A Bayesian Approach to Predictor Selection for Seasonal Streamflow Forecasting

DAVID E. ROBERTSON AND Q. J. WANG

CSIRO Land and Water, Highett, Victoria, Australia

(Manuscript received 28 October 2010, in final form 3 July 2011)

ABSTRACT

Statistical methods commonly used for forecasting climate and streamflows require the selection of appropriate predictors. Poorly designed predictor selection procedures can result in poor forecasts for independent events. This paper introduces a predictor selection method for the Bayesian joint probability modeling approach to seasonal streamflow forecasting at multiple sites. The method compares forecasting models using a pseudo-Bayes factor (PsBF). A stepwise expansion of a base model is carried out by including the candidate predictor with the highest PsBF that exceeds a selection threshold. Predictors representing the initial catchment conditions are selected on their ability to forecast streamflows and predictors representing future climate influences are selected on their ability to forecast rainfall. The final forecasting model combines selected predictors representing both initial catchment conditions and future climate influences to jointly forecast seasonal streamflows and rainfall. Applications of the predictor selection method to two catchments in eastern Australia show that the best predictors representing initial catchment conditions and future climate influences vary with location and forecast date. Antecedent streamflows are the best indicator of the initial catchment conditions. Predictors representing future climate influences are only selected for forecasts made between July and January. Indicators of El Niño dominate the selected predictors representing future climate influences. The skill of streamflow forecasts varies considerably between locations and throughout the year. Skill scores for the perennial streams of the Goulburn River catchment exceed 40% for several seasons, while for the intermittent streams in the Burdekin River catchment, the skill scores are lower.

1. Introduction

Forecasts of future seasonal streamflows provide valuable information to many water users and managers, including irrigators, hydroelectricity generators, rural and urban water supply authorities, and environmental managers (Plummer et al. 2009). Statistical techniques commonly used as a practical approach to produce forecasts of seasonal streamflows include regression models (Garen 1992; Kwon et al. 2009; Lima and Lall 2009; Pagano et al. 2009; Ruiz et al. 2007), linear discriminant analysis (Piechota et al. 2001), independent components analysis (Westra et al. 2008), and nonparametric statistical techniques (Sharma 2000; Sharma and Chowdhury 2011).

Recently the Bayesian joint probability (BJP) modeling approach to seasonal streamflow forecasting at multiple sites was developed (Wang and Robertson 2011; Wang et al. 2009). The BJP modeling approach has been adopted by the Australian Bureau of Meteorology for seasonal streamflow forecasting in Australia (Plummer et al. 2009). The approach has considerable flexibility for modeling a wide range of predictors and predictands. It produces joint probabilistic forecasts of streamflows at multiple sites that preserve intersite correlations. It allows the use of data that contains nonconcurrent and missing records in both parameter inference mode and forecasting mode. The model flexibility and data handling ability means that the BJP modeling approach has potential for wide practical application.

Statistical streamflow forecasting methods use predictors that represent two sources of streamflow predictability: the influence of initial catchment conditions and the effect of climate during the forecast period. In semiarid and temperate climates, the states of soil moisture and groundwater stores are the most relevant for seasonal streamflow forecasting. In cold climates, the depth and extent of snow cover can also be a useful initial catchment condition indicator (Garen 1992; Kwon et al. 2009; Pagano et al. 2009). Observed persistence in streamflows has led to the use of antecedent streamflows as an indicator of initial catchment conditions (Chiew and McMahon 2002; Chiew et al. 1998; Piechota et al. 2001;
Wang et al. 2009). Streamflow persistence arises because of soil and groundwater storages delaying rainfall-runoff responses (Chiew and McMahon 2002) and, therefore, antecedent streamflows can be considered as a variable that integrates the states of soil moisture and groundwater stores. Antecedent rainfall has also been used as an indicator of initial catchment conditions (Garen 1992; Pagano et al. 2009). In practice, antecedent streamflows and rainfall over a range of preceding periods have been used as predictors of streamflow from monthly through to seasonal totals (Garen 1992; Piechota et al. 2001; Souza Filho and Lall 2003).

The climate during the forecast period also has a major influence on streamflows. Future climate is dependent on the initial conditions of the ocean, land, and atmosphere over a large scale. A great number of climate indices representing the initial conditions of ocean, land, and atmosphere have been linked to future climate (and streamflows). These indices are typically derived from data analysis and their physical influence on the climate explained using climate models. Climate indices relevant to Australian conditions represent anomalies in the tropical Pacific Ocean [Niño-3, Niño-4, Niño-3.4, the ENSO Modoki index, and the Southern Oscillation index (SOI)], the tropical Indian Ocean [the eastern and western poles of the Indian Ocean dipole, the dipole mode index (Ashok et al. 2003; Saji et al. 1999) and the Indonesian index (Verdon and Franks 2005)], and extratropical zones [the Tasman Sea index (Drosdowsky 1993; Murphy and Timball 2008) and the southern annular mode (Hendon et al. 2007; Marshall 2003)]. The strength of relationships between these indices and climate has been shown to vary spatially and seasonally (Kirono et al. 2010; Risbey et al. 2009). For seasonal streamflow forecasting, it is reasonable to start with these climate indices at various lag times as predictor candidates.

It is clear that a large number of candidates can be considered as potential predictors of streamflows. A commonly used approach to dealing with a large set of candidate predictors is to select predictors and form just one model for forecasting streamflow or climate. The aim of the predictor selection is to detect the underlying relationships from historical data to give the highest possible skills in forecasting future events (Barnston and Smith 1996; DelSole and Shukla 2009; Drosdowsky and Chambers 2001; Garen 1992; Meier and Moradkhani 2009).

There are a number of pitfalls in predictor selection, many of which have not been considered in the establishment of statistical models for streamflow forecasting. When a predictor selection method is poorly designed, the results can be heavily influenced by chance features in available data. Such chance features in data may lead to an unacceptably high probability of detecting a relationship between streamflows and potential predictors that is not real. Forecasting models using predictors having relationships with streamflow that are not real will have little predictive power for future events (DelSole and Shukla 2009).

