From MOS to eMOS:
Generalising Model Output Statistics for Full Ensemble Forecasts

Devin ***Kilminster****, Liam* ***Clarke****, Mark* ***Roulston****, Christine* ***Ziehmann****, Jochen* ***Brocker****, Leonard* ***Smith***
Pembroke College, Oxford, OX1 1DW, United Kingdom, kilminst@maths.ox.ac.uk
www.lsecats.org

Abstract

It has long been known that the information content of weather forecasts extends beyond the model-variables that share the same name as the forecast target-variables of interest. Traditional model output statistics (MOS) algorithms extract information from any model-variable deemed relevant to estimating a given target-variable, especially when the "corresponding" model-variable, taken at face value, forecasts "poorly". Mathematically, this form of MOS can be seen as adopting a "projection" operator between model-state space and observations that is more complex than the identity operator. Ensemble forecasts allow the introduction of a new twist. Typically, one treats each individual ensemble member as a viable scenario, projecting it into observation space as a (dressed) forecast, and then combining all ensemble members; an alternative approach is to condition the probability forecast of the target value upon properties of the joint distribution of all the ensemble members (in a potentially multi-model ensemble). Thus eMOS goes beyond MOS in that it not only aims to locate information in each individual model run, but also considers the ensemble as a whole, not merely as a collection of scenarios. The approach is illustrated in precipitation forecasts, and more general interpretations relevant to THORPEX's core aims are noted.

1. Introduction

Models are not identical to reality; precipitation in a model, for example, is just not the same thing as real-world precipitation. There is a need, then, for interpretation of the results of numerical weather prediction. This need for interpretation is well known in the case of the correspondence between model-variables and the forecast target-variables: The practice of "bias correction" is widespread. Techniques such as model output statistics (MOS) (Glahn and Lowry 1972) go further, building a statistical relationship between the target-variable of interest, and any number of model-variables felt to provide relevant information to prediction of the target variable.

The purpose of this paper is to draw attention to another issue of interpretation: With the increasing availability of results from ensemble prediction systems, we are faced with not merely the problem of interpreting the model variables within each ensemble member, but also with the question of the appropriate interpretation of the variations *between* ensemble members. The ensemble members are a sample drawn from the ensemble generating process, but what relationship exists between this process and our real-world uncertainties?

We will consider the interpretation of ensembles in the realm of probabilistic forecasts, that is, the forecast should provide probabilistic information (say a distribution) for the target-variables. This question can be explored through the consideration of a simple taxonomy of ensemble interpretation methods. In the next section, we will look at these in the context of a simple precipitation prediction problem.

1. Direct Interpretation. Model-variables are treated as corresponding directly to target-variables. Furthermore, the distribution provided by the ensemble is treated as providing the appropriate forecast-distribution — if (for example) model-precipitation occurs in 50% of the ensemble members, then the probability forecast for (the target variable) precipitation is also 50%.
2. Scenario MOS. The direct correspondence between model and target variables is dropped. Ensemble members are assumed to correspond (individually) to plausible probabilistic *scenarios*. Quite a degree of interpretation can go into generating each scenario – a historical dataset can be used to produce a relationship between model-variables in each member and the appropriate corresponding scenario, hence "MOS". The

weightings used to combine the scenario distributions into the forecast distribution need not be constant – for example, the weight given to the control may differ from that given to the rest of the ensemble. When there is a symmetry between members, however, (say we consider the ECMWF ensemble *without* the control) it may be reasonable to give the individual scenarios equal weighting. This is the case we consider in what follows.

3. Ensemble MOS (eMOS). Any ensemble interpretation method can be viewed as a function from the information present in the ensemble to the probabilistic forecast for the target-variable. Viewed from this perspective, methods 1 and 2 encompass only a small proportion of such functions.

In the next section, we will see that the ability to make the forecast depend upon the ensemble taken as a whole can have definite advantages.

