Using Model Averaged Probabilistic Forecasts for Water Resources Management

A. Sharma[a] and U. Lall[b]

[a]School of Civil and Environmental Engineering, The University of New South Wales, Sydney, Australia

[b]Earth and Environmental Engineering and the International Research Institute for Climate Prediction, Columbia University, New York, USA

Abstract: A drawback of medium-to-long-term probabilistic forecasting methods is the relatively high uncertainty associated with model outputs, particularly when the models are used for prediction of future scenarios. This paper presents an extension to the probabilistic forecasting approach first presented in Sharma (2000b), that attempts to enhance the reliability of the model using an ensemble averaging approach. Each ensemble member or model is formulated using nonparametric statistical techniques and is restricted to have a relatively independent basis so as to represent the multiple mechanisms that influence the system being studied. The aim of using ensemble or model averaging is to reduce the chance of model misspecification, a common occurrence when the dependence is highly random and the system too complex to be explained by a limited number of predictors.

The usefulness of the procedure is demonstrated through an application to forecast the Southern Oscillation Index (SOI), the multiple models being formulated using predictors selected from prior lags of the SOI and globally distributed, gridded sea surface temperature anomaly data. The model is assessed by evaluating its performance both in cross-validation as well as by forecasting an entire period of the record that was left out in the model formulation process. The results indicate that the consideration of uncertainty in climatological observations and the use of an ensemble of model outputs results in probabilistic forecasts that are more reliable and accurate than is the case otherwise. The implications of using the probabilistic forecasts for water resources management are discussed.

Keywords: *Probabilistic forecasts, Southern Oscillation Index, Model averaging, Water resources management*

1. INTRODUCTION

Increasing water use and limited sources of supply are making water an increasingly precious resource for the world to share. Consequently, various alternatives for managing water are being evaluated and considered for use. This paper discusses one such option, namely, increasing supply reliability using probabilistic streamflow forecasts where knowledge of the climatic factors that drive flows is used to reduce forecast uncertainty. This paper presents a new approach for probabilistic forecasting. This approach is statistical in nature, and uses a system of carefully identified climatic predictor variables in formulating the forecasts. We use the quarterly Southern Oscillation Index time series to evaluate the utility of the probabilistic forecasting approach, using both a leave-one-out cross-validation formulation and a stand alone period that represents a pure forecast whose data is not used in formulating the probabilistic forecasting model.

Medium to long-term prediction methods are either dynamical, or statistically based, or a combination of the two. Dynamical methods attempt to simulate the physics governing the global ocean-atmosphere system in order to predict the state of the climate at the location of interest. In contrast to dynamical approaches, statistical approaches are empirical, using the observed historical record with carefully identified predictors to predict future values. Hence, as long as the right system of predictors have been used in defining the variability of the hydro-climatic variable, an assumption that the past is indicative of the future holds true. As a result, instead of using the physics, statistical models use data that are generally of the same kind that would be used as input for dynamical models, but extend far back in time. For statistical models to give reasonable answers, two main conditions must hold true: (i) the data must be long enough to represent the range of possibilities nature is likely to toss in the future, and (ii) the predictors used in the formulation of the approach must be legitimate, chosen based on a mix of our understanding of the physics that results in climate variability, as well as the empirical evidence that points towards their relevance.

An example of a probabilistic forecasting model used for reservoir inflows in select catchments in Australia is the Nonparametric Probabilistic Forecast Model (Sharma, 2000b). A statistical measure of dependence, the PMI or the Partial Mutual Information criterion (Sharma, 2000a) is used to select a subset of potential predictors. A limitation of such an approach can be its reliance on a handful of predictors, particularly when they are selected from a potentially large set. In such a case, the large number of available choices can lead to predictors that are simply a result of chance. Another problem with such an approach is an implicit assumption that a single model (or equivalently a single set of selected predictors) can attempt to explain the type of variability that is observed. This would be equivalent to assuming that variability in Australian rainfall (say) is a result of the fluctuations in the mid-Pacific sea surface temperatures alone, often used as an indicator of the strength of the El Nino Southern Oscillation (ENSO).

