The challenge of forecasting high streamflows 1-3 months in advance with lagged climate indices in south-east Australia

James C. Bennett[1], Q. J. Wang[1], Prafulla Pokhrel[1] and David E. Robertson[1]

[1]CSIRO Land and Water, Graham Road, Highett, Victoria, Australia 3190

Correspondence to:

James. C. Bennett

T: +61 3 9252 6229

E: james.bennett@csiro.au

A: CSIRO Land and Water, Graham Road, Highett, Victoria, Australia, 3190

Abstract

Skilful forecasts of high streamflows a month or more in advance are likely to be of considerable benefit to emergency services and the broader community. This is particularly true for mesoscale catchments (<2000 km^2) with little or no seasonal snow melt, where real-time warning systems are only able to give short notice of impending floods. In this study, we generate forecasts of high streamflows for the coming 1-month and coming 3-month periods using large-scale ocean/atmosphere climate indices and catchment wetness as predictors. Forecasts are generated with a combination of Bayesian joint probability modeling and Bayesian model averaging. High streamflows are defined as maximum single-day streamflows and maximum 5-day streamflows that occur during each 1-month or 3-month forecast period. Skill is clearly evident in the 1-month forecasts of high streamflows. Surprisingly, in several catchments positive skill is also evident in forecasts of large threshold events (exceedance probabilities of 25%) over the next month. Little skill is evident in forecasts of high streamflows for the 3-month period. We show that including lagged climate indices as predictors adds little skill to the forecasts, and thus catchment wetness is by far the most important predictor. Accordingly, we recommend that forecasts may be improved by using accurate estimates of catchment wetness.

Keywords: prediction, seasonal, Bayesian methods, high streamflows

1 Introduction

Skilful forecasts of high streamflows a month or more in advance have the potential to improve the management of floods. Flood warnings in Australia are presently derived from event-based forecast models that use real-time streamflow and rainfall observations to forecast floods with typical lead-times from hours to a few days, depending on flood travel time (Elliott et al., 2005). Real-time forecasts offer precise estimates of flood stage, but are only available around the time of the flood itself. This leaves emergency services a narrow window to prepare themselves and the community to mitigate flood impacts, particularly in mesoscale catchments that have little or no seasonal snowmelt. In these catchments flood warning systems can only give warning of floods from hours to one or two days in advance of an event. Ill-preparedness for floods can have serious implications. Pfister (2002) identified poor community preparedness to evacuate as the major cause of citizens' slow (and non-existent) responses to a flood evacuation order issued by emergency services. Australian emergency services rely heavily on volunteers for disaster response (Baxter-Tomkins and Wallace, 2009), and ensuring that sufficient volunteer-labour is available during emergencies is a challenge for flood-response agencies like the State Emergency Services (SES). Medium range forecasts (to forecast horizons of 3 months) of high streamflows are needed to enable both emergency services and the community to be better prepared for floods.

This study is a response to a request from the Australian Bureau of Meteorology to explore the skill of real-time high streamflow forecasts at medium range forecast horizons. The Bureau of Meteorology is the lead agency for flood warnings in Australia, and emergency services are important users of these flood warnings. While medium range forecasts of high streamflows cannot hope to be as precise as real-time flood models, forewarning of conditions that could result in large or frequent flooding in the next month or more could allow emergency services to better plan and prepare for the impacts of floods, for example by informing volunteer emergency services personnel of heightened flood risk in the coming month(s).

Several studies have described teleconnections between Australian runoff variability and large-scale oceanic and atmospheric climate indices (hereafter, *climate indices*), particularly climate indices describing the El Niño Southern Oscillation (ENSO) (Chiew et al., 1998; Verdon et al., 2004; Schepen et al., 2012a). These teleconnections have been used to produce forecasts of total seasonal streamflows that are skilful relative to forecasts derived from streamflow climatologies (Wang et al., 2009; Piechota et al., 1998; Sharma, 2000). Flood risk in south-east Australia has also been linked to ENSO (Kiem et al., 2003), but despite this no attempt has yet been made to use such a teleconnection to forecast high streamflows in Australia. Attempts to forecast high streamflows a month or more in advance are rarely reported for other continents, and the examples that exist focus on catchments where snowmelt makes a large contribution to seasonal floods (e.g. Kwon et al., 2009; Lindström and Olsson, 2011). Seasonal snow-melt is rarely an important feature of Australian rivers, and accordingly forecasts that rely on indicators of snow-melt have limited application in Australia.

The aim of this study is to apply a statistical technique, the Bayesian joint probability modelling approach (BJP), to the problem of forecasting high streamflows in mesoscale catchments over the coming 1-month and 3-month periods. The BJP was developed to forecast seasonal total volumes of streamflows (Wang et al., 2009; Wang and Robertson, 2011; Robertson and Wang, 2012) and is now used operationally by the Bureau of Meteorology to issue forecasts for more than 70 sites across Australia (forecasts available at http://www.bom.gov.au/water/ssf/). The BJP produces probabilistic streamflow forecasts that are more accurate than climatology, and, importantly, it is able to reliably estimate uncertainty in the streamflow forecasts. Knowledge of the amount of water held in storage in a catchment (in the soil, as ground water, in surface stores, or as snow/ice – collectively, *catchment wetness*) often contributes more skill to next-month/next-season forecasts of streamflow than climate forecasts (Shukla and Lettenmaier, 2011; Li et al., 2009; Koster et al., 2010; Mahanama et al., 2012). The BJP is able to use multiple predictors to generate forecasts, meaning forecasts can be constructed from both catchment wetness and predictors of climate. For example, Wang et al. (2009) used the BJP to pair the initial catchment wetness with the southern oscillation index (SOI) to forecast seasonal streamflow totals.

A number of sets of predictors can be used to construct different forecast models, and forecasts can be improved by selecting models with the best predictive power (Robertson and Wang, 2012) or by weighting models according to predictive power (Wang et al., 2012a). Wang et al. (2012a) showed that Bayesian model averaging (BMA) outperformed predictor selection methods for merging rainfall forecast models generated with the BJP. In addition, predictor selection can lead to artificially inflated estimates of cross-validation skill if the predictor selection is not included in the cross-validation (DelSole and Shukla, 2009; Robertson and Wang, 2013), a problem that is not present with the BMA method we use in this study.

Our study aims to test the ability of the BJP to forecast high streamflows up to three months in advance. To achieve this, we build a set of forecast models with the BJP by combining an estimate of initial catchment wetness with a suite of climate indices derived from oceanic and atmospheric variables. We combine the models with the BMA method described by Wang et al. (2012a) to maximise predictive power.

We next describe the study sites and give an overview of the forecast models. This is followed by descriptions of the verification measures we use to demonstrate the reliability and skill of the forecasts. We present the reliability and skill of these forecasts, and discuss the prospects for improving long lead forecasts of high streamflows. We conclude with a summary of the paper.

2 Data and methods

2.1 Study sites

Forecasts are generated for six catchments in south-east Australia shown in Fig. 1. Characteristics of the six catchments are summarised in Table 1 and Fig. 2. The catchments are selected as they have long (>40 year) streamflow records, are free of diversions or impoundments, and are minimally impacted by human activities. Streamflow data is taken from the quality controlled Catchment Water Yield Estimation Tool (CWYET) dataset (Vaze et al., 2011). All the catchments are of a size we describe as *mesoscale*, with drainage areas between 1000 km^2 and 2000 km^2. The catchments are large enough to minimise the influence of highly localised storms (e.g. localised convective storms) on the streamflow records. Conversely, catchments are small enough so that flood travel times extend no more than two days, making it difficult to get advance warning of floods of more than two days with a forecasting model that makes use only of observed rainfalls.

The catchments span a range of climate and hydrological conditions. Streamflows in the two north-eastern catchments, the Orara River (ORB) and the Nowendoc River (NOR), are only weakly seasonal, with the highest streamflows occurring in February and March (Fig. 2). The remaining catchments - Abercrombie River (ABH), Murray River (MUR), Mitta Mitta River (MMH) and Tarwin River (TAW) - have more strongly seasonal streamflow regimes, with high streamflows in the austral winter/spring, and low streamflows in the austral summer (Fig. 2). High-elevation areas in the MUR and MMH catchments often receive snowfalls in the Austral winter. However, even in these two catchments the contribution of seasonal snowmelt to streamflows is relatively small.

2.2 Forecast model

2.2.1 Overview

Forecasts are generated on the last day of each month for two periods: the coming month (Jan, Feb, ..., Dec), and the coming three months (JFM, FMA, ..., DJF). We refer to these as 1-month and 3-month forecast periods.

Fig. 3 gives a schematic overview of how forecasts are generated. Thirteen forecast models are generated with the BJP method (Fig. 3a) for each forecast period and for each predictand. Forecasts from these individual models are then merged using BMA (Fig. 3b). We now describe the components shown in Fig. 3 in detail.

2.2.2 Predictands

While we pursue forecasts of large streamflows in a bid to improve information available for the management of floods, we employ the term *high flows* rather than *floods* in this paper. This is because we seek to build monthly statistical models in catchments that often have highly seasonal flow regimes. We define high flows from each month by exceedance probability, and in months where mean flows are low these ‘high’ flows often do not constitute what would be considered flood flows in other months.

We investigate two predictands to represent high streamflows:

1. The maximum 1-day streamflow (mm/d) for each forecast period (Max1D).
2. The maximum 5-day aggregated streamflow (mm/d averaged across the 5 days) calculated for each forecast period (Max5D).

As already noted, neither Max5D nor Max1D is necessarily a large flood. For example, in the catchments with strongly seasonally delineated streamflows, Max5D streamflows in summer can be very low compared to Max5D winter streamflows. In low streamflow months, medians of both Max1D and Max5D streamflows are sometimes not much larger than average monthly streamflows (Fig. 2). For this reason, we also evaluate the performance of the forecasts in terms of probabilities of events exceeding larger thresholds (see Section 2.3.3).

The BJP is able to generate forecasts jointly for multiple predictands. In addition to either Max1D or Max5D, we also include total rainfall for the forecast period as a predictand (from the Australian water availability project (AWAP) gridded rainfall dataset; Jones et al., 2009). We jointly forecast rainfall and streamflow because the influence of lagged climate indices on streamflow occurs mainly through rainfall (Robertson and Wang, 2012). Statistically, the correlations between lagged climate indices and rainfall and between rainfall and streamflow tend to be stronger, and thus easier to capture from data, than the correlation directly between lagged climate indices and streamflow. By including rainfall as a co-predictand, the statistical model needs to satisfy three correlations, with the two stronger correlations providing some guidance on sensible values for the weaker correlation.

2.2.3 Predictors

We use lagged catchment wetness and lagged climate indices as predictors of high streamflows. We approximate catchment wetness with total streamflow in the previous month for both 1-month and 3-month forecast periods. Total streamflow can be a somewhat coarse measure of catchment wetness, and takes no account of differences in catchment wetness stores (e.g. snow *cf.* soil moisture). However, using total streamflow as an estimate of catchment wetness has the virtue of simplicity, and is adequate for this exploratory study.