In this study we seek to develop a rigorous predictor selection method. Rigorous methods of predictor selection seek to select only those predictors that are almost certain to have a real underlying relationship with streamflow. Uncertainty exists about model parameters that describe underlying relationships because of the finite available data (DelSole and Shukla 2009; Michaelsen 1987). Therefore, a good predictor selection method will consider model parameter uncertainty. Bayesian modeling, such as the BIP modeling approach to seasonal streamflow forecasting, explicitly accounts for parameter uncertainty (Gelman et al. 1995).

The purpose of forecasting is to predict unobserved events. Using cross validation for predictor selection allows predictors to be selected on the basis of their ability to predict events not included in model parameter inference (Michaelsen 1987). Predictor selection using cross validation implicitly imposes a penalty on model complexity, reducing the chance of overfitting the forecasting model (DelSole and Shukla 2009; Vehtari and Lampinen 2002). However, even when cross validation is used, there is still the possibility of selecting predictors where no real underlying relationship exists (DelSole and Shukla 2009). To overcome this problem, it is necessary to understand how the selection criterion responds to random predictors and establish a scale of evidence or significance for predictor selection.

Bayes factors are criteria commonly used for model selection in Bayesian modeling (Gelfand 1995; Vlachos and Gelfand 2003). Bayes factors compare the statistical evidence supporting alternative models using the prior predictive density. Jefferys established a scale of evidence for the comparison of models in pairs using Bayes factors (Kass and Raftery 1995). As an alternative to the traditional Bayes factor, the pseudo-Bayes factor (PsBF) compares the evidence supporting alternative models using the cross-validation predictive density (Gelfand 1995). The use of the PsBF is consistent with the selection of predictors for forecasting in that it assesses the performance of forecasts for events not included in the model parameter inference (Vehtari and Lampinen 2002). However, no standard scale of evidence exists for the PsBF (Vlachos and Gelfand 2003).

Models are often assessed and reported in terms of forecast skills. It has been widely recognized that model fitting skills are not a good indication of model predictive skills (Jolliffe and Stephenson 2003; Michaelsen 1987). To address this problem, the skill of cross-validation forecasts is often reported. However, if all the
data are used in predictor selection, then the forecast skill assessed through cross validation is likely to contain artificial skill. Double cross validation, where the cross-validation skill of forecasts independent of parameter inference and predictor selection is assessed, is necessary to get a more appropriate indication of forecast skill for future independent events (DelSole and Shukla 2009; Michaelsen 1987). However, double cross validation often requires a prohibitively large computational effort, and few, if any, studies attempt to assess the extent of artificial skill in statistical seasonal streamflow forecasts.

In this paper, we develop a new method of selecting predictors for seasonal streamflow forecasting using the BJP modeling approach that addresses many of the pitfalls of predictor selection. We adopt the PsBF as the predictor selection criterion and determine PsBF thresholds for selecting predictors so that any relationship detected from historical data has a high chance of being true. A stepwise predictor selection method is adopted that includes the candidate predictor with the highest PsBF that exceeds a selection threshold at each model expansion step. Predictors representing initial catchment conditions are selected on their ability to forecast rainfall. The final forecasting model combines predictors representing initial catchment conditions and future climate influences to jointly forecast rainfall and streamflow totals at multiple sites. We use cross validation after predictor selection to report skills of the final forecasting models, but we investigate the extent of artificial skill through double cross validation.

The predictor selection method is applied to two catchments in eastern Australia with contrasting hydrological conditions to produce joint multiple site forecasts of 3-month streamflow and rainfall totals starting on the first of each month. The BJP modeling approach is used to produce forecasts for all forecast dates by establishing 12 separate models—one for each month.

The mathematical formulation and implementation of the BJP modeling approach to streamflow forecasting at multiple sites is described in detail in Wang and Robertson (2011) and Wang et al. (2009). Only a brief overview is provided here. Arranging model predictors as a column vector \( y(1) \) and predictands as a column vector \( y(2) \), Yeo–Johnson transforms (Yeo and Johnson 2000) are applied to

\[
y = \begin{bmatrix} y(1) \\ y(2) \end{bmatrix}
\]

(1)

to normalize the variables and stabilize their variances. The transformed variables \( z \) are assumed to follow a multivariate normal distribution:

\[
z = \begin{bmatrix} z(1) \\ z(2) \end{bmatrix} 
\sim N(\mu, \Sigma) = N(\mu, \sigma \mathbf{R} \sigma^T).
\]

(2)

Model parameters, \( \theta \), include vectors of the Yeo–Johnson transform parameter \( \lambda \), the mean \( \mu \), and variance \( \sigma^2 \), and the correlation matrix, \( \mathbf{R} \). Thus, the BJP modeling approach assumes that the predictors and predictands are related linearly in the transformed space. Previous analysis suggests that this is a reasonable assumption (Wang et al. 2009).

The posterior distribution of the model parameters \( p(\theta | Y_{OBS}) \) is obtained through Bayesian inference using all the historically observed data of predictors and predictands \( Y_{OBS} = [y_{OBS}^1, y_{OBS}^2, \ldots, y_{OBS}^n] \), where \( y_{OBS}^t \) is for year \( t \), and \( t = 1, 2, \ldots, n \).

Markov Chain Monte Carlo (MCMC) sampling is used to draw a sample of \( m \) sets of parameters, \( \theta_k \), where \( k = 1, 2, \ldots, m \), that numerically represents the posterior distribution of model parameters.

Given predictor values \( y(1) \) for a new event, the probabilistic forecast of predictands \( y(2) \) is given by the predictive density

\[
f[y(2)] = \int p[y(2)|y(1); \theta]p(\theta|Y_{OBS}) \, d\theta
\]

\[\approx \frac{1}{m} \sum_{k=1}^{m} p[y(2)|y(1); \theta_k]. \]

(3)

Periods of zero streamflows and rainfall occur in many parts of Australia and in other arid and semiarid regions throughout the world. Within the BJP modeling approach, zero flow or rainfall data are treated as censored data that have unknown precise values but are known to be below or equal to zero (Wang and Robertson 2011). As data censoring may be applied with censoring thresholds other
than zero, a general censoring threshold \( y_c \) is presented in the mathematical formulations in this paper as in Wang and Robertson (2011).

