METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **15**: 103–111 (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/met.54

RMetS
ROYAL METEOROLOGICAL SOCIETY

Valuing information from mesoscale forecasts

Kees Kok,[a]* Ben Wichers Schreur[b] and Daan Vogelezang[a]
[a] *Royal Netherlands Meteorological Institute, De Bilt, The Netherlands*
[b] *Wageningen University and Research Centre, Wageningen, The Netherlands*

ABSTRACT: The development of meso-γ scale numerical weather prediction (NWP) models requires a substantial investment in research, development and computational resources. Traditional objective verification of deterministic model output fails to demonstrate the added value of high-resolution forecasts made by such models. It is generally accepted from subjective verification that these models nevertheless have a predictive potential for small-scale weather phenomena and extreme weather events. This has prompted an extensive body of research into new verification techniques and scores aimed at developing mesoscale performance measures that objectively demonstrate the return on investment in meso-γ NWP.

In this article it is argued that the evaluation of the information in mesoscale forecasts should be essentially connected to the method that is used to extract this information from the direct model output (DMO). This could be an evaluation by a forecaster, but, given the probabilistic nature of small-scale weather, is more likely a form of statistical post-processing. Using model output statistics (MOS) and traditional verification scores, the potential of this approach is demonstrated both on an educational abstraction and a real world example. The MOS approach for this article incorporates concepts from fuzzy verification. This MOS approach objectively weighs different forecast quality measures and as such it is an essential extension of fuzzy methods. Copyright © 2008 Royal Meteorological Society

KEY WORDS model output statistics; fuzzy verification; probabilistic verification

Received 13 September 2007; Revised 12 December 2007; Accepted 2 January 2008

1. Introduction

The failure of traditional objective verification methods when applied to high-resolution direct model output (DMO) generally is demonstrated using the double penalty that results from a displacement or phase error. Simply put, a high-resolution forecast that contrary to a low-resolution forecast produces a small-scale weather feature, but slightly displaced in space or time, will be penalized for not having the feature where it is observed and for having that feature where it is not observed. This double penalty fails to recognize the added value of the information in high-resolution forecasts that is often quite clear from a subjective evaluation.

Another major hurdle for mesoscale verification is what may be loosely termed the data problem. Verifying observations for small-scale weather phenomena are not always available, are unevenly distributed, provide incomplete coverage and sample at a different scale. The information contained in the observations may not give the same information as the model without some preconditioning (*cf* Cherubini *et al*., 2002). While mesoscale numerical weather prediction (NWP) models are by their formulation still deterministic, the small-scale weather phenomena they are predicting are not. Their predictability horizon generally does not exceed their short lifetime. This implies that DMO at the smallest scales should always be interpreted probabilistically. If, then, the information at these scales is essentially probabilistic, a deterministic interpretation and verification is inappropriate.

The analysis of this failure of traditional verification methods immediately points at ways of overcoming their inability to capture the essential characteristics of high-resolution forecasts. In recent years this has prompted an extensive body of research into alternative mesoscale verification methods. Much of this work has concentrated on some form of feature identification and associated performance measures that rate displacement, phase, volume and amplitude error (Davis *et al*., 2006a). Hoffman *et al*. (1995) coined the term distortion error for a combination of displacement and amplitude error. They suggested an application of this description of forecast error as a continuous field transformation to be used in a variational data assimilation framework. Du *et al*. (2000) applied this to ensemble forecasting. In application to precipitation verification, the contiguous rain area approach defined by Ebert and McBride (2000) has found wide application, (*cf* Davis *et al*., 2006a). This method can be applied not only to the rainfall distribution *per se*, but also to general statistical aspects of the rainfall pattern (Baldwin and Lakshmivarahan, 2003; Davis *et al*., 2006b). Other authors have concentrated on the scale issues associated with verification in general and mesoscale verification of precipitation in particular (Zepeda-Arce *et al*., 2000;

* Correspondence to: Kees Kok, Royal Netherlands Meteorological Institute, 3730 AE De Bilt, The Netherlands.
E-mail: kees.kok@knmi.nl

Tustison *et al*., 2001; Casati *et al*., 2004; Mittermaier, 2006). Roberts and Lean (2008) propose an evaluation of the scale dependence of forecast quality using the fraction skill score. The evaluation of NWP forecast using spectra (Skamarock, 2004) in a similar way deals with the concept of scale, though more as a model validation tool. Wavelet decomposition for forecast verification as proposed by Briggs and Levine (1997) can perhaps be seen as a cross between scale decomposition and feature extraction. Nachamkin (2004) proposes meteorological composites as a solution to the disparity between forecast and observational data, and applies this method to a distribution-oriented verification of heavy precipitation forecasts (Nachamkin *et al*., 2005). Theis *et al*. (2005) promote a pragmatic approach to deriving precipitation probabilities from spatial rainfall distributions. Mittermaier (2007), in a similar vein, proposes to use time-lag ensemble techniques to assess the behaviour of high-resolution precipitation forecasts.

Ebert provides a comprehensive overview of verification methods and literature on the WWRP/WGNE verification website (http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html). Recently, Ebert has reviewed mesoscale verification efforts (Ebert, 2008). In this article, Ebert (2008) proposes fuzzy verification as a general framework for evaluating mesoscale forecasts. As Ebert's review indicates there is a desire to consolidate the work on mesoscale verification. A great variety of methods hampers comparisons of verification and model results and introduces subjectivity in the choice of verification scores. These scores value different quality aspects of forecasts but provide no intrinsic weighting of these aspects. Of course, this variety is a consequence of the fact that mesoscale forecasts, more than synoptic forecasts, are situation dependent and application oriented. Often the conditions and applications are implicit in the design of the verification method. These methods, then, are not general, but have their value provided a skilled verification practitioner uses them prudently.