Eleven lagged climate indices are evaluated as potential predictors in this study, and these are listed in Table 2. We select these climate indices as they have been linked to rainfall in south-east Australia. The teleconnection between south-east Australian rainfall and ENSO has been extensively described (e.g. Schepen et al., 2012a; Chiew et al., 1998; Wang et al., 2009) including, as already noted, the link between flooding and ENSO (Kiem et al., 2003). We use five indices to describe ENSO: NINO3, NINO3.4, NINO4, the ENSO Modoki index (EMI) (Ashok et al., 2007) and the southern oscillation index (SOI) (Troup, 1965). The influence of Indian Ocean sea surface temperatures has also been linked to rainfall in south-east Australia, with the teleconnection being most evident in winter months (Verdon and Franks, 2005; Schepen et al., 2012a; Ashok et al., 2003). We use four Indian Ocean indices as predictors: the Indian Ocean west pole index (WPI), east pole index (EPI) and dipole mode index (DMI) (Saji et al., 1999), as well as the Indonesia index (II) (Verdon and Franks, 2005). Finally, extra-tropical sea surface temperatures and atmospheric features along Australia's east coast have been linked to south-east Australian rainfall (Murphy and Timbal, 2008; Risbey et al., 2009; Pook et al., 2006). We use the Tasman Sea index (TSI) (Murphy and Timbal, 2008) and an index of atmospheric blocking (BI140) (Risbey et al., 2009) to represent extra-tropical climatic features. The teleconnection between lagged atmospheric climate indices (e.g., the Antarctic Oscillation index describing the Southern Annular Mode; Schepen et al., 2012a) and Australian seasonal precipitation is often weak, as they show little persistence in comparison to SST-derived indices. We note that Schepen et al. (2012a) found no evidence of a relationship of lagged B140 and TSI with mean rainfall in any season. It is therefore unlikely that lagged TSI or B140 will contribute skill to high streamflow forecasts, however we have included them in case they have a relationship with high rainfall events. Atmospheric blocking, for example, has been correlated with larger rain storms (Pook et al., 2006).

We have not considered using multiple climate indices as joint predictors, which may describe the effects of interactions between climate indices on high streamflows. Some studies suggest that these interactions may be important in understanding concurrent relationships (e.g. Kiem et al., 2003), however results from our previous work demonstrates that adding a second joint predictor does not result in any improvement in forecast skill of seasonal total rainfalls or streamflows when using lagged climate indices (Robertson and Wang, 2012; Wang et al., 2012a).

Sea surface temperature climate indices are derived from the National Center for Atmospheric Research (NCAR) Extended Reconstruction of Sea Surface Temperature version 3 (Smith et al., 2008). B140 is derived from the National Centers for Environmental Prediction (NCEP)–NCAR reanalysis data (Kalnay et al., 1996). SOI is sourced from the Australian Bureau of Meteorology (BOM).

Mean monthly values of each climate index for the previous month are used for both 1-month and 3-month forecasts; accordingly we refer to these as *lagged* climate indices. Schepen et al. (2012a) showed that teleconnections between rainfall and lagged climate indices are strongest at short lags, and for this study we investigate only climate indices lagged by one month to establish forecast models. For example, for a 1-month forecast for June we use catchment wetness and NINO3 calculated for May as predictors, while for a 3-month forecast for January-February-March we use predictors calculated for December.

Catchment wetness is combined with each of the 11 climate indices to create 11 forecast models for each predictand and for each forecast period. In addition, one forecast model is developed using only catchment wetness as a predictor, and one forecast model is developed based only on climatology (using no predictors). This gives a total of 13 forecast models for each predictand and for each forecast period.

While the effect of snow on the two alpine catchments (MUR and MMH) is expected to be small, we investigated the use of snow accumulation as a predictor for these two snow-affected catchments. Including snow accumulation as a predictor in these two catchments resulted in no increase in forecast skill and is not presented here.

2.2.4 Bayesian joint probability modelling

The BJP is used to generate the 13 individual forecast models for each predictand and each forecast period (Fig. 3a), which we call *BJP forecast models*. Detailed mathematical formulations of the BJP are given by Wang et al. (2009), Wang and Robertson (2011) and Robertson and Wang (2012). In summary, the BJP is implemented as follows:

1. Predictands and predictors are transformed to normalise their distributions and stabilise their variances. Streamflow and rainfall are transformed with a log-sinh transform (Wang et al., 2012b), and climate indices are transformed with the Yeo-Johnson transform (Yeo and Johnson, 2000).
2. We assume that the set of transformed predictors and predictands can be described by a joint probability distribution – in this case a multivariate normal distribution.
3. The parameters of the log-sinh transform, the Yeo-Johnson transform, and the multivariate normal distribution are inferred jointly. Parameter inference is performed with Bayesian methods and Markov chain Monte Carlo (MCMC) sampling. Taken together, the parameters of the log-sinh transform, the Yeo-Johnson transform and the multivariate normal distribution define the statistical relationship between predictors and predictands, and allow us to generate forecasts.

Mathematically, if predictors are given by vector $y(1)$ and predictands by vector $y(2)$, the probabilistic forecast is given by

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

where M is the model used, and Y_{OBS} contains the historical data of both the predictors and the predictands used for model inference. θ is the vector of parameters for the log-sinh transform, the Yeo-Johnson transform, and the multivariate normal distribution.

2.2.5 Bayesian model averaging

Forecasts from the thirteen BJP forecast models are merged with BMA to produce one *BJP-BMA forecast* for each predictand and for each forecast period (Fig. 3b). The BMA method we use is described in detail by Wang et al. (2012a). For a set of models M_k, k=1, 2, ..., K, each model is assigned a weight, w_k. The forecasts are then merged by:

$$f_{BMA}(y(2)|y(1)) = \sum_{k=1}^{K} w_k \, f_k(y(2)|y(1)) \tag{2}$$

We calculate w_k by maximizing the posterior distribution of the weights, which is proportional to:

$$A = \prod_{k=1}^{K} (w_k)^{\alpha-1} \prod_{t=1}^{T} \sum_{k=1}^{K} w_k . p\left(y_{OBS}^{t}(2) \middle| y_{OBS}^{t}(1); Y_{OBS}^{(t)}, M_k\right) \tag{3}$$

where α is the concentration parameter, $y^t{}_{OBS}(1)$ and $y^t{}_{OBS}(2)$ are the predictors and predictands for events t=1, ...,T, and $Y^{(t)}{}_{OBS}$ is a matrix containing observed values of predictors and predictands for all the events except event t. $\prod_{k=1}^{K}(w_k)^{\alpha-1}$ is from the symmetric Dirichlet prior distribution used by Wang et al. (2012a). We use α values greater than 1 to distribute weights more evenly among models, which helps to stabilise the weights when there is significant sampling variability. Specifically, α=1+a/K with a=1. The remainder of the right side of Eq. 3 is the cross-validation likelihood function. By using the cross-validation likelihood function, we base each model weight on the predictive power of the model, rather than on the fitting ability of the model. A is maximised with an iterative expectation-maximization (EM) algorithm, as described by Wang et al. (2012a).

2.3 Forecast verification

Forecasts are verified using leave-one-out cross validation. Forecasts for events in year t=1, 2, ..., n are generated from all available historical data except those at year t. For each forecast variable y, this produces a series of forecast cumulative probability distributions $y^t \sim F^t(y^t)$. Forecasts are then verified against observations y^t_{OBS}.

Leave-one-out cross validation ensures that a forecast model is not validated against data used to build that model. We note that in this approach we use data after the forecast date to build the forecast model, data which would not be available to build operational real-time forecast models. The purpose of cross validation is to get an indication of model performance for future events. For future events, we would use all historical events to establish the model. The length of record used in model establishment in cross-validation is similar to (more precisely just short of) the full record length. In this sense, cross-validation gives a good indication of the skill of a true implementation for the future events.

Verifying the probabilistic forecasts is not straightforward, particularly when the aim is to forecast rare events. Here we evaluate forecast reliability to demonstrate that the probabilistic forecasts are neither too confident nor underconfident. We then assess forecast accuracy using three skill scores. We now describe each of the verification measures in detail.

2.3.1 Forecast reliability

For probabilistic forecasts to be meaningful, we must first demonstrate that the forecast probability distributions are reliable; that is, the uncertainty in the forecasts is reliably represented, and thus the forecast distributions are neither too wide (not confident enough) nor too narrow (overconfident). To achieve this, we present reliability diagrams. A reliability diagram plots the observed frequency against the forecast probability and shows how well the predicted probability of an event corresponds to its observed frequency (Wilks, 1995). We present reliability diagrams calculated from events that are larger than the 50% exceedance probability threshold of Max1D and Max5D streamflows.

2.3.2 Overall forecast accuracy: root mean square error in probability

The root mean square error in probability (RMSEP) works on the principle that if forecast and observed values are of similar exceedance probabilities then the forecast should be rewarded, even if the magnitudes of observed and forecast values are quite different (Wang and Robertson, 2011). RMSEP is calculated as follows:

1. We represent the observed historical distribution (climatology), y, in the form of non-exceedance probability, $F_{CLI}(y)$.
2. For events t=1, 2, ..., n, we take the median of the forecast distribution, y^t_{MED}.
3. RMSEP is then calculated as

$$RMSEP = \left[\frac{1}{n}\sum_{t=1}^{n}\left(F_{CLI}(y^t_{MED}) - F_{CLI}(y^t_{OBS})\right)^2\right]^{\frac{1}{2}} \quad (4)$$

4. We calculate RMSEP_{REF} by substituting the forecast median, y^t_{MED}, in Eq. 4 with the climatology median. We then calculate the RMSEP skill score:

$$SS_{RMSEP} = \frac{RMSEP_{REF} - RMSEP}{RMSEP_{REF}} \quad (5)$$

RMSEP (eq. 4) demonstrates the ability of the model to forecast the rank of a given event, ranked in relation to historical events (i.e., the ability to forecast an event's place on a cumulative distribution function generated from historical data). While this does not necessarily give an indication of how well the model is able to forecast the magnitude of an event, the ability to forecast an event's rank is likely to be very useful to users of the forecast, who could categorise an event as, for example, 'likely to exceed the 50 percentile of high flows' or similar. $\mathrm{SS}_{\mathrm{RMSEP}}$ (eq. 5) measures the ability of the forecasts to outperform a naive climatology forecast.

In addition, we calculate SS_{RMSEP} with $RMSEP_{REF}$ represented by the BJP forecast generated with only catchment wetness as a predictor (i.e., no climate information is used to generate $RMSEP_{REF}$). This allows us to show the relative contribution of catchment wetness and climate indices to forecast skill.

2.3.3 Accuracy of forecasts for large threshold events

For a given month, we consider a subset of larger 'high' streamflows to assess forecast performance. These larger streamflows are defined as having exceedance probabilities of 50% (Q_{50}), 25% (Q_{25}) and 10% (Q_{10}) for observed Max1D and Max5D. (These streamflows approximately correspond to annual exceedance probabilities (AEP) of 1:2 AEP, 1:4 AEP and 1:10 AEP. To keep the study as simple as possible, we have defined larger events on the basis of empirical exceedance probabilities rather than fitting an extreme value distribution, so we continue to refer to large streamflows in terms of exceedance probabilities.) We treat these large streamflows as thresholds (we term them *large threshold events*), and measure forecast skill by comparing the forecast probability of exceeding a large threshold event with the corresponding observation. Q_{50}, Q_{25}, and Q_{10} thresholds are shown for 1-Month Max1D and Max5D streamflows are shown in Fig. 2.

Use of multiple skill scores is recommended to demonstrate robustness in the results (e.g. Cloke and Pappenberger, 2008). We use two measures of skill to verify forecasts at larger streamflow thresholds: the Brier Score and the log-likelihood ratio.

Brier Score

The Brier score has been a staple for the verification of probabilistic forecasts since it was proposed by Brier (1950). We use the Brier score to verify forecasts of larger streamflows in order that our study can be compared to others.