Rearrange predictor vector \( y(1) \) into

\[
y(1) = \begin{bmatrix} y(1a) \\ y(1b) \end{bmatrix},
\]

(4)

where \( y(1a) \) consists of predictors whose values are precisely known and \( y(1b) \) predictors whose values are only known to be equal to or below the censoring thresholds \( y_c(1b) \). Rearrange also predictand vector \( y(2) \) into

\[
y(2) = \begin{bmatrix} y(2a) \\ y(2b) \end{bmatrix},
\]

(5)

where \( y(2a) \) consists of predictands whose values are precisely known and \( y(2b) \) predictands whose values are only known to be equal to or below the censoring thresholds \( y_c(2b) \). The conditional predictive density \( p[y(2)|y(1); \theta_k] \) in Eq. (3) then becomes

\[
p[y(2)|y(1); \theta_k] = p[y(2a), y(2b) \leq y_c(2b)|y(1a), y(1b) \leq y_c(1b); \theta_k].
\]

(6)

To evaluate this conditional predictive density, data augmentation is used to generate a random set of \( y_{AUG}(1b) \) that follows \( p[y(1b)|y(1a); \theta_k] \) and satisfies \( y_{AUG}(1b) \leq y_c(1b) \) (Wang and Robertson 2011). The corresponding conditional predictive density becomes

\[
p[y(2a), y(2b) \leq y_c(2b)|y(1a), y(1b) \leq y_c(1b); \theta_k] = p[y(2a)|y(1a), y_{AUG}(1b); \theta_k] \times \int_{y_c(2b)} \int_{y(2b)} p[y(2b)|y(2a), y(1a), y_{AUG}(1b); \theta_k] dy(2b).
\]

(7)

The last two terms in the above equation are further evaluated in terms of the corresponding Yeo–Johnson transformed variables that follow a multivariate normal distribution, giving

\[
p[y(2a)|y(1a), y_{AUG}(1b); \theta_k] = J_{\theta} y(2a) \rightarrow y(2a) \times p[z(2a)|z(1a), z_{AUG}(1b); \theta_k]
\]

(8)

\[
\int_{-\infty}^{y_c(2b)} p[y(2)|y(1a), y_{AUG}(1b); \theta_k] dy(2b)
\]

\[
= \int_{-\infty}^{y_c(2b)} p[z(2)|z(1a), z_{AUG}(1b); \theta_k] dz(2b),
\]

(9)

where \( z(\cdot) \) is the Yeo–Johnson transform of \( y(\cdot) \), and \( J_{\theta} y(2a) \rightarrow y(2a) \) is the Jacobian determinant for the transform from \( z(2a) \) to \( y(2a) \).

The conditional probability distributions of \( z(2a) \) and \( z(2b) \) in Eqs. (8) and (9) are multivariate normal, and their mean vectors and covariance matrices can be found using standard multivariate normal conditionalization (Gelman et al. 1995; Wang and Robertson 2011; Wang et al. 2009). Numerical integration of the multivariate normal distribution in Eq. (9) is performed in this study using the RANNRM algorithm (Genz 1993).

In Eq. (7), only one set of augmented data \( y_{AUG}(1b) \) is used in the approximation. Averaging over multiple sets of generated \( y_{AUG}(1b) \) would refine the estimate. However, in the context of evaluating the final predictive density of Eq. (3), we found that computational resources were better used to generate more sets of parameters than to refine the estimate of the conditional probability density through multiple sets of \( y_{AUG}(1b) \) for each parameter set \( \theta_k \).

The performance of the forecasting models is assessed using leave-one-out cross validation. Leave-one-out cross validation involves establishing the model using all historical data except one event and then producing a forecast for the left-out event (Gelfand 1995; Lockwood and Schervish 2005). This procedure is repeated for all events in the historical record to provide an overall assessment of model performance. For each event, the cross-validation predictive density is given by

\[
f'[y(2)] = p[y(2)|y(1) = y_{OBS}(1); Y_{OBS}^{(i)}] = \int_{\theta} p[y(2)|y(1) = y_{OBS}(1); \theta] p(\theta) Y_{OBS}^{(i)} d\theta.
\]

(10)

where \( Y_{OBS}^{(i)} \) denotes the observed predictors and predictands for all the historical events \( [y_{OBS}^{(1)}, y_{OBS}^{(2)}, \ldots, y_{OBS}^{(n)}] \) except \( y_{OBS}^{(i)} \), and \( y_{OBS}^{(i)} \) is the observed predictors for event \( t \).

The evaluation of the cross-validation predictive density of Eq. (10) follows the same method as for the general predictive density of Eq. (3). The leave-one-out cross-validation posterior parameter distribution is needed to compute the leave-one-out cross-validation predictive density. For each event in the historical record,
we use MCMC sampling to draw a sample of 1000 sets of parameters that numerically represent the posterior parameter distribution based on all data except the event of interest. Forecasts are numerically represented by a sample of 1000 sets of values—one generated for each of the 1000 sets of parameter values.

b. Predictor selection method

The predictors of streamflows relating to the two sources of streamflow predictability are selected using a stepwise forward selection process. Predictors representing initial catchment conditions are selected separately from those representing future climate influences. The selection of predictors representing the two sources of streamflow predictability is undertaken separately to ensure that the selected predictors explicitly represent each source of predictability. We anticipate that future research will lead to the identification of alternative candidate predictors for both sources of predictability. Therefore, we seek to produce flexible forecasting models that can include future research results without the need to repeat the full predictor selection process. The limitation of independently selecting predictors representing the different sources of predictability is that interactions between the two sets of predictors are ignored. The predictor selection process is also taken separately for each month as previous studies have found that the strength of the relationships between streamflows and rainfall and their potential predictors varies seasonally (Piecchota et al. 2001; Switanek et al. 2009).