It is unsatisfactory that such mesoscale methods do not provide a clear indication of the intrinsic value of the considerable investment in the development of mesoscale models. It may be argued, following Murphy (1993), that value and information are not intrinsic properties of a forecast, but depend on the application and on the availability of other information. Mesoscale forecasts should thus always be verified in comparison to other forecasts available to the forecaster or customer. Callies (2000) has presented a method for comparative forecast verification. The application dependence of forecast value seems to prohibit any general method of verification. However, given that there is a system that extracts information from the DMO, it is an obvious choice to use the same system in comparative forecast verification.

In recent years, the use of ensemble prediction to derive probabilistic information based on mesoscale models has attracted much attention. Molteni *et al*. (2001) and Marsigli *et al*. (2001) used a mesoscale model to downscale medium range coarse resolution forecasts. They showed that in the medium range, where synoptic predictability is predominantly determined by the uncertainty in initial conditions, downscaling precipitation over orography can be successful because of the improved resolution in the small-scale forcing. Orrell *et al*. (2001) have shown that in the short range, model error is a major component of forecast uncertainty. The approach to the ensemble prediction of model uncertainty is still very much an issue in the successful application of ensemble prediction to short-range forecasting. Ensemble prediction is not a method that is further explored in this article. Instead we focus on the more traditional method of model output statistics (MOS; Glahn and Lowry, 1972)

MOS is a widely accepted and generally applicable tool for the extraction of probabilistic information from DMO. Expressing the uncertainty in model forecasts, provided it can be done skilfully, can be of great value to professional users. They can use probabilistic information in their own decision making in accordance with their own cost-loss ratio (Murphy, 1977; Katz and Murphy, 1997). While MOS is a general method, it can be tailored to a particular application through a skilled choice of the predictands.

In this article, it is proposed that using MOS in the context of comparative forecast verification provides a general method for valuing information from mesoscale forecasts. This line of reasoning is demonstrated in the next section using an abstract educational example: the cosine model. To illustrate this theory in practice a real world example is presented in Section 3. In this section, the precipitation at a station is modelled with MOS using model precipitation in the neighbourhood in a manner similar to fuzzy verification methods. This demonstrates that this method is an, probably essential, addition to such methods. In particular, it is shown that it provides an intrinsic weighting of different quality aspects of a numerical weather forecast and an intrinsic way of dealing with scale issues. Precipitation is chosen as predictand in the example for reasons of comparison, giving that most mesoscale verification research deals with precipitation. In the verification, the familiar Brier score (BS) is used as an example of probabilistic verification (Wilks, 2006). The BS and its decomposition are widely used and are well-understood measures of forecast quality (Brier, 1950; Murphy, 1973). In Section 4 the findings are summarized and the merits, pitfalls and shortcomings of the proposed method are discussed. Section 5 concludes the article with the main findings.

2. The cosine model: an educational example

In this article, the attention is focussed on differences between high-resolution and low-resolution models. To illustrate the deterioration of deterministic verification scores if the additional high-resolution information is not perfect, this section describes a theoretical example. It follows ideas outlined in Kok (2002). This example explores the impact of known error statistics of amplitude and phase (intensity and position) of small-scale features

DOI: 10.1002/met

and how this knowledge can be incorporated into a probabilistic forecasting system.

To simplify things consider a predictand in only one dimension. Look for instance at two model versions, their only difference being the resolution. In this section they are referred to as high-resolution model (HRM) and low-resolution model (LRM). The HRM can deal with one additional small-scale cosine-like wave that cannot be resolved by the LRM. So the only difference is that of a single small-scale wave. The observations (OBS) have exactly the same scales as HRM, so no smaller scales are present in the OBS. To focus on the effects of the differences in resolution, further assume that the larger scales are always perfectly predicted. So the LRM gives perfect forecasts for the scales it can resolve. The verification results of the HRM depend on the correctness of the forecasts of the smallest scale wave. The HRM, LRM and OBS are defined as follows:

$$\begin{aligned} HRM &= A_1 \cos x + A_2 \cos(2x + \alpha_2) \\ LRM &= A_1 \cos x \\ OBS &= A_1 \cos x + C_2 \cos(2x) \end{aligned} \tag{1}$$

Now look at the consequences of various *prescribed* (known) uncertainties in phase and amplitude of the smaller scale wave. Suppose that the predictions of phase and amplitude are correct to within a certain range. More specifically, suppose that phase predictions α_2 are correct to $\pm$ phase uncertainty c, and further assume that every value within this interval is equally probable. In Figure 1 the expected correlation coefficient is plotted as a function of phase uncertainty for the case $A_1 = 4$ and $A_2 = C_2 = 2$. Unless there is a considerable degree of certainty, the high-resolution part of the forecast will deteriorate the verification. Other deterministic metrics show similar behaviour.

Now consider the probability of a precipitation amount above a certain threshold. As an example, this threshold is chosen in such a way that the observed frequency for arbitrary forecasts will be 1/3. This can be regarded as the climatology of the event. This is an arbitrary number that does not influence the conclusions of the experiment. The results for a 'high-resolution forecaster' in terms of the expected BS are shown in Figure 2. It is assumed that both LRM and HRM are well calibrated.

