TECHNICAL MEMORANDUM

393

Effects of observation errors on the statistics for ensemble spread and reliability

Øyvind Saetra, Jean-Raymond Bidlot, Hans Hersbach and David Richardson

Research Department

December 2002

This paper has not been published and should be regarded as an Internal Report from ECMWF.
Permission to quote from it should be obtained from the ECMWF.

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen terme

For additional copies please contact

The Library
ECMWF
Shinfield Park
Reading
RG2 9AX
library@ecmwf.int

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, RG2 9AX, England

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error, omission and for loss or damage arising from its use.

Abstract

Effects of observation errors on rank histograms and reliability diagrams have been investigated, using a perfect model approach. The Lorenz model (Lorenz 1963) was used to simulate an idealised ensemble prediction system with 50 ensemble members and one control forecast. Observation errors at verification time were introduced by adding normally distributed noise to the final true state of the system. One of the major finding was that rank histograms are very sensitive to the presence of observation errors, leading to over-populated upper and lower ranks. Reliability diagrams were far less sensitive in this respect. The resulting u-shaped rank histograms can easily be misinterpreted as too low spread in the ensemble prediction system. To account for this effect when real observations are used to assess an ensemble prediction system, Gausian noise based on the verifying observation error can be added to the ensemble members before the statistics are calculated. The method has been tested for the ECMWF ensemble forecasts of ocean waves and forecasts of the geopotential at 500 hPa. The EPS waves were compared with buoy observations from the Global telecommunication System (GTS) for a period of almost three years. When the buoy observations were taken as the true value, more than 25 % of the observations appeared in the two extreme ranks for the day 3 forecast range. This number was reduced to less than 10 % when an estimate of the observation errors were added to the ensemble members. Ensemble forecasts of the 500 hPa geopotential are verified against the analysis. When observation errors are neglected, the maximum numbers of outliers are more than 10 % for most areas except for Europe, where the observation errors are relatively smaller. Introducing noise on the ensemble members, based on estimates of analysis errors, remarkably reduced the number of outliers. Particularly in the short range, where a peak around day 1 more or less vanished.

1 Introduction

Presently, ensemble forecasting is an important product delivered by a number of operational meteorological centers around the globe. By perturbing the initial state of a high-resolution deterministic forecast, a number of ensemble members is integrated at a usually lower resolution. From the spread in such ensembles, an estimate on the quality of the high-resolution deterministic forecast is made, and in the medium range, clusters of members are a guideline to the likelihood of alternative weather scenarios. Examples of meteorological centers that produce ensemble systems on an operational basis are the US National Center for Environmental Prediction (NCEP), the Canadian Meteorological Centre (CMC) and the European Centre for Medium-Range Weather Forecasts (ECMWF).

An important aspect of a forecasting system in general and for an ensemble system in specific, is the assessment of performance. Due to its probabilistic nature, verification is for ensemble systems more complex than it is for deterministic forecasts. Popular tools that measure the performance of probabilistic forecasts are rank histograms, reliability diagrams, Brier scores, ROC curves and cost-loss analyses (a detailed description may be found in Strauss and Lanzinger 1996, Wilks 1995 and Richardson 2000). For all of these tools, ensemble forecasts are compared to the verifying weather pattern. This truth is measured either directly by using observations, or by using verifying analysis fields. The errors in these quantities are usually neglected. Argument is that they are small compared to the errors in the forecasts, and therefore, will have a negligible effect on the results. However, in the short range, where forecast errors are still small, this argument may not be justified. For instance, at ECMWF, the accuracy of the one-day forecast in the geopotential height at 500 hPa is comparable to the observation error of radio sondes.

In this paper the impact of observation errors at verification time on verification tools will be studied. Focus will be on rank histograms and reliability diagrams, i.e., tools that measure the statistical aspects of ensemble forecasts. The proper way of doing this is to transform the probability density function (PDF) for the truth, to a PDF for the verifying observations or analysis. This means that the PDF created by an ensemble system is to be convoluted with the verifying observation error before it is to be presented to the verification tool under consideration. In practice, this can be accomplished by adding Gaussian noise to the ensemble members with a standard deviation given by these observation errors. It will be shown that in the short-range rank histograms are quite sensitive to observation errors, and that there is even a non-negligible effect in the medium range. The effect on reliability diagrams will be found to be limited.

The fact that models may have systematic errors or biases can both obscure, or make it difficult to interpret the results of verification. As a method for the testing of the pure impact of the sensitivity of observation errors on verification tools, a perfect model approach may be used. In such a study, the initial state of one of the ensemble members is defined as the truth and, because of the perfect model approach, so will be its trajectory. Observations may be created by perturbing the truth according to their uncertainties. This approach will be followed in Section 2 by using the Lorenz model (Lorenz 1963) as a simple version of a non-linear dynamical system. In Section 3 the effect of the incorporation of verifying observation errors will be considered for the ensemble prediction system (EPS) that is operational at ECMWF. Both the case of verification on the basis of observations (ocean wave heights are taken as an example) and on the basis of verifying analyses (geopotential at 500 hPa) is considered. The paper will end with some concluding remarks.