The stepwise forward predictor selection process involves initially establishing a base model, $M_0$, that contains no predictors. The base model, $M_0$, is then expanded in steps to new models, $M_1$, each of which includes an additional candidate predictor. At each model expansion step, the best new model, $M_t$, becomes the base model, $M_0$, for the subsequent expansion step. $PsBF$ (Gelfand 1995) is used to assess the best new model at each expansion step. The $\log_e(PsBF)$ is defined as

$$\log_e PsBF = \log_e \prod_{t=1}^{n} \frac{f_{M_t}(y'(2) = y'_{OBS}(2))}{f_{M_{t-1}}(y'(2) = y'_{OBS}(2))},$$  

(11)

where $f_{M_t}(y'(2) = y'_{OBS}(2))$ is the model $M_t$ cross-validation predictive density of Eq. (10) evaluated at the observed predictand values.

1) SELECTION OF PREDICTORS REPRESENTING INITIAL CATCHMENT CONDITIONS

The selection of predictors representing the initial catchment conditions is based on the ability of the candidate predictors to forecast streamflows at multiple sites. In this study, the candidate predictors representing the initial catchment conditions are antecedent catchment rainfall and antecedent streamflow totals for up to the preceding 3 months.

A streamflow forecasting model for the sites of interest may be incrementally expanded by including at each step an additional predictor that has a $PsBF$ value that is the highest among all candidate predictors and exceeds a specified selection threshold. However, we only undertake one model expansion step because our experience suggests that a second expansion step often does not lead to the selection of second predictors, and when it does, the selected second predictors often do not show a coherent seasonal pattern and do not lead to improved forecasts. This will be further discussed in section 4b(1).

2) SELECTION OF PREDICTORS REPRESENTING FUTURE CLIMATE INFLUENCES

The selection of predictors representing future climate influences on streamflows is based on the ability of candidate predictors to forecast rainfall—an indicator of the future climate. In this study, 13 indices based on sea temperature and atmospheric pressure anomalies in the Pacific, Indian, and Southern Oceans—the three dominant influences on Australian climate—are considered potentially useful predictors (Table 1). Specifically, these indices with lags of up to 3 months are included as candidate predictors. A model forecasting catchment average rainfall may be incrementally expanded by including at each step an additional predictor that has a $PsBF$ value that is the highest among all candidate predictors and exceeds a specified selection threshold. For the same reason as previously discussed in relation to selecting predictors representing initial catchment conditions, we also only undertake one model expansion step in selecting predictors representing future climate influences.

As discussed in the introduction, the best predictors representing future climate influences are expected to vary at large spatial scales and with season. Locally between neighboring catchments, however, some continuity in the selected predictors is still expected. To reduce the likelihood of inconsistencies due to data used in predictor selection, the predictors representing future climate influences are selected by using a spatially consistent dataset from a national rainfall analysis and using a consistent period of record for all catchments. Many of the climate indices are produced from sea surface temperature anomalies. Sea surface temperature analyses are most reliable after 1950 (Smith and Reynolds 2003) and, therefore, selection of the predictors representing future climate influences is performed using data for the period 1950–2008.
3) ESTABLISHING SELECTION THRESHOLDS

The lengths of streamflow, rainfall, and predictor data records are limited. The available observations of these data also contain random variation. The random variation in data combined with the limited record length can introduce variation in the $P_{SBF}$. Therefore, the $P_{SBF}$ values obtained using the available (limited) data may be different than the true value that would be obtained if data were unlimited. This can potentially lead to the selection of predictors where no real underlying relationship exists and the $P_{SBF}$ obtained is solely an artifact of the combination of random variation and a limited record length.

To limit the chance of selecting a predictor when there is no real underlying relationship, we establish $P_{SBF}$ thresholds for predictor selection. A distribution of the highest $P_{SBF}$ of all the relevant candidate predictors is approximated from randomized data that should not contain any real underlying relationships between predictors and predictands. The distribution is then used to establish a selection threshold.

A real set of data is randomized by randomly reordering the year sequence of the set of candidate predictor data to produce a mismatch with the year sequence of the predictand data. The $P_{SBF}$ is calculated for each randomized candidate predictor and the highest $P_{SBF}$ is recorded. The randomization of candidate predictors is repeated to produce a set of 100 highest $P_{SBF}$ values. The set of highest 100 $P_{SBF}$ values are then used to form an empirical distribution of the highest $P_{SBF}$ that could be obtained when there are no real underlying relationships between the predictors and predictands. The advantage of using randomized data, as opposed to using purely random data, is that the marginal distribution of each of the candidate predictors and predictands is maintained and the correlations between candidate predictors are preserved.

Separate empirical distributions of the highest $P_{SBF}$ of candidate predictors using randomized data are established for the selection of predictors representing initial catchment conditions and for the selection of predictors representing future climate influences because the number of candidate predictors and length of record is different in each situation. Selection thresholds are then determined from the empirical distributions that will give a relatively small chance of selecting a predictor when there is no real underlying relationship.

c. Forecasting model skill assessment

1) FORECASTING MODEL SKILL

To produce the final forecasting models, we combine the predictors representing the initial catchment conditions and the ultimate forecast skill for each time step is evaluated. The skill assessment is performed by comparing the forecasted values with the observed values for the verification period.
and future climate influences to jointly forecast 3-month catchment rainfall and streamflow totals at multiple sites. One forecasting model is produced for each month, resulting in 12 models.

In the final forecasting models, we set the correlations between forecast rainfall and predictors representing the initial catchment conditions to zero. The initial conditions of the land surface over a large region can influence the climate during the forecast period. However, the selected predictors representing initial catchment conditions are catchment-scale indicators of the land surface conditions. While a relationship between the initial catchment conditions and forecast rainfall may still exist, the relationship is expected to be too weak to be detected from such local data. Setting the correlations to zero eliminates any effect that data noise might have on the correlation parameters if they were inferred from data. Our experience confirmed that rainfall forecasts decoupled from the predictors representing initial catchment conditions were more stable.