Given forecast distributions y^t at events t=1, 2, ..., n, and streamflow thresholds $Q_{P,}$ with exceedance probabilities P=50%, 25%, 10%, the forecast is presented as the probability of exceeding the streamflow threshold:

$$1 - F^t = p(y^t > Q_P) \tag{6}$$

We calculate the Brier score as:

$$BS = \frac{1}{n}\sum_{t=1}^{n}(1 - F^t - O^t) \tag{7}$$

where O^t takes the value of 1 if the threshold is exceeded, and 0 if it is not exceeded. We calculate BS_{REF} by substituting F^t with a forecast calculated from climatology, F^t_{REF}. We then calculate the Brier skill score:

$$SS_{BS} = \frac{BS_{REF} - BS}{BS_{REF}} \tag{8}$$

Log-likelihood ratio

The Brier score has been subject to criticism, particularly for producing unintuitive results for rare (and in our case, large) events when assessing very sharp forecasts (i.e., forecast probabilities of 100% or 0%) (Jewson, 2008; Benedetti, 2010). We adopt the recommendations of Benedetti (2010) and Jewson (2008), who both advocate variations on the likelihood to assess probabilistic forecasts. We term this measurethe log-likelihood ratio (LLR).

The LLR is based on the likelihood ratio described by Jewson (2008). For all exceedance forecasts 1-F^t, let all the cases of t where 1-F^t exceeds a streamflow threshold Q be given by the set A, and all cases of t where the streamflow threshold is not exceeded be given by B. The log-likelihood for a forecast is calculated by:

$$LL = \log_e\left(\prod\nolimits_A (1 - F^t) \prod\nolimits_B F^t\right) \tag{9}$$

The log-likelihood of the reference forecast, LL_{REF}, is calculated by substituting F^{t}_{REF} (again, based on climatology) for F^{t} in Eq. 9. The LLR is then calculated by:

$$LLR = LL - LL_{REF} \tag{10}$$

The LLR differs from skill scores like RMSEP or the Brier score in that it does not show proportional improvement over a reference forecast on a normalised scale (often -∞% - 100%), making direct comparisons to other skill scores difficult. However, the LLR is essentially identical to the natural logarithm of the pseudo Bayes factor ($\log_e$(PsBF)) presented by Robertson and Wang (2012) and Schepen et al. (2012a). Robertson and Wang (2012) showed that values of the $\log_e$(PsBF) up to 2 are indistinguishable from statistical noise, while there is a 95% chance that the relationship between a forecast model and observations is true if the $\log_e$(PsBF) is greater than 4. We adopt the qualitative categories for the LLR presented by Schepen et al. (2012a) for our study: little evidence of skill where $LLR<2$; positive evidence of skill where $2<LLR<4$; strong evidence of skill where $4<LLR<6$; very strong evidence of skill where $LLR>6$.

3 Results

3.1 Suitability of BJP for modelling high streamflows

The log-sinh transform used to normalise streamflows has been shown to be well-suited to hydrological data in general (Wang et al., 2012b; Del Giudice et al., 2013), but its ability to adequately describe high streamflows needs to be established. In Fig. 4 we show the log-sinh transformed normal distributions fitted to observed Max1D values for two example months, February and September (other months give very similar results). These two months represent low and high streamflow regimes: February is a month of low mean streamflows in MMH, MUR, ABH and TAW, and a month of high mean streamflows in ORB and NOR, while September is a month of high mean streamflows in MMH, MUR, ABH and TAW and a month of low mean streamflows in ORB and NOR. In general, the assumed log-sinh transformed normal distributions appear to adequately represent the marginal distribution of observations. Almost all observations fall within the confidence bounds of the fitted distributions, including large Max1D events. The log-sinh transformed normal distributions represent observed events well even in catchments with highly variable streamflows, such as ORB and ABH. In summary, the log-sinh transform is flexible enough to normalise the events we are attempting to forecast.

3.2 Forecast reliability

In general, forecast uncertainty is reliably represented by the forecasts after cross-validation. Fig. 5 shows reliability diagrams for the NOR and MUR catchments for Max1D 1-Month forecasts (the other catchments, not shown, produce similar results). In these diagrams, forecast probabilities are divided into five bins (see inserts). The [0.05, 0.95] uncertainty interval of the observed relative frequency is calculated through bootstrap resampling of the forecasts and observed streamflows. For the majority of forecast probability ranges, the uncertainty interval of the observed relative frequency intersects the theoretical 1:1 line, indicating that the forecasts of high streamflows are reliable. Similar results are obtained for the other catchments for all predictands and forecast periods (not shown). These results support the findings of Wang et al. (2009) and Wang and Robertson (2011), who showed the BJP produces reliable forecasts of seasonal streamflows.

3.3 Overall forecast skill

Fig. 6 shows BJP-BMA cross-validated hindcasts of Max1D for an example 20-year period for all catchments. Visual inspection of the hindcasts shows that the credible prediction intervals largely encompass the range of observations. In catchments with strongly seasonal streamflows (e.g. MUR, MMH), the mean of the ensemble forecast often gives realistic predictions of Max1D streamflows during seasons of high streamflows. Accuracy of forecasts in more variable catchments (e.g. NOR, ABH) is much more difficult to ascertain from these time series, and we now turn to formal measures of skill to assess these.

RMSEP skill scores are positive for Max5D forecasts for the 1-month forecast period for most months and catchments (Fig. 7b). Skill in Max5D 1-month forecasts is particularly strong in the winter-spring months (June-November). Skill in Max1D 1-month forecasts is generally lower than for Max5D 1-month forecasts (Fig. 7a, 7b). Max1D streamflows are inherently more variable than Max5D streamflows, as Max5D streamflows are smoothed by the greater number of data included in their calculation. This makes forecasting Max1D streamflows more challenging. Nonetheless, RMSEP skill scores for Max1D 1-month forecasts are positive for most catchments and seasons (Fig. 7a). Max1D 1-month forecast skill is strongest in the winter-spring months. For the 3-month forecast period, RMSEP scores are generally lower for both Max1D and Max5D forecasts, although positive skill scores occur in winter-spring for the MUR, MMH, and ABH catchments, and the NOR catchment shows skill intermittently through the year (Fig. 7c, 7d).

The reason for the reduced performance of the 3-month forecasts becomes evident when we review the contribution of climate indices to forecast skill. Fig. 8 shows RMSEP skill scores calculated relative to BJP forecasts generated using only streamflow as a predictor. The plot shows the skill gained by the inclusion of climate indices for Max1D 1-month forecasts. Fig. 8 shows that almost no skill is gained in any month or catchment by including climate indices, meaning the forecasts depend heavily on catchment wetness for skill. Results are similar for Max5D (not shown). This finding is also supported by Robertson and Wang (2013), who found that climate indices made only weak contributions to the skill of forecasts of seasonal streamflow totals in the MMH and MUR catchments. The contribution of catchment wetness to forecast skill declines over longer forecast periods (Mahanama et al., 2012; Shukla and Lettenmaier, 2011; Li et al., 2009). Thus forecasts for longer periods are less accurate than for shorter forecast periods. This effect is also evident in individual catchments. The TAW catchment, for example, has the lowest autocorrelation of monthly streamflows of the six catchments (not shown), and forecasts for this catchment show poor skill in relation to streamflow climatology.

Nonetheless, 3-month forecasts can be skilful in certain catchments at times of the year when the influence of catchment wetness on high streamflows is strong. The influence of catchment wetness on streamflows is generally strongest on the receding limb of the annual hydrograph (Robertson and Wang, 2013). For the ORB and NOR catchments the annual hydrograph recedes in March-May, while in the ABH, MMH and MUR catchments the annual hydrograph recedes in August-November. This results in positive RMSEP skill scores for 3-month forecasts of these catchments during these months (Fig. 7c, 7d).

Overall, RMSEP generally shows positive skill scores for 1-month forecasts for both Max1D and Max5D streamflows, while 3-month forecasts are substantially less skilful. However, the positive RMSEP skill scores may be the result of good agreement of forecasts with lower ‘high’ streamflows, and not reflect forecasts at larger streamflows. We now turn to forecast skill at higher streamflows to determine the size of streamflows for which forecasts are skilful.

3.4 Forecast skill for large threshold events

In general, forecast skill declines as streamflows get larger (Figs. 9-12). Brier scores show more instances of positive skill than LLR scores, particularly for streamflows larger than Q_{10}. Because the Brier score has known problems with infrequent events (Benedetti, 2010), we focus on the LLR score to discuss forecast skill at larger streamflows.

Substantial skill is evident in forecasts where observed Max1D streamflows are larger than Q_{50} for 1-month forecasts, in both the Brier score (Fig. 9) and the LLR (Fig. 10). LLR scores are higher for Max5D streamflows than for Max1D streamflows, and the highest LLR scores generally occur in July-November. Skill is not related to seasonal changes in high or low Max1D/Max5D streamflows. The ARB, MUR, MMH and catchments show high skill during months of high streamflow (winter-spring, Fig. 10, Fig. 2) while the ORB and NOR catchments only exhibit skill during months of low streamflow (Jul-Nov, Fig. 10, Fig. 2). As with the RMSEP scores, the TAW catchment shows the lowest skill. Four of the six catchments show positive LLR scores in 6 or more months of the year for 1-month forecasts of Max5D streamflows above Q_{25} (Fig. 10). For Max1D streamflows greater than Q_{25}, three catchments show positive LLR scores in six or more months of the year (Fig. 10). Little skill is evident in any catchment or season for either Max1D or Max5D streamflows above Q_{10}.

Skill for 3-month forecasts of larger streamflows is generally low (Figs. 11, 12). Except for one catchment (MUR), catchments show little forecast skill in the majority of months for any of the streamflow thresholds tested for either Max1D or Max5D streamflows. We find positive skill scores for 3-month forecasts in the MUR catchment of Max5D streamflows above Q_{50} and Q_{25} for six or more months, and also for Max1D streamflows above Q_{50} (Fig. 12). Indeed, forecasts for MUR performed best in most measures and skill scores. It is not clear why this should be so. MUR receives reliable rainfall in the winter and spring, resulting in relatively low variability and strong autocorrelation in monthly streamflows. However these characteristics also apply to the nearby MMH catchment, for which forecasts perform no better than for ABH, ORB or NOR in a number of measures (e.g. Fig. 10).

Overall, forecast skill is positive to very strong for 1-month exceedance forecasts of streamflows exceeding Q_{50} for a majority of months in all but the TAW catchment. Skill is not related to seasonal cycles of high and low streamflows. Positive skill scores are also found in several catchments for 1-month exceedance forecasts of streamflows exceeding Q_{25}. The remaining large streamflow forecasts tested here showed little skill in most catchments.

4 Discussion

RMSEP skill scores reported here show the 1-month forecasts to be superior to climatology in forecasting high streamflows. Further, the skill in forecasts is not limited to the lowest of the 'high' streamflows - forecasts of the probability of exceeding Q_{50} Max1D streamflows one month in advance show robust skill in a number of catchments. We note, however, that the Q_{50} Max1D streamflows are still not necessarily very large streamflows. Skill in forecasting large threshold events in two catchments, ORB and NOR, is restricted to months where 'high' streamflows are small, and in which damaging floods are unlikely to occur. Conversely, skill in the MUR, ABH and MMH catchments is evident during periods of high streamflow. Accordingly, forecast skill in these catchments may be valuable to the Bureau of Meteorology when they are seeking to answer more general questions about the risks of high streamflows in a coming month. We note that the usefulness of the forecast is likely to vary with catchment in any case, both because forecast skill varies between catchments and because the prospect of flood damage varies greatly between catchments (i.e., in one catchment a common high streamflow event may damage property or have other deleterious impacts, in another catchment large floods may be of little consequence).