2. Simplified Example

We will examine the three approaches to ensemble interpretation in the context of the prediction of 12 hour precipitation measured at a certain weather station (specifically, Helgoland; WMO10015), at given lead times. In this section, we consider a vastly simplified version of this problem — we predict the (target) probability of precipiation greater than 0 mm using only the number of members in the ECMWF ensemble (excluding the control) in which the (model) precipitation fell above 0 mm. This simplification is illuminating as it allows us to, relatively easily, determine the optimal application of each approach — thus we will not be comparing "strawmen". We will return to more realistic versions of the problem in the next section.

For all results in this paper, we show the performance of our predictions over the test period 1 May 1998 to 31 April 1999. Where a training period is required, we used only historical data between 1 May 1997 and 31 April 1998. To evaluate the performance of predictions, we use the *ignorance* (Roulston and Smith 2002), a skill score for probabilistic forecasts defined by the negative logarithm (in base 2, say) of the predicted probability for the actual outcome. Thus, if a 75% chance of precipitation is predicted, and precipitation actually occurs, we score –log(0.75), or approximately 0.415; if precipitation did not occur, we would score –log(0.25)=2 — smaller ignorance is better. Ignorance should be considered as a relative score — we will always consider the difference in ignorance between two forecasts. (One of the forecasts might be "climatology", in Figures 1 and 4, for example, the performance of climatology defines the zero-line.)

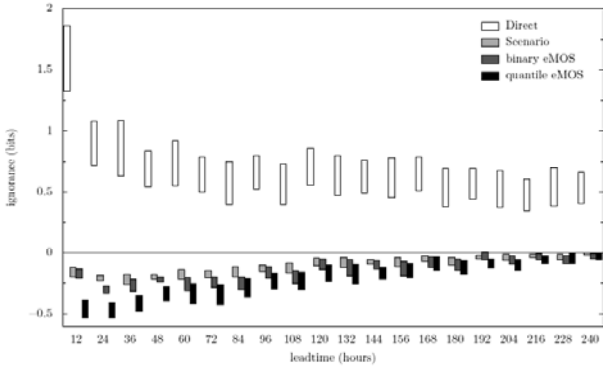

Figure 1. Comparison of the ignorance of different ensemble interpretation methods with respect to climatology over the test period for different leadtimes. The intervals shown are plus and minus one standard deviation of the bootstrap (Efron and Tibshirani 1986) average ignorance relative to climatology. Lying above the zero-line indicates performance worse than climatology, lying below indicates better performance.

In Figure 1, we see the performance of the direct interpretation of the ensemble with respect to climatology. Climatology takes the proportion of observed "wet" instances in the training period as the predicted probability of precipitation.

An examination of Figure 1 indicates that the direct interpretation of the ensemble is not only significantly worse than climatology, but the forecasts exhibit "anti-skill" — performance with respect to climatology actually *improves* with leadtime. One would reasonably expect that precipitation would generally become harder to predict rather than easier at longer leadtimes. This merely confirms that model-precipitation is just not the same as real precipitation, and that a more sophisticated interpretation is required.

Under the scenario interpretation of an ensemble, one maps each individual member into a probabilistic forecast of the target, and then combines these into a final forecast. In our simplified setting, the possible scenario interpretations are very limited. We have only two types of ensemble members ("model-wet" and "model-dry"), and we wish only to forecast a binary event ("target-wet" or "target-dry"). The scenario interpretation is completely determined by just two parameters — the forecast probability for target-wet given model-wet and the forecast probability for target-wet given model-dry. These are the parameters of the "scenario MOS". For a given leadtime, these parameters may be fit in order to minimise ignorance over the training period. Again, Figure 1 shows the results: Performance is now generally better than climatology, and (sensibly) prediction becomes generally more difficult with increasing leadtime.

It is important to note that in this simplified setting, we have been able to cover, with our two parameters, essentially every possible scenario interpretation. In a more realistic setting (such as will be considered in the next section), we would have only been able to cover a subset of the possible interpretations. The results shown in Figure 1, therefore, represent the best that can be expected of the scenario interpretation in this case. It is natural to ask, then, if by relaxing the assumptions of the scenario interpretation even better performance might be possible?