The skill of forecasts produced using the final forecasting models is assessed using leave-one-out cross validation. We assess the overall skill of streamflow and rainfall forecasts at each site using three skill scores based on percentage reduction in forecast error relative to a reference forecast. The error measures used in the three skill scores are the root mean squared error in probability (RMSEP) of the forecast median, the mean squared error (MSE) of the forecast median, and the continuous ranked probability score (CRPS) of the forecast distribution. The reference forecasts for the three skill scores are the observed historical (climatology) median for the RMSEP and MSE skill scores and the full distribution of the observed historical (climatology) events for the CRPS skill score.

Each skill score assesses different aspects of the forecast distribution. The RMSEP skill score gives all forecast events similar opportunity to contribute to the overall assessment of forecast skill. The MSE skill score is similar to the Nash–Sutcliffe efficiency commonly used in hydrology and can be overly sensitive to a few events with large forecast errors. Both the RMSEP and MSE skill scores assess the forecast skill using a point representation of the forecast distribution. The CRPS skill score, on the other hand, assesses the reduction in error of the whole forecast probability distribution; however, like the MSE skill score, it can be sensitive to a few events with large forecast errors.

2) ASSESSMENT OF ARTIFICIAL SKILL THROUGH DOUBLE CROSS VALIDATION

The skill of the final forecasting models assessed through (single) cross validation may contain artificial skill because all available data are used in the selection of predictors in the first place. Double cross validation is necessary to assess the “true” skill of the forecasts. In leave-one-out double cross validation, one event from the historical dataset of predictors and predictands is left out. Predictors are selected and model parameters are inferred using the rest of the data. A forecast is then made from the established model for the left-out event. This is repeated for each of the historical events, and the skill of the forecasts is then assessed by comparing the forecasts with corresponding observations.

The double cross-validation process generates forecasts for events that have not been used in the predictor selection. As the predictors can vary for each forecasting event, double cross validation assesses the skill of forecasts derived from the predictor selection process as well as model inference. The difference between the skill assessed through double cross validation and the skill assessed through single cross validation is considered artificial skill. In this study, double cross validation is only carried out for one of the catchments because of the large computational effort involved.

3. Data

Two locations in eastern Australia were selected to demonstrate the predictor selection method and assess the skill of subsequent forecasts. The locations were chosen in catchments with considerable consumptive water use, where streamflow forecasts are potentially of significant value, and with contrasting hydrologic conditions. The two selected locations are the Goulburn River catchment in Victoria and the Burdekin River catchment in northern Queensland. Figure 1 shows the locations of the selected catchments.

At each location, three gauging stations were identified that have unregulated catchments and long, relatively complete records (Table 2). Gauging stations in close proximity to each other were chosen to ensure that they respond to the same climate influences and may be expected to have similar hydrological behavior (Fig. 2).

Figure 2 also presents the annual hydrographs for the study gauging stations. The streamflows display a distinct seasonal cycle in both the magnitude and variability. In the Goulburn River catchment, the streams are perennial and flows peak during August and September and recede to low levels between February and May. In the Burdekin River catchment, the streams are intermittent with peak streamflows occurring during February and March and little or no flow between August and October. The months with the largest streamflows also display the greatest variability for both locations.
Monthly catchment average rainfall data for each catchment were computed from 5-km gridded data available from the Australian Water Availability Project (Jones et al. 2009). Monthly values of the 13 climate indices were obtained from a range of data sources (Table 1).

4. Results

4.a. Forecasts using randomized predictor data

We established cumulative distributions of the highest $\log_e(PsBF)$ values of candidate predictors representing initial climate conditions and of candidate predictors representing future climate influences obtained using randomized data. For predictors representing initial catchment conditions, there is a 30% chance that the highest $\log_e(PsBF)$ is greater than zero and about a 5% chance the highest $\log_e(PsBF)$ is greater than two. Therefore, if $\log_e(PsBF) = 0$ is used as a threshold for predictor selection, one of the candidate predictors will nearly always be selected despite the fact that no real relationships are expected from the randomized data. To correctly select the no-predictor model with a 95% chance, a threshold value of $\log_e(PsBF) = 4$ will be needed. Compared with predictors representing initial catchment conditions, a higher threshold value is required for selecting predictors representing future climate influences to achieve the same chance of correctly selecting the no-predictor model. This is mainly due to the much larger number of candidate predictors representing future climate conditions (39) than candidate predictors representing initial catchment conditions (6), although other factors, such as the marginal distributions of predictors and predictands and correlation structure of the candidate predictors, may also have some influence.

We adopt a threshold value of $\log_e(PsBF) = 2$ for selecting predictors representing initial catchment conditions and $\log_e(PsBF) = 4$ for selecting predictors representing future climate influences, so that there is only a 5% chance of selecting a predictor when no underlying relationships are expected from the data used. For datasets with different marginal distributions of predictors and predictands and correlation structure of the candidate predictors, the chance may vary from the 5%, but for expediency we adopt these thresholds as a rule of thumb to save the large computational effort required for deriving the distributions of highest PsBF of the candidate predictors for each application. The use of these thresholds may potentially exclude predictors that do have true, but weak, underlying relationships with the predictands. However, from our experience, predictors that produced $\log_e(PsBF)$ less than these thresholds make only marginal contributions to forecast skill for independent events.

<table>
<thead>
<tr>
<th>Catchment</th>
<th>Gauging station</th>
<th>Station name</th>
<th>Catchment area (km²)</th>
<th>Available record</th>
<th>Mean annual flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goulburn River</td>
<td>405205</td>
<td>Murrindindi River at Murrindindi</td>
<td>108</td>
<td>1950–2008</td>
<td>476 mm (0.051 km³)</td>
</tr>
<tr>
<td>Goulburn River</td>
<td>405209</td>
<td>Acheron River at Taggerty</td>
<td>619</td>
<td>1950–2008</td>
<td>487 mm (0.301 km³)</td>
</tr>
<tr>
<td>Goulburn River</td>
<td>405214</td>
<td>Delatite River at Tonga Bridge</td>
<td>368</td>
<td>1950–2008</td>
<td>306 mm (0.112 km³)</td>
</tr>
<tr>
<td>Burdekin River</td>
<td>120002C</td>
<td>Burdekin River at Sellheim</td>
<td>36 260</td>
<td>1968–2008</td>
<td>117 mm (4.24 km³)</td>
</tr>
<tr>
<td>Burdekin River</td>
<td>120303A</td>
<td>Sutter River at St. Anns</td>
<td>50 291</td>
<td>1967–2008</td>
<td>28 mm (1.41 km³)</td>
</tr>
<tr>
<td>Burdekin River</td>
<td>120302B</td>
<td>Cape River at Taemas</td>
<td>16 074</td>
<td>1968–2008</td>
<td>38 mm (0.61 km³)</td>
</tr>
</tbody>
</table>
b. Goulburn River catchment