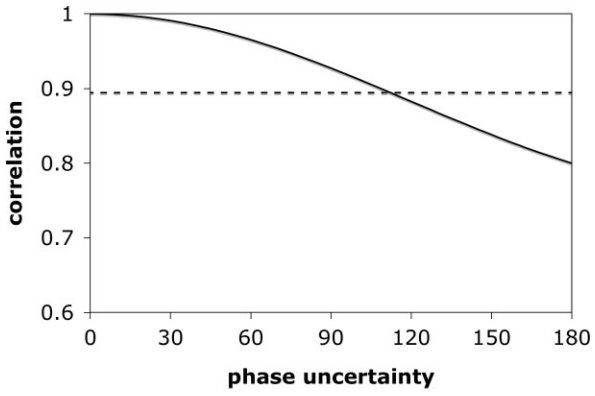

Figure 1. Expected correlation coefficient as a function of phase uncertainty c of the smallest scale waves for the high-resolution model, HRM (full line) and low-resolution model, LRM (dashed line).

First look at the uncertainties in phase without incorporating amplitude uncertainties. The bottom thick line gives the results. When the position is exactly known and the phase uncertainty is zero, the BS is zero. On the other hand, if there is no skilful information in the prediction of the phase this information can be ignored and predictions are simply determined by the 'climatological' frequency resulting in an expected BS of 2/9. This situation corresponds exactly to the situation of the 'low-resolution forecaster' who does not have any information at all about the smallest scales and will 'always' issue the climatological probability of the event.

However, the figure also indicates that as soon as there is information contained in the prediction of the phase, no matter how small, the expected BS will only improve. In terms of the comparison between the HRM and LRM, this means that as soon as there is skilful information in the prediction of the smallest wave then the expected BS of HRM is better than the one for LRM. This is quite in contrast with deterministic verification of the DMO: phase predictions have to be significantly better than random before deterministic verification results are better for the HRM than for the LRM.

The inclusion of uncertainties in the amplitude of the smallest cosine wave yields the dashed and dotted lines. These have higher expected BS for all values of the phase uncertainties (and higher for increasing uncertainty about the value of the amplitude).

Figures 1 and 2 contrast the deterministic verification of DMO with the probabilistic interpretation and verification of the information contained within that output. The latter appears to be a more appropriate tool for valuing information from mesoscale forecasts, with the added bonus that uncertainties in phase and amplitude can be evaluated together in their combined effect on the probability of an event to occur.

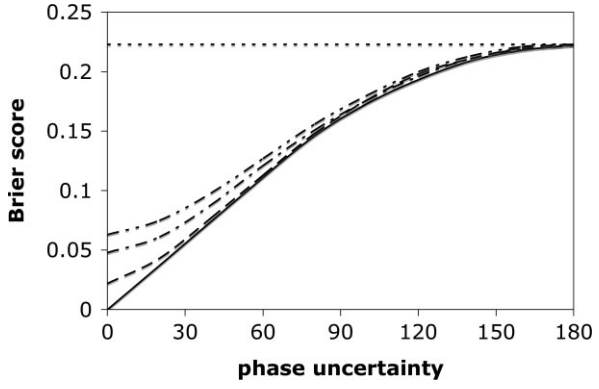

Figure 2. Expected Brier score for the high-resolution model (HRM) as a function of phase uncertainty c and for different degrees of amplitude uncertainty A_2: amplitude known (full line); by a factor $\sqrt{2}$ (dashed line); by a factor 2 (dash-dot-dashed line); by a factor 4 (dash-dot-dotted line). The expected Brier score for LRM (dotted line) is 2/9.

DOI: 10.1002/met

This abstract example has been constructed to elucidate the strength of probabilistic verification applied to mesoscale forecasts. Reality is often resilient to such abstract notions. To verify the proposed approach its implications in a real world application are explored.

3. Thresholds of precipitation at a single station: a real world example

The theoretical example shows that as soon as there is knowledge about the uncertainty of the not fully predictable scales, this knowledge can in principle be used to improve the forecast. The real world experiment this section describes, does not explicitly deal with the different scales that are present in a predicted precipitation field but instead deals with the total forecast precipitation field and searches for any information relevant for the predictand at hand. This information may be different of course for different choices of the predictand. The information is obtained from the forecast field using statistical post-processing, more specifically MOS.

MOS predictions are routinely used to post-process and enhance the results of NWP forecasts at operational weather forecasting centres around the world, and are considered essential guidance products to aid weather forecasters. An important feature of statistical forecasting methods is the capacity to produce probability forecasts. They provide explicit expressions of the inherent uncertainty that is present in weather forecasts. It seems an obvious technique, therefore, to assess and objectively quantify the probabilistic information present in deterministic model output. Owing to the intrinsically probabilistic nature of the smaller length and timescales it seems even more appropriate in high-resolution forecasting. Another advantage of probabilistic forecasts is that they allow users to extract more value from them when making decisions (e.g. Thompson, 1962; Murphy, 1977; Katz and Murphy, 1997).

3.1. Setup of the experiment

The concept outlined above is tested on atmospheric data. The numerical outputs of two models whose grid distances differ by a factor of two are compared. The property investigated is the predictive potential of predicted total precipitation fields for observed precipitation in a single station, in particular for the probability of exceeding certain amounts of precipitation. The thresholds considered are >0, ≥ 1.0, ≥ 2.5 and ≥ 4 mm in 3 h. It is important to note that no numerical information other than precipitation forecasts is used. The models used in the comparison are the operational European Centre for Medium-Range Weather Forecasts (ECMWF) model and the control model of the ensemble prediction system (EPS) of ECMWF on which since February 2006, precipitation is calculated on a N400 and N200 reduced Gaussian grid, respectively. This corresponds to a grid distance of 0.225° and 0.450° in the north–south direction, which is close to 25 and 50 km respectively. The experiment uses extracted forecast data from the two models of 3-h accumulated precipitation for lead times from +3 to +72 in steps of 3 h on latitude–longitude grids closely resembling the Gauss grids of the models (Figure 3). The extent of the grids in the east–west direction is approximately 500 km, in the north–south direction it is 600 km but only neighbourhoods with a radius of 250 km are used (indicated by asterisk in Figure 3). The centre of the Gauss grids is 52.2 °N, 4.95 °E very close to the observing station De Bilt, located in the centre of the Netherlands. The data set contains 1200 UTC forecasts only, for the period February 2006 through to July 2007. No stratification into seasons is performed.