2 Idealised experiments with the Lorenz model

To produce a synthetic data set under idealised conditions, where both initial spread, uncertainties and observation errors could be manipulated, the non-linear dynamical system originally suggested by Lorenz (Lorenz 1963) has been used. Here, the dependent variables X, Y and Z are determined by the equations

$$\begin{aligned}\frac{\mathrm{d}X}{\mathrm{d}t} &= -\sigma X + \sigma Y \\ \frac{\partial Y}{\partial t} &= -XZ + rX - Y \\ \frac{\partial Z}{\partial t} &= XY - bZ,\end{aligned} \tag{1}$$

where σ, r and b are constants. To solve these equations numerically the double approximation procedure suggested by Lorenz was used. The constants were taken to be: $\sigma = 10$, r=28 and b=8/3, and the dimensionless time step $\Delta t = 0.001$. In order to generate a data set that could later be used as initial conditions, the model was run for a period of 100,000 time steps. The result is shown as the grey lines in Figure 1, and yields the well known Lorenz attractor.

When generating the data set used in the statistical analysis, randomly chosen points from the Lorenz attractor were used as the true state of the system. By assuming that the observations were subject to errors, the analysis was found by adding normally distributed noise to the true initial states. This analysis was used to initialise the control forecasts. For the ensemble members, the initial conditions were calculated by adding normally distributed noise to the analysis. The numerical model was then used to propagate the true state, analysis and the ensemble members forward in time. To make this resemble the EPS currently running operationally at the ECMWF, 50 ensemble members in addition to the control forecast were used. In each case, the forecast range was 2000 time steps. An example of the trajectory of one such true state is shown as the dotted line in Figure 1. If the standard deviation used for calculating the initial conditions for the ensemble members are smaller than the value used for the analysis, the ensemble spread will be too low, and vice versa. A system with perfect spread is obtained if the same value is used for both.

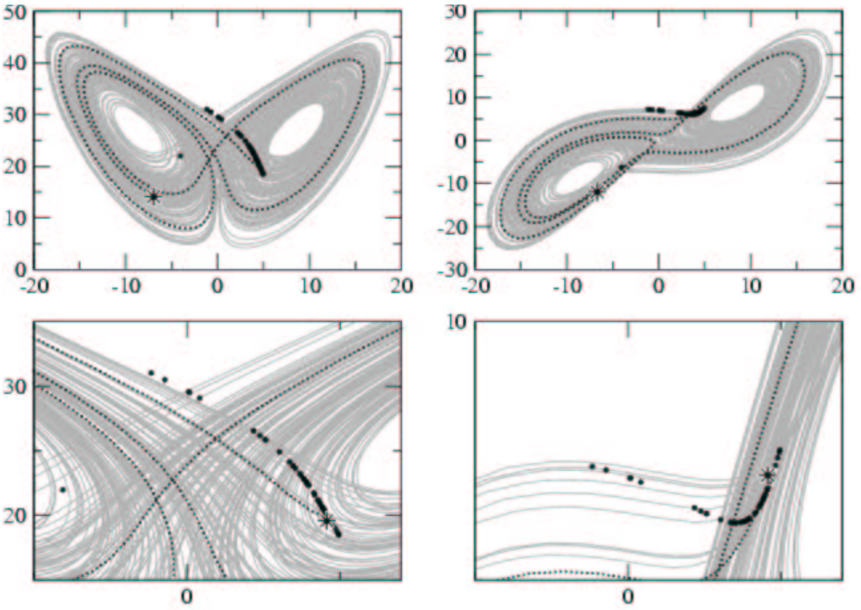

Figure 1: Example of the trajectory in phase space of the "true" solution (dotted line) together with the final states of the ensemble forecasts of this (small stars). Upper left corner shows the xz-projection, and the xy-projection is given in the upper right corner. The lower left and right corners are close up of the final state. In the two uppermost plots, the large star denotes the initial state. In the two close-ups the large star shows the final true state.

Verification is usually performed with respect to one-dimensional quantities such as 2 metre temperature. Here we will use the absolute distance from the origin, normalised by the sample mean value as a measure of the system state:

$$E = \frac{\sqrt{X^2+Y^2+Z^2}}{\overline{E}} \tag{2}$$

where $\overline{E} \approx 26.44$ is the approximate average Euclidian distance from the origin in phase space. In a similar way as for the initial states, observation errors were simulated by adding normally distributed noise to true value at the final states. Note that the same values of this parameter could be obtained for more than one point in phase space. Hence, the value is not unique to a particular state of the system. Nevertheless, the same is true also for parameters such as the significant wave height, which is a measure of the total wave energy at one specific location. The same value could be attained either by waves travelling in various directions, or by different combinations of wind-sea and swell. Figure 1 shows an example of an ensemble forecast using this model together with a close-up of the final state for all the ensemble members together with the true state.