Model averaging (Hoeting et al., 2000) is one way to get around the difficulties noted above. This involves formulating multiple models of the system under study, each model having a basis that is relatively independent of that being used in the other models considered. For instance, one model for predicting Australian rainfall could have a basis in the mid-Pacific region often associated with the ENSO, while another could have a basis in the Indian Ocean that has a prominent impact on rainfall in western and southern Australia. The nonparametric probabilistic forecasting approach presented in later sections makes use of the model averaging rationale outlined above. An application to the prediction of the quarterly Southern Oscillation Index time series is used to illustrate the usefulness of the approach.

2. PREDICTOR SELECTION USING PARTIAL MUTUAL INFORMATION

The partial mutual information (PMI) (Sharma, 2000a) between the dependent variable *y* and the independent variable *x*, for a set of pre-existing predictors ***z***, can be estimated as:

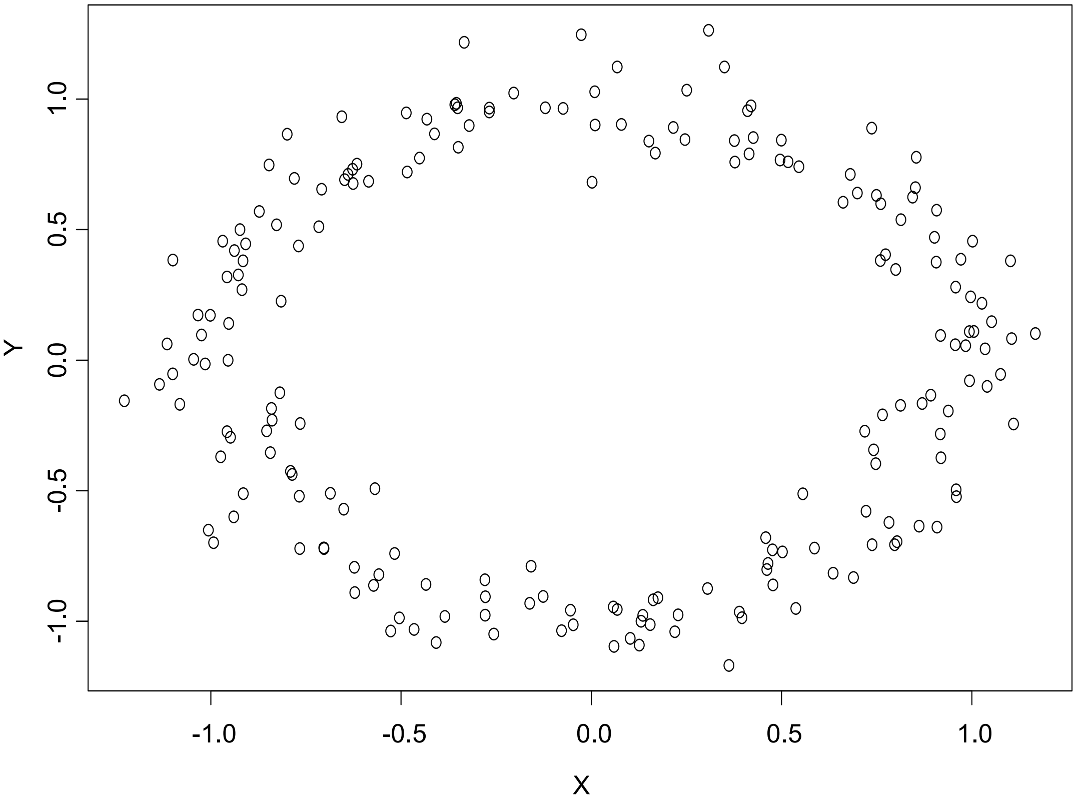

Figure 1 – Synthetically generated sine-wave sample lagged by half phase length. The coefficient of correlation for this sample equals –0.008 whereas the corresponding PIC value is 0.97.

$$PMI = \frac{1}{n}\sum_{i=1}^{n}\log_e\left[\frac{f_{X',Y'}(x_i', y_i')}{f_{X'}(x_i')f_{Y'}(y_i')}\right] \tag{1}$$

where

x'_i and y'_i are the i'th residuals in the sample data set of size n, the residuals being estimated conditional to the pre-identified predictor vector z, and,

$f_X'(x'_i)$, $f_Y'(y'_i)$, and $f_{X',Y}'(x'_i,y'_i)$ are the respective marginal and joint probability densities, estimated using nonparametric kernel density estimation methods as described in Sharma (2000b).