3.2. Potential predictors

On these grids, a large number of potential predictors are calculated. They are constructed in such a way as to incorporate all the relevant features that can be

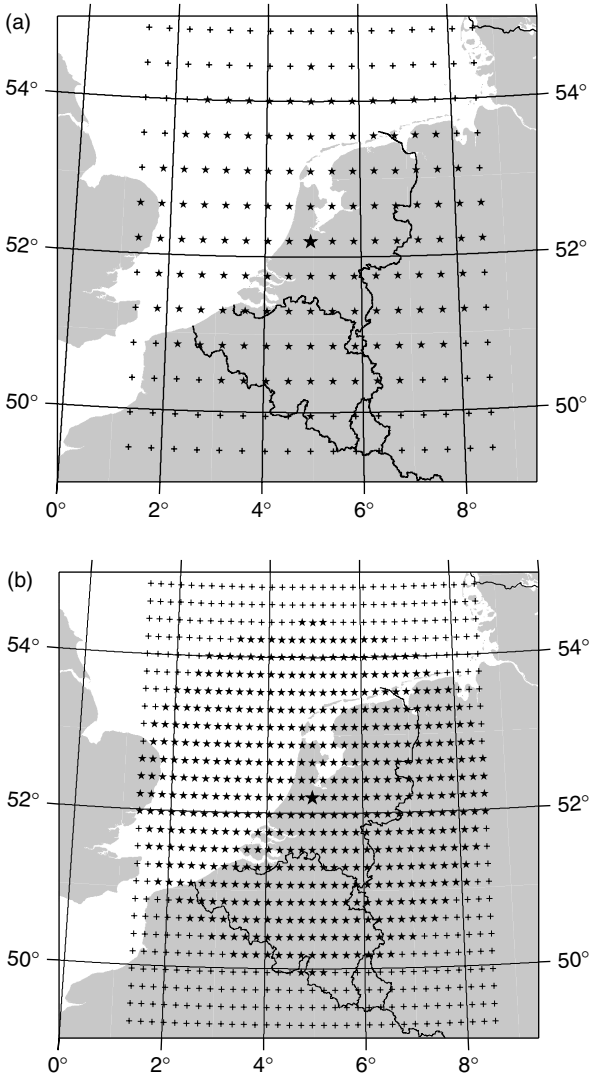

Figure 3. Grids of the low (a) and high-resolution model (b) on which the predictors are calculated. Only the points indicated by asterisk are used. Near the centre these grids resemble the Gaussian grids of the EPS and of the operational model of ECMWF respectively. Station De Bilt is indicated by a large asterisk.

DOI: 10.1002/met

derived from precipitation fields that may be of relevance for exceedance probabilities at the observing site. A statistical analysis is performed on the total data set and yields a set of selected predictors that 'explains' the occurrence of the event as best as possible. This does not mean of course that predictors not selected are not good predictors, but it simply means that they do not provide additional information over the combination of the selected ones.

The potential predictors can be divided into the following four categories:

The central grid box value. This is usually referred to as the DMO.

Predictors regarding general features of the forecast field. A first predictor is the vicinity of predicted precipitation to the central gridpoint. This is calculated by the distance to the first gridpoint with 3-h precipitation. In addition, the extent of the rain area is considered (if there was rain predicted in the central gridpoint) and the square root of the total amount of precipitation in this area. The distance to the first dry gridpoint determines the extent. The square root is taken because precipitation forecasts tend to be non-normally distributed. Sometimes even the cube root or the fourth root is used (Hamill *et al*., 2004).

Predictors defined on different-sized neighbourhoods. Five neighbourhoods are considered defined as circular areas with radii of 50, 100, till 250 km. These steps are approximately equal to the grid spacing of the lower resolution model (control). The maximum extent of the neighbourhood is limited to 250 km because it is expected that at greater distances from De Bilt, forecasts do not bear skilful (additional) information. The predictors that are considered on each neighbourhood are the mean precipitation, the square root of the maximum precipitation amount and the fraction of gridpoints with precipitation.

Distance-weighted predictors. A number of predictors are constructed by weighing the forecast precipitation with distance to the central grid box. This construction rates the general notion that a large precipitation amount close to the predictand location probably may yield a higher probability of occurrence than the same amount predicted further away from the station. Or likewise, a small precipitation amount predicted close to the predictand location might have the same effect as a larger amount predicted at larger distances.

The area on which these predictors are calculated is a circular area with radius of 250 km. On this area, the weighted average and the weighted maximum precipitation (anywhere within the area) is calculated. Three weighting functions have been applied, one linear and two exponential ones, all going to zero at 250 km distance.

In addition, all these weighted predictors were also calculated on elliptical regions oriented in an east–west direction with a width of 250 km in the east–west and 125 km in the north–south direction. The idea behind this is that most of the precipitation in The Netherlands comes from the west and positional errors in the forecast are therefore larger in the east–west direction.

All the above predictors are calculated on the output of both models. There is only one additional set of predictors that can only be calculated on the high-resolution grid. This set of predictors is defined on a neighbourhood of 25 km radius and is the same as on the other neighbourhoods. This set is included to investigate whether the smallest scales not present in the lower resolution model bear additional information. All other predictors are calculated exactly in the same manner on the two models and therefore differences in resolution are included in the predictors only implicitly.

In the experiment, only information that is contained in the precipitation fields at verification time is considered. This was done to mimic the way a forecaster subjectively uses a predicted precipitation field. Furthermore, in this manner it is made to resemble the way fuzzy verification is usually performed.