The 1-month forecasts rely heavily on catchment wetness for skill. This supports the many studies that have demonstrated the preeminent contribution of catchment wetness to the skill of seasonal streamflow forecasts for catchments (or seasons) where seasonal snow-melt does not occur (e.g. Mahanama et al., 2012; Shukla and Lettenmaier, 2011; Li et al., 2009; Koster et al., 2010; Robertson and Wang, 2013). Accordingly, improving estimates of catchment wetness is likely to be a simple way of improving forecasts. Accumulated streamflow for a month can be a poor measure of catchment wetness. For example, a high value of total streamflow may be caused by a single intense rainfall event that causes infiltration-excess overland flow, resulting in a large streamflow but little infiltration. In this example the catchment wetness is overestimated by total streamflow. Catchment wetness can be modelled more effectively for forecasting with so-called ‘dynamical’ approaches (Rosenberg et al., 2011; Robertson et al., 2013a) that use soil-moisture accounting models (e.g. conceptual rainfall-runoff models forced by observed rainfall and evaporation) to improve estimates of catchment wetness and thereby improve forecasts.

The ability of the BJP-BMA models to forecast high streamflows a month or more in advance is limited by knowledge of climate during the forecast period. This problem is not likely to be easily surmountable. The high variability of larger rainfall events makes their prediction inherently difficult. In addition, climate indices that have the potential to forecast particular types of rain-bearing weather patterns may have little persistence from month to month. This is particularly so for climate indices calculated from atmospheric variables, which tend to be less persistent than oceanic variables. For example, we have used the atmospheric blocking index (B140, see Table 2) to attempt to account for atmospheric blocking and associated cutoff lows in our forecasts. Cutoff lows associated with atmospheric blocking bring a substantial proportion of rainfall to south-east Australia (Pook et al., 2006), and may counteract the drying associated with very strong El Niño years (Brown et al., 2009). However, we find that B140 adds little skill to forecasts of high streamflows, supporting Schepen et al. (2012a) who showed that lagged B140 had no significant statistical relationship to mean rainfall anywhere in Australia. Similarly, this would very likely apply to other atmospheric indices, e.g. those used to describe the Southern Annular Mode or the Subtropical Ridge of high pressure (position or intensity).

As we noted in the introduction, several studies have shown positive relationships between climate indices and streamflow/rainfall in south-east Australia. However, our work shows that the benefit of using lagged climate indices to forecast high streamflows in south-east Australia is negligible. This can be explained in four ways:

1. Many studies examine teleconnection between concurrent climate indices and streamflow/rainfall (e.g. Verdon and Franks, 2005; Ashok et al., 2003; Pook et al., 2006). The teleconnection between lagged climate indices and rainfall may be weaker than for concurrent indices as implied by the often weak relationships between lagged climate indices and Australian rainfall found by Schepen et al. (2012a).
2. Even if a significant teleconnection exists between a lagged climate index and high streamflows, this information may still not contribute skill to forecasts of high streamflows when we include catchment wetness as a predictor because:
    a. even if the teleconnection between high rainfalls and lagged climate indices is strong, the influence of catchment wetness on high streamflows is so much more powerful that the predictive information provided by lagged climate indices is rendered negligible;
    b. the catchment wetness predictor implicitly contains information about the current state of the climate (e.g., a very wet October), and any information provided by lagged indices may be subsumed by the climate information implicit in catchment wetness.
3. Even in areas where lagged climate indices show a significant teleconnection to seasonal rainfalls (Schepen et al., 2012a), the high variability of large rainfalls associated with high streamflows means that any positive relationships that have been shown to exist between lagged climate indices and seasonal rainfall totals may not apply to high rainfall events.

4. Some studies (e.g. Kiem et al., 2003) use an index describing the Interdecadal Pacific Oscillation (IPO) to relate rainfall/streamflow to climate indices. If we limit our assessment of forecasts only to periods where IPO was in the negative phase, it is possible that ENSO SST indices may add more skill to the forecasts (as suggested by Kiem et al., 2003). However, we sought to assess forecast skill in the context of generating forecasts in real-time. Describing the IPO is not particularly useful for real-time forecasting because it is only possible to define an IPO phase with certainty in retrospect (although informed speculation about the present IPO phase is possible; see, e.g., Cai and van Rensch, 2012). That is, it is often not possible to know with certainty which IPO phase we are in at the present time, so it cannot be used to inform real-time forecasts.

Using conceptual rainfall runoff models forced by rainfall forecasts from dynamical climate models to forecast high streamflows at long lead times is an attractive alternative to the statistical models we have presented here. Statistical models require large volumes of data to characterise relationships between predictors and predictands, and this is particularly important when forecasting rare events. If dynamical climate and hydrological processes can be accurately simulated, fewer data may be required to generate skilful forecasts. Further, dynamical climate models should, in theory, be able to account for complex interactions between different climate drivers, which may influence rainfall. At present dynamical climate models do not necessarily exhibit more skill than statistical forecasts of seasonal precipitation (e.g. Schepen et al., 2012b). Future improvements in dynamical climate models used for forecasting weeks to months advance (e.g. Marshall et al., 2011) may ultimately improve forecasts of high rainfalls. In addition, we note that the skill of statistical forecasts may complement that of dynamical rainfall forecasts (e.g. the statistical rainfall forecasts may exhibit skill in different seasons or locations to dynamical forecasts; Schepen et al., 2012b), and that merging forecasts of high rainfalls from dynamical and statistical models may improve overall skill. Using climate indices derived from SST forecasts from coupled ocean-atmosphere dynamical climate models shows promise in improving forecasts of monthly rainfall totals at lead-times of more than six months (Hawthorne et al., 2013), and avoids the use of lagged climate indices for forecasting.

Our forecast method could be adapted to catchments in different regions by including predictors that are relevant to a given region. In colder regions, seasonal snow melt has been shown to be a very important predictor of seasonal streamflows (e.g. Mahanama et al., 2012), and indicators of future snowmelt (e.g. temperature) could be included as predictors in this model. In addition, climate indices that are important to a given region may also be included, although their utility for forecasting high streamflows may be negligible, as we have shown here.

The high streamflow forecasts we have developed here may be bolstered in future by the inclusion of Numerical Weather Prediction (NWP) models in hydrological forecasting. The Australian Bureau of Meteorology does not presently use NWP forecasts to quantify flood forecasts, although they are used qualitatively to inform flood warnings (Elliott et al., 2005). Very high resolution NWP forecasts have been shown to improve flood forecasts (Roberts et al., 2008). At present, however, NWP forecasts are skilful only for a few days (typically <6 days); and even skilful NWP forecasts are often not accurate enough for use in hydrological forecasting systems, even in catchments substantially larger than those tested here (Cloke and Pappenberger, 2009; Shrestha et al., 2013; Cuo et al., 2011). As NWP models and post-processing of NWP forecasts improve (e.g. Robertson et al., 2013b), NWP forecasts may complement the simpler forecasts we have generated in this study.

5 Summary and conclusions

We have explored the ability of existing statistical forecasting methods to produce forecasts for high streamflows for the coming month and the coming three months. Forecast models are built from a combination of climate predictors and catchment wetness. Models are constructed with a Bayesian joint probability method, and the models are then weighted based on their predictive power using Bayesian model averaging.

Skill is clearly evident in forecasts of high streamflows for the coming 1-month period. Forecasts of larger events, including maximum 1-day streamflows of exceedance probabilities as low as 25%, are also skilful in comparison to long-term climatologies. Our 1-month high streamflow forecasts have the potential to complement existing real-time flood warnings currently used in Australia, to give emergency services and the community more warning of impending high streamflows.

Almost all forecast skill derives from the catchment wetness predictor. If the forecasts are to be extended to additional catchments, they are likely to be poor in catchments that have little month-to-month memory in streamflows. Forecasts in skilful catchments may be improved somewhat by using more refined estimates of catchment wetness.

We find substantially lower skill in forecasts of high streamflows for the coming 3-month period. The influence of catchment wetness on streamflows diminishes over longer periods, and climate predictors add little skill to the forecasts. Future improvements in forecasts of extreme rainfalls from dynamical climate models may be able to improve longer range forecasts of high streamflows.

Acknowledgements

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. Thanks to Yong Song (CSIRO Land and Water), Christopher J. White (Bureau of Meteorology) and Senlin Zhou (Bureau of Meteorology) for their comments on earlier drafts. Thanks to Ben Livneh and two anonymous reviewers for comments that have improved this manuscript.

References

Ashok, K., Guan, Z., and Yamagata, T.: Influence of the Indian Ocean dipole on the Australian winter rainfall, Geophysical Research Letters 30, 1821, 10.1029/2003GL017926, 2003.

Ashok, K., Nakamura, H., and Yamagata, T.: Impacts of ENSO and Indian Ocean dipole events on the southern hemisphere storm-track activity during austral winter, Journal of Climate, 20, 3147-3163, 10.1175/jcli4155.1, 2007.

Baxter-Tomkins, T., and Wallace, M.: Recruitment and retention of volunteers in emergency services, Australian Journal on Volunteering, 14, 1–11, 2009.

Benedetti, R.: Scoring rules for forecast verification, Monthly Weather Review, 138, 203-211, 10.1175/2009MWR2945.1, 2010.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Monthly Weather Review, 78, 1-3, 10.1126/science.27.693.594, 1950.

Brown, J. N., McIntosh, P. C., Pook, M. J., and Risbey, J. S.: An investigation of the links between ENSO flavors and rainfall processes in southeastern Australia, Monthly Weather Review, 137, 3786-3795, 10.1175/2009MWR3066.1, 2009.

Cai, W., and van Rensch, P.: The 2011 southeast Queensland extreme summer rainfall: a confirmation of a negative Pacific Decadal Oscillation phase?, Geophysical Research Letters, 39, L08702, 10.1029/2011GL050820, 2012.

Chiew, F. H. S., Zhou, S. L., and McMahon, T. A.: Use of seasonal streamflow forecasts in water resources management, Journal of Hydrology, 270, 135–144, 10.1016/S0022-1694(02)00292-5, 1998.

Cloke, H. L., and Pappenberger, F.: Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures, Meteorological Applications, 15, 10.1002/met.58, 2008.

Cloke, H. L., and Pappenberger, F.: Ensemble flood forecasting: a review, Journal of Hydrology, 375, 613–626, 10.1016/j.jhydrol.2009.06.005, 2009.

Cuo, L., Pagano, T. C., and Wang, Q. J.: A review of quantitative precipitation forecasts and their use in short-to medium range streamflow forecasting, Journal of Hydrometeorology, 12, 713–728, 10.1175/2011JHM1347.1, 2011.

Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., and Rieckermann, J.: Improving uncertainty estimation in urban hydrological modeling by statistically describing bias, Hydrology and Earth System Sciences, 17, 4209-4225, 10.5194/hess-17-4209-2013, 2013.

DelSole, T., and Shukla, J.: Artificial skill due to predictor screening, Journal of Climate, 22, 331–345, 10.1175/2008JCLI2414.1, 2009.

Elliott, J., Catchlove, R., Sooriyakumaran, S., and Thompson, R.: Recent advances in the development of flood forecasting and warning services in Australia, International conference on innovation, advances and implementation of flood forecasting technology, Tromsø, Norway, 2005, 1-10, 2005.