If z represents an empty set, the PMI collapses to the Mutual Information criterion (Fraser & Swiney, 1986).

The PMI can be expressed on a [0,1] scale using the following expression:

$$PIC = \sqrt{1 - \exp(-2 \times PMI)} \tag{2}$$

where *PIC* refers to the Partial Information Correlation, which collapses to the absolute value of partial correlation if all variables involved follow a Gaussian distribution. The usefulness of the above measures is more apparent when the variables are not Gaussian. Consider the relationship between x and y in Figure 1. The two variables are highly dependent, but still record a sample coefficient of correlation equal to -0.008. On the other hand, the *PIC* for the same two variables is equal to 0.97, representing more appropriately the nonlinear nature of dependence that exists between the variables.

An issue that arises when dealing with climate data is representation of unequally uncertain periods in the data. This is more apparent when using sea surface temperature information that was sparsely recorded in the early twentieth century and is measured more accurately using satellite imagery in current times. The PMI was modified to account for such variations in uncertainty, the rationale being to use the standard error associated with each observation as the basis for ascertaining its contribution in calculating the sample estimate. Further details on this modification will be presented at a later date.

3. A NONPARAMETRIC PROBABILISTIC FORECASTING APPROACH

The nonparametric probabilistic forecasting model is formulated as follows:

1. Predictor identification: The PMI criterion is used to select a suitable number of predictors, these predictors being selected from a range of climatic indices as described in the next section. The number of predictors depends on their predictive performance measured in a leave-one-out cross-validation setting.

2. Forecasting approach: The predictors identified in the earlier step are used to formulate a conditional prediction model using nonparametric kernel density estimation methods as has been described in Sharma (2000b). The output from this model is a conditional probability density function estimated based on specified predictor values.

3. Formulation of multiple models – Steps 1 and 2 are repeated a sufficient number of times to formulate multiple conditional prediction models. The rational behind formulating multiple models is to represent the multiple mechanisms that may be introducing variability in the variable being predicted. To ensure that different mechanisms are represented in each constituent model, the cross-dependence between leading predictors for each model as measured by the PMI criterion, is kept below a specified threshold. As a result, the leading predictors, and consequently their successors form relatively independent bases, ensuring that each individual model has a basis that is relatively independent of the other constituents.

4. Estimation of model weights – Once the multiple model constituents have been formulated, the next step is to ascertain the weights that result in the best predictive linear weighted averaged model output. These weights are ascertained based on a constrained optimisation approach, the objective being to maximise the predictive performance of the averaged model in a leave-one-out cross-validation setting. As the weights are constrained to sum to unity, the resulting averaged model output is also a legitimate conditional probability density function.

It should be noted that one of the reasons behind the use of model averaging in the probabilistic forecasting model described above, is to reduce the uncertainty introduced by selecting predictors from a possibly large set of variables. Another reason is to incorporate multiple processes that lead to variability in the predicted variable, as compared to formulating an approach that attempts to represent a single although dominant mode of variability. We illustrate the probabilistic forecasting methodology through an application to forecast the quarterly Southern Oscillation Index next.

4. CASE STUDY – SOUTHERN OSCILLATION INDEX

A quarterly time series of the Southern Oscillation Index was used to illustrate the usefulness of the probabilistic forecasting approach outlined in the previous section. The Southern Oscillation Index data used here represents the period 1866 to 2002. The 1866 to 1992 data was used for developing the model and for ascertaining model performance in a leave-one-out cross-validation setting, while the 1993-2002 period was used to evaluate the performance of the model in a pure-forecast setting. Separate models were formulated for each quarter and for lead times ranging from one quarter (3 months) to four quarters (one year).