If we assume that the ordering of the members of the ensemble contains no predictive information, then the relevant information in the ensemble can be summarised by a single integer in the range 0 to 50 — the number of model-wet members. *Any* interpretation is then just a function from the number of model-wet members (or equivalently the proportion) into the predicted probability for target-wet. The direct interpretation (viewed as a function on the proportion of model-wet members) is essentially the identity. The scenario interpretations can also be simply described: Let *x* be the number of model-wet members of the ensemble, *a* be the scenario-MOS parameter for the predicted probability of target-wet given model-wet, and b be the parameter for target-wet given model-dry. Then the scenario forecast for the probability of target-wet is $xa + (50 - x)b = (a - b)x + 50b$. This is just a linear function. Requiring linearity is a very strong constraint on a function. The idea of eMOS is to appropriately allow a wider class of functions (interpretations). By breaking the scenario assumption, eMOS, is able to interpret the ensemble as a whole. In this simplified setting, this is simply to allow the use of a non-linear function between the ensemble summary and the probability of precipitation.

What class of non-linear functions should we pick from? It is possible to parameterise all the functions in the simplified setting, but it requires 51 parameters, and there is simply is not enough training data to fit them all. Instead, we content ourselves with choosing a smaller set of functions with only a few parameters. Unlike the case with scenario-MOS, we are not guaranteed to cover all possible variations of eMOS. Our aim here is merely to show eMOS sometimes outperforms the scenario approach. Rather than discuss the most appropriate functional forms for eMOS, we make three observations:

1 The interpretation will not depend upon the ordering of members in the ensemble.
2 The function should be reasonably "smooth": small changes in the ensemble should not make large changes in the forecast.
3 The function should be monotonic: as the number of model-wet members increases, the forecast probability of target-wet should not decrease.

In any case, we fit the parameters of our chosen eMOS by minimising its ignorance across the training period. (An alternative method is presented in Roulston et. al. (2001), in which the function is buit up by the method of "analogues" — the proportion of historical cases "near-by" the current case is used as the function value. This method does not guarantee monotonicity.) A typical result is shown in Figure 2. Here we plot the functions corresponding to the the three types of interpretations for a leadtime of 108 hours. One can clearly see the non-linear nature of the eMOS interpretation, and the scenario interpretation can be fairly easily imagined to be the "best" linear approximation to the non-linear eMOS. Clearly the direct interpretation is far from optimal.

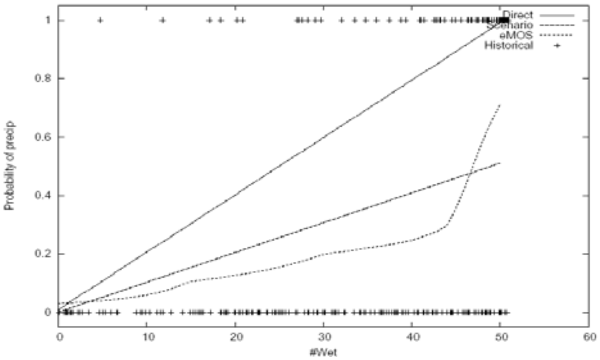

Figure 2. Plots of functions corresponding to the three types of interpretations for a leadtime of 108 hours. The direct interpretation corresponds to the identity, the scenario interpretation is linear, and the eMOS interpreation is non-linear. Also plotted are the historical cases in the training period, their position being given by the number of historically wet ensemble members — at the top of the plot are the cases that were actually wet, the dry cases are on the bottom.

The performance of eMOS with respect to climatology is displayed in Figure 1 — it appears eMOS is out-performing the scenario interpretation. To be sure we should directly compare the two methods. Such a comparison is provided by Figure 3. For most leadtimes, the best scenario interpretation performs significantly worse than the eMOS.

