness and trusted the “chaotic” system most.

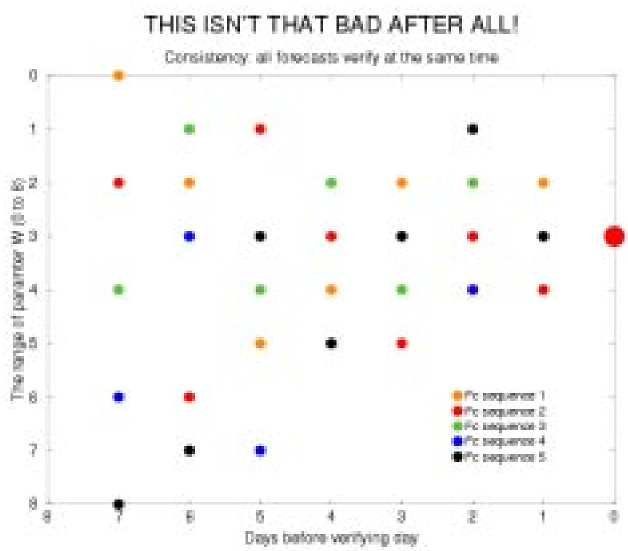


Figure 34 : The same as figure 32 with the connecting lines removed.

But the impression of “jumpy” and thus less reliable forecasts is misleading. A verification will actually show that the “jumpy” forecasts score better in terms of RMSE (and are equal in mean absolute error). A deeper analysis shows that the conceived relation between “jumpiness” and skill from a mathematical-statistical point of view is an artifice (see Appendix B). purely a political or psychological concept, a consequence that most end-users are human beings and not computers.

8.4.2. Consistency-skill?

Numerous statistical verifications have shown a correlation between D+5/ D+6 forecast consistency and D+5 accuracy of around 30%. It is not much, it only explains 9% of the variance. But the correlation is not zero and there has been suggestions that it could be enhanced in one way or the other.

But all attempts failed. It has turned out that the only way to increase the correlation would to make the forecast system worse, and then the correlation would anyhow never exceed 50%(see appendix B for more details).

If there is any relation between consistency and accuracy, it relates to the *preceding* forecast, not the current one (Persson and Strauss, 1995, Persson, 1997).

Verifications have shown that the correlation between the D+5/D+6 consistency and the error of the earlier D+6 is around 70%! Unfortunately, this can not be used for operational purposes.

A D+6 forecast from yesterday which is consistent with today's D+5 is most likely of higher quality than a normal forecast at this range. But what use can be made of this? The D+6, will on average not be better than the last D+5 - which it is similar to anyhow.

While objective verifications of the consistency-skill correlation at least have produced values between 20% and 40%, subjective verifications have failed to do so. During several ECMWF Training Courses the attendees have as exercises made subjective evaluations of forecast consistency and skill and found correlations between 20% and -20%.

8.4.3. Beware of consistent forecasts!

Experience shows that forecasters, in spite of all the difficulties with "jumpy" forecasts, manage to handle those situations well. The reason might be that the "jumpiness" urges him not necessarily to follow the latest NWP forecast, but seek out alternative information in previous forecasts or from other models (the "poor man's eps approach", see 8.5).

Paradoxically, it is in cases with several days of *consistent* forecasts, when the forecasters can find themselves in great difficulty. Several days with high consistency might have lulled him into a false feeling of reliability. The forecaster might be tempted to trust the NWP also with respect to scales which are not normally predictable. The lack of forecast alternatives makes it difficult for him to prepare his mind for possible alternative developments. Consequently he is taken by surprise when the NWP suddenly changes direction.

8.4.4. Instead of complaining about jumpiness - make use of it!

The "jumpiness" is normally caused by important changes in the analysis from one cycle to the other. It therefore serves as a simple ensemble forecast system. Like the real EPS it alerts the forecaster to possible forecast problems. The last 3-4 forecasts will therefore often help him to identify those scales which, in spite of all the "jumps", remain consistently forecast and therefore ought to be more predictable. The deterministic forecast in such situations should not necessarily be the last one, but a "consensus forecast", a weighted average of the last 3-4 forecast. This weighting can be define objectively from the inverse of the Mean

Square Error. *Any* subjective weighting is of course better than just following one of the individual forecasts, even the latest one.

The inconsistent parts of the forecast will provide information about possible alternative developments. The limited sample will of course make estimations of quantitative probabilities impossible. However, a mention of possible developments will automatically increase the value of the main “consensus forecast”. This “poor man’s ensemble forecast” can involve other models.

To sum up:

-Jumpiness is an unavoidable consequence of imperfect skill with a realistic model

-Avoid over-interpreting details when the forecasts are consistent

-When inconsistent - go for the latest, or a synthesis of the two - but do not over-interpret and never abstain from issuing a forecast

-Treat the inconsistency as a source of extra information, "what might happen"

-Do not be over-confident when the forecast are consistent, do not be under-confident when they are inconsistent.

So far our discussion has been rather qualitative. Now it will be shown how the approach outlined above will provide the most accurate deterministic forecasts and be a fruitful stepping-stone to the *rea*l EPS.

So while a good NWP should preserve a realistic variance around climate a skilful forecaster should do the opposite and gradually reduce the variability around climate for increasing forecast ranges.

Although the mathematics involved is simple, it has taken some time to establish in the meteorological community the necessity of combining measures of forecast accuracy with the condition of constant model activity. One reason for this difficulty is that what we demand of NWP products is not what we demand of products to end-users: *in contrast to a NWP model a synoptic forecaster should not have a constant forecast variability.*

8.5 The poor man’s ensemble approach

A combination of deterministic forecasts, from the same model or from different models, will yield more accurate forecasts than just relying on one model. The spread or “jumpiness” will provide important additional information.

8.5.1. What is right for a NWP model is not right for a forecaster

A good NWP model is able to develop synoptic features with the same overall frequency as in the atmosphere. At a range when the forecast skill is high the forecasts are close to the verification and the day-to-day change normally small. For forecast ranges when the forecast skill is low, the model is more free to make make significant changes from one day to another.

Deficient NWP models, such as a an old type quasi-geostrophic model or a primitive equation model with too much diffusion, are not able to simulate the full atmospheric variability. The limited climatological range will make the models unable to forecast large anomalies, on the other hand they will not display large "jumps" from one run to the next. Paradoxically, some "jumpiness" is a sign that the NWP is a realistic model of the atmosphere.

However, what is a good sign with a NWP model, is not necessarily good if it applies to a forecasters. In contrast to a numerical model, a forecaster should not be "jumpy", often changing his forecast from one occasion to the next. Even if the changes, like the NWP updates, on average improve the performance, it will have an adverse psychologically effect on the public.

"Jumpy" forecasts will reduce the public's confidence in the forecasts overall. It is bad enough to make a five day forecast that is wrong. But it takes five days to realise that, and by then most people might have forgotten what was said anyhow; *a radical forecast change will be spotted immediately.*

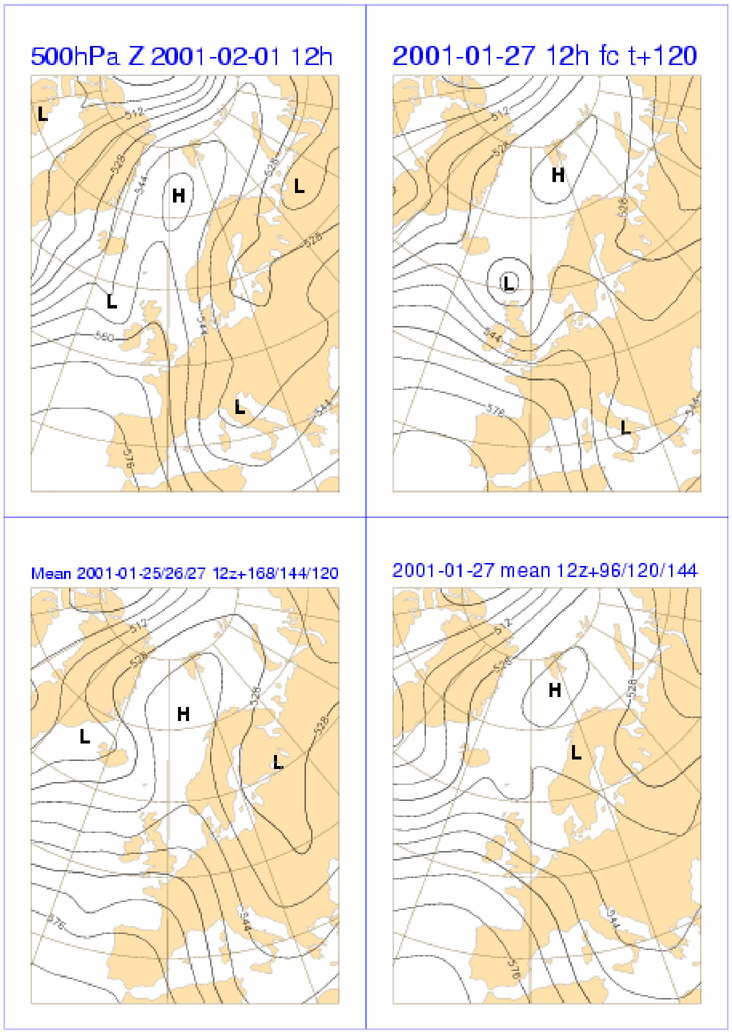


Figure 35 : The same situation as figure 30. The verifying analysis at 1 February 2003 in the upper left and the “jumpy” D+5 forecast from 27 January 2001 in the upper right. The lower left forecast is an average of the D+7 from the 25, D+5 from the 26 and D+5 from the 27 January, all verifying on the 1 February. The lower right is the average of the D+4, D+5 and D+6 from the 27 January forecast. The spatial and temporal averaging has smoothed the non-predictable scales.

Forecasts to the end-users should therefore not be similar in appearance to the NWP output. They should for example not necessarily try to cover the full atmospheric variability. Just because the NWP forecast output always appear realistic, this should not be the case for forecasts to the end-users. It is not possible, just be a look, to identify a NWP forecast as being for the short or medium range - *with the forecasts to the end-users this should be the case*

8.5.2. The better the NWP model - the worse the forecast?

We all know how the NWP models have improved over the last 30-40 years, so it would be surprising to know that in some respects they are performing worse than a forecaster from the pre-NWP era. But this is partly true.

A forecaster in the pre-NWP era mainly made forecasts 1-2 days ahead. For very short forecasts he relied on the last observations, so in some sense they were cleverly extrapolated persistence forecasts. For longer lead-times he could do nothing more than rely on climatological information.

It can be shown, empirically and theoretically (Appendix B) that a forecast system which, like persistence forecasts, reflect the full atmospheric variability, will for increased forecast lead times, converge to an *error saturation level* 41% above the errors of climatological statements (actually $\sqrt{2}$ times the climate error).

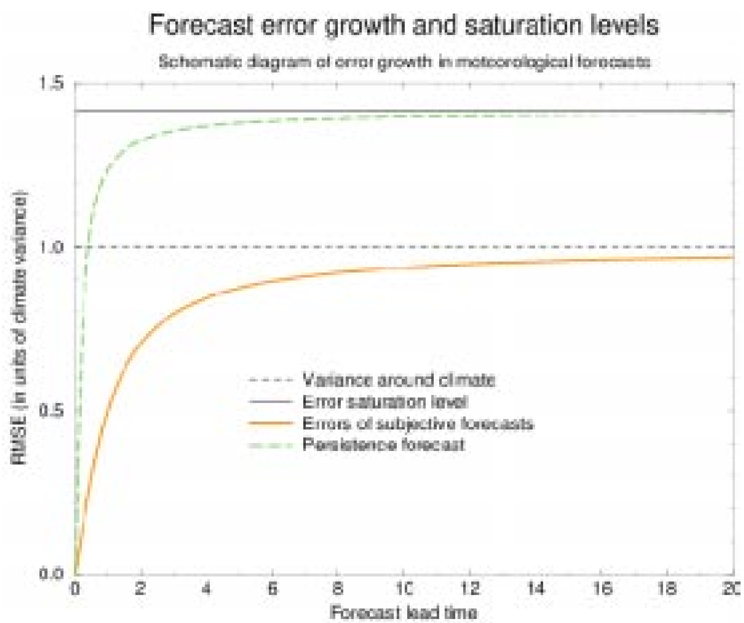


Figure 36 : A schematic representation of the pre-NWP accuracy of weather forecasts measured in RMSE with the climatological variance (the accuracy of a climatological statement) as a norm. The time unite at the x-axis is arbitrary. While the subjective forecaster was careful to make less and less detailed forecasts, the persistence forecasts became less accurate than climate in quite early forecast ranges converging to a 41% higher error level.

Good NWP forecasts display the same variability around climate as the observations, and will therefore, for increasing forecast ranges, approach the same high error saturation level as persistence forecasts, but of course more slowly. However. quite early on the NWP accuracy will be less than for a pure climatological statement - and for a pre-NWP forecaster.

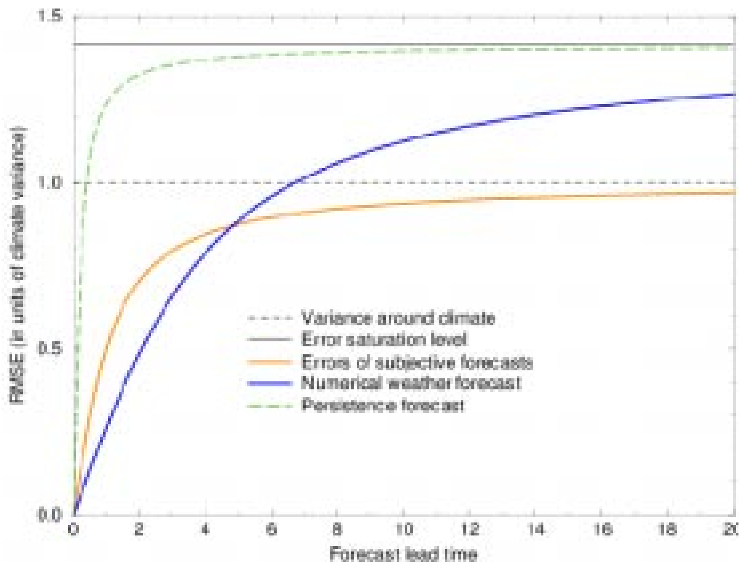


Figure 37 : Same as figure but with the error growth of a NWP forecast model schematically indicated. At some stage the NWP is less accurate than climatological information and before that worse than a pre-NWP forecaster.