1) PREDICTORS REPRESENTING INITIAL CATCHMENT CONDITION

The \( \log_e(PsBF) \) values of the candidate predictors for the selection of the first predictor representing initial catchment conditions for the Goulburn River catchment are presented in Fig. 3. There is a distinct seasonal pattern in the \( \log_e(PsBF) \) of the best first predictor representing initial catchment conditions, peaking in January and dipping to the lowest in April. The best first predictor representing initial catchment conditions is antecedent streamflows for all forecast dates. For the majority of forecast dates, the total streamflow for the previous month is the best first predictor representing initial catchment conditions.

We investigated whether a second predictor representing initial catchment conditions would contain additional independent information about the states of catchment groundwater and soil water stores. The \( \log_e(PsBF) \) values for the second predictor representing initial catchment conditions are presented in Fig. 4. The highest \( \log_e(PsBF) \) values are an order of magnitude smaller than for the best first predictor. The best second predictor has \( \log_e(PsBF) \geq 2 \) for only the months of May, September, and October. More importantly, the highest \( \log_e(PsBF) \) values show little seasonal pattern and are likely to contain considerable noise. Results for
the second predictor representing future climate influences and for the Burdekin River catchment also show a lack of coherent seasonal patterns. Therefore, a second step of model expansion that would include a second predictor was not undertaken, as discussed in section 2b(1).

2) PREDICTORS REPRESENTING FUTURE CLIMATE INFLUENCES

The log$_e$(PsBF) of the candidate predictors for the selection of the first predictor representing future climate influences for the Goulburn River catchment are presented in Fig. 5. The log$_e$(PsBF) for the best first predictor representing future climate influences is greater than four for five forecast dates. These forecast dates occur between July and November. During this time, many of the predictors representing future climate influences that have log$_e$(PsBF) $\geq$ 4, including the best predictors for forecasts made in July, August, and November, are indices describing ENSO anomalies in the Pacific Ocean. However, for forecasts made in September and October, the best predictors representing future climate influences is the Indian Ocean dipole mode index.

3) SELECTED PREDICTORS OF STREAMFLOW AND RAINFALL

Table 3 summarizes the selected predictors for forecasting models for the Goulburn River catchment. Antecedent streamflows are selected as the predictors representing initial catchment conditions for all months.

4) FORECAST SKILL

The skill scores for streamflow forecasts and rainfall in the Goulburn River catchment are presented in Fig. 6. For streamflows, forecast skill is highest for forecasts made between October and February. These forecast seasons contain the receding limb of the annual hydrograph (see Fig. 2). The lowest skill scores are for forecasts made in April. The April season forecasts contain the start of the rising limb of the annual hydrograph (see Fig. 2). At this time of the year in the Goulburn River catchment, the predictors representing initial catchment conditions have the lowest log$_e$(PsBF) (Fig. 5) and no predictors representing future climate influences are selected (Fig. 6 and Table 3).

The skill scores for rainfall forecasts in the Goulburn River catchment are near zero for all months except for forecasts made in October and November when some small skill is obtained. The result demonstrates the difficulty in producing skillful seasonal forecasts of rainfall for the Goulburn River catchment.

c. Burdekin River catchment

1) PREDICTORS REPRESENTING INITIAL CATCHMENT CONDITIONS

The log$_e$(PsBF) of the candidate predictors for the selection of the first predictor representing initial catchment conditions for the Burdekin River catchment are presented in Fig. 7. There is a seasonal pattern in the log$_e$(PsBF) of the best first predictor of initial catchment conditions and for the Burdekin River catchment also show a lack of coherent seasonal patterns. Therefore, a second step of model expansion that would include a second predictor was not undertaken, as discussed in section 2b(1).
conditions. The $\log_e(\text{PsBF})$ is less than two for forecasts made between September and January.

Between February and August, the $\log_e(\text{PsBF})$ is greater than two for some candidate predictors. The best first predictor representing initial catchment conditions during this period is antecedent streamflows for the majority forecast dates, with total rainfall for the previous 3 months being the best predictor for forecasts made in March.

2) PREDICTORS REPRESENTING FUTURE CLIMATE INFLUENCES

The $\log_e(\text{PsBF})$ of the candidate predictors for the selection of the first predictor representing future climate influences for the Burdekin River catchment are presented in Fig. 8. The $\log_e(\text{PsBF})$ for the best predictor representing future climate influences is greater than four for six forecast dates. All the best first predictors representing

<table>
<thead>
<tr>
<th>Forecast date</th>
<th>Predictor representing initial catchment conditions</th>
<th>Predictor representing future climate influences</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>Total streamflows for preceding 1 month</td>
<td>SOI lag 1</td>
</tr>
<tr>
<td>February</td>
<td>Total streamflows for preceding 1 month</td>
<td>SOI lag 1</td>
</tr>
<tr>
<td>March</td>
<td>Total streamflows for preceding 2 months</td>
<td></td>
</tr>
<tr>
<td>April</td>
<td>Total streamflows for preceding 3 months</td>
<td></td>
</tr>
<tr>
<td>May</td>
<td>Total streamflows for preceding 1 month</td>
<td></td>
</tr>
<tr>
<td>June</td>
<td>Total streamflows for preceding 1 month</td>
<td></td>
</tr>
<tr>
<td>July</td>
<td>Total streamflows for preceding 3 months</td>
<td></td>
</tr>
<tr>
<td>August</td>
<td>Total streamflows for preceding 3 months</td>
<td>SOI lag 1</td>
</tr>
<tr>
<td>September</td>
<td>Total streamflows for preceding 1 month</td>
<td>Indian Ocean dipole mode index lag 1</td>
</tr>
<tr>
<td>October</td>
<td>Total streamflows for preceding 1 month</td>
<td>Indian Ocean dipole mode index lag 1</td>
</tr>
<tr>
<td>November</td>
<td>Total streamflows for preceding 2 months</td>
<td>Niño-3 lag 1</td>
</tr>
<tr>
<td>December</td>
<td>Total streamflows for preceding 1 month</td>
<td></td>
</tr>
</tbody>
</table>
future climate influences that have \( \log_{10}(PsBF) \geq 4 \) are indices related to ENSO anomalies in the Pacific Ocean.