3. Using More Information

In the preceding section, only the binary information of model-wet or model-dry was used. Once this is extended to using the model-*amount* of precipitation in each member, it is no longer possible to summarise the ensemble by a single number: the ensemble can now be viewed as a one-dimensional distribution of model-precipitations. This could be approximately summarised by the selection of a number of its quantiles, say the 10%, 50%, and 90%. We now build our *quantile eMOS* as a function from these 3 quantities into the predicted probability for target-wet. We fit by minimising ignorance on the training period. The results are shown in Figures 1 and 3. The quantile eMOS generally outperforms the eMOS of Section 2, especially for leadtimes of up to about 3 days.

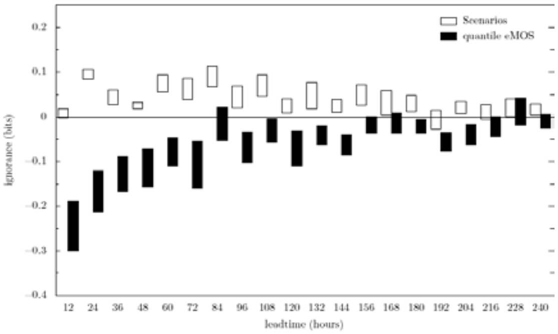

Figure 3. Comparison of the scenario interpretation and the quantile eMOS of Section 3 versus the eMOS of Section 2. Plotted are bootstrap intervals for the ignorances relative to the eMOS of Section 2 — the zero line represents the performance of this eMOS.

4. Other Thresholds

Figure 4 displays the result of applying the quantile eMOS to the forecasting of a range of other thresholds. Generally, some skill over climatology is maintained even for target-thresholds of up to 10 mm, although as the threshold increases skill tends to decrease — it should be pointed out that only about 6 instances of precipitation at a rate greater than 10 mm in 12 hours occur in the training period. Better performance might be expected here if a larger historical archive of comparable quality was available.

5. Conclusions

There is useful information in ensembles that can be exploited only by interpreting the ensemble as a whole, rather than as a sampling of individual scenarios. This fact has implications for the design and use of operational ensemble prediction systems.

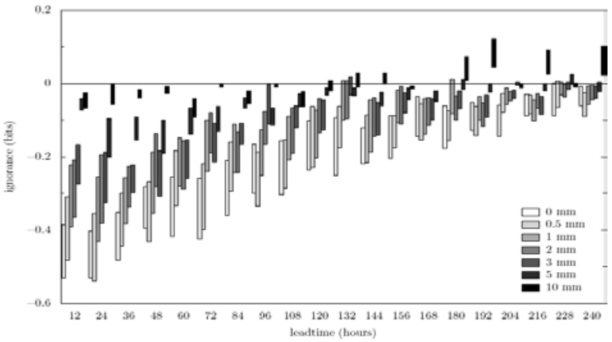

Figure 4. Ignorance of quantile eMOS vs climatology for a range of thresholds.

As with any statistical procedure, the quality of the training set is of paramount importance, the quality of the results are limited by the number of examples of the event one wishes to predict in the training set. This argues in favour of the generation of a large forecast-verification archive for any operational ensemble forecasting system to better meet THORPEX goals of socio-economic application.

Future work will involve extension of these methods to the interpretation of multiple sources of forecast information, and the consideration of targets of more direct economic importance. Also, by combining information from a number of different threshold targets (such as were produced in Section 4), more continuous forecasts of targets such as temperature could be made. In general, interpretation of the joint distribution of multi-model multi-initial condition ensembles could lead to a significant increase in the information content of current forecast products.

Acknowledgement:
Support for this work was provided by US Weather Research Program funds administered by the US National Oceanic and Atmospheric Administration. We also thank Deutscher Wetterdienst for the provision of the data used in this study.

References:

Efron, B., Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science 1*, 54-77, 1986.

Glahn, H. R., Lowry, D. A. The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor. 11*, 1203-1211, 1972.

Roulston, M. S., Smith, L. A. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review 130 (6)*, 1653-1660, 2002.

Roulston, M. S., Ziehmann, C., Smith, L. A. A Forecast reliability index from ensembles: A comparison of methods. Tech. Report for Deutscher Wetterdienst, 2001.