A forecaster who just reads off the NWP output would, at some forecast range be less accurate than a a pre-NWP forecaster and at some range score worse than a climatological forecast.

The forecast range when a modern NWP forecast becomes less accurate than climate depends on the parameter. For weather phenomena coupled to small scale systems, like precipitation and cloudiness, it might be at a two day range, for temperature and pressure in the free atmosphere it might be 5-8 days. Forecasts which are less accurate than climate score ACC < 50%.

8.5.3. The poor man's ensemble approach

The "art" of weather forecasting with the help of modern NWP systems is thus to combine their high degree of realism and short-range accuracy, with the pre-NWP convention of approaching climate for longer forecast ranges. This is, as we will see in ch.9, most consistently accomplished in the EPS system. However, relying only on forecasts from deterministic models, it is possible to improve the accuracy considerably:

-Use of the relation between scale and predictability (8.3.1.)

-Use of spatial and/or temporal smoothing (8.3.2.)

-Use of statistical interpretation (ch. 10)

The result might not be as accurate as when using the EPS, partly because the limited number of "ensemble members" in the poor man's approach. However, due to the coarser resolution in the EPS compared to the operational TL511, the degree of geography related details might be better described.

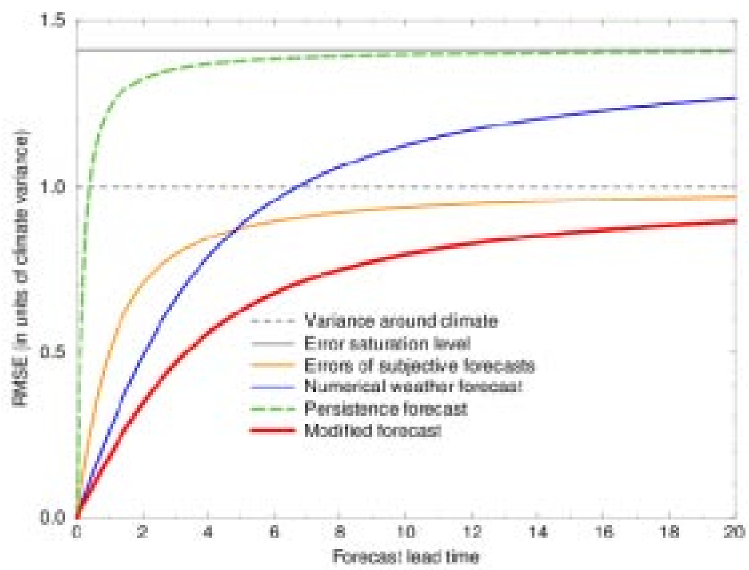


Figure 38 : Same as figure but with the error growth of a an ensemble mean forecast included. It incorporates the accuracy of the short range NWP with a convergence toward the errors saturation level of a climate statement. A poor man's ensemble mean is likely to lie between the red and blue lines, dependent on the number of "ensemble members" and the quality of the deterministic models.

8.4.6 Anomalous and extreme weather events

It is important to realize that some non-systematic errors can easily be misinterpreted as being "systematic". This relates to anomalous and extreme events.

This misinterpretation of "systematic errors" comes back in many disguises. If, for a certain location, the number of *forecasts* of heavy rain equals the number of cases *observed*, there is obviously no systematic model error. But since the heavy rain forecasts do not always verify it might give an impression that the model is forecasting heavy rain "too often". On the other hand judging from the cases of observed heavy rain, which are not always forecast, the model instead appears to forecast these event "not often enough.

.

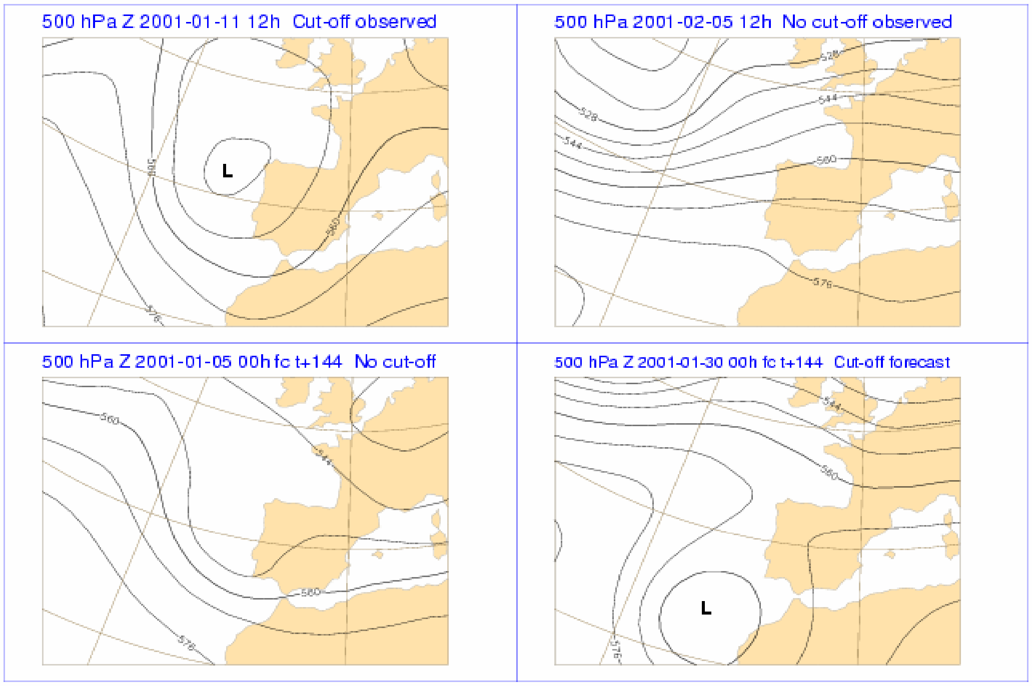


Figure 39 : The maps to the left show a case when the forecast failed to forecast an existing cut-off, the maps to the right a case when a cut-off was forecast which never occurred. Verifications over a long periods show that approximately the same number of cut-offs are forecast as are observed - they just do not always match at a D+6 range.

The difficulty with extreme weather events is that the forecasting system should on one hand detect as many of these as possible, without too many false alarms. What is “too many” is ultimately a matter of subjective choice (See. ch. 6.2 and 6.5). The ECMWF has chosen to maintain a uniform variability in the model over the ten day forecast range with neither over- nor under-forecasting (see. ch. 6.1.4). This is important to make it possible for the EPS to forecast extreme developments. It also the EPS that provides the most consistent treatment of extreme event forecasting (see 9.5).

9. The use of the Ensemble Prediction forecasts (also needs some more work and illustrations!)

The ensemble prediction system (EPS) has introduced new possibilities and challenges, but also technical and scientific complications which are increasingly researched and debated.

9.1. Introduction

The EPS can be used at different levels of complexity, from categorical, single-value forecasts, over probabilistic multi-value forecasts to providing direct input into elaborate decision making systems.

9.1.1. From a priori forecast skill to probabilities

When the EPS started in 1992 the ambition was initially to assess the skill of the operational forecast by relating it to the spread or variance of the ensemble. This ambition was soon abandoned. One immediate obvious reason was that the operational forecast was run at a more than three times higher resolution than the EPS. But, more importantly, the only accuracy that could be mathematically assessed by the spread was the accuracy of the mean of the ensemble.

Instead of trying to assess the skill of the deterministic forecast the aim soon became one of estimating the probability distribution (or probability density function, pdf) for important weather parameters. From these pdf:s it was possible to get quantitative estimates of the probability of certain thresholds being exceeded. A categorical value was provided by the mean or median of the ensemble.

9.1.2. The main objectives of the EPS

The EPS has mainly three purposes:

-To provide categorical forecasts with the highest possible accuracy;

-To provide a measure of the accuracy or reliability of this categorical forecast, and to present alternative developments and their likelihood;

-To provide a quantitative basis for reliable and useful probability forecast, in particular of extreme weather events;

For most purposes the output from the EPS must go through some post-processing of statistical or graphical nature to make it easy to use by the forecaster or end-user.

9.1.3. The challenges ahead

The output of the EPS represents an enormous quantity of forecast meteorological fields and offers an almost unlimited combinations of products. An important task is to find methods to condense the information. While the forecasters in the past had to make sense out of thousands of SYNOP, SHIP, TEMP and PILOT reports, coming out from the telecommunication lines to provide the best weather forecast service, they will in the future have to make sense out of 10-100 times this amount of information. It is a challenge to the meteorologists to convey the relevant parts of the EPS information to the end-customers or the public.

9.2. Deterministic use of the EPS

For several practical, traditional and psychological reasons categorical (single-value) forecasts are the most requested from the end-users. Deterministic statement like the mean (the best estimate), the most likely and the median can easily be extracted from probability distributions.

9.2.1. The ensemble mean

The *ensemble mean* is obtained by averaging all ensemble forecasts. This has the effect of filtering out features of the forecast that are less predictable. These features might differ in position, intensity and even presence among the members. The averaging retains those features that show agreement among the members of the ensemble. This is also, but to a lesser extent, the case with the central cluster in the tubing.

The averaging technique works best some days into the forecasts when the evolution of the perturbations are dominantly non-linear. During the initial phase, when the evolution of the perturbations has a strong linear element, the ensemble average is almost identical to the control because of the “mirrored” perturbations (added to and subtracted from the Control).

9.2.2. Reduced forecast “jumpiness”

As discussed in the previous chapter (8.4) the day-to-day forecast inconsistency (“jumpiness”) is as damaging to the public’s confidence in the forecasts as bad forecasts, sometimes even more. Thanks to the generally larger consistency of probabilistic forecasts in general, it follows that the pfd will display a high degree of consistency compared to single deterministic forecasts. As a consequence also the mean or median from the EPS will display a high degree of consistency.

9.2.3. Interpreting mean fields

The averaging leads to a smoothing of the forecast fields. The degree of smoothing depends on the spread of the ensemble: when the spread is small it will be possible to follow individual synoptic systems, sometimes even fronts, into the medium range. When the spread is large, only the largest atmospheric scales, the planetary (Rossby) waves remain.

During the late medium range, well defined intensifying synoptic system in the individual ensemble member forecasts might, when averaged in a mean cluster, indicate a *weakening* of the system. This paradoxical results occurs when the position of the system displays a geographical spread.

9.2.4. Unrealistic forecast fields?

An unavoidable side effect of the averaging, in particular of flow patterns, is that the resulting mean field often looks physically unrealistic: a type of flow that can never verify. So for example, the average of 50 small, equally shaped and intense baroclinic vortices in different geographical locations will not yield some “middle of the road” vortex, but most likely a shallow low pressure system.

But this is an unavoidable statistical effect of trying to condense all information into a single number. It is similar to the case of a region having on average 1.8 children per family, although every family has 1, 2, 3 or more children. It is a matter of judgement if the practical advantages justify any possible misunderstanding due to the “un-realism” of the result. The solution in many cases is to provide additional information, such as the spread, alternative developments and probability statements.

9.3. Variance measures

If the 50 member ensemble forecasts are quite different from each other, it is obvious that many of them are wrong. If there is a food agreement among the members, there is more reasons to be confident about the forecast and that most of them are close to the truth.

9.3.1. The ensemble spread

The ensemble spread measures the differences between the members in the ensemble forecast. Small spread indicates *low* forecast uncertainty, large spread *high* forecast uncertainty. It indicates how far into the forecast the ensemble mean forecast can carry informative value and helps the forecaster to express appropriate uncertainties.

Large spread should not be taken as a reason not to issue a forecast. The best strategy is then to issue a forecast based on the ensemble mean but to be careful in the formulations and try to indicate possible alternatives. The spread will also indicate what is *not* likely to happen, which at times might be as important as knowing what is likely to happen. Only when the spread might cover most of the climatological range nothing can be deduced from the forecast about the significant deviations from the climatological normal. Then the attention should change to possible extreme events and their probabilities (see below 9.4.6.).

Although the predictability decreases with forecast time, there are many occasions when this is not the case. Frequently the largest spread is found about half-way through the ten day forecast. Then the developing of a cyclone, for example at D+4, might be very uncertain, but not the formation of a blocking high some days later. Even if the forecast is uncertain in absolute terms, it may be quite accurate in relative. The actual temperature a week or so ahead might be difficult to specify; but for many applications a confident forecast of the *trend* might be quite useful.

9.3.2. Standard deviation fields

The ensemble standard deviation field, 500 hPa geopotential height, superposed on the ensemble mean field of the same parameter, allows to identify the meteorological features which are most affected by forecast uncertainty. However, there is not always a strong relation between standard deviation and what is regarded as synoptic spread.

Two similarly looking forecast maps might still display large differences if they contain strong gradients in slightly out of phase. On the other hand, two synoptically completely different maps with weak gradients will display small differences. The reason is similar to the one that makes RMSE and ACC sometimes convey different impressions of forecast quality: RMSE is sensitive to differences in magnitudes, the ACC to differences in shapes (see 7.1).

9.3.3. Conflicting spread indications

The spread-skill inter-relationship is complicated by the fact that the spread often varies considerably from one parameter to another. A low confidence in the temperature forecast does not exclude that the confidence in the precipitation forecast might be high.

During a blocking event there can be a large spread in the upper air fields, but small spread in the weather elements. Conversely, in a zonal regime, with small spread in the upper air fields, the differences in the track and timing of a baroclinic wave might yield large spread in the weather parameters.

9.3.4. Epsgrams

The ensemble information at an individual grid–point location may be displayed through a probabilistic meteogram, which indicates the time evolution of a given parameter for all ensemble members. The spread is indicated by the range of forecast values. 50% of the members are distributed evenly around the median to define a vertical rectangle. The remaining members define the extreme 25% “spikes”. The epsgram thus provides a discrete probability information in the intervals 0-25%, 25-50%, 50-75% and 75-100% which is sufficient for many applications.

The deterministic T_L511 and T_L255 forecasts may be included as a reference, as could the ensemble mean. To help the assessment of useful predictability the typical climatological variance might also be included. It is often useful to know when the spread might cover most of the climatological range.

As stressed in chapter 6.7 users should be careful when interpreting direct model output. Only the temperature forecasts in the epsgrams can be corrected for differences between true height above the sea level and the model height (according to the Standard atmosphere).