To use this system to test the effects of observation errors, 12 different experiments were defined. These experiments can be sorted into 3 groups as; experiments with perfect spread, too low initial spread and too large initial spread. For the cases with too low initial spread, the standard deviation used for the ensemble members were $1/4$ of the value used for the analysis. The standard deviation for the ensemble members in the experiments with too large spread, was taken to be 4 times larger than the standard deviation used to pick the analysis. A summary of all the values used to determine the spread together with the standard deviation used for measuring the final state for each experiment is given in Table 1. For each experiment, 60,000 observations and forecasts were generated.

Ideally, the ensemble members and the verifying observations are random draws from the same probability

	Initial spread		Noise final state		
Exp	ENS	ANA	ENS	OBS	Rank hist.
1	0.01	0.01	0	0	flat
2	0.0	0.0	0.25	0.25	flat
3	0.01	0.01	0	0.02	u shape
4	0.01	0.01	0.02	0.02	flat
5	0.0025	0.01	0	0	u shape
6	0.0025	0.01	0.02	0	u+bell shape
7	0.0025	0.01	0	0.02	u shape
8	0.0025	0.01	0.02	0.02	u shape
9	0.04	0.01	0	0	bell shape
10	0.04	0.01	0	0.02	u+bell shape
11	0.04	0.01	0.02	0.02	bell shape
12	0.04	0.01	0.02	0	bell shape

Table 1: Specification of the twelve experiments for the Lorenz model

distribution. This means that the ranks of an observation when mixed with the ensemble members, should be random (Hamill 2001). If one such a set, containing n ensemble members and 1 observation, is ranked from the lowest value to the highest value, any rank of the observation has the same probability, namely $p = 1/(1+n)$. Hence, in our case with 51 ensemble members (including the control forecast), the probability that the observation falls outside the range of the ensemble forecast, being either the lowest or the highest, is $2/52 \approx 0.038$. The frequencies of the observed ranks can be illustrated graphically by using so-called rank histogram (Anderson 1996; Talagrand and Vautard 1997; Hamill 2001). The result from the experiment where the same standard deviation was used for both analysis and ensemble members, and no errors were added to the observations, is shown in the rank histogram in figure 2 (experiment 1). As expected for this case, the rank histogram is almost flat.

To test the ability of the ensemble to correctly forecast probabilities of a certain event, reliability diagrams may be used (Strauss and Lanzinger 1996; Wilks 1995). Here, we will consider the event that $E > 1$. The forecasted probabilities are split into discrete bins ranging from zero to one. For each probability class, the number of times the event is observed with respect to the total number of ensemble forecasts in that class, defined as the observed frequency, is plotted against the corresponding probability class. For a perfectly reliable forecasting system, these points lie on the diagonal line. The reliability diagram for the case with perfect spread and no observation errors is shown in Figure 3 (experiment 1).

An evenly distributed rank histogram, without any forecasting value, can be constructed in several ways. In an example given by Straus and Lanzinger (1996), an ensemble composed by picking fields at random from the analysis over the past 10 years, would yield a correct spread, but with no skill above climatology. Similarly, choosing evolved ensemble members randomly from pure Gausian noise will also produce a nearly flat histogram. This is demonstrated by experiment 2, which is plotted in Figure 2. Here, the ensemble forecasts and observations are pure random draws from a normal distribution. To obtain this, both ensembles and observations were given the same value, before normally distributed noise with standard deviation of 0.25 were added. Although the rank histogram is almost flat, the reliability diagram in Figure 3 (experiment 2) reveals no forecasting skill at all.

In the first experiment, perfect ensemble forecasts were compared with observations that were taken to be the true state of the system. To simulate the effects of observation errors, normally distributed noise can be added to this final state. The effect of this on the rank histogram is shown in Figure 2, experiment 3. As can be seen,