3) SELECTED PREDICTORS OF STREAMFLOW AND RAINFALL

The selected predictors for forecasting models for the Burdekin River catchment are summarized in Table 4. No predictors representing initial catchment conditions are selected for forecasts made between September and January and no predictors representing future climate influences are for forecasts made between February and June. No predictors at all are selected for forecasts made in December.

4) FORECAST SKILL

The skill scores for streamflow and rainfall forecasts in the Burdekin River catchment are presented in Fig. 9. The forecast skill is the highest for gauge 120002C for all months, while the skill scores for the other two streamflow gauges are considerably lower. The period of high skill occurs between April and October when seasonal streamflows are receding (Fig. 2d) and when predictors representing future climate influences have the highest \( \log_{10}(PsBF) \) (Fig. 8). Rainfall forecasts show some skill between July and November, when predictors representing future climate influences are selected (Fig. 9).

Between December and May, two of the gauges have small negative skill scores, suggesting that the forecasts made using the selected predictors do not perform as well as a climatology forecast. Predictors are selected because they resulted in better forecasts than a zero predictor BJF model, which is equivalent to a climatology forecast that allows for parameter uncertainty. The period of low skill corresponds to when streamflow and rainfall are at their highest and also when both the initial
catchment conditions and predictors representing future climate influences have the lowest log$_e$(PsBF).

d. Artificial skill assessed through double cross validation

The skill scores for the Goulburn River catchment were obtained using both single and double cross validation. The difference between the skill scores obtained through single and double cross validation represents the artificial skill of the predictor selection method. The artificial skill in RMSEP for streamflow and rainfall forecasts varies between −1% and 13%, and is generally close to zero. This means that forecasts with skill scores, obtained through single cross validation, of just over 10% or less need to be treated with caution as they may have little, or no, skill for independent events.

5. Discussion

At both locations, the seasonal pattern in the skill of streamflow forecasts appears to mirror that of the log$_e$(PsBF) of the best predictors representing initial catchment conditions. This suggests that the majority of streamflow forecast skill arises from knowledge of the initial catchment conditions, principally antecedent streamflows. In some isolated seasons, a second predictor representing initial catchment conditions produces a log$_e$(PsBF) greater than two (see Fig. 4). We do not select these second predictors because they display little seasonal pattern and do not necessarily lead to improved forecasts. However, the existence of second predictors with a log$_e$(PsBF) greater than two suggests that more refined indicators of initial catchment condition may exist. Many authors have suggested that physically based or conceptual rainfall-runoff models are useful for forecasting seasonal streamflows (Shi et al. 2008; Wood et al. 2005). Future work will investigate using the output from

FIG. 7. Same as Fig. 3 but for the Burdekin River catchment.

FIG. 8. Same as Fig. 5 but for the Burdekin River catchment.
rainfall-runoff models to provide more refined indicators of the initial catchment condition.

At both sites, many of the predictors representing future climate influences that produce a $\log_e(PsBF)$ greater than four are indicators of ENSO. The predictors have their highest $\log_e(PsBF)$ values for forecasts made between August and October. These results are consistent with the findings of other authors who have established

![Figure 9](image-url)

**FIG. 9.** Same as Fig. 6 but for cross-validation forecasts in the Burdekin River catchment at gauges (a) 120002C, (b) 120303A, (c) 120302B, and (d) catchment rainfall.

<table>
<thead>
<tr>
<th>Forecast date</th>
<th>Predictor representing initial catchment conditions</th>
<th>Predictor representing future climate influences</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>SOI lag 3</td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>Total streamflows for preceding 1 month</td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>Total rainfall for preceding 3 months</td>
<td></td>
</tr>
<tr>
<td>April</td>
<td>Total streamflows for preceding 1 month</td>
<td></td>
</tr>
<tr>
<td>May</td>
<td>Total streamflows for preceding 1 month</td>
<td></td>
</tr>
<tr>
<td>June</td>
<td>Total streamflows for preceding 2 months</td>
<td></td>
</tr>
<tr>
<td>July</td>
<td>Total streamflows for preceding 1 month</td>
<td>Niño-3.4 lag 1</td>
</tr>
<tr>
<td>August</td>
<td>Total streamflows for preceding 3 months</td>
<td>SOI lag 1</td>
</tr>
<tr>
<td>September</td>
<td></td>
<td>Niño-4 lag 1</td>
</tr>
<tr>
<td>October</td>
<td></td>
<td>Niño-3.4 lag 1</td>
</tr>
<tr>
<td>November</td>
<td></td>
<td>Niño-4 lag 2</td>
</tr>
<tr>
<td>December</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
that ENSO is the dominant driver of monthly rainfall variability over large parts of eastern Australia and that the strongest influence is during the Australian spring (Drosdowsky and Chambers 2001; Kirono et al. 2010; McBride and Nicholls 1983; Murphy and Timbal 2008; Risbey et al. 2009). However, the Indian Ocean dipole mode index is the best predictor representing future climate influences for September and October forecasts in the Goulburn River catchment. For some parts of southeastern Australia, the Indian Ocean dipole appears to have a stronger relationship with spring rainfall than indicators of ENSO, particularly in recent decades (Risbey et al. 2009; Ummenhofer et al. 2009). Recent research suggests that the influence of ENSO on southeastern Australia is exerted through the Indian Ocean (Cai et al. 2011). Therefore, while the Indian Ocean dipole is a better predictor of southeastern Australian rainfall, the variation in the dipole is believed to originate from ENSO forcing.