Locations along seas borders often suffer from large systematic errors, in particular with respect to temperature. The reason is that the strong temperature differences between land and sea may be poorly resolved by the T_L255 resolution. If the values are interpolated the effect will spread even further into land.

9.3.5. Spaghetti diagrams

So called *“spaghetti diagrams”* are actually maps where a certain isoline for each of the members is plotted. It provides an efficient way to summarize the EPS information. But care must yet be exercised in the interpretation since these "spaghetti diagrams" are sensitive to the gradients of the field. In areas where the gradient is weak they easily show large isoline spread, even if the situation is highly predictable. On the other hand, in areas of strong gradients they have a tendency to have small isoline spread, even if there are important forecast variations.

9.3.6. Clusters

Clusters, to some degree also tubes, are useful for *qualitative* synoptic risk assessments. The number of members in each cluster give its "weight" or probability, since all members are a priori regarded as equally likely.

As mentioned earlier, the "tubing" does not provide any explicit probabilities. Synoptic experience suggest, however, that every "tube" has a 10% chance of verifying, which leave the central cluster with a typical probability of 60-90% to verify, depending on the number of "tubes".

The clustering depend to some extent on the area. A blocking event might figure prominently in the "European" cluster. However, in a sub-area, less affected by the blocking, the clustering might focus on differences in the forecast of a certain cut-off development.

When estimating risks from the clusters, users should remember that, for a specific location, different clusters might have the same consequences in terms of weather, temperature and wind. The user might therefore have to make his own "clustering of the clusters".

Sometimes different clusters can look quite similar. This is mostly because they might differ in the *overall* level of geopotential height due to different temperatures in the forecast air masses. The clusters should therefore be treated as separated since for example a *cold* zonal flow will not yield the same weather as a *warm* zonal flow.

9.3.7. Postage stamp maps

An approach, which seems to be popular with many forecasters, is to look at the individual EPS members on a so called "postage stamp charts". Doing so he can roughly check the relevance of the current clustering, in particular with respect to his own area. However, doing so one must keep in mind that the clustering covers a time interval of 72 hours during which the members are on average "similar" according to some norm. This does not of course exclude that they at the beginning or end of the time interval might have similarities with members in other clusters.#

The user is also reminded that, because influences travel fast in the atmosphere, the performance of any member the first 12 or 24 hours of the forecast does not correlate locally more than a day or so into the future. The factors which determines the skill of a D+3 forecast are found in the initial performance far upstream, typically 80-100˚ to the west. The performance of the EPS over Europe at D+3 is related to initial conditions and early performance over the west Atlantic.

9.3.8. Guidelines for synoptic use of the EPS

There are at least two principle ways of working with the synoptic products from the ensemble forecast system.

The most common is to start by inspecting the last days' T_L511 forecasts to establish for how long into the forecast there is a reasonable consistency, and what alternative developments are indicated after that. Then the last days' EPS clusters are examined to determine if the inconsistencies in the operational model are reflected in the clusters, but also to determine if the EPS itself appears to be reasonably consistent. *This will establish which is the most likely synoptic development, and the main alternatives.* When this is done the Epsgrams and the probability maps are consulted in order to establish if the spread in the weather parameter forecasts correspond with the synoptic spread in the EPS clusters.

A more direct approach is to do the opposite: to start with the local weather parameter information, both in the Epsgrams and the probability maps. Only after that the clusters and deterministic forecasts are consulted to find the synoptic background to the weather forecast and their probabilities. The advantage with this method is that when there are weak relations between the spread in the large scale synoptic flow pattern and the local weather evolution the forecaster does not necessarily have to spend time finding out which flow scenarios are more or less likely.

9.4. Probability forecasts

If the purpose of the EPS just was to produce accurate categorical forecasts there would be need for only 10-15 members to define a sufficiently accurate ensemble mean. The reason why the EPS has 50 members (a number that might be increased in the future) is the need to make accurate probabilistic estimates of the risk of extreme and rare events. These tend mostly to be low risk (<10%) in particular in the medium range. Since the political, economic and human consequences of extreme weather can be large even small risk assessments may be significant.

9.4.1. Probability of weather events

If all ensemble members are a priori assumed equally likely, the probability of a weather event is simply the proportion of EPS members forecasting this event. From this set of distributions the probability of virtually any parameter, which are forecast by the model, can be computed. Maps of such probability distributions normally show great consistency from one forecast to the next.

When consulting the probability maps it is important to be aware of the time interval; is it a instantaneous probability like the risk of winds >15 m/s at 00 or 12 UTC, or does the probability refer to >5 mm over a 24 hour interval? A 10% probability of strong winds has quite different significance if the time interval is 10 minutes, 12 hours or one day.

Note that a 25% probability of >5 mm/24h can be related both to a showery regime where 25% of the area is expected to have substantial rain and equally well to the uncertainty of the arrival of a rain band. A 25% risk forecast for temperatures < 0˚C can relate to the possible early morning clearing of a low cloud cover, or the possible arrival of cold air from Greenland.

9.4.2. Probabilities over longer time intervals or large areas

The longer the time interval over which the probabilities are calculated, the higher but also more skilful they are. The confidence in the individual rain forecasts for days 5, 6 and 7 separately, is always lower than for the whole 72 hour interval.

Calculating probabilities with respects to several grid points defining a certain geographical area, have many advantages apart from increasing the skill of the forecasts. Certain extreme events like heavy rainfall have hydrological consequences far away from the immediate location. Since area probabilities will yield higher risk values than for local probabilities care must be exercised when conveying this probability information.

9.4.3. Probabilities of combined events

The EPS is also suitable to calculate probabilities of combined events like <6/8 cloud cover *and* temperatures >20˚C. Combinations of temperature and wind can define a “wind chill” index, temperature and humidity a “comfort index”, both examples of products that can be derived from the EPS output and given probability formulations.

9.4.4. Guidelines for combining categorical and probabilistic information

Since both deterministic and probabilistic information is useful and easily available, there might be reasons to find ways to combine them in a way that helps the forecaster to get a broad overview.

One way to achieve this is to present the most significant probability and categorical information together in a map, in the same way as the traditional composite synoptic maps provide a simultaneous overview of the flow pattern, air masses and local weather.

So for example, the 500 hPa geopotential ensemble forecast could provide the background for the 850 hPa temperature anomalies, expressed as probabilities. These anomalies correspond roughly to the notion of cold and warm air masses, with the 500 hPa mean flow as the “steering level”.

In a similar way the 1000 hPa ensemble mean field could provide the background to two probability fields, each depicting important weather features: the probability of >5mm/24h and >15 m/s. The mean flow and the gale probability will complement each other in a valuable way. So for example when the gradients in the ensemble mean weakens for longer forecast ranges, indicating a *decreasing* mean wind speed, the gale probabilities might very well *increase*.

9.5. What value can the forecaster add to the EPS?

It is quite possible to run an operationally well designed EPS with post-processing providing tailor suited forecasts in tabular or digital form to end users. So what role will the forecaster play in the future with respect to the EPS? The forecasters still have an important role to play: a) to provide crucial input in cases of extreme weather, b) to detect and take systematic errors in the EPS system into account and c) to help the end-user make the optimum decision.

9.5.1. Extreme weather events

Extreme weather events are coupled both to the small and large atmospheric scales. The large scale extremes can be long periods of anomalous temperatures or rainfall over large areas. These situation, often coupled to persistent blocked flow, are skilfully forecast, five days or more in advance. For small scale extreme events, like heavy rainfall, strong winds and rapid changes in temperature the forecast skill decreases from day 3 onwards (see table 10).

The cascading process for providing guidance on severe events, long as well as small scale, proposed by WMO/CBS, aligns well with traditional forecast practises:

72 hours and earlier: Preliminary indicative guidance, based on probabilistic EPS material. The forecaster is advised to consider deviating from the EPS in the rare event of the T_L511 operational model has during the last 2-3 days indicated a risk of a severe event, which has not been included in the EPS. The forecaster should in particular adding his own input when the EPS has forecasted a large scale synoptic flow regime where severe weather, which it is not possible for a T_L255 resolution to describe, can not be excluded.;

24 to 72 hours in advance: more specific warning guidance based on a mixture of probabilistic EPS and deterministic material. The EPS should not be disregarded at this time range, but in a more qualitative way since the number of extreme forecasts might not be a reliable indication of the probability.

< 24 hours in advance: warnings issued by the responsible centre, based on detection and tracking of the severe weather system. Awaiting EPS based high-resolution, limited area forecast systems, the forecaster is advised to focus his attention to the categorical forecast information from the last 6-12 hours.

9.5.2. The EPS system is not perfect

The EPS has a small tendency to underestimate the atmospheric variability, i.e. to identify all possible weather regimes that can occur in a given situation. This can sometimes be seen when the deterministic T_L511 provides a forecast that is not covered by the ensemble. It is difficult to estimate how this should be interpreted, but a tentative suggestion is that the T_L511 should be regarded as likely as a handful members. *It should not, unless there are strong reasons, be regarded as the most likely solution.* Remember that the T_L511, like the Control and the analysis for purely statistical reasons, should be outside the ensemble around 4% of the time.

There should, in principle, not be any inconsistent “jumps” in the EPS forecasts from one run to the next. But sometimes, when there is dynamic activity over many regions on the Northern Hemisphere, it might not be enough with 50 analysis alternatives. In those cases it can happen that the perturbations one day are more concentrated in one sensitive region than another, only to change the distribution slightly the next day. In such cases the forecaster is wise to consider *both* days’ ensemble output.

9.5.3 The end-user is not perfect either!

Even if the forecasters cannot modify the probability values coming out from the EPS (in particular if it has been subjected to some post-processing) they can still add extra value by helping the end-user rely on relevant information and assist him in the decision making process.

There are broadly speaking two categories of users: those who know precisely what weather information they need, and those who do not. The first group, which is small, is in principle satisfied with automatically prepared information in tabular or digital form. The second group, which constitutes most of the general public, is erratic in their concerns: cloud cover, temperature variations, occurrence of rain and strong winds; everything might have an impact on their decisions.

The main requirement of the general public is a forecast that emphasizes the “important” aspects of the meteorological future. At the same time, they want to be alerted to any risk of severe weather. Probabilistic statements have to be expressed carefully so that subjective judgements of end users eventually reflect the actual risks. Stating that “we cannot exclude the possibility of hurricane gusts” might be better understood and lead to relevant protective actions taken, than a statement that “there is a 10% risk of temporary >30 m/s winds between 18-24 UTC”.

Operational forecasters have a crucial role to play in interpreting the probabilistic guidance and communicating to the public their own subjective assessment of weather related uncertainties.

10. Comments on the use of statistical interpretation It needs some more work. Note that although the examples are from Kalman filtering (in Sweden) the point I want to make concerns all kind of evaluations, both in NWP and MOS (in the world!)

There is a wide literature on the subject of statistical interpretation and post-processing of NWP. Here only some topics will be discussed which are particularly relevant to medium-range forecasts.

10.1. Statistical interpretation of deterministic forecasts

A dynamical–statistical interpretation of the NWP output can be produced for any particular weather parameter (predictand) e.g. precipitation, cloud, visibility, temperature, provided that historical data for the location exists.

10.1.1. Non-adaptive interpretations

There are two traditional statistical interpretation methods: the Perfect Prog Method (PPM) and the Model Output Statistics (MOS) technique. In the PPM a statistical relationship is established between observed values of the predictand and *analysed* predictors from the free atmosphere; in the MOS a statistical relationship is established between observed values of the predictand and *forecast* predictors, both from the surface and the free atmosphere.

The MOS techniques also partly compensate for the model's systematic errors. If the model has a tendency to under– or over–forecast any predictor, this will be compensated for (Murphy and Katz, 1985, Glahn et al, 1991).

10.1.2. Adaptive interpretative methods

Adaptive methods, in particular the Kalman filter, shares MOS' advantage of being able to compensate for model errors while at the same time being able to continue to work despite changes in the model characteristics. In contrast to MOS and PPM, the adaptive filter does not need any long historical data base. It can start to provide skilful interpretations 1or 2 weeks after the start. If the model changes in any significant way, the adaptive system will notice it and gradually, normally within a week, adjust the statistical relationship (Persson, 1991, WMO, 1992, Cattani, 1994).

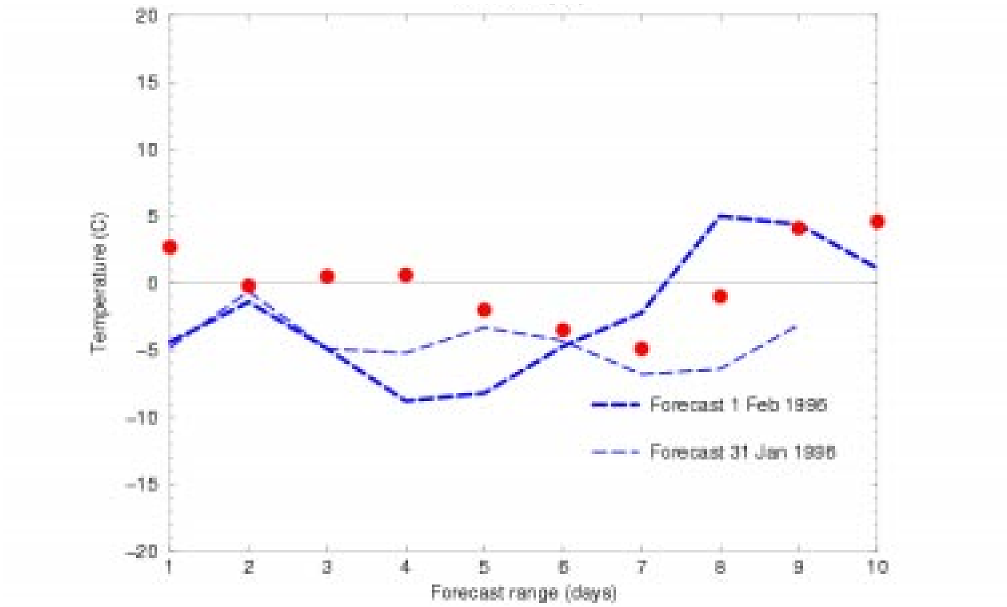


Figure 40 : The same as figure 1, but with verifying 12 UTC observations. Both forecasts are systematically too cold and when they verify on day 6, it is probably due to and additional error, a wrongly forecast synoptic pattern, which happens to compensate the systematic error.