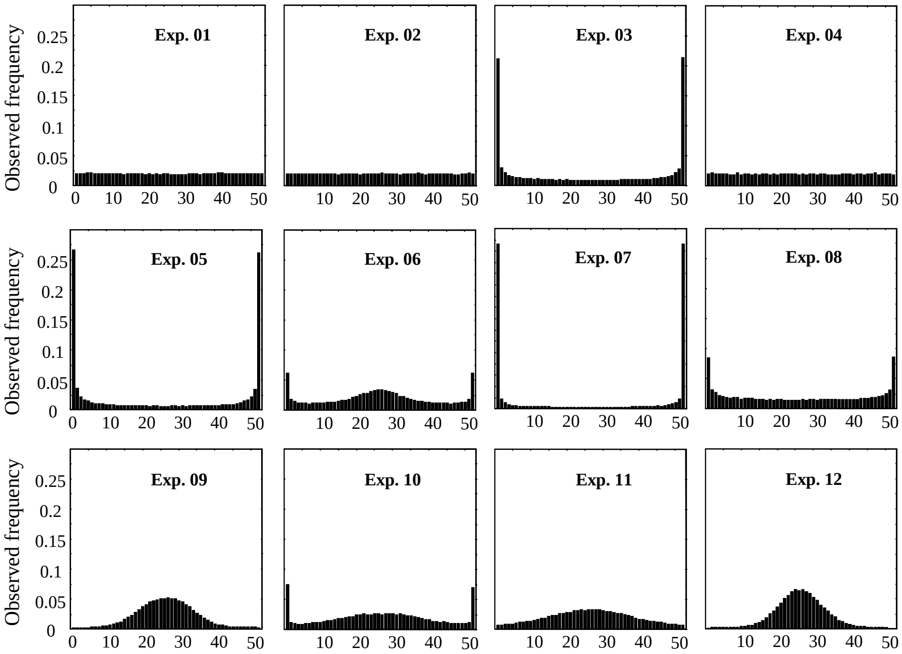

Figure 2: Rank histograms based on artificial data sets generated using the Lorenz model. The experiments 1, 3 and 4 in the upper row are the cases with perfect spread in the ensemble forecasts. Experiment 2 represents pure normally distributed noise. The experiments in the second row are cases with too low ensemble spread. the results for too large ensemble spread are given by the four cases shown in the third row.

this has a rather dramatic effect on the frequencies in the two extreme ranks, and can easily be misinterpreted as too low spread in the ensemble forecasts. However, in this case the ensemble spread is perfect, and the strongly u-shaped rank histogram is solely caused by observation errors. The corresponding reliability diagram seems to be much less sensitive in this respect (Figure 3, experiment 3).

As argued in the introduction, the proper verification is obtained by adding noise to the ensemble members as well. The rank histogram method does not distinguish between large and small differences. It is of course meaningless to rank the data according to differences that are much smaller than the observation errors. Hence, this effect must be filtered out before calculating the frequencies of observations in the different ranks. In experiment 4, this has been done by adding the same amount of noise to the ensemble members as was used for the observations. The effect on rank histograms is striking (see Figure 2): the u-shape for experiment 3 has been removed completely, leaving a flat histogram, as it should. Again, the effect on the reliability diagram (Figure 3) is much less apparent.

In Figure 2, experiment 5, the effect of too small initial spread on the rank histogram is demonstrated. Here, the true values are used for the observations. In this case the rank histogram is also strongly u-shaped, and looks very similar to the result obtained in experiment 3, although in that case, this was caused by observation errors. Based on this, it is hardly possible to distinguish between the effects of observation errors or too low ensemble spread. When noise was added to the evolved ensemble member for this last case, the observed frequencies in the two extreme ranks were reduced (Figure 2, experiment 6). However, in contrast to the case with true measurement errors, a bell shape also appeared in the centre of the histogram. The combined effect of too small ensemble spread and observation errors can be seen in Figure 2, experiment 7. In this case, the u-shape of

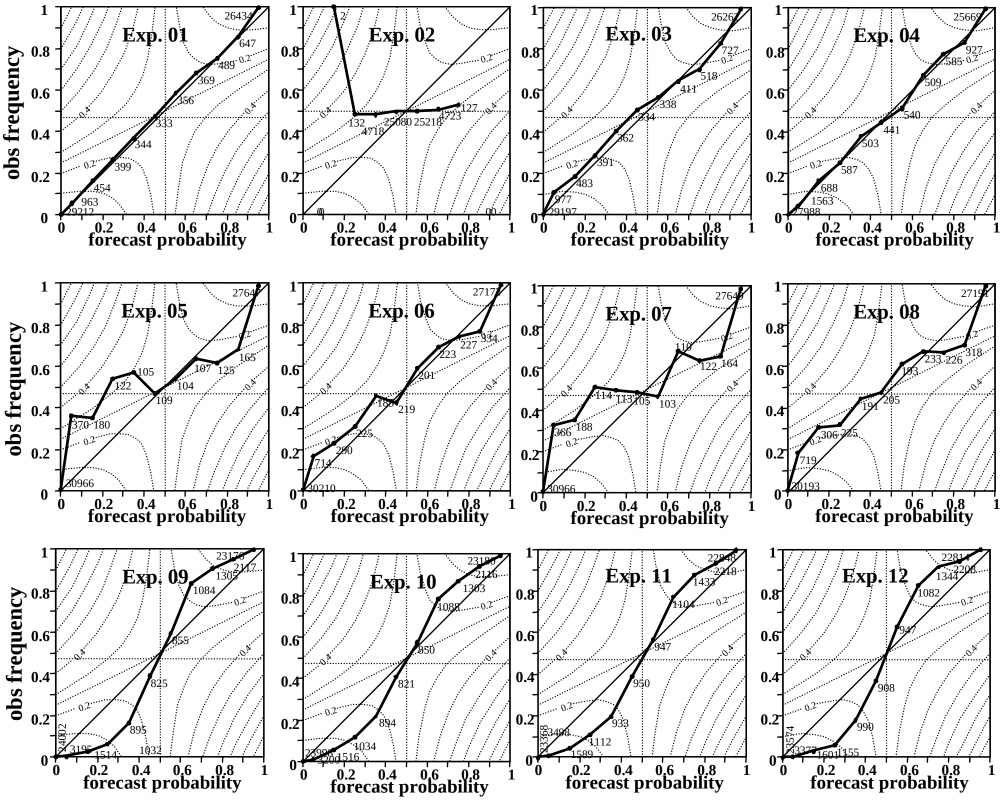

Figure 3: Reliability diagrams based on artificial data sets generated using the Lorenz model. The experiments 1, 3 and 4 in the upper row are the cases with perfect spread in the ensemble forecasts. Experiment 2 represents pure normally distributed noise. The experiments in the second row are cases with too low ensemble spread. the results for too large ensemble spread are given by the four cases shown in the third row.

the rank histogram is even more pronounced. Experiment 8 shows the result for this last case when observation errors are filtered by adding noise to the ensemble members.