Hawthorne, S., Wang, Q. J., Schepen, A., and Robertson, D. E.: Effective use of GCM outputs for forecasting monthly rainfalls to long lead times, Water Resources Research, 49, 5427–5436, 10.1002/wrcr.20453, 2013.

Jewson, S.: The problem with the Brier score, arXiv: physics/0401046v1 [physics.ao-ph], available at http://arxiv.org/abs/physics/0401046v1 (last accessed June 2013), 2008.

Jones, D. A., Wang, W., and Fawcett, R.: High-quality spatial climate data-sets for Australia, Australian Meteorological and Oceanographic Journal, 58, 233-248, 2009.

Kalnay, E., M. Kanamitsua, R. Kistlera, W. Collinsa, D. Deavena, L. Gandina, M. Iredella, S. Sahaa, G. Whitea, J. Woollena, Y. Zhua, A. Leetmaaa, R. Reynoldsa, M. Chelliahb, W. Ebisuzakib, W. Higginsb, J. Janowiakb, K.C. Mob, C. Ropelewskib, J. Wangb, R. Jennec, and Joseph, D.: The NCEP/NCAR 40-year reanalysis project, Bull. Amer. Meteor. Soc.,, 77, 34, 1996.

Kiem, A. S., Franks, S. W., and Kuczera, G.: Multi-decadal variability of flood risk, Geophysical Research Letters, 30, 1035, 10.1029/2002GL015992, 2003.

Koster, R. D., P., S. P., Mahanama, Livneh, B., Lettenmaier, D. P., and Reichle, R. H.: Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow, Nature Geoscience, 3, 613-616, 10.1038/NGEO944, 2010.

Kwon, H.-H., Brown, C., Xu, K., and Lall, U.: Seasonal and annual maximum streamflow forecasting using climate information: application to the Three Gorges Dam in the Yangtze River basin, China, Hydrological Sciences Journal, 54, 582-595, 10.1623/hysj.54.3.582, 2009.

Li, H., Luo, L., Wood, E. F., and Schaake, J.: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting, Journal of Geophysical Research: Atmospheres, 114, D04114, 10.1029/2008jd010969, 2009.

Lindström, G., and Olsson, J.: A systematic review of sensitivities in the Swedish flood-forecasting system, Atmospheric Research, 100, 275-284, 10.1016/j.atmosres.2010.09.013, 2011.

Mahanama, S., Livneh, B., Koster, R., Lettenmaier, D., and Reichle, R.: Soil moisture, snow, and seasonal streamflow forecasts in the United States, Journal of Hydrometeorology, 13, 189-203, 10.1175/jhm-d-11-046.1, 2012.

Marshall, A. G., Hudson, D., Wheeler, M. C., Hendon, H. H., and Alves, O.: Assessing the simulation and prediction of rainfall associated with the MJO in the POAMA seasonal forecast system, Climate Dynamics, 37, 2129–2141, DOI 10.1007/s00382-010-0948-2, 2011.

Murphy, B. F., and Timbal, B.: A review of recent climate variability and climate change in southeastern Australia, International Journal of Climatology, 28, 859-879, 10.1002/joc.1627, 2008.

Pfister, N.: The case of an evacuation from Grafton, The Australian Journal of Emergency Management, 17, 19-29, 2002.

Piechota, T. C., Chiew, F. H. S., Dracup, J. A., and McMahon, T. A.: Seasonal streamflow forecasting in eastern Australia and the El Niño–Southern Oscillation, Water Resources Research, 34, 3035–3044, 10.1029/98WR02406, 1998.

Pook, M. J., McIntosh, P. C., and Meyers, G. A.: The synoptic decomposition of cool-season rainfall in the southeastern Australian cropping region, Journal of Applied Meteorology and Climatology, 45, 1156-1170, 10.1175/JAM2394.1, 2006.

Risbey, J. S., Pook, M. J., McIntosh, P. C., Wheeler, M. C., and Hendon, H. H.: On the remote drivers of rainfall variability in Australia, Monthly Weather Review, 137, 3233–3253, 10.1175/2009MWR2861.1, 2009.

Roberts, N. M., Cole, S. J., Forbes, R. M., Mooreb, R. J., and Boswell, D.: Use of high-resolution NWP rainfall and river flow forecasts for advance warning of the Carlisle flood, north-west England, Meteorological Applications, 16, 23-34, 10.1002/met.94, 2008.

Robertson, D. E., and Wang, Q. J.: A Bayesian approach to predictor selection for seasonal streamflow forecasting, Journal of Hydrometeorology, 13, 155-171, 10.1175/JHM-D-10-05009.1, 2012.

Robertson, D. E., Pokhrel, P., and Wang, Q. J.: Improving statistical forecasts of seasonal streamflows using hydrological model output, Hydrology and Earth System Sciences, 17, 579-593, 10.5194/hess-17-579-2013, 2013a.

Robertson, D. E., Shrestha, D. L., and Wang, Q. J.: Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting, Hydrology and Earth System Sciences, 17, 3587–3603, doi:10.5194/hess-17-3587-2013, 2013b.

Robertson, D. E., and Wang, Q. J.: Seasonal Forecasts of Unregulated Inflows into the Murray River, Australia, Water Resour Manage, 27, 2747-2769, 10.1007/s11269-013-0313-4, 2013.

Rosenberg, E. A., Wood, A. W., and Steinemann, A. C.: Statistical applications of physically based hydrologic models to seasonal streamflow forecasts, Water Resources Research, 47, W00H14, 10.1029/2010WR010101, 2011.

Saji, N. H., Goswami, B. N., Vinayachandran, P. N., and Yamagata, T.: A dipole mode in the tropical Indian Ocean, Nature, 401, 360-363, 1999.

Schepen, A., Wang, Q. J., and Robertson, D.: Evidence for using lagged climate indices to forecast Australian seasonal rainfall, Journal of Climate, 25, 1230–1246, 10.1175/JCLI-D-11-00156.1, 2012a.

Schepen, A., Wang, Q. J., and Robertson, D. E.: Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall, Journal of Geophysical Research, 117, D20107, 10.1029/2012JD018011, 2012b.

Sharma, A.: Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: part 3 — a nonparametric probabilistic forecast model, Journal of Hydrology, 239, 249-258, 10.1016/S0022-1694(00)00348-6, 2000.

Shrestha, D. L., Robertson, D. E., Wang, Q. J., Pagano, T. C., and Hapuarachchi, H. A. P.: Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose, Hydrology and Earth System Sciences, 17, 1913-1931, 10.5194/hess-17-1913-2013, 2013.

Shukla, S., and Lettenmaier, D. P.: Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill, Hydrology and Earth System Sciences, 15, 3529–3538, 10.5194/hess-15-3529-2011, 2011.

Smith, T. M., Reynolds, R. W., T.C.Peterson, and Lawrimore, J.: Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006), Journal of Climate, 21, 2283–2296, 2008.

Troup, A. J.: The southern oscillation, Quarterly Journal of the Royal Meteorological Society, 91, 490–506, 10.1002/qj.49709139009, 1965.

Vaze, J., Perraud, J., Teng, J., Chiew, F., Wang, B., and Yang, Z.: Catchment Water Yield Estimation Tools (CWYET), 34th World Congress of the International Association for Hydro-environment Research and Engineering and 33rd Hydrology and Water Resources Symposium and the 10th Conference on Hydraulics in Water Engineering, Brisbane, 2011, 1554-1561, 2011.

Verdon, D. C., Wyatt, A. M., Kiem, A. S., and Franks, S. W.: Multidecadal variability of rainfall and streamflow: Eastern Australia, Water Resources Research, 40, W10201, 10.1029/2004WR003234., 2004.

Verdon, D. C., and Franks, S. W.: Indian Ocean sea surface temperature variability and winter rainfall: Eastern Australia, Water Resources Research, 41, W09413, 10.1029/2004WR003845, 2005.

Wang, Q. J., Robertson, D. E., and Chiew, F. H. S.: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites, Water Resources Research, 45, W05407, 10.1029/2008WR007355, 2009.

Wang, Q. J., and Robertson, D. E.: Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences, Water Resources Research, 47, W02546, 10.1029/2010WR009333, 2011.

Wang, Q. J., Schepen, A., and Robertson, D. E.: Merging seasonal rainfall forecasts from multiple statistical models through Bayesian model averaging, Journal of Climate, 25, 5524-5537, 10.1175/JCLI-D-11-00386.1, 2012a.

Wang, Q. J., Shrestha, D. L., Robertson, D. E., and Pokhrel, P.: A log-sinh transformation for data normalization and variance stabilization, Water Resources Research, 48, W05514, 10.1029/2011WR010973., 2012b.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Elsevier, New York, 648 pp., 1995.

Yeo, I. K., and Johnson, R. A.: A new family of power transformations to improve normality or symmetry, Biometrika, 87, 954–959, 10.1093/biomet/87.4.954, 2000.

Table 1 Charactericstics of catchments used in this study.

Name	Short name	Streamflow record used	Fraction of record missing	Area (km²)	Annual rainfall (mm)	Annual runoff (mm)	Runoff coefficient
Orara River at Bawden Bridge	ORB	1956-2006	4.2%	1823	1396	407	0.29
Nowendoc River at Rocks Crossing	NOR	1950-2006	3.9%	1898	1155	258	0.22
Abercrombie River at Hadley No. 2	ABH	1960-2005	0.5%	1626	842	117	0.14
Murray River at Biggara	MUR	1950-2005	2.5%	1254	1178	446	0.38
Mitta Mitta River at Hinnomunjie	MMH	1950-2006	2.6%	1528	1343	297	0.22
Tarwin River at Meeniyan	TAW	1955-2006	3.1%	1066	1084	233	0.21

Table 2 List of oceanic and atmospheric climate indices used as predictors.

Index	Description
Southern Oscillation Index (SOI)	Troup (1965)
NINO3	Mean SST anomaly over 150–90°W and 5°N-5°S
NINO3.4	Mean SST anomaly over 170–120°W and 5°N–5°S
NINO4	Mean SST anomaly over 150–160°E and 5°N–5°S
ENSO Modoki Index (EMI)	Ashok et al. (2003)
Indian Ocean Dipole Mode Index (DMI)	Saji et al. (1999)
Indian Ocean West Pole Index (WPI)	Saji et al. (1999)
Indian Ocean East Pole Index (EPI)	Saji et al. (1999)
Indonesia Index (II)	Verdon and Franks (2005)
Tasman Sea Index (TSI)	Murphy and Timbal (2008)
140°E Blocking Index (B140)	Risbey et al. (2009)

Fig. 1 Catchments (shaded) and streamflow gauge sites (black dots) used in this study.