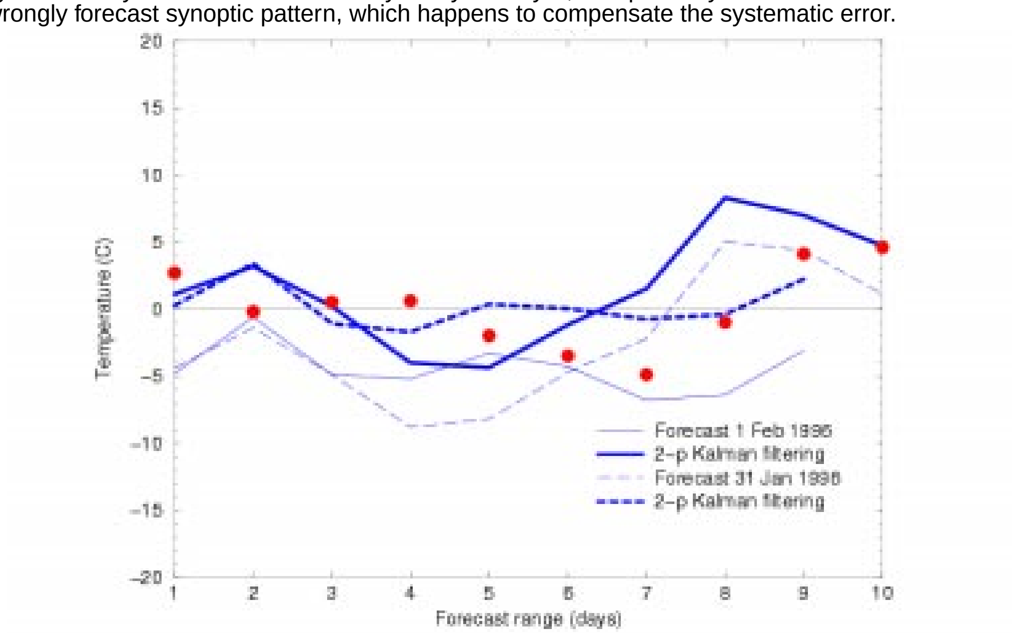


Figure 41 : Being subjected to statistical interpretation removes to a large extent the systematic error, leaving the synoptic, non-systematic error to dominant. Note that the correction is slightly larger for cold forecasts than for warm.

The most simple task is to modify the 2 meter temperature or the 10 meter wind speed which mostly have convenient statistical structures. MOS and PPM allows interpretation of NWP to thunderstorms, visibility and other parameters not available directly from the NWP.

10.2. Problems of validating a statistical scheme

As with interpreting scores of weather forecasts, manual as well as from NWP, it is not trivial to evaluate the performance of a statistical interpretation scheme. What looks good might quite often be rather bad, what looks bad can be good.

10.2.1. Does statistical interpretation dampen extremes?

It is sometimes said that statistical interpretation tends to smooth out extremes.This depends on the particular scheme and is not necessarily the case. The following example will show that a statistical scheme might very well improve the forecasts of extremes.

During the cold season the ECMWF model has, like many other NWP models, difficulties to simulate the extreme cold temperatures which may occur in northern and eastern Europe. The times series below show the ECMWF +24 hour forecasts of the 2 m temperature for a location in Lapland. The RMSE=5.0˚ K..

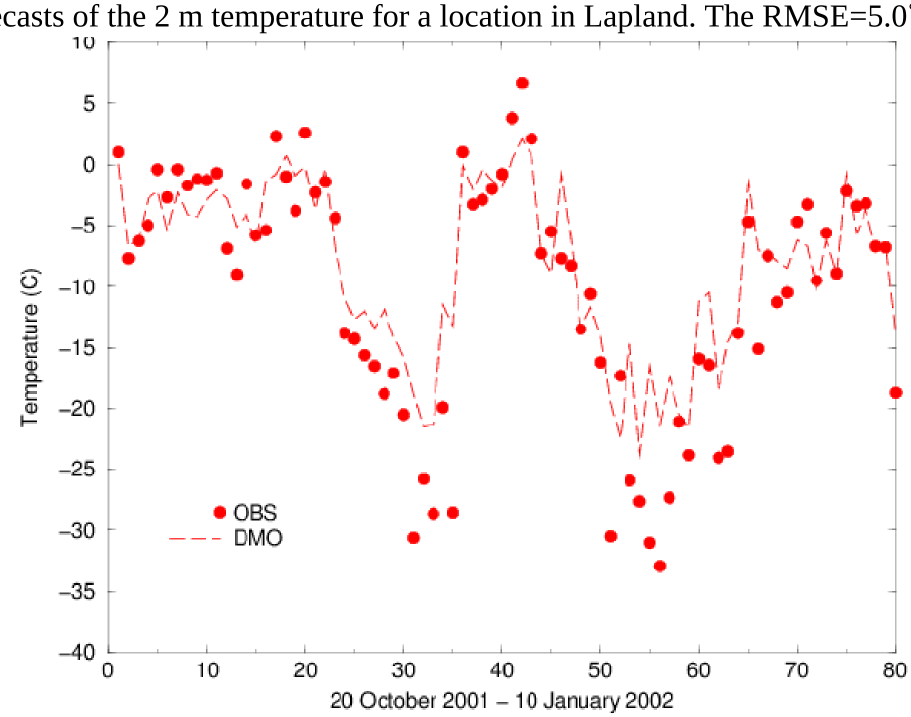


Figure 42 : Observed and 24 h forecast 2 m temperatures for Kiruna in Lapland in early winter 2002-2003. The extreme cold events are too warm by about 10˚ K the whereas the mild periods are rather well forecast.

The systematic error can then be given a mathematical form like

Error = A + B• Forecast temperature

where A and B in a MOS system would be constant, in an adaptive system vary in time. Because the expected error, and thus the correction, are also dependent on the forecast, albeit in a linear way, it will enable the statistical system to make different correction for different regimes. While a simple bias correcting scheme can only correct for mean errors, a scheme with more parameters or dimensions, can also correct for systematic errors in the variance.

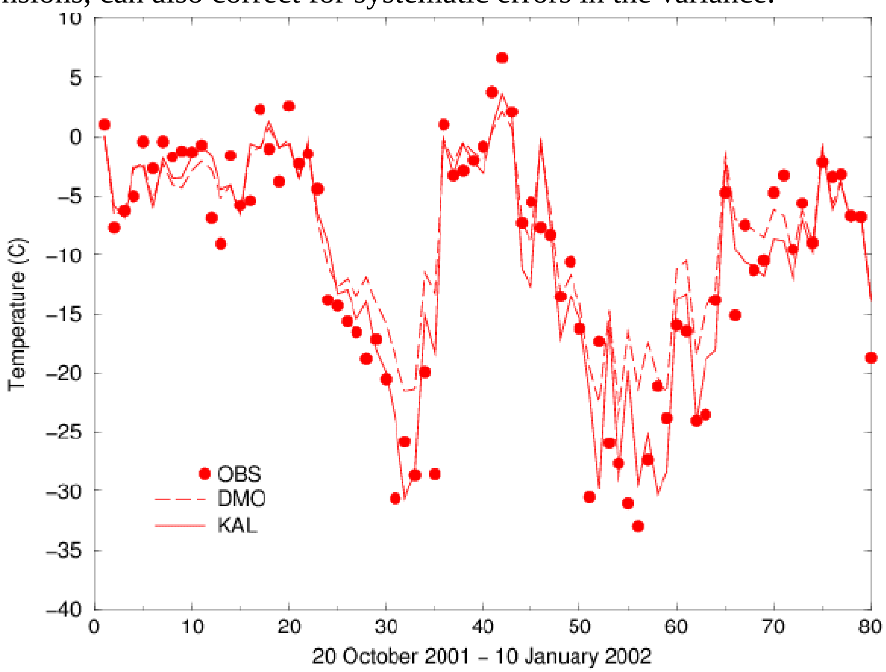


Figure 43 : The statistical interpretation has modified the NWP output mainly for the cold regimes where the systematic error was large.

This is not possible in simple bias correcting schemes when the correction is independent of the regime.

10.2.2. What looks bad might be good...

The statistical interpretation really managed to improve on the forecasts of extreme temperatures. While the NWP hardly went below -20˚ C the statistical interpretation managed to reach -30˚C.

Developing statistical systems will introduce the same paradoxical problem as his colleagues on the NWP modelling side: what looks like a definite improvement of the forecast system does not show up as clearly in the standard scores.

However, the RMSE scores actually indicated that the statistical scheme, by increasing the variance, even to the observed level, had made the forecasts *worse* and increased the RMSE to 5.1˚K.

With increased variance the forecasts are more likely to be *very wrong* than when the variance was underestimated. In other words: do not try to forecast extremes unless you are very certain. (See discussion in ch. 7 and in more detail in Appendix B.

In a forecast system which only aims at producing categorical forecasts with the highest accuracy, this is a sensible attitude. However, it makes it very difficult to issue any *risk forecasts* of extreme events.

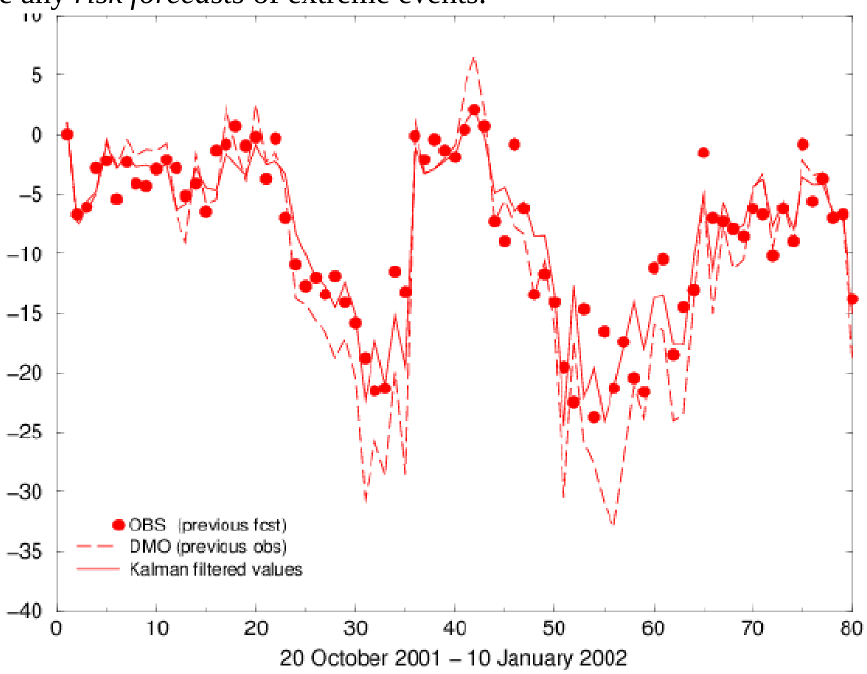


Figure 44 : In the case the forecast variance is larger than the observed the statistical interpretation can equally well correct for this.

The opposite happens when the NWP over-estimates the variability. The statistical schemes are not only able to correct for that, it will also show up clearly in the standard verification scores. The improvement appears to be as significant as in the previous example, but now the RMSE has fallen to 2.9.

10.3. Statistical interpretation of the EPS output

The usefulness of the EPS output can be improved by different forms of statistical post-processing, both of local character and regional.

10.3.1. Local interpretation

Since the EPS products are the same as from a NWP model it can easily be applied to any statistical post–processing like MOS, PPM or Kalman filtering. The error correcting equations are calculated from the EPS control and applied for each EPS member, after which plumes, histograms or probability charts can be made.

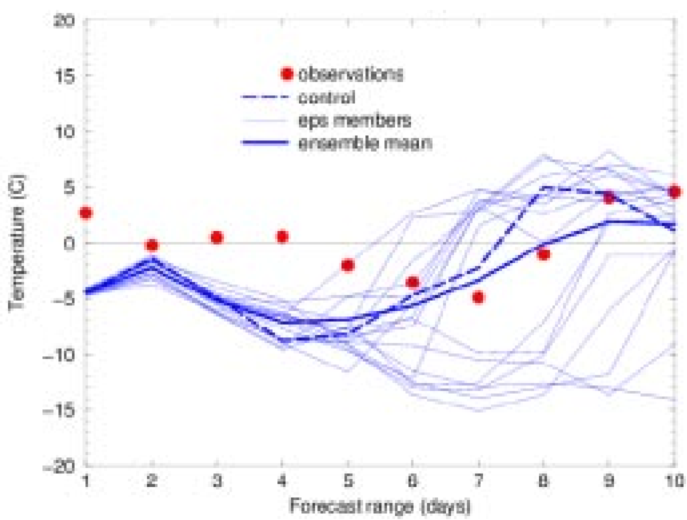


Figure 45 : The same as figure x, but now with the output from the EPS. The non-perturbed Control forecast from 1 February is supported by more than half of the members, indicating a return to milder conditions after about five days. A substantial minority support the forecast from the previous day, 31 January, that it will stay cold during most of the ten-day period. However, the verifying observations show that the whole forecast is rather poor. The spread during the last days was almost 25˚K.

Every forecast range is corrected with a specific equation, taking the systematic and non-systematic errors into account. Mean errors are normally eliminated, but due to the non-systematic errors the approach tend to reduce the variance of the post-processed products. This is avoided with an approach where the error correcting equation if calculated from the analysis or a short range forecast only (assuming the synoptic pattern is perfectly forecast).