From Figure 3, (experiment 5 to 8) it is clear that too low ensemble spread has a stronger impact on reliability diagrams, than observation errors have. For the two cases without noise added to the ensemble members, the reliability curves are s-shaped, under-forecasting low probabilities, and over-forecasting high probabilities. When noise is added to the ensemble members, the curves are improved.

The effect of too large ensemble spread is demonstrated by experiment 9 to 12. The rank histogram for experiment 9 shows the Gausian shape obtained when the true values are used for the observations. The combined effect of observation errors and too large initial spread is shown by experiment 10. The result when noise is added to the evolved ensemble members is given by experiment 11. This removes the over-representation in the extreme ranks, while a dome shape remains. Finally, in experiment 12, the result of adding noise to the ensemble members in the case where the observations are perfect, is shown.

The reliability diagrams for the cases with too large spread (Figure 3 experiment 9 to 12), also indicate that these

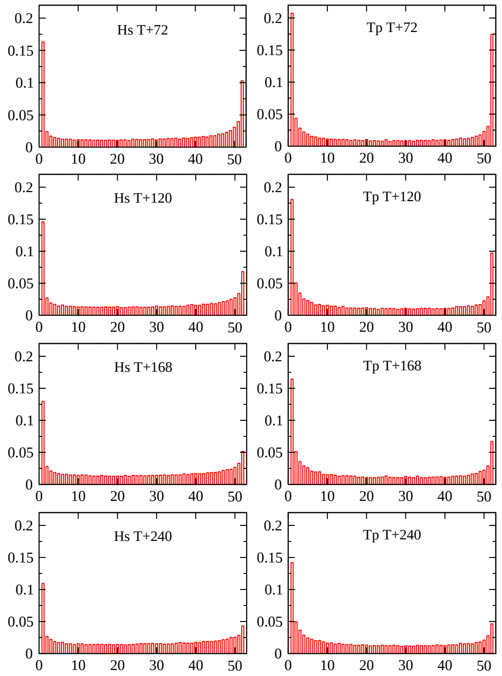

Figure 4: Rank histogram for wave height (Hs, left) and peak period (Tp, right) when compared with buoy and platform data. Here, the observations are taken as the true value. Note that Tp is only reported by US and Canadian buoys. Forecast steps are 72, 120, 168 and 240 hours. The period covered spans from September 1999 to March 2002.

are more sensitive to errors in the ensemble spread than to observation errors. The reliability curves for the four cases looks almost similar, with a pronounced s-shape, over-forecasting low probabilities and under-forecasting high probabilities.

3 The ensemble prediction system at ECMWF

In this section the impact of errors in the verifying observations on rank histograms will be studied for the EPS running operationally at ECMWF (Molteni et al. 1996; Buizza et. al 2000). First the situation in which verification is performed with respect to observations is considered. As an example, the focus will be on ocean wave heights. Then the case in which rank histograms are based on verifying analyses will be studied. This will be done for the most commonly used parameter, in this respect, which is the geopotential height field at 500 hPa.

3.1 Ocean waves

To see how observation errors may affect the interpretation of the ensemble spread for the EPS, ensemble forecasts of ocean wave will be considered. In June 1998, the ocean wave model WAM (Komen et al. 1994) was coupled to the atmospheric circulation model (Janssen et al. 2002). From then on, ensemble forecasts of ocean waves have been available on a daily basis.

Saetra and Bidlot (2002) used buoy and platform observations for the assessment of the ECMWF wave ensembles. The model was compared with wave data obtained via the global Telecommunication System (GTS) for the period from October 1999 to March 2002 for off-shore locations around the US and Canadian coasts and

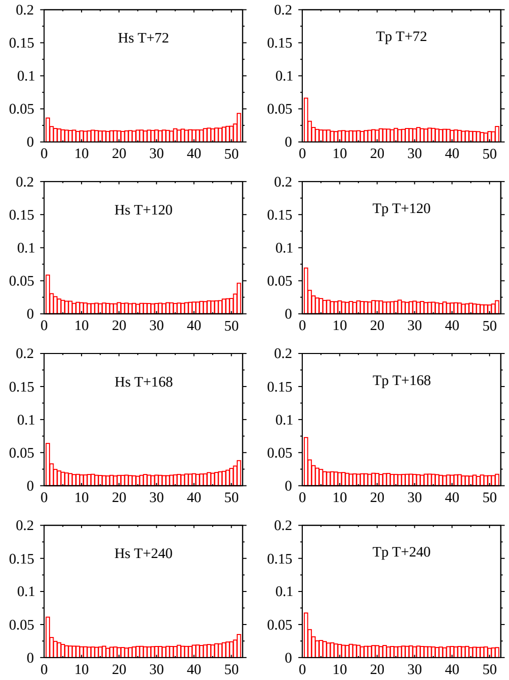

Figure 5: Rank histogram for wave height (Hs, left) and peak period (Tp, right) after noise have been added to the ensemble members. Note that Tp is only reported by US and Canadian buoys. Forecast steps are 72, 120, 168 and 240 hours. The period covered spans from September 1999 to March 2002.