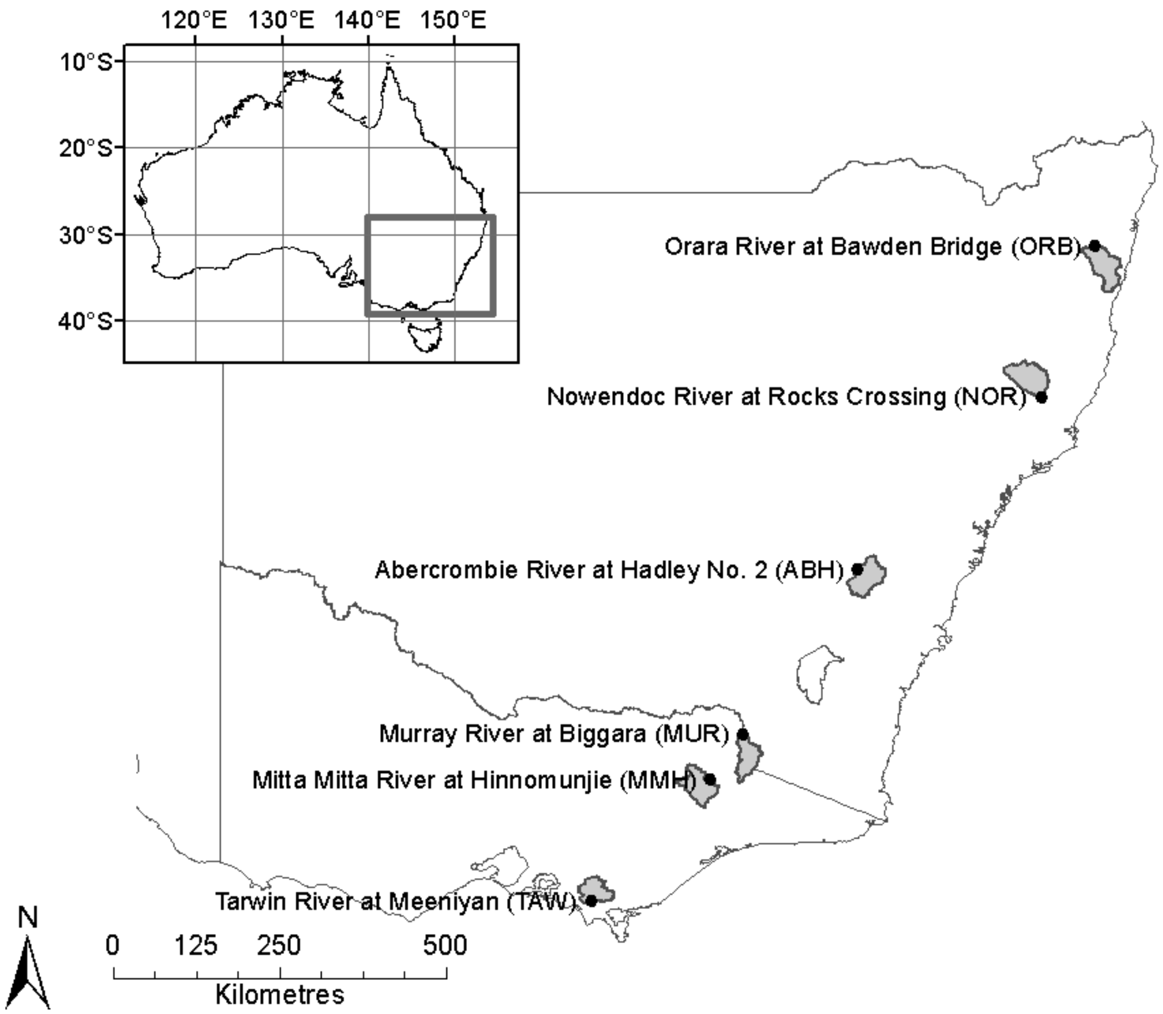

Fig. 2 Catchment streamflow characteristics. Black dots show average monthly streamflows. Boxes show maximum five-day streamflow (Max5D - blue) and maximum 1-day streamflow (Max1D - red) occurring during each month for exceedance probabilities of 50% (Q_{50}, bottom edge) to 10% (Q_{10}, top edge), with box centreline showing Max5D/Max1D streamflows of exceedance probability of 25% (Q_{25}).

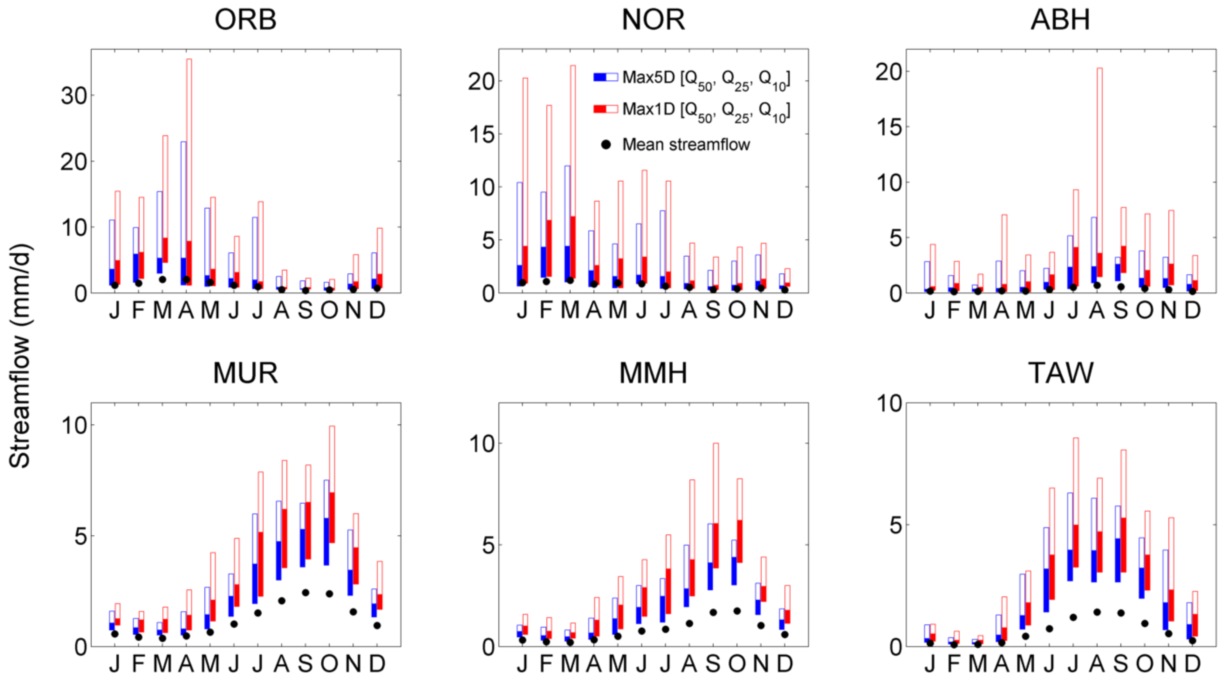

Fig. 3 Schematic of forecast model. (a) Example of individual forecast model generated with the Bayesian joint probability method. In this example, catchment wetness (CW) and NINO3.4 predictors are used to predict Max1D streamflows. Rainfall is included as a joint predictand to elicit more information from the climate indices. Parameters for the transforms and joint probability distribution are inferred jointly. This process is repeated for thirteen different predictor-sets. (b) The forecasts from thirteen BJP models are weighted based on cross-validated predictive performance with Bayesian model averaging (BMA) to produce a merged BJP-BMA forecast. The use of a symmetric Dirichlet prior encourages even weights in instances of high sampling uncertainty. See text for details.

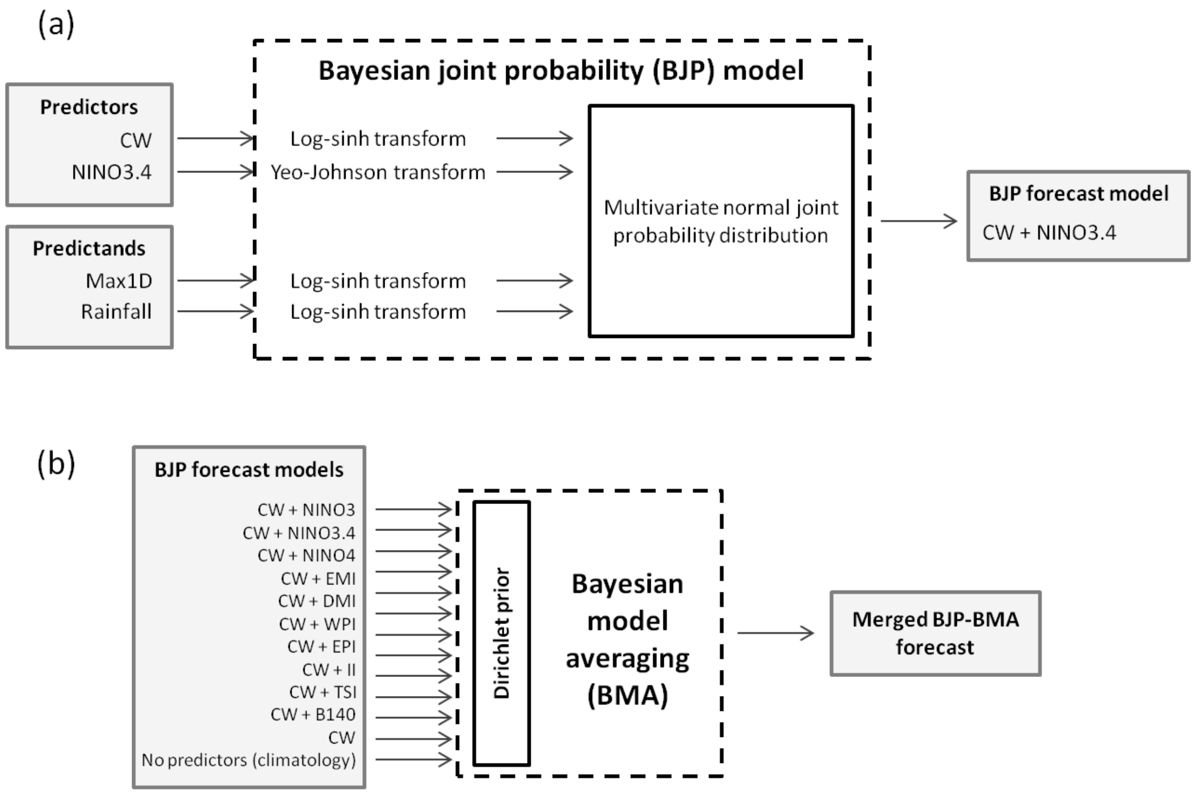

Fig. 4 Fit of log-sinh transformed normal distributions to Max1D values for two months. Red circles show actual values, black solid line shows fitted log-sinh tranform, dashed lines show [0.1, 0.9] confidence intervals.