In the Burdekin River catchment, the skill of streamflow forecasts is close to zero for many months, mainly between December and May. During this period, the predictors representing initial catchment conditions that are selected have relatively low \( \log_e(PsBF) \) values. Therefore, streamflows during the forecast period are primarily related to rainfall during the forecast period. Between December and May, the skill of rainfall forecasts is also low. The skill of rainfall forecasts from the Australian Bureau of Meteorology’s dynamical seasonal forecasting model [the Predictive Ocean Atmosphere Model for Australia (POAMA)] appears to be greater than both climatology and existing statistical forecasting methods (Lim et al. 2009). Therefore, the use of rainfall forecasts from POAMA as a predictor representing future rainfall and streamflows may provide an opportunity to increase forecast skill during some times of the year.

Results suggest that the artificial skill due to predictor selection is typically very low but may be as high as over 10% in RMSEP skill score. For streamflow forecasts in the Goulburn River catchment where skill is relatively high, the skill of operational forecasts for independent events are expected to remain high. However, where skill is low, as in the case of the Burdekin River catchment, it is possible that the artificial skill will be similar in magnitude to the single cross-validation skill estimates. In this case, the skill of operational forecasts for independent events may be negligible and, therefore, forecasts need to be treated with caution.

The use of model averaging or hierarchical modeling would eliminate the need for predictor selection and, therefore, cross-validation skill estimates would not contain artificial skill. These techniques allow for model uncertainty as well as parameter uncertainty to be included in forecasts and, therefore, the cross-validation skill of forecasts will be closer to the double cross-validation skills reported in this paper. However, model averaging and hierarchical modeling techniques have substantial additional mathematical and computational requirements that are potentially prohibitive for operational forecasting. A future study will investigate developing computationally efficient methods of model averaging or hierarchical modeling that eliminate the need for predictor selection.

6. Conclusions

Forecasts of future seasonal streamflows are valuable to a range of users. Statistical methods that are commonly used to forecast streamflows require the selection of appropriate predictors representing initial catchment and future climate conditions using historical data. However, the large number of available candidate predictors and limited historical data can lead to the selection of predictors because of chance features of the available data, which subsequently produce poor forecasts for independent events.

This paper introduces a rigorous predictor selection method for the Bayesian joint probability modeling approach to seasonal streamflow forecasting at multiple sites that addresses the pitfalls of selecting predictors for seasonal forecasting. The stepwise predictor selection method uses the pseudo-Bayes factor (PsBF) as the predictor selection criterion. At each model expansion step, the predictor with the highest PsBF is selected, provided that the PsBF exceeds a selection threshold. Predictors representing the initial catchment conditions are selected on their ability to forecast streamflows and predictors representing future climate influences are selected on their ability to forecast rainfall. The final forecasting model combines selected predictors representing both initial catchment conditions and future climate influences to jointly forecast rainfall and streamflow totals at multiple sites.

A numerical investigation was undertaken to determine appropriate predictor selection thresholds. The distributions of the highest PsBF were empirically estimated for predictors representing initial catchment conditions and for predictors representing future climate influences using randomized predictor data. PsBF selection thresholds corresponding to a 95% chance of correctly selecting a no-predictor model were estimated from the empirical distributions. A threshold of four was adopted for the selection of predictors representing future climate influences and a threshold of two was adopted for the selection predictors representing initial catchment conditions. These thresholds balance the probability of rejecting a predictor having a true underlying relationship with the probability of selecting a predictor because of chance features in the data.
The predictor selection method was applied to two catchments in eastern Australia with contrasting hydrological conditions. The results illustrate that the best predictor representing initial catchment conditions varies with location and forecast date. For the two study catchments, antecedent streamflow is a better indicator of the initial catchment conditions than antecedent rainfall. There is some consistency in the predictors representing the influence of future climate on future streamflows. During the austral winter through to the early summer period, indicators of the ENSO process in the Pacific Ocean are the best predictors of future climate conditions, with the exception of September and October forecasts in the Goulburn River catchment where the Indian Ocean dipole mode index is selected. At other times of the year, the climate indices included in this study do not appear to be useful predictors of seasonal rainfall and streamflows.

The skill of streamflow forecasts varies considerably between locations and throughout the year. Skill scores of greater than 40% in RMSEP skill score are achieved in the Goulburn River catchment where streams are perennial. Skill scores for the intermittent streams in the Burdekin River catchment are much lower than for the Goulburn River catchment. In both catchments, the highest skill scores are for forecast seasons that contain the receding limb of the annual hydrograph, while the lowest skill scores are for forecasts made for the first few months of the wetting up season.

The extent of artificial skill of the predictor selection method was assessed for the Goulburn River catchment by comparing skill scores obtained through single and double cross validations. Generally, the artificial skill is very low, but was found to be as high as 13% in one instance. This suggests that the forecasts with single cross-validation skill scores of about 10% or less need to be treated with caution as they may have little or no skill for independent events.

Future work will investigate ways to improve the forecast skill when catchments are wetting up and when future climate is least predictable from indices of large-scale climate anomalies. More specifically, rainfall-runoff modeling and dynamical seasonal climate forecasting modeling will be used to produce improved predictors representing initial catchment conditions and future climate influences.

Acknowledgments. This research has been supported by the Water Information Research and Development Alliance between the Australian Bureau of Meteorology and CSIRO Water for a Healthy Country Flagship, the South Eastern Australian Climate Initiative, and the CSIRO OCE Science Leadership Scheme. We thank Neil Plummer, Jeff Perkins, Dr. Senlin Zhou, Andrew Schepen, Trudy Peatey, and Dr. Daeyok Shin from the Australian Bureau of Meteorology for many valuable discussions as well as providing the streamflow and rainfall data for this study. Tom Pagano and Prasanth Hapuarachchi have contributed to the quality of the publication through their review of an early version of this manuscript.

REFERENCES