This approach does not use all the information contained in the model output that is relevant to observed precipitation. Additional information comes from instability measures derived from the model's vertical structure, from (larger scale) circulation features like convergence and from forecasts for different verification times. These are therefore generally included in objective probability forecast equations.

3.3. Selection procedure

The predictors are selected using logistic regression with a forward stepwise selection method (Brelsford and Jones, 1967; Wilks, 2006). At each step, a predictor is chosen that produces the best regression in conjunction with the predictors chosen on previous steps; thereby, a significance threshold of 0.05 is specified. Each chosen predictor is kept in the equation unless the specified significance threshold of 0.10 is exceeded at a following step. The regression coefficients are determined using the maximum likelihood method. This is an iterative method that maximizes the product of all computed probabilities of the (non) occurrence of the event in the dependent data set. Overfitting is avoided as much as possible, e.g. by limiting the number of predictors selected from a single predictor category defined in Section 3.2.

3.4. Results

Although it is often said that the small-scale information from high-resolution forecasts is very valuable it is hard to show that in ordinary verification (scores). This is again illustrated by comparing the central grid box values of the control and operational forecasts against station observations at De Bilt. In Figure 4, this is done in terms of the root mean square (RMS) error but other metrics show the same behaviour. The smaller scale operational model performs much less than the control forecasts even at the earlier forecast ranges. The large fluctuations of the RMS error from one lead time to the next are strongly

DOI: 10.1002/met

related to differences in observed frequencies. The misses of the higher precipitation events dominate the RMS errors. The high-resolution model suffers more from the double penalty than the low-resolution model.

The two central grid box forecast values can be regarded as dichotomous (or categorical) forecasts, i.e. an event will or will not happen. In Figure 5, the results are shown for the events of having any precipitation (Figure 5(a)) and of having at least 4 mm of precipitation in 3 h (Figure 5(b)). For the lowest threshold there is hardly any difference between the two models, whereas for the relatively extreme case the operational model is clearly much worse. This is true for the shorter lead times already. Once again the small-scale information of the high-resolution model deteriorates the skill in terms of deterministic (or categorical) verification scores. Note the large dependence on the time of day with highest values during daytime. This is probably related to the higher frequency of convective events.

The apparent smaller skill for the high-resolution DMO completely vanishes after statistical post-processing has extracted as much information as possible from the precipitation forecast fields. This is shown in Figure 6 in which the BS is presented for the probabilities of exceeding each of the four thresholds. The operational model seems to contain at the least the same amount of skilful information over most of the lead times for all thresholds. A possible exception may be the shorter lead times at the two highest thresholds. This may be due to spin-up problems and it might also be due to the fact that the analysis has failed to include the most suitable predictor(s). The BSs are of course much smaller than in the categorical case as can be seen in Figure 5 for the corresponding thresholds. Note that the predictive potential of the post-processed forecast equations has not been tested on independent data. Exceedance of the highest threshold of 4 mm in 3 h is a rather extreme event in The Netherlands (occurring in the data set in about 3% of the cases) and therefore the statistical significance of the results for this threshold is less than for the other thresholds.

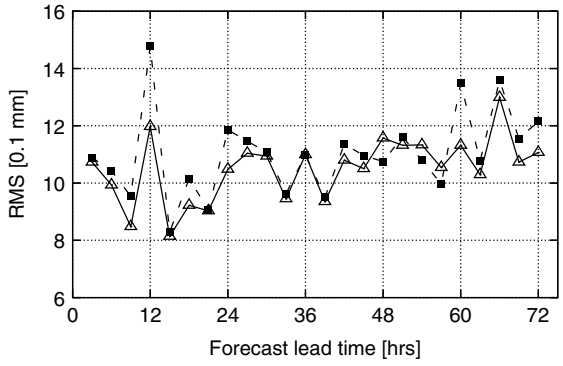

Figure 4. Root mean square difference as a function of lead time between the observed precipitation at De Bilt and the direct model output (the central gridpoint value) of the operational (dotted line, closed squares) and control forecast (full line, open triangles).

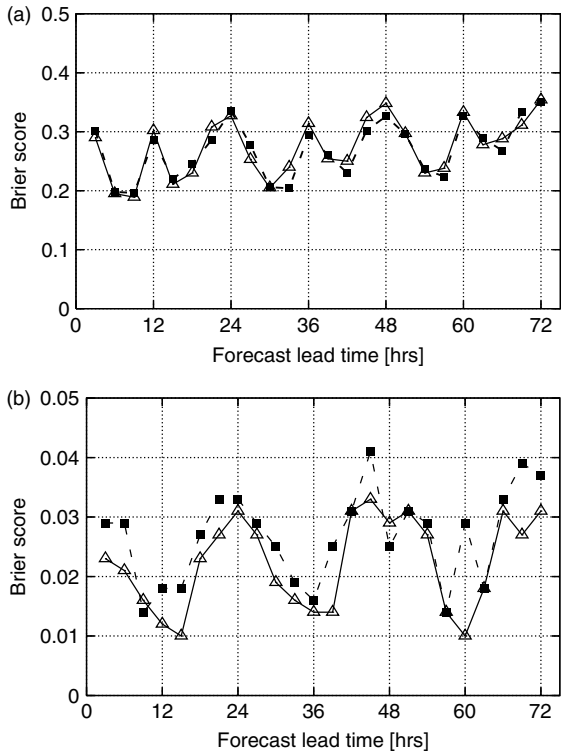

Figure 5. Brier score as a function of lead time for the categorical DMO forecasts of the operational (dotted line, closed squares) and control model (full line, open triangles) for the dichotomous events with thresholds (a) >0 mm and (b) ≥4 mm of precipitation in 3 h at station De Bilt.