Figure 46 : The relation between the +24 hour 2 metre temperature forecast and the corresponding error for Volkel during January 1996. The larger the circle, the more recent the verification. Although there is an over-all bias of about -3 K, there is also a linear relationship which can be described two-parameter error equation. In this case the error is approximately described by the equation -2 + 0.4

forecast temperature. For forecasts around zero the correction will be around +2 K, for forecasts around -10 the correction is more likely to be around -6 K

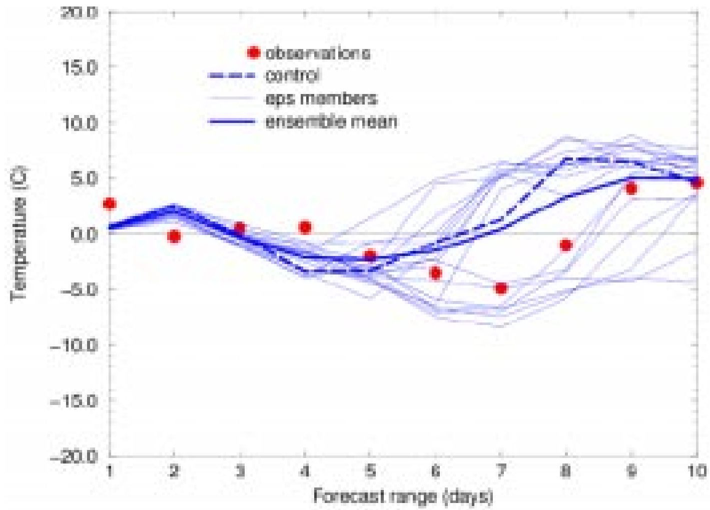


Figure 47 : The same as figure x2, but after statistically derived corrections have been applied to all the EPS members. Now the whole EPS plume was very realistic and that the substantial minority which indicated a prolonged cold period verified. Note that the statistical correction has reduced the spread of the plume to less than 15 K.

10.3.2. Calibrating probabilities

The probabilities themselves can be artificially improved statistically. The verifications tend to show that the forecasts are over-confident with low probabilities verify too frequently, high probabilities verify not often enough. A statistical correction can be imposed which will “upgrade” low probabilities and “downgrade” high probabilities. This kind of calibration tends to decrease the range of forecast probabilities so that very low or very high probabilities are never present.

10.3.3. Climatological clustering

An alternative form of clustering is to calculate a number of climatologically pre-defined flow patterns. The clustering algorithm then put every member in the typical cluster. It has the advantage of making it easy to synoptically assess the EPS members, but has the advantage to represent extreme or unusual flow patterns.

10.3.4. Clustering and statistical interpretation

With climatological clustering it is possible to provide statistical information about the probability of certain weather events, providing this pattern verifies. At a certain location, the probability for an event becomes the combination of the probability that the event will take place provided the flow type will occur, multiplied with the probability, derived from the EPS, that it will occur. This type of post-processing is in particular valuable in mountainous regions where there is low skill in the weather parameters.

10.4. The Extreme Forecast Index (EFI)

There is more work to do to de-mystify the mathematics behind the EFI. An Appendix D is suggested at the end.

10.4.1. What is “extreme”?

Extreme or anomalous events can be of mainly two types: large or medium scale persistent anomalies like cold outbreaks or heat waves lasting for more than a week. The EPS is well equipped to forecast this type. But they can equally well be in the small scale with heavy rain and/or strong winds. They might on such a small scale not even be possible to simulate by the model. T. With increasing resolutions the quality is steadily improving.

10.4.2. The idea behind EFI

What constitutes an extreme event depends on location or season. In winter-time 15 m/s in Brest or -5˚C in Berlin is not extreme, but the opposite, 15 m/s in Berlin or -5˚C in Brest, would be. To quantify this notion of extreme weather events an Extreme Forecast Index has been developed. It measures the difference between the probability distribution from the EPS and the model climate distribution.

The assumptions is that what is an “extreme” event in the models climate also should might an extreme event in the real atmosphere.

The underlaying assumptions is that if a forecast is extreme relative to the model climate, the real weather is also likely to be extreme vs the real climate.A common problem i statistics is to make sure that a given an empirical sample represents the total population. What we want to achieve is in a way the opposite: we want to know whether or not the EPS forecast is very different from the usual (climate) distribution of events. Rather than a yes/no answer we would like to have an index that gives continuous values. We would also want it to tell us if the deviation from climate is in a direction that may be dangerous for human activity. To have calm weather in Brest is indeed extreme, but it is unlikely to cause any problems.

10.4.3. The alarm bell map

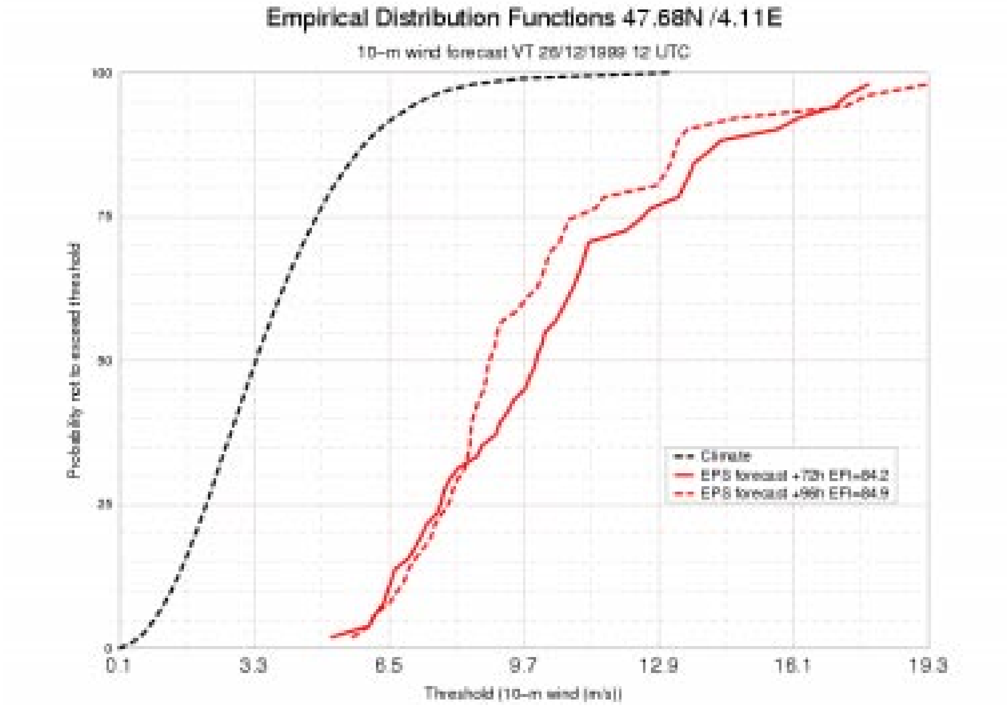
To facilitate *warnings of extreme weather* a new experimental product, an "Alarm bell" map has been developed. It maps the geographical distribution for every 12 hour period of the "Extreme Forecast Index" (EFI). This is an extension of the pure probabilities which do not take the climatological probability into account.

10.4.4. The problem of the reference climate

The choice of model climate instead of the real climate has been made to take proper account of the limitations of the model characteristics. Only meteorological systems which can be resolved by the model is taken into account. Excludes mesoscale heavy precipitation and wind.

The climate reference, which is the TL255 model climate has been evaluated using analyses D+5 and D+10 forecast from the period 1997-99. At this time the EPS was run at TL159. Tests have shown that the EFI is not too sensitive to changes in climate reference. Plans are under way to use the ERA-40 analyses and forecasts to build up a more realistic pseudo-climate.

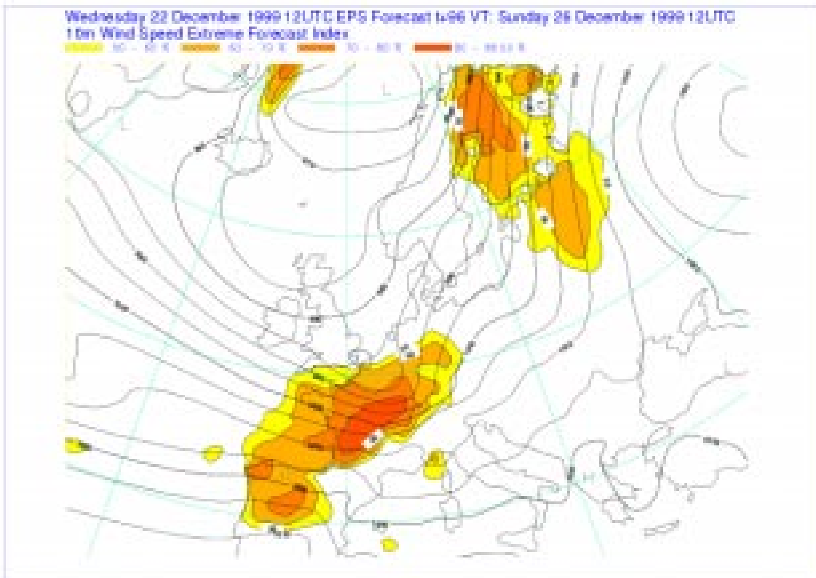
10.4.5. The interpretation of the EFI



The EFI takes values between -1 if all EPS members forecast values that are below the absolute minimum, +1 if they all forecast above the absolute maximum. EFI=0 if all members forecast the climatological mean.

USE: The primary thing to keep in mind that the extremes depicted in the EFI are with respect to location and season. A North European forecaster who in spring sees the EFI warn about extremely cold temperatures in the Mediterranean area should of course realize that the weather might not be extremely cold from his Scandinavian point of view.

The EFI cannot replace probabilities, just put them into perspective. The EFI is a parameter to alert the forecaster.



The probabilities in the EPS is therefore compared with the climatology. A shift in the predicted frequency distribution away from the climatological norm is used to provide a EFI. The reference climatology is the T_L255 model climate.

11. Forecasts beyond ten days (my favourite!)

11.1 Why seasonal and monthly forecasts?

In December 1994 the ECMWF Council approved a programme in atmosphere-ocean coupled modelling to address seasonal prediction. A group was assembled in July 1995 and seasonal prediction began in earnest in 1997. In March 2002 a programme for *experimental* monthly forecasts was started. At the time of writing it is not operationally available, but might be so in the near future.

11.1.1 Are forecast beyond ten days possible?

There is no doubt that forecasts of some skill beyond ten days would be beneficial for energy, agriculture, insurance, health and emergency relief. Many of these applications will be even more important outside Europe. But judging from the daily verification scores, where the ACC falls below 60% after about a week, it looks as if skilful deterministic forecasts beyond ten days are a long way off.

But, as we have seen in the previous chapter, averaging deterministic forecasts in time or space will tend to increase their skill and usefulness. We “buy” extra predictability and usefulness by putting less importance on the exact timing or position of a weather event. With the EPS technique we “buy” extra predictability and usefulness by quantifying the level of uncertainty.

With these techniques the limit of statistical skill and usefulness can be extended up to two weeks. But we may go even further by exploiting a forcing mechanism which we so far have not made full use of: the changes in the surface of the earth, in particular that part that covers two thirds of its surface - *the oceans*.

11.1.2 The physical forcing from the oceans

Medium-range weather forecasts are essentially an *atmospheric initial value problem*: the three-dimensional initial state of the atmosphere must be defined in greatest detail and accuracy. Since the state of the earth’s surface, in particular the sea surface temperature (SST) of the oceans, changes slowly, it is justified to have the SST simply persisted during a ten day integration.

However, with respect to the atmosphere, the seasonal forecast is an *oceanic boundary value problem*; it has been known for very long times that the SST is crucial for forecasting the weather on a seasonal basis. But during a full season, defined as three months, the SST, in particular in the tropics, may undergo significant changes, that will have a profound impact on the atmospheric circulation. That is why a forecast of the oceanic state must be an integrated part of the seasonal forecast system. And this oceanic forecast is essentially an *oceanic initial value problem*: the three-dimensional initial state of the ocean must, like the atmosphere, be defined as accurately as possible.

During the last decade a combination of scientific, observational and computational advances have provided motives why the long-term changes of SST and its global impact on the weather might be predicted with numerical methods months ahead.

11.1.3 Advances in scientific understanding

The most important cause of large scale changes in the weather is the irregular variations of equatorial Pacific SST. Warm El Nino and cold La Nina episodes represent opposite extremes of the El Nino Southern Oscillation (ENSO) cycle, which affects the SST, rainfall, vertical motion and air pressure over an area spanning more than half of the circumference of the earth. These tropical phenomena are also known to influence the weather conditions in the extra-tropics, in particular from the Pacific over North America to the North Atlantic. This “teleconnection” might explain why some areas in the extra-tropics are less difficult to predict than others.

But there are also other causes of seasonal climate variability. Unusual cold or warm SST in the tropical Atlantic or Indian Ocean can cause major shifts in the seasonal climate in nearby continents, like the rainfall in northeastern Brazil and tropical east Africa.

Like the ocean SST, the extent of snow covered ground and the amount of soil moisture varies over months and seasons. It is their “long memory”, which opens up possibilities to calculate their effect on the atmospheric circulation far beyond ten days.

Another source of predictability is believed to be the Madden Julian Oscillation (MJO), a 30-60 day tropical oscillation which starts over the Indian Ocean and propagates slowly eastwards. Since there are some indications that these oscillations might to be due to an ocean-atmosphere interaction, the use of a coupled system may help to capture some aspects of this variability.

11.1.4 Computational and observational advances

Just as important as the scientific advances are the technical advantages in numerical modelling, computational resources and observational availability.

Increased computer power and the development of numerical models of the ocean circulation have provided the technical feasibility of extended weather forecasts. Naturally, at this early stage, the ocean models still require extensive further refinement, but they are already good enough to use in a forecast system where the atmosphere and ocean are coupled. Further increases in computer power will improve their spatial and temporal resolution and dynamic-physical realism.

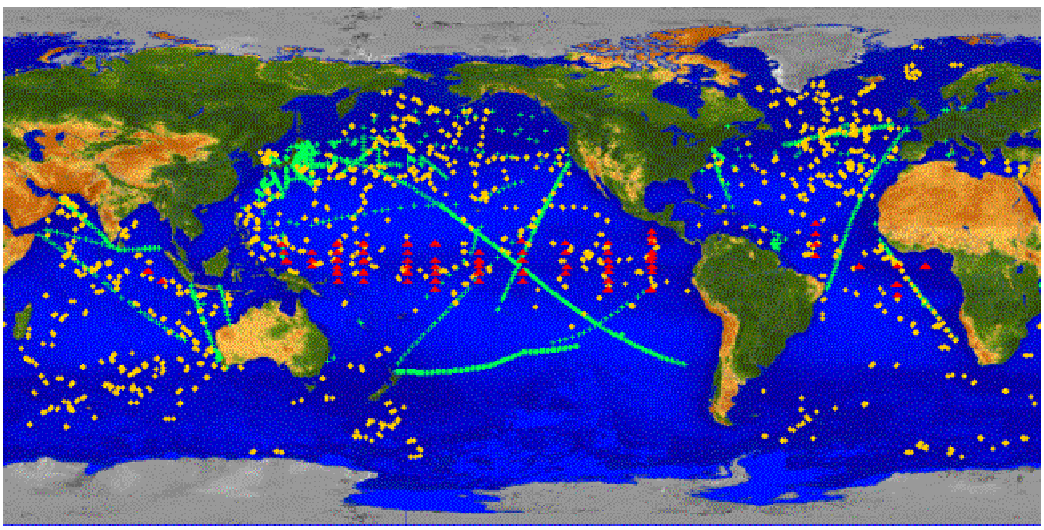


Figure 48 : The data coverage for May 2003. Red dots are moored buoys, yellow are ARGO floats, green XBT (expandable bathy thermographs). During the last years there has been a dramatic increase of real-time oceanographic observations.