The PMI criterion was used to select predictors for the various models from selected climate indices and sea surface temperature anomaly (SSTA) data. The climate indices considered included the Southern Oscillation Index and the NINO3, while the SSTA data was reconstructed from shipping and remote measurements as described in Kaplan (1997). This data is available over a 5dx5d latitude-longitude grid covering much of the earth's ocean surface. Predictors were selected from lagged variable values, the maximum lag being considered extending to 24 quarters (six years). Given the large number of variables the predictors were identified from, the potential for selecting incorrect or spurious predictors was significant. The use of model averaging in such a setting was considered as an efficient means of reducing the predictive uncertainty that such spurious predictor choices could introduce.

Table 1 – Predictive performance of the probabilistic forecasting model. "mam" refers to March-April-May, "jja" to June-July-August, "son" to September-October-November, and "djf" to December-January-February.

Q	L	Correlation		Likelihood Ratio	
		Cross-Validation 1866-1992	Pure-forecast 1993-2002	Cross-Validation 1866-1992	Pure-forecast 1993-2002
mam	1	0.71	0.67	1.72	1.51
mam	2	0.69	0.48	1.65	1.24
mam	3	0.59	0.58	1.33	1.51
mam	4	0.47	0.35	1.09	1.36
jja	1	0.62	0.63	1.66	1.52
jja	2	0.62	0.33	1.42	1.10
jja	3	0.57	0.75	1.21	1.08
jja	4	0.43	0.16	1.20	1.13
son	1	0.81	0.92	2.04	2.12
son	2	0.64	0.69	1.32	1.77
son	3	0.44	0.57	1.25	1.28
son	4	0.38	0.47	1.13	0.76
djf	1	0.83	0.94	2.29	3.13
djf	2	0.76	0.71	1.82	1.54
djf	3	0.68	0.76	1.43	1.54
djf	4	0.68	0.38	1.47	1.20

5. RESULTS

As mentioned in the previous section, the performance of the probabilistic forecasting approach was evaluated in a leave-one-out-cross-validation setting for the 1866-1992 period, and in a pure-forecast setting for the 1993-2002 period. A summary of the results obtained for all quarters and lead times is presented in Table 1. Two measures of performance are used. These are: (a) correlation between observed and the predicted expected value, and (b) likelihood ratio which represents the average ratio of the conditional (predicted) probability density at the observation being predicted, and the marginal (or unconditional) probability density at the same observation. While the first statistic represents the accuracy with which an expected or single valued forecast can be issued, the second represents increased probability with which each observation can be predicted in a leave-one-out cross-validation setting.

Some observations from the results in Table 1 are:

1. The results for the leave-one-our cross-validation (representing the period 1866-1992) are statistically indistinguishable from those for the pure-forecasts (representing the period 1993-2002). One could thus expect leave-one-out cross-validation to represent the type of results that would be obtained in future applications of the model.

2. The results are especially accurate for the September-October-November and the December-January-February quarter when the El Nino state is full developed in the oceans. This result is in accordance with the results obtained using most statistical and dynamical models for predicting the ENSO.

3. While the results for the March-April-May and the June-July-August quarters are not as accurate as that for the later quarters, they represent a high accuracy on part of the model. This is especially important given that most dynamical and other approaches have difficulty predicting the actual onset of El Nino, which starts developing during the March-April period. It is also interesting to note that the model is able to sustain a relatively high accuracy even at a lead of one year.

4. It is reassuring to note that probabilistic forecasting model performs well both in terms of the prediction of the expected value (represented by the correlation results in Table 1) as well as the conditional distribution (represented by the likelihood ratio results). Likelihood ratios consistently greater that one suggest an improved representation of the conditional probability distribution over climatology and hence a high potential for use of model outputs in risk-based water management applications.