The number of predictors selected by the regression is in the order of two to four. There is a considerable variation in the specific choice due to the fact that a large number of highly correlated expressions of the predictors are included. However, a few tendencies can be observed. First of all, for both of the models the central grid box value was never selected, not even in the first step of the selection procedure. This means that the DMO is not the best indicator of precipitation accumulation exceedances, not even at the earlier lead times.

In all cases a combination of predictors from at least two of the other categories is selected. Wider neighbourhoods become important with increasing forecast lead time, in accordance with a decrease in deterministic predictability. Predictors on the largest defined neighbourhood are increasingly important beyond lead times of 48 h. This is in agreement with findings of Theis *et al.* (2005) who constructed from deterministic +48-h forecasts an ensemble of forecasts by taking into account also precipitation amounts that were predicted at gridpoints in a neigbourhood of the observation. Verification of their ensemble on three different neighbourhoods showed that the best results were obtained on the largest neighbourhood of about 140 km.

There is also a slight shift from distance-weighted predictors to predictors defined on neighbourhoods with increasing lead time indicating that precipitation predicted

DOI: 10.1002/met

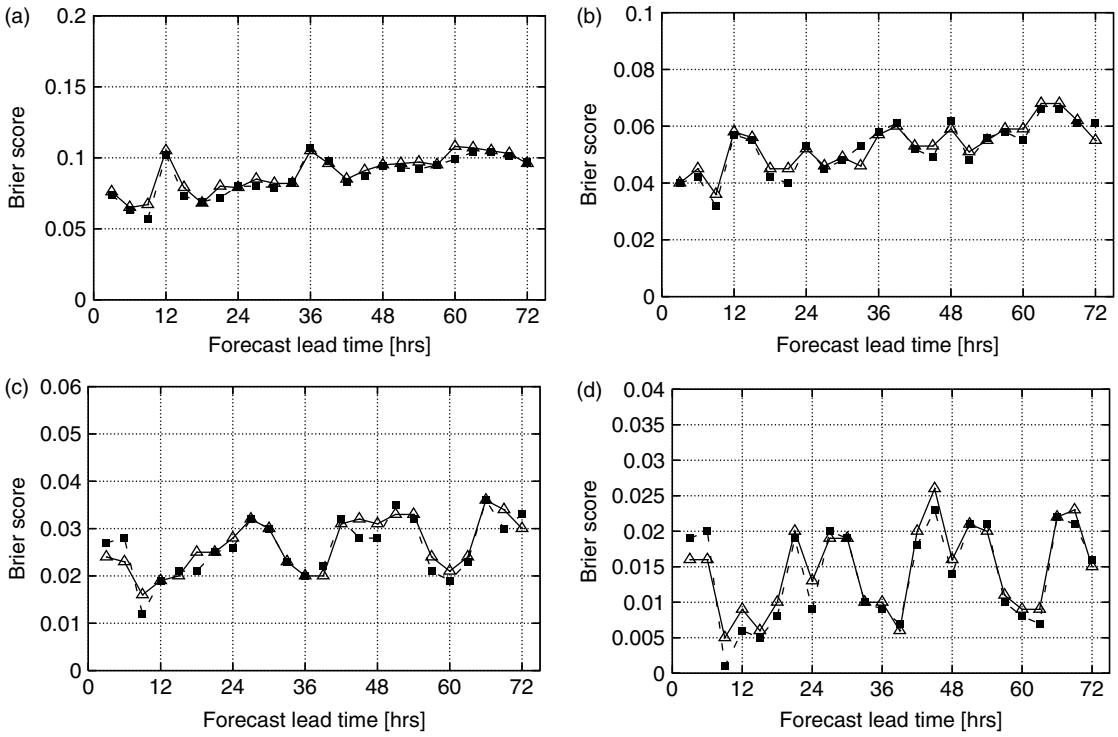

Figure 6. Brier score as a function of lead time for the post-processed forecasts of the operational (dotted line, closed squares) and control model (full line, open triangles) for 4 dichotomous events with thresholds (a) >0 mm, (b) ≥1.0 mm, (c) ≥2.5 mm and (d) ≥4 mm in 3 h at station De Bilt.

anywhere within a neighbourhood becomes more important than the closeness to the observing station. No distinction can be made between the importance of predictors defined on circular areas and those on elliptical areas. Finally, the explicit high-resolution predictors defined on a 25 km circular area are never selected, so all differences between post-processed operational and control forecasts are due to differences in the information contained in the predictors with the same formulation.

4. Discussion

A probabilistic interpretation of DMO from NWP systems is an essential step in creating valuable information from mesoscale forecasts. This is due to the uncertainty that is an inherent trait of the weather and it is due to the uncertain nature of decision processes of the end user, in terms of the precision with which these processes are known, the undetermined variation among users in the sensitivity of their decisions to weather conditions and their willingness to comply with the outcome of a cost-loss analysis (Roulston and Smith, 2004). This is particularly true of precipitation forecasts. Indeed, the analysis of the real world example presented shows that DMO at an observation point does not feature as a predictor for precipitation thresholds.

Consequently the evaluation of the (added) value of high-resolution forecasts should use probabilistic verification techniques.

Fuzzy verification methods in one way or another attempt to take into account the probabilistic nature of precipitation forecasts by using all precipitation amounts that are predicted in the vicinity of an observation. Usually the extent of the neighbourhood is not objectively determined. Sometimes the fraction of grid boxes inside a predefined neighbourhood with predicted precipitation greater than a given exceedance threshold is interpreted as probability of exceeding the threshold at an observing site (Theis *et al.*, 2005). In scale-recursive methods, varying sizes of neighbourhoods are used to determine the scale dependence of forecast quality. The choice of verification method, scoring rule and most importantly the interpretation and weighting of verification results require a good understanding of the implicit statistical assumptions inherent in the verification method and the intended application of the forecasts and the verification results. As such fuzzy methods depend on skill and subjective preference.