on both sides of the British Isles. This data set will be used here. The focus will be exclusively on the effect of observations errors and the interpretation of the number of outliers in rank histograms as a measure of the ensemble spread. A detailed description of the observations used, is given in the paper referred to above. In Figure 4, the rank histograms for significant wave height and peak period is shown for four different forecast ranges. At day 3 ($T+72$), more than 25 % of the significant wave height observations are outside the ensemble. For the peak period, this number is more than 35 %. At later forecast steps, the number of outliers is reduced. But even at day 10, the percentage of observations in the two extreme ranks are still about 15 % for the significant wave height.

Similar results have been obtained by several authors. Strauss and Lanzinger (1996) compared the global temperatures at 850-hPa with the analysis, and obtained the percentage of outliers to be approximately 22 % at day 6, compared to the theoretical value of 6 % for the 32-member ensemble used in their study. Evans et al. (2000) give rank histograms for the 500-hPa height at T+156 (day 6.5) over the North Atlantic-European region. Here, about 15 % of the observations are in the two extreme ranks. Buizza et al. (2000) show the difference between the percentage of outliers and the reference value for the 500-hPa height over Europe for the 51-member ensembles. For the period after 1998, they found that for the day 5 forecasts, the number of outliers were approximately 10 % larger than the theoretical value of 3.8%. In all cases, the observations were taken as the true value. In the light of the result obtained in the previous section, the relatively large fraction of outliers may be explained, at least to some extent, by the presence of observation errors.

For wave height observations, the mean error has been estimated to be approximately 12 % of the mean wave height (Saleh Abdalla, personal communication). In an attempt to filter out the effects of variability smaller than the uncertainties in the observations, errors for the significant wave height are assumed to be normally distributed with standard deviation of 12 % of the true value. The same amount of noise will therefore be added to each ensemble member. For the peak period, we used a standard deviation for the errors of 1 second. This

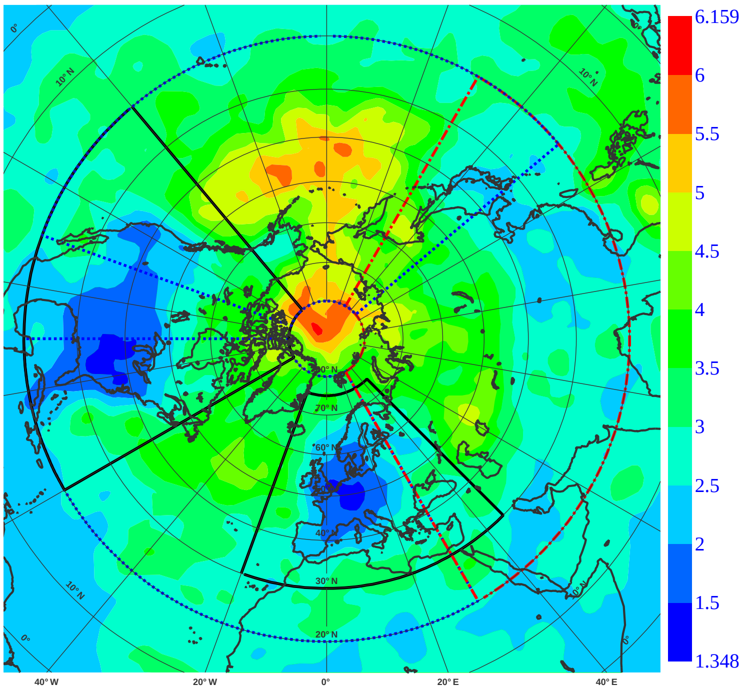

Figure 6: Estimate for the error covariance of the analysis in Z500 for the period between 2 and 31 October 2000. Average values for North America and Europe (black solid boxes), for the North Atlantic plus Europe and North Pacific (blue dotted boxes) and for Asia (red dashed box) are 2.86 m (NAmerica), 2.68 m (Europe), 2.90 m (AtWEur), 3.73 m (Pacific), and 3.19 m (Asia) respectively. The average over the Northern Hemisphere (20N$\leq$lat$\leq$80N) is 3.24 m.

is based on the fact that this corresponds approximately to the average resolution of the wave spectra, and that the observations are truncated to the nearest second. The results of this are given in Figure 5. Although the rank histograms still indicate that the ensembles have too small spread, this is far less pronounced than if the observations are taken to be the true value. For the significant wave height, the number of outliers is reduced to less than 10 % for all forecast ranges. Also for the peak period, the number of outliers is about 10 % in all cases. One interesting feature is that for the wave height, the number of outliers is more or less the same for all forecast ranges. This is contrary to what was obtained before observation errors were taken into account, and demonstrates that the system is more stable than earlier results have indicated.