Space limitations prevent us from presenting the full range of predictor choices used in formulating the models summarised in Table 1. However, predictors for all constituent models for the September-October-November quarter for a lead of 3 months are presented in Table 2. The weights and predictors associated with each constituent model are listed. It is interesting to note that the dominant mode being represented is the Markovian dependence associated with the state of the ENSO system (represented by the SOI and the SSTA for the 2-7° latitude and 218° longitude range). However, additional SSTA locations, corresponding to the eastern and western edges of Australia (-27/178, lag 1 and -42/133, lag 8) are also dominant choices in the predictors identified. Were a single set of predictors used in formulating the probabilistic forecasting model, it is likely that the variability characterised through these additional predictor choices would not have been adequately represented.

In order to ascertain the improvement in the model as a result of the model averaging, a stepwise addition of the model constituents was performed for the September-October-November quarter, lead-1 probabilistic forecasting model. Results from this assessment are presented in Table 3. It is interesting to note that there is an increase in the forecast performance for the 1866-1992 period, as well as the pure forecast period 1993-2002. While the improvements are not highly significant in the results presented, similar improvements were noted for most of the other quarters and lead times. What is reassuring about this and the other results is that these improvements are present in the pure forecast period, indicating that the process of averaging model outputs reduces the uncertainty that would be present if a single model were used instead.

Table 2 – Constituent models and their respective weights for the September-October-November quarter, lead-1 model averaged probabilistic forecasts

		Predictor 1		Predictor 2		Predictor 3		Predictor 4	
Model	Weight	Lat/Long	Lag	Lat/Long	Lag	Lat/Long	Lag	Lat/Long	Lag
1	0.25	SOI	1	7/228	1	-42/133	8		
2	0.18	7/218	1						
3	0.26	-27/178	1	SOI	1	-42/133	8	42/298	28
4	0.13	17/253	1	SOI	1	2/218	1		
5	0.17	32/178	1	SOI	1	7/218	1		

Table 3 – Improvements in probabilistic forecasting as a result of model averaging

# of	Correlation	
PredSets	1866-1992	1993-2002
1	0.76	0.81
2	0.77	0.85
3	0.80	0.91
4	0.81	0.91
5	0.81	0.92

6. DISCUSSION

Existing approaches for probabilistic forecasting suffer from several serious limitations. While many of the limitations relate to the simplistic distributional and dependence assumptions implicit in their formulation, a more serious limitation is increased predictive uncertainty which is a result of the limited ability of such models at representing the secondary modes of variability in the system. A new probabilistic forecasting approach was presented in this paper that addressed some of the limitations noted above. This approach was novel in two main respects: (a) it was cognisant of the varying levels of uncertainty present in the climatic data being used in the modelling, and (b) it attempted to capture both the dominant as well as the less dominant modes of variability in the system through use of multiple models whose outputs were averaged using a carefully formulated linear weighting scheme.

While the above application presents results for the forecast of the Southern Oscillation Index, the use of the method for more general water resources applications is straightforward. Two areas where such an approach could be used in water management is water allocation (where sequences of the year-ahead flows are forecast and used to augment the current storage levels using appropriately formulated water demand forecasts), and reservoir operation (where the operation is optimised using pre-specified risk-based operational objectives, the system performance being simulated using the probabilistic forecasts as representative sequences that can be expected as inflows into the system). Details on these applications will be presented in a separate paper.

REFERENCES

Fraser, A.M. and Swinney, H.L., Independent coordinates for strange attractors from mutual information. Phys. Rev. A, 33(2): 1134-1140, 1986.

Hoeting, JA., Madigan D. Raftery AE. Volinsky CT., Bayesian model averaging: A tutorial, Statistical Science, Volume 15, Issue 3, Pages 193-195, 2000.

Kaplan, A., Kushnir, Y., Cane, M.A. and Blumenthal, M.B., Reduced Space Optimal Analysis For Historical Data Sets - 136 Years of Atlantic Sea Surface Temperatures. Journal of Geophysical Research Oceans, 102(C13): 27835-27860, 1997.

Sharma, A., Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - A strategy for system predictor identification, Journal of Hydrology, Volume 239, Issues 1-4, Pages 232-239, 2000a.

Sharma, A., Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3 - A nonparametric probabilistic forecast model, Journal of Hydrology, Volume 239, Issues 1-4, Pages 249-258, 2000b.