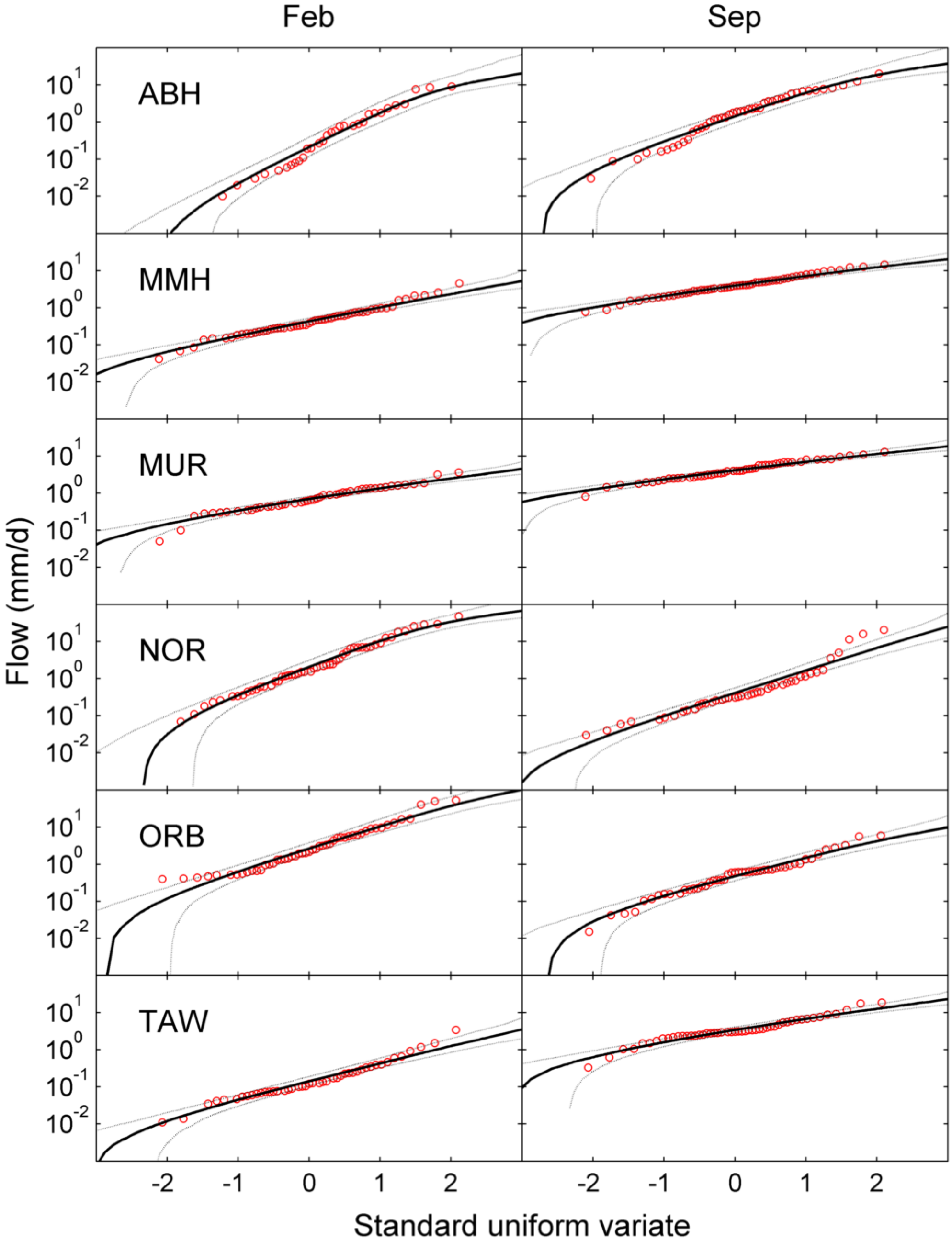

Fig. 5 Forecast reliability diagrams at two catchments for Max1D streamflows of exceedance probability ≤50%. (Forecasts are divided into five bins. 1:1 dashed lines, perfectly reliable forecast; circles, observed relative frequency; vertical lines, [0.05, 0.95] uncertainty interval of observed relative frequency; inserts, number of events in the different forecast probability bins.)

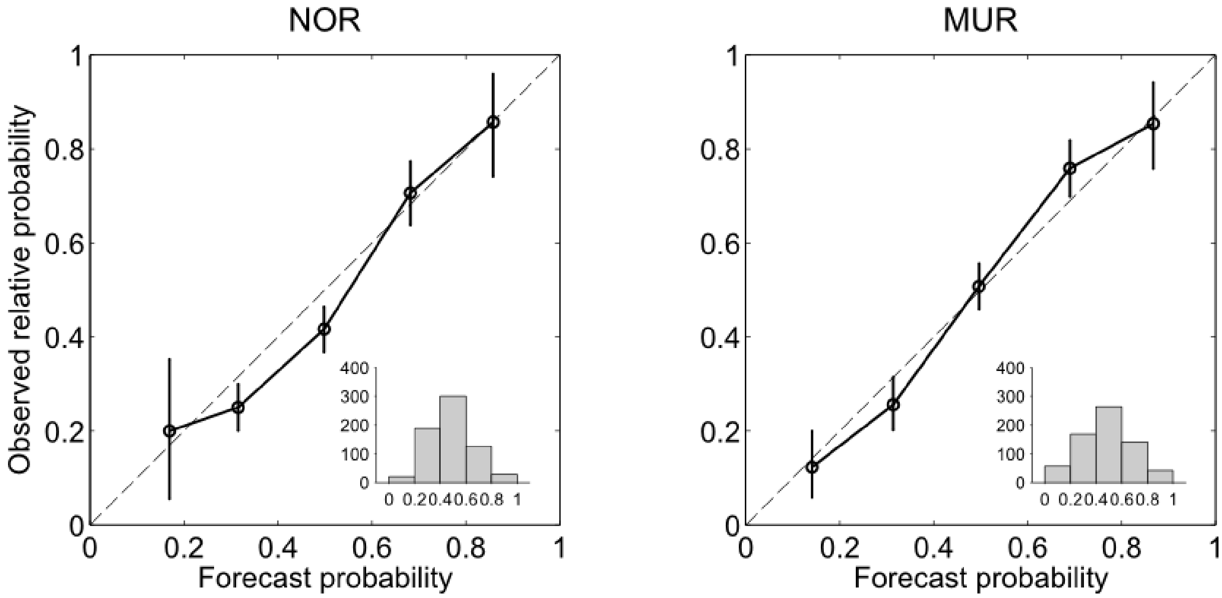

Fig. 6 Example forecast time series of cross-validated BJP-BMA for Max1D. Red circles show observed Max1D values, black points and lines show mean forecast and [0.1, 0.9] credible prediction intervals.

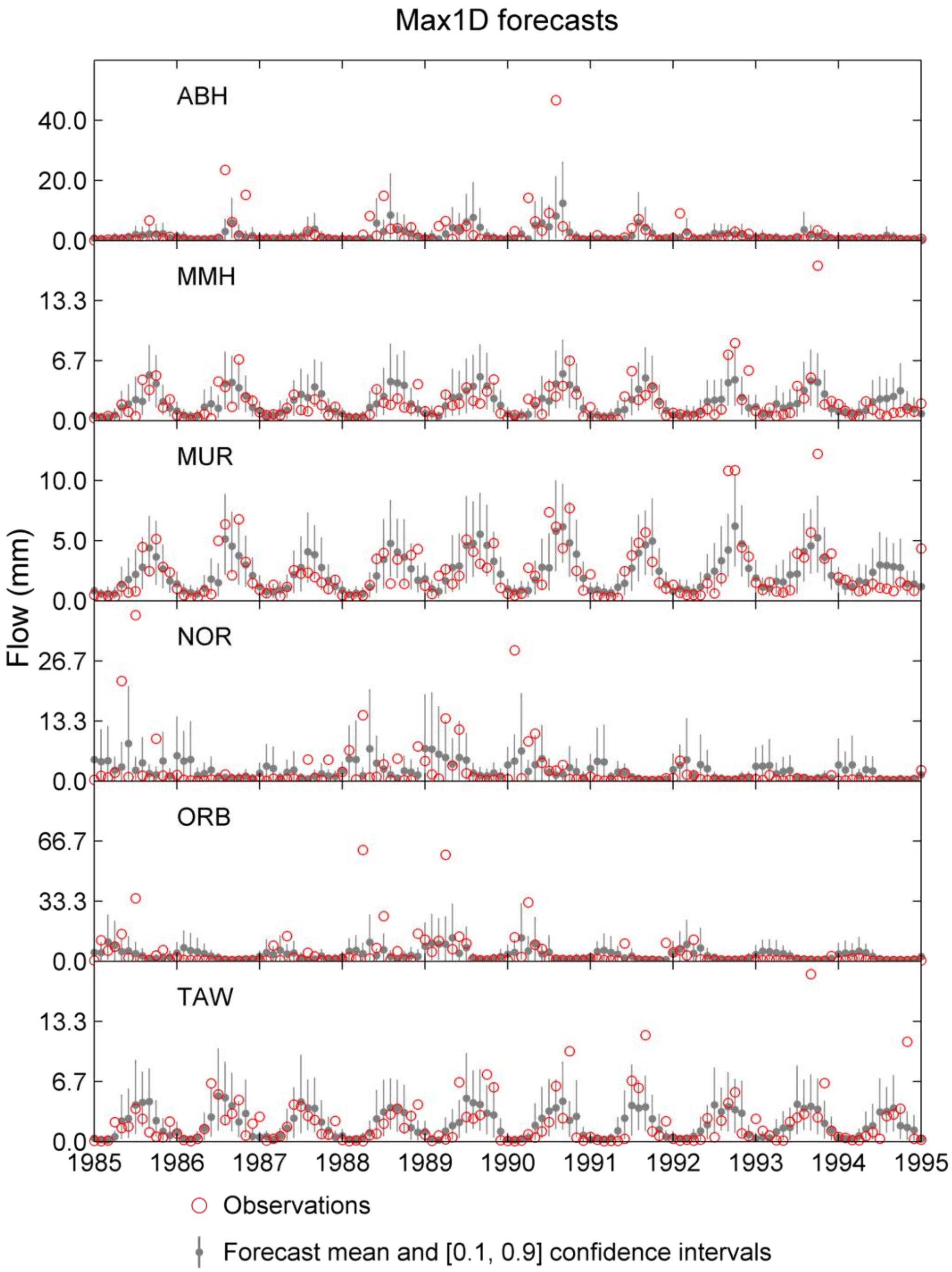

Fig. 7 RMSEP skill scores. Catchments are ordered by their location, from northernmost (top) to southernmost (bottom). (a) Max1D streamflows for 1-month forecasts, (b) Max5D streamflows for 1-month forecasts, (c) Max1D streamflows for 3-month forecasts, and (d) Max5D streamflows at 3-month forecasts. Scores show proportional improvement of forecasts over climatology forecasts.