Using MOS on a real world example of precipitation at an observation site has demonstrated the possibility of assessing objectively the predictive potential of high-resolution models in a comparative verification with low-resolution models. This confirms the advantage of probabilistic interpretation and verification over deterministic model output verification that is found in a constructed theoretical example. A large number of quantities resembling those that are used in fuzzy verification are used as predictors in the given example. This shows MOS to be an essential extension of fuzzy methods. The MOS method objectively weighs and combines the effect of general features of the forecast field, the size of neighbourhoods and the effect of distortion

DOI: 10.1002/met

errors on forecast quality. Predictors other than the precipitation field itself may contain useful information on the exceedance probability of rainfall thresholds (Lemcke and Kruizinga, 1988; Glahn *et al*., 1991). Such predictors were excluded from the presented analysis. Nevertheless the precipitation field itself contains sufficient information to demonstrate the value of a resolution increase on precipitation forecasts.

The real world example uses forecasts on 25- and 50 km grids. The resolution of advanced mesoscale models nowadays is much higher, in the order of 1 km. The challenge of mesoscale verification is to value information at this resolution. The limited deterministic predictability of weather phenomena at this scale and thus the need for probabilistic verification is an even bigger issue than it is at the meso-β scale. There are no principal objections to applying the methodology of this article to kilometre-scale forecasts. The application of MOS is not restricted to station forecasts. When MOS is applied to areal forecasts the availability and conditioning of verifying observations increasingly becomes a problem as resolution goes up.

The present analysis is based on a smaller data set than would be used in operations. The analysis has been economical with the data by not testing the forecast equations on independent data. However, this lack of normal rigor should not affect the main conclusions.

To get the most from MOS both in forecast mode and in a comparative valuation of information content can be a laborious task. It requires a wider search for relevant predictors in the spatio-temporal domain, data stratification to incorporate seasonal variation of the relationships between predictors and predictand, the pooling of stations to achieve statistical significance especially in the case of rare predictands and, last but not least, a thorough testing of the derived forecast equations on independent data. Even then there is no mathematical proof that the selected predictors are in any sense the optimal description of the information in the model. The same, however, may be said of any form of forecast interpretation or post-processing.

In addition, the application of MOS usually requires a large set of data, preferably from a model with a formulation and resolution that is unchanged over the training period. For rare events such as exceedances of large precipitation thresholds and for predictands with a strong seasonal dependence training periods of several years are required. Operationally several techniques are employed to overcome these drawbacks: predictor selection on pooled station or spatial data (Glahn *et al*., 1991; Schmeits *et al*., 2005), updating schemes, such as recursive regression, Kalman filtering and weighted blending of old and new model data (Wilson and Vallée, 2002) and reforecasting with a new model (Hamill *et al*., 2004). Probabilistic forecast equations have to be developed for each predictand, i.e. for each station or area, threshold, accumulation period and forecast range.

These characteristics of the MOS scheme seem unappealing if it is used purely for the purpose of verification, in particular in the case of the introduction of a new model. The basic premise behind this article's advocacy of this method for verification is that MOS is already available as the operational scheme for producing probabilistic forecasts. Put alternatively, it is the way information is extracted from deterministic model data and using it as a method to evaluate this information then seems only natural. The requirement of statistical significance and thus the requirement of large sets of training data are explicit in the design of a MOS scheme. These are just as much requirements of any form of forecast evaluation, be it subjective or objective. Not realizing this, or even ignoring this, may lead to large verification data sets, but little information on forecast value.

5. Conclusion

In this article it has been demonstrated that probabilistic verification of mesoscale forecasts, based on the same statistical post-processing that is used to derive probabilistic information from deterministic NWPs, is able to capture the added value of increased model resolution, in contrast to traditional deterministic verification.

This concept has been explored in a theoretical example and its feasibility has been demonstrated in a real world application to the prediction of precipitation at a station. This application uses the general method of MOS. The analysis has made use of fuzzy verification concepts such as neighbourhoods and distance weighting as predictors. It was found that MOS is able to combine and weigh these predictors objectively. The argument has been put forward that therefore it is an essential extension to fuzzy verification methods.

The finding that DMO is not selected by the regression as a predictor can be related to the probabilistic nature of precipitation as a prime example of the limited predictability of small-scale weather and supports the contention that probabilistic forecasting and subsequently probabilistic verification are essential in the extraction and valuation of mesoscale forecasts.

The discussion of the practical implications, the merits and drawbacks of applying MOS to mesoscale verification has led to the conclusion that the considerable effort required to implement MOS for verification essentially comes free if MOS is already used as a forecasting tool. Moreover, this effort is intrinsic to all mesoscale verification, if its aim is not just to produce vast amounts of verification data, but a proper valuation of the information contained in mesoscale forecasts.

Acknowledgements

The authors thank Maurice Schmeits, Seijo Kruizinga and Gerrit Burgers for their helpful comments on an earlier version of this article. Also we thank all anonymous reviewers for their careful revisions, which have helped to improve the manuscript.

DOI: 10.1002/met

References

Baldwin ME, Lakshmivarahan S. 2003. Development of an events-oriented verification system using data mining and image processing algorithms. *Preprints, Third Conference on Artificial Intelligence Applications to Environmental Science*. AMS: Long Beach, CA, CD-ROM, 4.6.

Brelsford WM, Jones RH. 1967. Estimating probabilities. *Monthly Weather Review* **95**: 570–576.

Brier GW. 1950. Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review* **78**: 1–3.