Also of importance is the gradual build up of a fairly comprehensive *in situ* global ocean observing system, which provides temperature observations down to 500 m depth. European, US and Japanese satellites provide, apart from the upper-air meteorological observations, also SST data and observations of the wind surface stresses and sea surface heights.

This is the scientific and technical rational for the ECMWF endeavours to explore the possibilities for forecasting out to six months deviations from the climatological normal conditions. As an intermediate step between the medium-range and the seasonal forecasts, also 32 day integrations are made to explore the possibility of *monthly forecasts*.

11.2 The ocean forecast system

Since the ocean forecast system constitutes the core of the monthly and seasonal forecast systems, it is natural to start with a brief description of its characteristics. It is more or less the same for the monthly and seasonal forecasts, and consists of a dynamic ocean model and data assimilation system.

11.2.1 The ocean model

The ocean model is based on the same principles as its atmospheric counterpart: integrating the equations of motion forward in time. It also uses different kinds of parametrizations, although the role of moisture is replaced with salinity.

More specifically, the model is based on HOPE (Hamburg Ocean Primitive Equation model) version 2. It is global with 29 vertical levels. In the east-west direction the resolution is 1.4˚ (150 km), in the north-south it varies with latitude. In order to resolve ocean baroclinic waves which are trapped at the equator the resolution is relative high, 0.3˚ (33 km) in a band 10˚ on either side of the equator. Polewards of 10˚, the resolution smoothly decreases to 1.4˚ (150 km) at 30˚ lat. Adding to the dynamics, a parametrization scheme for sub-grid vertical and horizontal mixing is in operation.

The ocean model can reproduce the general features of the circulation and the thermal structure of the upper layers of the ocean and its seasonal and interannual variations. It has, however, systematic errors, some of which are caused by the coarse resolution: the model thermocline is too diffuse, the Gulf Stream does not separate at the right location. Other shortcomings, like a too shallow Pacific thermocline and a generally too warm SST along the Pacific side of the south American coast, might depend on both resolution and forcing errors.

11.2.2 The sub-surface ocean analysis

Like the ocean dynamical model, the ocean data assimilation bears resemblance to its atmospheric counterpart.

Between the ocean surface and 400 m depth a conventional OI analysis is performed on horizontally overlapping sub-domains of the model. Observational input comes from observations all over the globe, but mostly from the tropical Pacific, the tropical Atlantic and, to an increasing degree, from the Indian Ocean.

The observations are not only provided by stationary (moored) buoys, but also from so called Expendable Bathythermographs (XBT). More recently the observing system has been expanded by drifting ARGO floats and the extension of TAO-type moorings into the tropical Atlantic (PIRATA) and the west Pacific (TRITON). Climatology is used for both salinity and fresh water influx from river runoff.

The correlation length scale near the equator is 1000 km in E-W direction, 150 km in the N-S direction, polewards of 15˚ 300 km in all directions.

The OI analysis is made every ten days, with a window five days on either side of the model “first guess”. The increments, i.e. the effects of the observations, are gradually added in small portions, to allow the model dynamics to adjust gradually to changes in the density field and thereby avoid gravity waves to form.

11.2.3 No data assimilation below 400 m

Below 400 m there is no data assimilation at all. In an early version of the system the initial conditions were purely derived from the integration of the dynamical model, just as atmospheric analyses over data sparse areas to a large extent are based on forecasts: the information from data-rich areas can be propagated by the model into data-sparse regions. However, in the case of the ocean model, the propagation was unrealistically slow and allowed two “separate” ocean layers to develop with dynamical and thermal instabilities forming in the transition zone. Presently the information from above 400 m is “propagated” by statistical vertical influence functions, similar to those in atmospheric data assimilation.

11.2.4 The ocean surface analysis

There is no real temperature assimilation at the ocean surface. Instead SST generated by the model dynamics is relaxed towards the NCEP SST fields.

There are, as a matter of fact, two categories of NCEP SST data: one arriving daily and one weekly (centred on Wednesdays). The weekly is an average over the preceding seven day period Monday-Monday, and is of higher quality than the daily data set. On balance, the high-quality weekly SST data has been found to be the most suitable for monthly and seasonal forecasting. The time delay, however, causes substantial practical problems, which are solved differently for the monthly and seasonal forecast systems.

Seasonal forecasts are run on an initial state from the 1st in every month. The available weekly SST analysis is then normally out of date by several days. The actual production of the seasonal forecast is therefore delayed until the *next* weekly

SST analysis becomes available. From these two analyses, one valid before and one valid after the 1st, an interpolated SST analysis is computed. The seasonal forecast is then run with a delay of about 11 days, which is quite acceptable.

Monthly forecasts are run every 14 days. They can not sustain a delay of almost two weeks waiting for a more recent SST update, this has to be created by the analysis system itself. The anomalies of the last high quality weekly SST analysis is brought forward to the initial time.To provide initial conditions for a specific monthly forecast, the ocean model is integrated forward up to the required date (at most 11 days), forced by the analysed wind stress, heat flux and net precipitation (precipitation minus evaporation). Here the relaxation to SST is less strong so the final analysis to some degree is also affected by the model generated values.

11.2.5 Generation of ensembles of analyses

During the assimilation the ocean model is forced by analyzed meteorological variables. These are of course not known perfectly, in particular not in data sparse ocean areas. To account for this *five* alternative ocean analyses are driven by five slightly different meteorological fields (mainly based on statistical estimates). Due to the relaxation to the NCEP fields the five SST analyses are very similar.

To account for the uncertainties in the SST initial conditions a large number of additional SST perturbations are calculated (using statistics from historical data sets). For the *seasonal* forecasts 8 SST perturbations are added to each of the five analyses, yielding a total of 40 different SST analyses. For the *monthly* forecasts, to each of the five analyses 10 SST perturbations are added, yielding a total of 50 different SST analyses. All perturbations are linearly interpolated down to 40 m

11.2.6 Quality of the ocean analysis

Although the ocean data assimilation still has many limitations, it tends to correct some systematic biases such as the equatorial thermocline being too diffuse. Comparisons with independent observations shows that the analysis follows closely the observed variations in the Pacific ocean.

11.3 The atmospheric forecast system

Both the seasonal and monthly forecasts are run in ensemble mode on downscaled versions of the ECMWF operational forecast model.

11.3.1 The atmospheric forecast models

For the *seasonal forecast model* an older version of the ECMWF operational model from early 2001 (23r4) is used. It has 40 levels up to 10 hPa and a horizontal resolution of T_L95 which corresponds to 1.875˚ in N-S direction (200 km). The time step is set to 60 minutes.

For the *monthly forecast model* the current version of the operational model is used (at the time of writing 26r1). It has a spectral resolution T_L159, corresponding to 1.125˚ in the N-S direction (135). There are 40 levels in the vertical up to 10 hPa and a 60 minutes time step.

The *ocean-atmosphere coupling* is achieved by a two-way communication: the atmosphere affects the ocean through its wind, heat and net precipitation (precipitation-evaporation), whilst the ocean affects the atmosphere through SST. The frictional effect due to the ocean waves is accomplished in the same way as with the operational ten-day forecast system through the ocean wave model.

For the seasonal forecasts the interaction is once a day, while the monthly forecast is every hour. This high frequency coupling may have some impact on the development of some synoptic scale systems, such as tropical cyclones.

11.3.2 The ensemble forecasts

The integration of the ensemble members in monthly forecasts follow closely the operational EPS described earlier. Atmospheric perturbations are computed using the singular vector method. These include perturbations in the extra-tropics as well as perturbations in some tropical areas by targeting tropical cyclones. This is not the case for the seasonal forecasts where, as mentioned above, only the ocean initial conditions are perturbed.

However, just as in the EPS, for both models the tendencies in the atmospheric physics are randomly perturbed during the integration in order to take into account the uncertainties in the model formulation (stochastic physics).

11.4 The problem with model errors

After about ten days into the forecast, the model output from the monthly and seasonal forecast models starts to show signs of “drift”, i.e. displaying systematic model errors. The ECMWF has chosen to not introduce any “artificial” terms to remove or reduce any imbalances in the equations to try to reduce the drift. No steps are taken during the integration. Rather *a posteriori* corrections are made.

11.4.1 Corrections for the monthly forecasts

For the monthly forecasts the effect of the model drift is estimated from previous integrations of the same model from previous years. A five member ensemble is integrated from the same day and month as the real time forecasts over a long period (1990-2001). This results in a 60 member ensemble from which systematic errors can be calculated. With these “back statistics” it is possible to remove a large part of the drift during the post-processing. This is done every two weeks, in connection with the real-time forecast.

11.4.2. Correction for the seasonal forecasts

For seasonal forecasts the effect of the model drift is estimated from previous integrations of the same model from previous years. A five member ensemble is integrated from the same day and month as the real time forecasts over a long period (1987-2001). This results in a 75 member ensemble from which systematic errors can be calculated. With these “back statistics” it is possible to remove a large part of the drift during the post-processing. However, for the seasonal forecast the limited number of past forecasts in combination with low predictability makes validation a difficult task, especially in mid-latitude.

The forecast seasonal anomalies and the model drift are of the same magnitude, which is often less than 1˚ C but can reach 4˚ C in some parts of the oceans.

The forecasts of El Nino indices (see below) are bias corrected and given as anomalies with respect to long-time climatological data. For all other variables only forecast anomalies relative to the model climate for the 1987-2001 period are considered, assuming that they correspond to similar anomalies in the atmosphere. Corrections for shortcomings in the model variance are not made, nor are the derived probabilities corrected.

11.5 Monthly forecast products

The monthly forecast is run every 14 days. The output is available from the MARS archive, the web MARS, and as ready-to-use graphical products on the ECMWF web site.

11.5.1 Monthly forecast products in MARS

Upper-air fields are archived every 12 hours, whereas surface fields, including wave and ocean forecasts, are archived every 3, 6,12 or 24 hours depending on the parameter. Weekly means, weekly maximum and minimum, and standard deviations are also calculated and archived. The weeks are defined in the following weekly intervals: day 5-11, day 12-18, day 19-25 and day 26-32.

Ensemble means and standard deviations are calculated and archived for a limited number of fields: temperature at 850 and 500 hPa, and geopotential at 1000 and 500 hPa. These fields are also available as monthly means and anomalies.

In order to avoid retrieving 51 members to create EPSgrams, several fields have been reordered and the minimum, 25%, median, 75% and maximum of the ensemble distribution have been archived. The fields for which this EPSgrams is made are temperature at 850 hPa, total cloud cover, 2-metre temperature, total precipitation and 10 metre wind speed.

11.5.2 Monthly forecast products available on the ECMWF web site

Ensemble mean anomaly maps are computed for surface and 2-metre temperature, total precipitation and MSLP. They are averaged over the weekly periods defined above and the plots display the difference between the ensemble-mean of the real-time forecast and the ensemble-mean of the "back statistics" (the forecast shift from the estimated model climatological mean). In addition a test has been applied to estimate whether this shift is statistically significant.

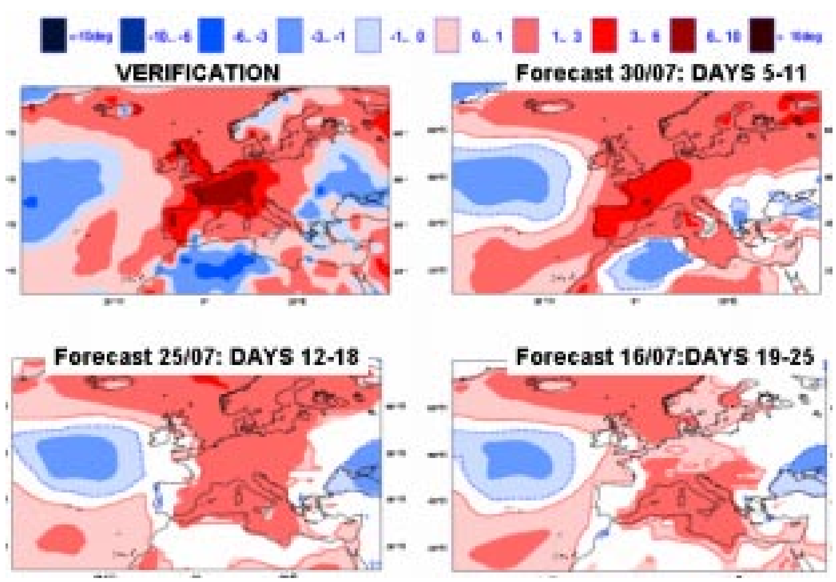


Figure 49 : The extreme heat wave that affected large parts of western and southern Europe in August 2002 was well indicated by the monthly forecast system. The figure shows the ensemble mean of the anomalies (relative to the model climate) for three monthly forecasts initialized on 16, 25 and 30 July valid 3-9 August.

Probability maps display the probability that the verification value will be above normal (with respect to the model climate). Only areas where the statistical test shows that the anomaly is significant are shaded.

Tercile maps are maps displaying the probability that the verification will fall into any of three equally probable (33%) intervals: below normal (“lower tercile”), normal and above normal (“upper tercile”). If there is no signal, the weather will be “normal”, and the tercile maps will display probabilities close to 33%. Therefore, the probability range 20-40% has been blanked out, in order to highlight areas of significant information.

Plumes display for several European cities, the evolution of the ensemble forecast of geopotential at 500 hPa, 12-hour accumulated precipitation and 850 hPa temperature.