3.2 Geopotential at 500 hPa

Most commonly, ensemble systems are verified with respect to verifying analysis fields, rather than with respect to observations. The advantage of using verifying analysis fields instead of observations is that they are easily available. In addition, these fields represent the same scales as the forecast fields of the ensemble members, and, therefore, avoid representative and/or collocation errors that would be introduced when using observations (like the buoys in section 3.1).

Before the impact of the errors in these fields on rank histograms can be assessed, an estimate for their typical magnitude is to be made. For October 2000, such an estimate for Z500 is displayed in Figure 6. It is based on the spread within 10 analysis experiments averaged over the specified period (Mike Fisher, personal communication). For each such experiment, observations were randomly perturbed according to their uncertainty, before they were assimilated into the ECMWF 4D-Var system. Analysis errors are found to be small over Western Europe and over the Eastern part of the U.S., and large over the Pacific and the Arctic. It reflects

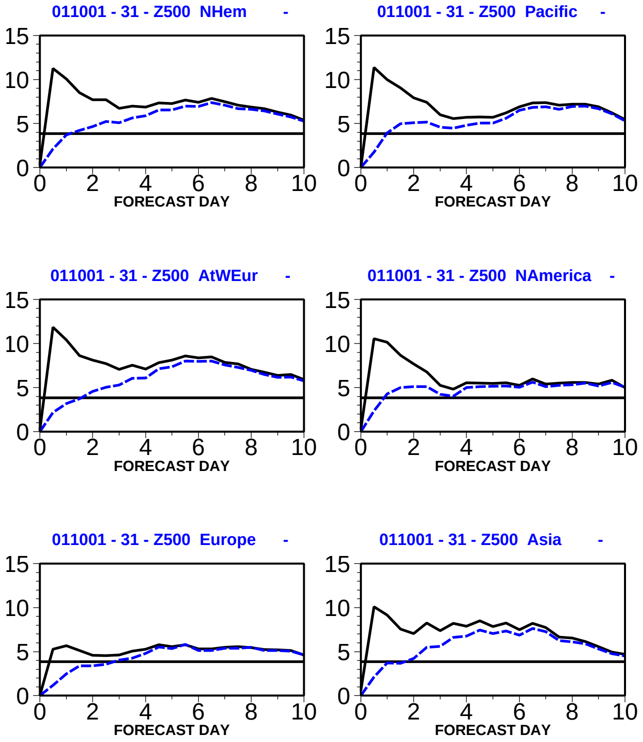

Figure 7: Frequencies of outliers as function of forecast time, averaged over October 2001 for six different areas. Solid curves are for the (usual) case in which observation errors at verification are neglected, dashed curves for the case in which estimates of such errors (given in Figure 6) have been included.

the high, respectively low density of the conventional observational network (mainly radio sondes and aircraft measurements) in these areas.

The in this way determined estimates for the Z500 analysis error for October 2000 were used for the verification period of October 2001. In the case that the analysis error at verification time was incorporated, ensemble members were perturbed according to these estimates. So on average, larger perturbations were applied in the Pacific, than for the European area. For six different areas for the October 2001 period, the percentage of outliers for Z500 is shown in Figure 7. These outliers are the sum of the two outer bins of the rank histogram. In case the system has the correct statistical properties, this quantity should be 3.8% for EPS ($N = 51$). The Northern Hemisphere is defined by all points with a latitude between 20N and 80N, the five other areas are determined by the boxes in Figure 6.

As was seen for the other cases regarded in this paper, the impact of taking the ’observation’ error into account is considerable. In the (usual) case when they are not incorporated (black solid curves) the frequency of outliers is far too high in the short range. For this range the inclusion of errors in the verifying analysis (blue dashed curves) has the largest effect. The frequency of outliers becomes much lower, and more constant as function of forecast time. The impact is smallest for Europe, because here (see Figure 5), average analysis errors are smaller than in the other five areas. It is also exactly this region for which the frequency of outliers was smallest (7% compared to 10-12% for the other areas) in case the error in the verifying analysis were neglected. Both the largest impact in the short range and the smallest mismatch for the unperturbed method in regions were the analysis error are smallest, favour the conjecture that the effect of the inclusion of observation errors in the verification is realistic.

Note that at analysis time the frequency of outliers is zero, regardless whether errors in the (verifying) analysis are taken into account or not. It results from the fact that the initial ensemble members are symmetrically perturbed around the analysis. By definition, the analysis is located in the middle of the ensemble, and, therefore, can never be an outlier. This effect forms an extra complication that obscures the assessment of the statistical quality of an ensemble system. The time-scale until which it has an effect is connected to the decorrelation time between the forecast and analysis. From Figure 7 it may be suggested that rank histograms are affected up to day 1. This undesirable feature mainly applies when verification is performed with respect to the verifying analysis fields. It is less present when the comparison with 'real' observations is performed (like was the case for the previous subsection).