Briggs WM, Levine RA. 1997. Wavelets and field forecast verification. *Monthly Weather Review* **125**: 1329–1341.

Callies U. 2000. Comparative forecast evaluation: graphical Gaussian models and sufficiency relations. *Monthly Weather Review* **128**: 1912–1924.

Casati B, Ross G, Stephenson DB. 2004. A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorological Applications* **11**: 141–154.

Cherubini T, Ghelli A, Lalaurette F. 2002. Verification of precipitation forecasts over the Alpine region using a high-density observing network. *Weather and Forecasting* **17**(2): 238–249.

Davis CA, Brown B, Bullock R. 2006a. Object-based verification of precipitation forecasts. Part I: methodology and application to mesoscale rain areas. *Monthly Weather Review* **134**: 1772–1784.

Davis CA, Brown B, Bullock R. 2006b. Object-based verification of precipitation forecasts. Part II: application to convective rain systems. *Monthly Weather Review* **134**: 1785–1795.

Du J, Mullen SL, Sanders F. 2000. Removal of distortion error from an ensemble forecast. *Monthly Weather Review* **128**: 3347–3351.

Ebert EE. 2008. Fuzzy verification of high resolution gridded forecasts: a review and proposed framework. *Meteorological Applications* **15**: 53–66.

Ebert EE, McBride JL. 2000. Verification of precipitation in weather systems: determination of systematic errors. *Journal of Hydrology* **239**: 179–202.

Glahn HR, Lowry DA. 1972. The use of Model Output Statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology* **11**: 1203–1211.

Glahn HR, Murphy AH, Wilson LJ, Jensenius JS. 1991. *Lectures and Papers Presented at the WMO Training Workshop on the Interpretation of NWP Products in Terms of Local Weather Phenomena and their Verification*. Wageningen: The Netherlands.

Hamill TM, Whitaker JS, Wei X. 2004. Ensemble re-forecasting: improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review* **132**: 1434–1447.

Hoffman RN, Liu Z, Louis J-F, Grassotti C. 1995. Distortion representation of forecast errors. *Monthly Weather Review* **123**: 2758–2770.

Katz RW, Murphy AH. 1997. *Economic Value of Weather Forecasts*. Cambridge University Press: Cambridge; 222.

Kok CJ. 2002. Using probabilities as a means to evaluate high resolution forecasts. SRNWP Mesoscale Verification Workshop 2001, De Bilt, The Netherlands, 68–72.

Lemcke C, Kruizinga S. 1988. Model output statistics forecasts: three years of operational experience in The Netherlands. *Monthly Weather Review* **116**: 1077–1090.

Marsigli C, Montani A, Nerozzi F, Paccagnella T, Tibaldi S, Molteni F, Buizza R. 2001. A strategy for high-resolution ensemble prediction. II: limited-area experiments in four Alpine flood events. *Quarterly Journal of the Royal Meteorological Society* **127**: 2095–2115.

Mittermaier MP. 2006. Using an intensity-scale technique to assess the added benefit of high-resolution model precipitation forecasts. *Atmospheric Science Letters* **7**(2): 35–42.

Mittermaier MP. 2007. Using time-lag ensemble techniques to assess the behaviour of high-resolution precipitation forecasts. *Geophysical Research Abstracts* **9**: 08457.

Molteni F, Buizza R, Marsigli C, Montani A, Nerozzi F, Paccagnella T. 2001. A strategy for high-resolution ensemble prediction. I: definition of representative members and global-model experiments. *Quarterly Journal of the Royal Meteorological Society* **127**: 2069–2094.

Murphy AH. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology* **12**: 595–600.

Murphy AH. 1977. The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review* **105**: 803–816.

Murphy AH. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting* **8**: 281–293.

Nachamkin JE. 2004. Mesoscale verification using meteorological composites. *Monthly Weather Review* **132**: 941–955.

Nachamkin JE, Chen S, Schmidt J. 2005. Evaluation of heavy precipitation forecasts using composite-based methods: a distributions-oriented approach. *Monthly Weather Review* **133**: 2163–2177.

Orrell D, Smith L, Barkmeijer J, Palmer TN. 2001. Model error in weather forecasting. *Nonlinear Processes in Geophysics* **8**: 357–371.

Roberts NM, Lean HW. 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review* **136**: 78–97.

Roulston MS, Smith LA. 2004. The boy who cried wolf revisited: the impact of false alarm intolerance on cost-loss scenarios. *Weather and Forecasting* **19**: 391–397.

Schmeits MJ, Kok CJ, Vogelezang DHP. 2005. Probabilistic forecasting of (severe) thunderstorms in the Netherlands using model output statistics. *Weather and Forecasting* **20**: 134–148.

Skamarock WC. 2004. Evaluating mesoscale NWP models using kinetic energy spectra. *Monthly Weather Review* **132**: 3019–3032.

Theis SE, Hense A, Damrath U. 2005. Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorological Applications* **12**: 257–268.

Thompson JC. 1962. Economic gains from scientific advances and operational improvements in meteorological prediction. *Journal of Applied Meteorology* **1**: 13–17.

Tustison B, Harris D, Foufoula-Georgiou E. 2001. Scale issues in verification of precipitation forecasts. *Journal of Geophysical Research* **106**: 11,775–11,784.

Wilks DS. 2006. *Statistical Methods in the Atmospheric Sciences*, 2nd edn. Academic Press; 627.

Wilson LJ, Vallée M. 2002. The Canadian updateable model output statistics (UMOS) system: design and development tests. *Weather and Forecasting* **17**: 206–222.

Zepeda-Arce J, Foufoula-Georgiou E, Droegemeijer KK. 2000. Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *Journal of Geophysical Research* **105**(D8): 10,129–10,146.

DOI: 10.1002/met