Instead of showing the time evolution of each ensemble member, which may look noisy after about ten days, in particular for precipitation, the time series are also displayed in the form of quartiles. The plots display the time evolution of the ensemble mean (red), 25-75% domain (cyan) and extremes (dark blue area), the same colour code as the EPS plumes. The precipitation plots display the time evolution of accumulated precipitation, instead of 12-hour accumulated values.

“Stamp maps”, similar to the EPS “stamp maps”, display all the 51 member forecasts for every five days of the forecast, as well the weekly means. The levels are 500 hPa and MSLP (with the -6˚ and 16˚ isotherms at 850 hPa added).

Flow maps depict the anomaly of the 500 hPa ensemble mean over the Northern Hemisphere for each separate weeks.

Cluster maps cover seven day intervals starting at day 5. They indicate the number of members falling into six pre-defined and climatologically equally probable 500 hPa flow patterns (based on re-analysis data). The maps display the average of these members and the variability in the population. Note that this clustering is based on other principles than the EPS clustering.

Hovmöller diagrams (or trough-ridge diagrams) provide a very handy summary of the large scale evolution in the monthly forecast. The x-axis display the longitude and the y-axis the time evolution of a north-south averaged scalar, in this case the forecast 500 hPa geopotential *anomaly* between 35˚ and 55˚ N.

11.6 Skill of the monthly forecast system

Preliminary verifications carried out on the first 30 forecasts (March 2002 to May 2003) point to skill up to about two, perhaps two and half weeks. The last 19-32 days integrations may display some signal for high threshold events, like the probability that the 2-meter temperature will exceed 2K.

11.7 Seasonal forecast products

The seasonal forecast is run with the 1st of every month as the initial date. The output is available from the MARS archive and as ready-to-use graphical products on the ECMWF web site depicting monthly averages of temperature and rainfall, both as anomalies and probabilities.

11.7.1 Seasonal forecast products in MARS

A wide range of products is available in MARS: ocean, atmospheric and wave analysis and forecast data. Upper air fields of temperature, specific humidity, geopotential and wind are available for each member for every 12 hour time step at the pressure levels 1000, 925, 850, 700, 500 and 200 hPa. A large selection of atmospheric variables and wave forecasts are archived at intervals of 6, 12 or 24 hours. Monthly mean, maximum, minimum and standard deviation of fields are also available. Widely used products are the SST anomalies in different parts of the tropical Pacific. In particular in the NINO3 area 5N-5S, 90-150W the value of SST is often used as an indicator of El Nino activity.

11.7.2 Seasonal forecast products on the ECMWF web site

Ensemble mean anomaly maps global, as well as regional, are computed for surface and 2-metre temperature, total precipitation and MSLP. The maps are labelled with the period for which they are valid, e.g. DJF02 for the three month period December 2001 - February 2002. As with the monthly maps, areas where the forecast anomaly is significantly shifted from the normal are highlighted. The maps are based on daily means. Only areas where the statistical test shows that the anomaly is significant are shaded.

Probability maps display the probability that the verification value will be above normal, with respect to the model climate. As with the monthly forecasts, probabilities away from 50% are highlighted, with the maps only displaying probabilities above 60% or below 40%.

Tercile maps, similar to the monthly forecast tercile maps, display the probability that the verification will fall into any of three equally probable (33%) intervals. The probability range 20-40% (“around normal”) has been blanked out, in order to highlight areas of significant information.

El Nino forecasts are presented as plumes of the SST anomaly. The plots show the forecast values of monthly-mean anomalies for individual ensemble members, together with a verifying analysis where available. A typical six month forecast has six forecast values on its plotted trajectory, additional to the observed value at the first point. A separate plot is made for each calendar month, showing the trajectories of all forecasts with initial dates in that month.

Since the probabilities are calculated on only 40 members, there is some sampling uncertainty. The TL95 model is also affected by systematic errors which lead to systematic errors and underestimation of variance.

11.8 Skill of the seasonal forecast

As with calibration (see 10.4 above) verification of the seasonal forecasts is a difficult task due to a combination of low predictability and limited number of past forecasts.With only 15 years of forecasts (run in real time and in back-log from analyses 1987-2001), each combination of location, season and lead time only has 15 independent verifications.

To illustrate the difficulty, assume that the system were totally lacking skill, i.e. the correlation between forecast and observed anomalies is zero. Still a global map of anomaly correlation would have 5% of its area yielding correlations of 44% or more.

In order to reduce the number of areas where skill might arise by chance, studies have been made at specific El Nino teleconnections. El Nino is known to be the largest single source of predictable internannual variability. Verifications show indeed that during El Nino (and La Nina) the skill of the precipitation forecasts in regions and seasons known to have a teleconnection with the El Nino is much higher than during neutral conditions.

Recent studies have shown that the seasonal forecast system is superior to statistical systems in forecasting the onset of El Nino or La Nina. But once an event has started statistical systems have comparable skill. The dynamical model is also better than the statistical models in forecasting Atlantic and Indian Oceans SST, but is hampered by model errors.

In many parts of the tropics, where changes such as those associated with El Nino can have a large impact on global weather patterns, a substantial part of the year-to-year variation in seasonal-mean rainfall and temperature is predictable. In mid-latitudes, the level of predictability is lower, and Europe in particular is a difficult area to predict. Information about the seasonal forecast reliability based on its past performance is available at http://www.ecmwf.int/products/forecasts/d/charts/seasonal/verification.

11.9 Recommendations to users of the monthly and seasonal forecasts

As with all output from a numerical weather prediction system simplistic use of the direct model output is not recommended. Actual forecasts for end-users should be carefully prepared, perhaps combining data from several empirical and/or numerical sources. It is good practice to compare the forecast charts for a given target period at different lead times as they become available. Simply trying to read off local values from the latest available maps could be very misleading.

The seasonal and monthly models are global, and can only hope to represent the large scale of weather patterns. Local weather and climate can be much influenced by features (hills, coastlines, land surface) too small to be included in the relatively low resolution model. Local knowledge and expertise, if possible helped by statistical interpretation schemes, will be important in assessing model output, and translating it into realistic statements about local and regional prospects.

11.10 Prospects for the future

Using numerical models of the ocean and atmosphere to calculate seasonal forecasts is a challenging problem. The point at which seasonal forecasts become good enough to be useful to a particular user will depend very much on his requirements. From a practical perspective, today’s system might already be of use for some application, but not for others.

A mature seasonal forecasting capability will take many years to develop. While an operational schedule may be maintained, the science of seasonal forecasting is still under development. From a scientific perspective today’s system is still experimental.

Epilogue

The ECMWF was set up in 1975 with the aim of providing 10 day forecasts of economic value for the European area. The first target was to provide 5 day forecasts which had the same skill as 2 day forecasts before the “computer age”. This has been achieved and the deterministic forecast now have a skill up to 8 days. The skill varies considerably with sometimes useful forecasts up to 10 days, sometimes hardly beyond 4days. The EPS provides a measure of the accuracy of an ensemble mean and probabilities of possible alternatives and extremes.

Used in this way the forecasts, either as an ensemble mean or in a probabilistic sense, already has useful skill up to day 10. The continued work at the ECMWF is to develop this skill further. The resolution of the deterministic model will increase to T_L799, the ensemble system will be run on T_L399. The use of satellite data will increase in quantity and quality. The consequences will not only be a continued increase in skill of large scale weather systems, but also of small scale, in particular in the EPS.

This provides the meteorological services with an even more increased potential to serve a wide range of needs in the society, since in particular the EPS is suited for tailor suited forecast production. The challenge for today’s meteorologists is not only to make use of the current and future skill of the ECMWF forecasts, but also to develop new products and reach new sectors of society and satisfy new demands.

This will unavoidably involve an increased proportion of automatic or computer to computer generated products. The best experts to do this work are meteorologists with forecast experience and skill in computer based systems. The design, maintenance and upgrade of computer based post-processing software is already, and will increasingly become, an important task.

Forecasts generated in this way, perhaps with computer-to-computer access, will free the forecasters from some routine work and enable them to concentrate on situations where their personal intervention and interpretation is needed. To fulfil their task as presenters of information, the forecasters must not only be familiar with the way the atmosphere works, but also how the NWP models work and how their products can be interpreted and post-processed. Hopefully this User Guide has provided a useful basis for this with respect to the ECMWF forecast system.

12. References and further literature

12.1 ECMWF documentation and publications

12.1.1 Newsletter

A quarterly ECMWF Newsletter is distributed to national weather services in the Member States and users of the GTS products worldwide. It deals with topics in meteorology and the operational activities at the Centre and provides short descriptions of operational changes to the analysis and forecasting system. The newsletter also deals with computing topics.

A collection of important articles related to the development of the ECMWF analysis and forecast system, together with relevant references, can be found in *ECMWF Data Services, 1999: ECMWF/WCRP Level III-A Global Atmospheric Data Archive, The description of the evolution of the ECMWF forecasting system and corresponding archive.*

12.1.2 Bulletins and memoranda

Comprehensive documentation of the analysis and forecasting system, the archiving and dissemination is given in the Meteorological Bulletins. The Computer Bulletins provide the guidance to the Centre's computing facilities. Scientific and technical aspects of the Centre's work are discussed in informal ECMWF Technical Memoranda. A limited distribution within the ECMWF Member States applies to these three types of documentation. Individual copies are available from the Centre's library on request.

12.1.3 Proceedings and reports

Proceedings from the Centre's annual seminar and workshops are distributed widely to the national weather services and scientific institutions of the meteorological community.

ECMWF publishes reviewed papers of results in its own series of Technical Reports, available in the libraries of most national weather services and scientific institutions.

12.1.4 Documentation

A documentation of the analysis and forecast model can be found in the ECMWF Research Manuals:

Data assimilation - scientific documentation (Meteorological Bulletin 1.5/1)
Forecast model - adiabatic part (Meteorological Bulletin 1.6/3)
Forecast model - physical parametrization (Meteorological Bulletin 1.6/2)

12.2 User Guide references

12.2.1 Analysis system

Introductory note: The ECMWF implementation of four dimensional variational assimilation is covered by three papers (Rabier et al, 2000; Mahfouf et al, 2000 and Klinker et al, 2000)

Andersson, E. and H. Järvinen, 1999: Variational quality control, Q. J. R. Meteorol. Soc.,1999, vol 125, pp. 697-722

Bouttier, F. and P. Courtier, 1998: Data assimilation concepts and methods. ECMWF training course lecture notes, Mar 98, 64pp.

Courtier, P., J-N. Thépaut and A. Hollingsworth, 1994: A strategy for implementation of 4DVAR using an incremental approach, QJRMS, vol, 120, pp.1367-88.

Bouttier F. and F. Rabier, 1998: The operational implementation of 4DVAR, ECMWF Newsletter Number 78, reprinted in ECMWF Data Services, pp.72-76.

Derber,J. and F.Bouttier, 1999; A reformulation of the background error covariance in the ECMWF global data assimilation system.Tellus 51A,195--221

Fisher, M. 1998: Development of a simplified Kalman filter. ECMWF Tech. Memo. 260.

Gérard, E, and R. Saunders, 1999: 4D-Var assimilation of SSM/I total column water vapour in the ECMWF model. Q. J. Roy. Meteor. Soc.vol. 125, pp.3077-3101.

............to be updated

Appendix A: 4DVAR - an elementary introduction

Assume that we have two reports, $+10^o$ and $+15^o$ C, of the temperature for a certain location. The first has an assumed error of 2 K, the second 3 K. The true value is therefore probably closer to $+10^o$ than to $+15^o$. A traditional statistical least-square technique weights together the two observations, with weights proportional to the “precision” or accuracy of the measurements defined as the inverse of the variances of the assumed errors. In our case these variances are 4 and 9, the weights become 9/(4+9) ≈0.7 and 4/(4+9)≈0.3,which yields an “analysed” value of

10 * 0.7 + 15 * 0.3 = 11.5 C.

Let us now assume that of the two values one is an observation (O), the other a background field value (F) with accuracies σ_O and σ_B respectively. In accordance with what stated above the weighting formula may then be written

$$A = O\frac{\sigma_B^2}{\sigma^2{}_B + \sigma^2{}_O} + F\frac{\sigma_O^2}{\sigma^2{}_B + \sigma^2{}_O}$$

Most meteorological objective analysis techniques are further developments of this simple least-square approach. They have, however, certain weaknesses. So for example, the method is local: only observations within a limited area could be considered at each time to influence a given grid point. They could not cope well with non-conventional data. That is why a variational technique was introduced.

This technique starts by introducing a so called “cost function” J(S) which measures the sum of the squares of the distances (or misfit) of different atmospheric states S to the observation O and the background F.

$$J(S) = \frac{1}{2}\left[\frac{(F-S)^2}{\sigma_B^2} + \frac{(O-S)^2}{\sigma_O^2}\right]$$

The observation O is known, and we are looking for a value of A that will make the cost function J(S) as small as possible. We do this by differentiating J with respect to S. This will result in the same value of A as for the least square approach above, which shows the internal consistency of the two approaches.

One may therefore ask the motivation for defining this cost function? The reason is that it opens up a door to a mathematical formalism which will provide a powerful tool, in particular when the scalars in the right hand equation are replaced by vector and matrices.

Let us start by trivially re-arranging the equation above

$$J(S) = \frac{1}{2}[(F-S)\frac{1}{\sigma_B^2}(F-S) + (O-S)\frac{1}{\sigma_O^2}(O-S)]$$

We now introduce an operator H which can be anything from a simple interpolation of the analysis to the exact position of an observation, to a complicated equation that converts temperature and moisture information into radiance values. Instead of using the difference between analysis and observation, we use the difference between the converted value H(S) and the observation Y:

$$J(S) = \frac{1}{2}[(F-S)\frac{1}{\sigma_F^2}(F-S) + (Y-H(S))\frac{1}{\sigma_O^2}(Y-H(S))]$$

The next step is to consider not one, not many, but *all* observations made simultaneously over the globe. The scalars A, F and Y now become vectors of enormous dimensions since all the global observations enter as elements. So for example when there is “only” one million observations globally available at a certain time interval the vector **Y**=(obs_1, obs_2, obs_3,................obs_{999999}, $obs_{1000000}$).

The error $\sigma_O{}^2$ and $\sigma_B{}^2$ variances turn from scalars into the error covariance matrices **R** for the observations. To match the vectors they are formally of the size 1000000 x 1000000. The matrice.**B** for the background or first guess is approximately of the same dimension.