4 Conclusions

Based on the results obtained by using the Lorenz model in section 2, it was found that observation errors may have a rather dramatic effect on rank histograms, leading to the increase in the number of observations in the lower and upper ranks. Ranking data obtained with a perfect model for both forecasts and observations, resulted in u-shaped histograms when normally distributed noise was added to the observations. The results were almost identical to those obtained with perfect observations but too low ensemble spread. This false impression of too low ensemble spread is caused by the fact that the method does not distinguish between large and small errors. Only significant differences, i.e. differences larger than the observation errors, should contribute to the upper and lower ranks when observations are outside the ensemble range. To account for this, the probability distribution from the ensemble forecast must be convoluted with the observation errors. By adding normally distributed noise to each ensemble member, using the same standard deviation as for the observation errors, the flat rank histogram were restored for the perfect model case.

The reliability diagrams are less sensitive in this respect. For the case with perfect spread, the reliability diagrams were almost identical to cases with perfect observations and observations with noise added to them. Errors in the ensemble spread on the other hand, had a stronger impact on the reliability curves. Too low ensemble spread resulted in under-forecasting of low probabilities. Again, adding normally distributed noise to the ensemble members slightly improved the results. Too large ensemble spread resulted in s-shaped reliability curves, over-forecasting for low probabilities and under-forecasting for high probabilities.

Many investigations have pointed out that the ECMWF EPS do not have enough spread (Strauss and Lanzinger 1996; Evens et al. 2000; Buizza et al. 2000). Most studies of ensemble spread are based on either rank histograms, or counting the number of outliers. Since usually the observations or verifying analysis fields are taken to be the truth, these results can to some extent be explained by the presence of errors in these quantities. When comparing the wave ensembles with buoy observations, we have demonstrated that the total number of outliers for the day 3 forecasts are reduced from more than 25 % to less than 10 % when a reasonable estimate of the observation errors are taken into account. Although the spread is still too low, the performance of the EPS is in this respect much better than often suggested.

Also for the Z500 forecasts, the impact of analysis errors on ensemble spread was considerable. For the short range, the ensemble spread has traditionally been considered far too small. Usually, a peak in the number of outliers has been found in the short forecast range. The fact that the results have usually been better over Western-Europe and the U.S. where observation networks are more dense, may possibly be an indication that the results have been dependent on analysis errors. When noise based on the estimated of analysis errors were added to the ensemble members, the peak in the number of outliers in the short range more or less vanished. Also, the numbers of outliers became more constant in time.

Acknowledgements

This research is partly funded by the European Research Council through the SEAROUTES project. We like to thank Peter Janssen, Saleh Abdalla and Mike Fisher for support and valuable discussions.

References

Anderson, J. L., 1996: A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations. *J. Climate,* **9**, 1518-1530.

Buizza, R., J. Barkmeijer, T. M. Palmer and D. S. Richardson, 2000: Current status and future developments of the ECMWF Ensemble Prediction System. *Meteor. Appl.*, **7**. 163-175.

Evans, R. E., Harrison, M. S. J., Graham, R. J. and K. R. Mylne: Joint Medium-Range Ensembles from The Met. Office and ECMWF Systems. *Mon. Wea. Rev.*, **128**, 3104-3127.

Hamill, T. M., 2001: Interpretation of rank histogram for verifying ensemble forecasts, *Mon. Wea. Rev.*, **129**, 550-660.

Janssen, P. A. E. M., J. D. Doyle, J. Bidlot, B. Hansen, L. Isaksen and P. Viterbo, 2002: Impact and feedback of ocean waves on the atmosphere. Advances in Fluid Mechanics, Atmosphere-Ocean Interactions, Vol. I, WITpress, Ed. W.Perrie. 155-197.

Komen, G. J. Cavaleri, M. Doneland, K. Hasselmann, S. Hasselmann, and P. A. E. M. Janssen, Eds., 1994: Dynamics and Modelling of Ocean Waves. *Cambridge Univeristy Press*, 533 pp.

Lorenz, E. N., 1963: Deterministic Nonperiodic Flow. *J. Atmos. Sci*., **20**, 130-141.

Molteni, F., Buizza, R., Palmer, T. N. and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation, *Q.J.R. Meteorol. Soc.* **122**, pp. 73-119.

Richardson, D. S. 2000 Skill and relative economic value of the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **126,** 649–668

Saetra, Ø., and J-R. Bidlot, 2002: Assesment of the ECMWF ensemble prediction system for waves and marine winds, ECMWF Research Department Tech. Memo. 388. 29 pp.

Strauss, B., and A. Lanzinger, 1996: Verification of the Ensemble Prediction System (EPS). *ECMWF Newsletter*, 72, 9-15.

Talagrand, O. and R. Vautard, 1997: Evaluation of probabilistic prediction systems. *Proceedings of the ECMWF Workshop on predictability, Reading, England* 20-22 October 1997, 1-25.

Wilks, D. S., 1995: Statistical Methods in the Atmospheric Sciences: An Introduction. *Academic press*, 467 pp.