This is the basic equation for the 3DVAR assimilation system. In spite of the enormity of the numerical calculations, it still preserves the structure of our first formula above.

What characterizes 4DVAR is that *time* enters as an additional element. The aim is to find the A_0 in the beginning (t=0) of a 12-hour time interval, which minimizes J. The minimization will not only be dependent on the conditions at t=0, but to the whole dynamic evolution during the 12-hour interval.

So while the first term on the right hand side is kept alone, the second term will be evaluated for every single time step (t=1 -> N). The operator **H** is as before the operator which converts model states into observation states. But we now also introduce operator $\mathbf{M_n}$, which is the forecast model at time step n.

By integrating the atmosphere forward by $\mathbf{M_n}$ and “backward” by the so called adjoint in the 12-hour window, 4DVAR process ensures that the initial conditions at 09 and 21 UTC are worked out in a way that provides optimum fit to the observations throughout the 12 hour window.

So mathematically we are looking for an atmospheric state S_0=A at time n=0 which minimizes the cost function J(S).

$$J(s) = \frac{1}{2}[(\mathbf{F_0} - \mathbf{S_0})^T \mathbf{B_0^{-1}} (\mathbf{F_0} - \mathbf{S_0}) +$$

$$\sum_{n=1}^{N} ((\mathbf{Y}_n - \mathbf{HM}_n(\mathbf{S}_n))^T) \mathbf{R}_n^{-1} (\mathbf{Y}_n - \mathbf{HM}_n(\mathbf{S}_n))$$

The first term: the difference between the first guess and the initial state determines only partly the size of J.

The second term: the Σ–term, sums up all the differences between the evolving forecast and the *n* number of observations of varying kind.

Appendix B: The mathematics of forecast errors and “jumpiness”

The Root Mean Square Error (RMSE), here denoted **E**, is the root mean square difference between forecast (f) and the verifying analysis (a)

$$E = \sqrt{(\overline{(f-a)^2})}$$

Assuming there is no bias in the forecasts, the square of RMSE, the MSE, can be written

$$E^2 = \overline{(f-a)^2} = \overline{(f+c-c-a)^2} = \overline{(f-c)^2} + \overline{(a-c)^2} - 2\overline{(f-c)(a-c)}$$

where the overbar denotes averages in space and time, i.e. over a large number of forecasts. This can be re-written:

$$E^2 = A_f^2 + A_a^2 - 2cov((f-c)(a-c))$$

Of the three terms on the r.h.s. the first two the variance around climate of the forecast (A_f^2) and analysis (A_a^2). The former depends on the realism of the atmospheric model, the latter by the observed characteristics of the atmospheric flow. The former can be affected by human intervention, not the latter. Both vary strongly with seasons with maximum in winter and minimum in summer.

For $A_f^2 = A_a^2$ the model’s synoptic-dynamic activity is at the same level as in the real atmosphere, f.or $A_f^2 < A_a^2$ it is less than in the real atmosphere, which will contribute to lower the MSE. For the, unusual, case when $A_f^2 > A_a^2$ the model’s excessive synoptic-dynamic activity will contribute to increase the MSE[1]. Comparing verifications between different models or different versions of the same model, is only possible when the general variability of the atmosphere has been on the same level.

The third term measures the covariance between forecast and observed anomalies, and represents, in some sense, the “skill” of the forecasts. It can be shown that the cosine of the angle between the vectors corresponds to the ACC.

1. Since the general level of RMSE depends on the range of atmospheric variability, the RMSE displays strong seasonal variations since the winter anomalies normally are larger than the summer ones. Changes in the RMSE level, from one season or year to another, might therefore not necessarily be due to changes in the model characteristics, but to the nature of the atmospheric flow.

.

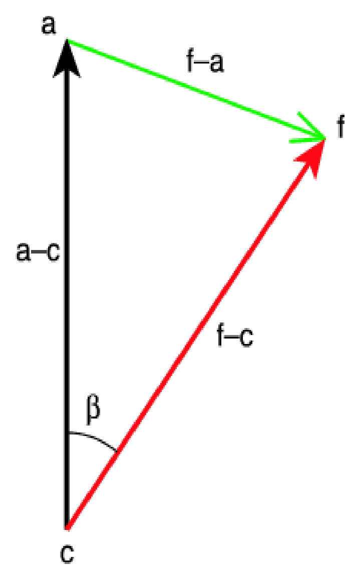


Figure 50 : The same analysis can be conducted in graphical form in a phase–space, using vector algebra where the observed anomaly **a-c** and the forecast anomaly **f-c** are represented by vectors of length $\mathbf{A_f}$ and $\mathbf{A_a}$ separated by an angle **ß**. It can be shown that the ACC=cos(**ß**) and that the distance **a–f** is proportional to RMSE.)

For forecast ranges with no predictive skill, this last term is zero. For the normal case of A_f=A_a this yields an upper limit of the average RMSE values, the so called “error saturation level”, $E_{saturation}= A_a \sqrt{2}$

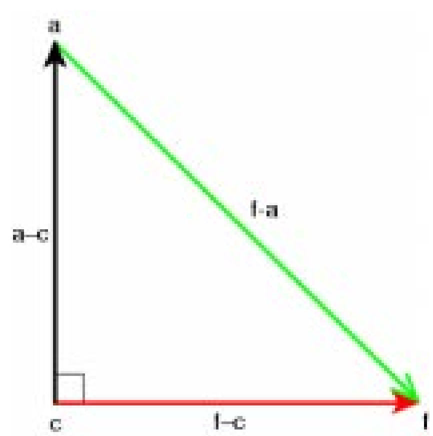


Figure 51 : For the case when β=90˚the forecast and analysis are uncorrelated and ACC=cosβ

This formalism also allows us to define a consistency-skill relationship as the covariance between (f-a), the forecast error of model and the forecast difference (f-g). Applying the same expansion around the climate above we have:

$$C = \overline{(f-a)(f-g)} = \overline{(f-c)}^2 - \overline{(f-c)(g-c)} - \overline{(f-c)(a-c)} + \overline{(g-c)(a-c)}$$

For completely unskilled and uncorrelated forecasts $C=(f-c)^2 = A_f^2$ which corresponds to a correlation of 50%. This can be shown in a geometrical form:

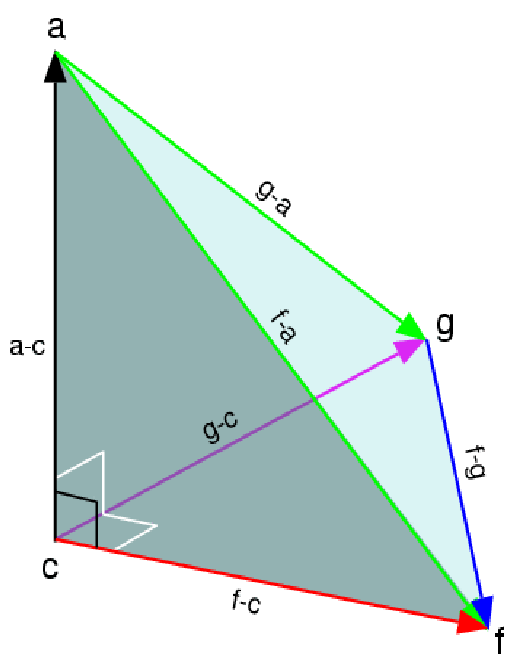


Figure 52 : The case with two realistic, but skill-less forecast systems (f) and (g), can be illustrated by an extension to the concept in figure 1, as a three-dimensional vector diagram. The “consistency” or “jumpiness” is represented by the vector (f-g). By “watching” the geometrical figure from the upper right (facing down to the left) yields the image in figure 4 below

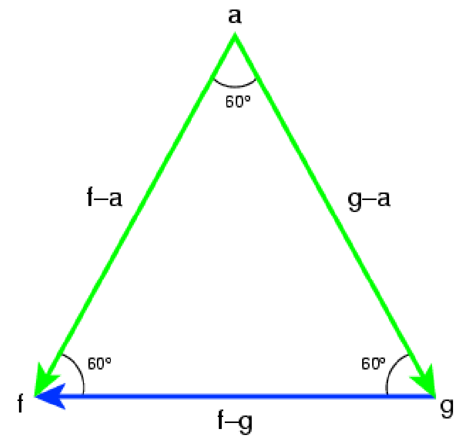


Figure 53 : Figure 3 "watched" from the upper right displays the three vectors (f), (g) and (f-g) and their mutual correlations, which is the cosine of their intermediate angles 60°, which yields corrections of 0.5.

When the skill of the forecasts increases, for example because shorter forecast rangers are considered, (f) and (g) will become more correlated and the term -(f-c)(g-c) will reduce the correlation between consistency and skill, C.

If (f) is a later forecast run than (g) it is on average more skilful and the sum of the two last terms -(f-c)(a-c)+(g-c)(a-c) is negative and will further decrease C. Gradually the consistency/skill correlation will drop from 50% to 20-30% which are typical values for the consistency-skill correlation around day 5 and 6.

The same formalism and graphics also make us understand why forecast errors from different models tend to look similar. The expression for the covariance of the forecast errors (f-a) and (g-a) for two different forecast systems (f) and (g):

$$\overline{(f-a)(g-a)} = \overline{(a-c)}^2 + \overline{(f-c)(g-c)} - \overline{(f-c)(a-c)} - \overline{(g-c)(a-c)}$$

shows that when the forecast systems lack any skill and are mutually correlated

$$\overline{(f\text{-}a)\,(g\text{-}a)} = \overline{(a\text{-}c)}^2 = A_a^{\,2}$$

with a corresponding correlation of 50%.

Appendix C: The Brier Scorer (more work is needed to convey a feeling for the decomposition of the Brier score)

The most common verification method for probabilistic forecasts, the Brier score BS is similar to the RMSE, measuring the difference between a forecast probability of an event p and its occurrence o, expressed as 0 or 1 depending on if the event has occurred or not. As with RMSE, the lower the Brier score the "better"

$$BS = \overline{(p-o)^2}$$

A Brier Skill Score (BSS) is conventionally defined as the relative probability score compared with the probability score of a reference forecast

$$BSS = (BS_{ref} - BS)/(BS_{ref})$$

The BS score can be decomposed in a similar way as the RMSE, yielding three terms which help to explain different aspects of the scoring system

$$BS = \overline{(p_k - c)^2} + \overline{(c_k - c)^2} + (1-c) \cdot c$$

where a sample of N forecasts has been divided into categories (k=1,2,3...T) each comprising n_k forecasts of average probability p_k. The observed frequency in each frequency is c_k and c the observed frequency of the whole sample. The first term expresses the degree of reliability, the second the resolution and the third the uncertainty of the forecasts (Atger, 1999).The decomposition of the Brier score (David S. Richardson, personal communication) for a set of N probability forecasts is given by

$$b = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2$$

where p_i is the forecast probability for case i, and o_i is the observed outcome (1 if event occurs; 0 otherwise). If the forecast probability is limited to a finite set of K possible values (p_1,p_2,p_3,....p_k), then b can alternatively be written as

$$b = \frac{1}{N}\sum_{k=1}^{N_k}\left(\sum (p_k - o_{kj})^2\right)$$

where M_k is the number of forecasts where the forecast probability is p_k, and is o_{kj} again 0 or 1 depending on the actual outcome for each case.

Now group together the cases in each category k depending on whether the event occurred or not

$$b = \frac{1}{N}\sum_{k=1}^{K}[M_k^{\ 1}(p_k-1)^2 + M_k^0(p_k-0)^2]$$

Here $M_k^1$is the number of cases where the forecast probability was p_k; and M_k^0 the event was observed; and is the number of cases where the forecast probability was and the event was not observed. ($M_k^1 + M_k^0 = M_k$).

To put everything in terms of frequencies, let
$g_k = M_k/N$

and
$o_k = M_k^1/M_k$

so o_k is the relative frequency of occurrence of the event given the forecast probability is p_k. Substituting these into Eq 3. gives

$$b = \sum_{k=1}^{K}[g_k o_k(p_k-1)^2 + g_k(1-o_k)p_k^{\ 2}]$$

which expands to

$$b = \sum_{k=1}^{K} g_k[p_k^2 - 2o_k p_k + o_k]$$

or

$$b = \sum_{k=1}^{K} g_k[(p_k-o_k)^2 + o_k(1-o_k)]$$

The overall (climatological or sample) observed frequency of the event is given by

$$\bar{o} = \frac{1}{N}\sum_{k=1}^{K} M_k^{\ 1} = \sum_{k=1}^{K} g_k o_k$$

So Eq. 8 becomes

$$b = \sum_{k=1}^{K} g_k (p_k - o_k)^2 + \bar{o} + \sum_{k=1}^{K} g_k o_k{}^2$$

Then rework the last term in Eq 10 to get

$$\sum_{k=1}^{K} g_k o_k{}^2 = \sum_{k=1}^{K} g_k [(o^2{}_k - 2o_k\bar{o} + \bar{o}^2)^2 + 2o_k\bar{o} + \bar{o}^2]$$

$$= \sum_{k=1}^{K} g_k (o_k - \bar{o}) + 2\bar{o} \sum_k g_k o_k - \bar{o}^2 \sum_k g_k$$

$$= \sum_{k=1}^{K} g_k (o_k - \bar{o})^2 + \bar{o}^2$$

Finally put this back into Eq. 10 to give the decomposition as

$$b = \sum_{k=1}^{K} g_k (p_k - o_k)^2 + \sum_{k=1}^{K} g_k (o_k - \bar{o})^2 - \bar{o}(1 - \bar{o})$$

The first term is the weighted (by forecast frequency) difference between forecast probability and observed relative frequency. This is a measure of how much the forecast probabilities can be taken at face value (reliability). On the reliability diagram this is the weighted sum of the distance (vertical or horizontal) between each point and the diagonal (from (0,0) to (1,1).

The second term is the weighted difference between observed relative frequency for each class k and the climatological frequency. This is the resolution. While reliability, like bias, can be compensated for, there is nothing that can be done about the resolution – in a sense this is the more important measure of the potential usefulness of the forecasts. On the reliability diagram this is the weighted sum of the distance (vertical or horizontal) between each point and the horizontal line

The last term is the variance of the observations (the uncertainty) and is purely dependent on the observations. The Brier skill score effectively removes this term and so gives a score that is more related to the forecast performance.

Appendix D: The mathematics of the Extreme Forecast Index (EFI)

The great challenge!