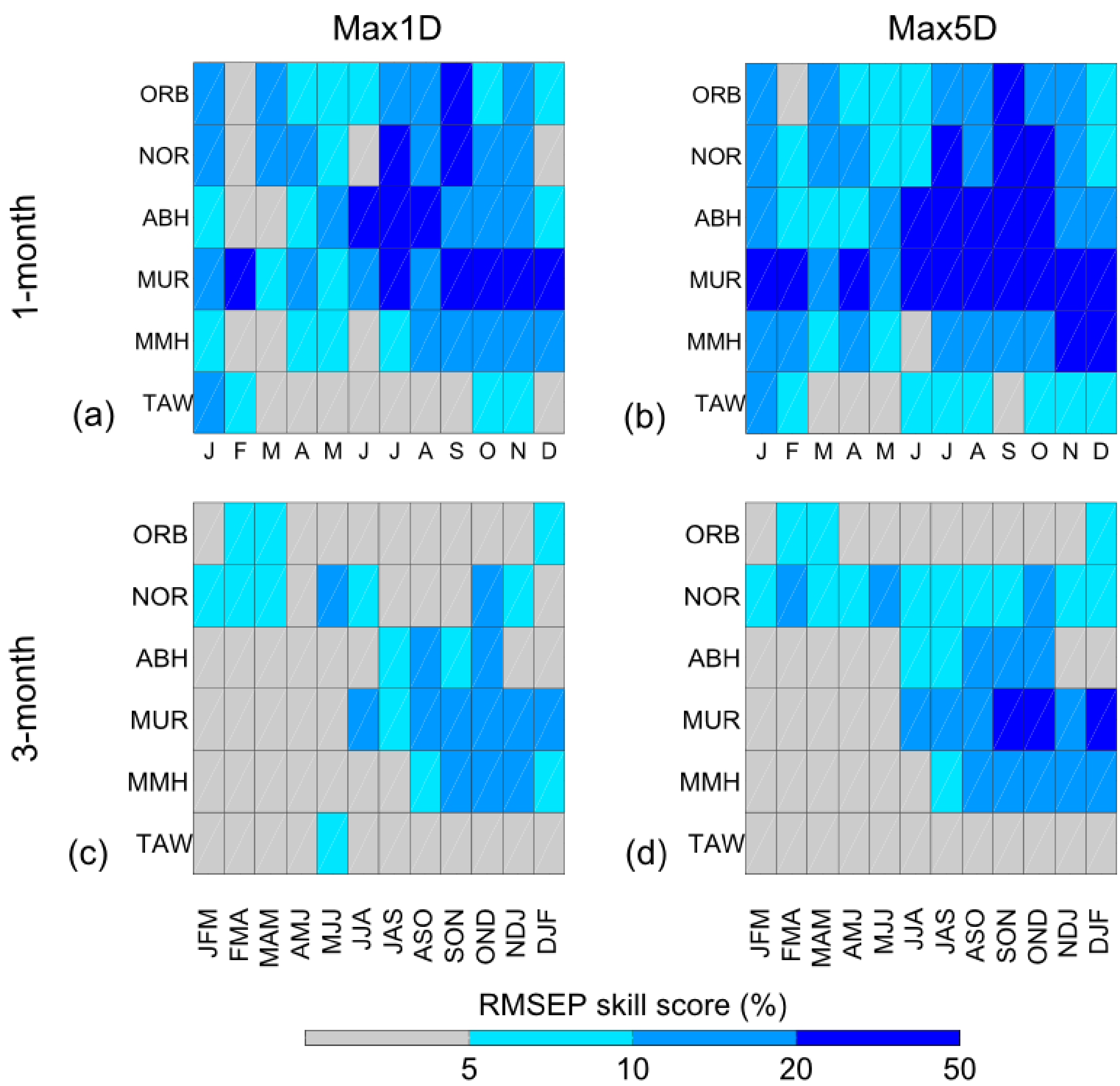

Fig. 8 Skill added by climate indices to forecasts. Plot shows RMSEP skill scores for Max1D 1-month forecasts calculated with respect to BJP forecasts generated with only catchment wetness as a predictor. Scores show proportional improvement of BJP-BMA forecasts over BJP forecasts generated with only catchment wetness as a predictor.

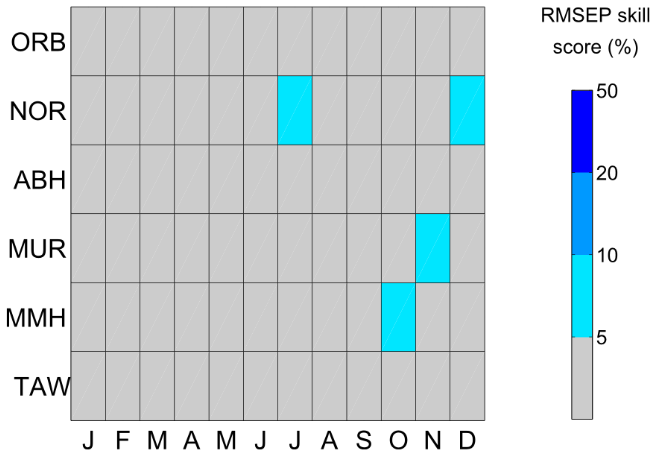

Fig. 9 Brier skill scores calculated at three streamflow thresholds for 1-month forecasts. Scores show proportional improvement of BJP-BMA forecasts over climatology forecasts.

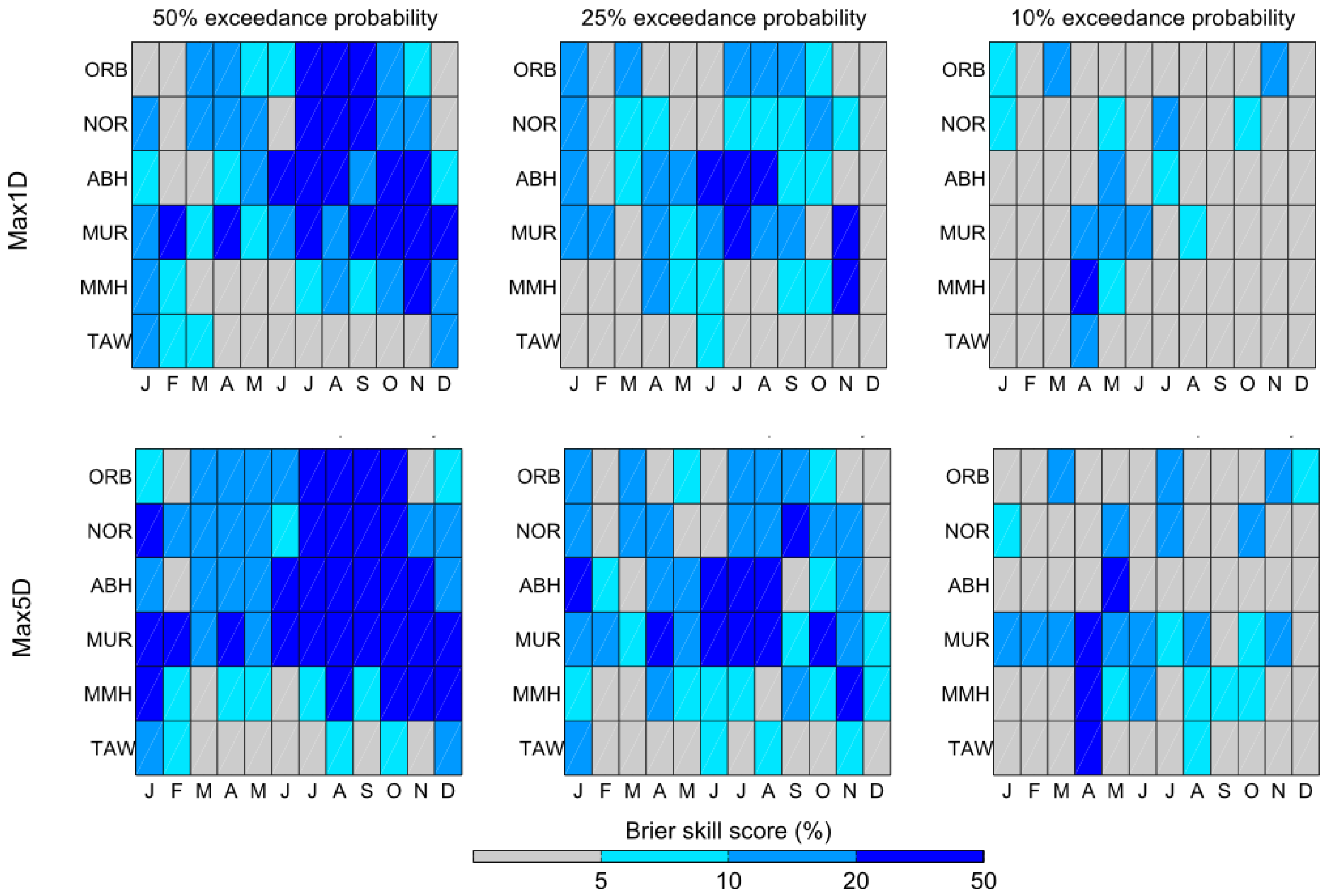

Fig. 10 Evidence of skill from the log-likelihood ratio (LLR) at three streamflow thresholds for 1-month forecasts. Scores show evidence of skill of BJP-BMA forecasts over climatology forecasts. Categories are taken from Schepen et al. (2012a): little evidence of skill where LLR<2; positive evidence where 2<LLR>4; strong evidence where 4<LLR>6; very strong evidence where LLR>6.

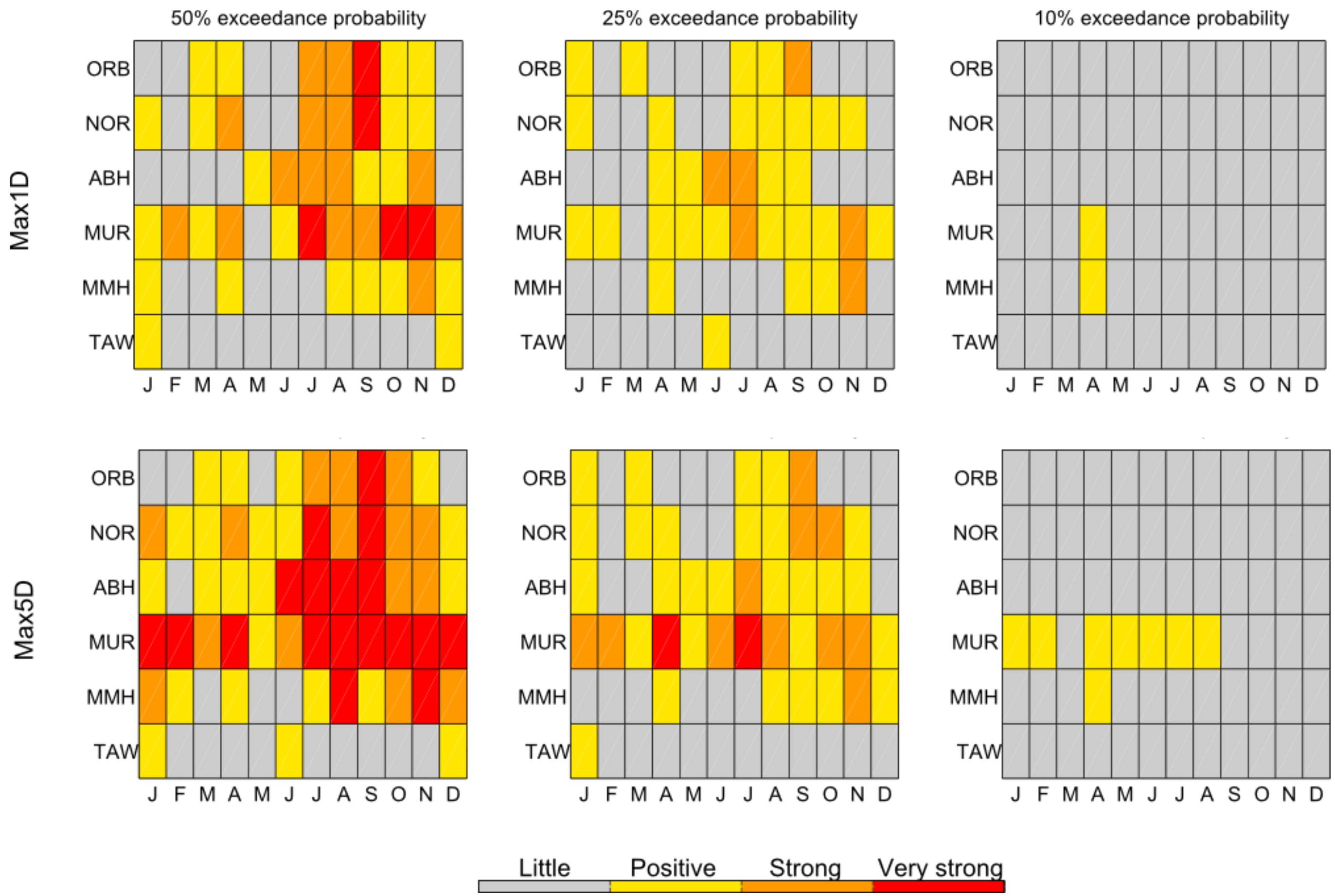

Fig. 11 Brier skill scores calculated at three streamflow thresholds for 3-month forecasts. Scores show proportional improvement of BJP-BMA forecasts over climatology forecasts.

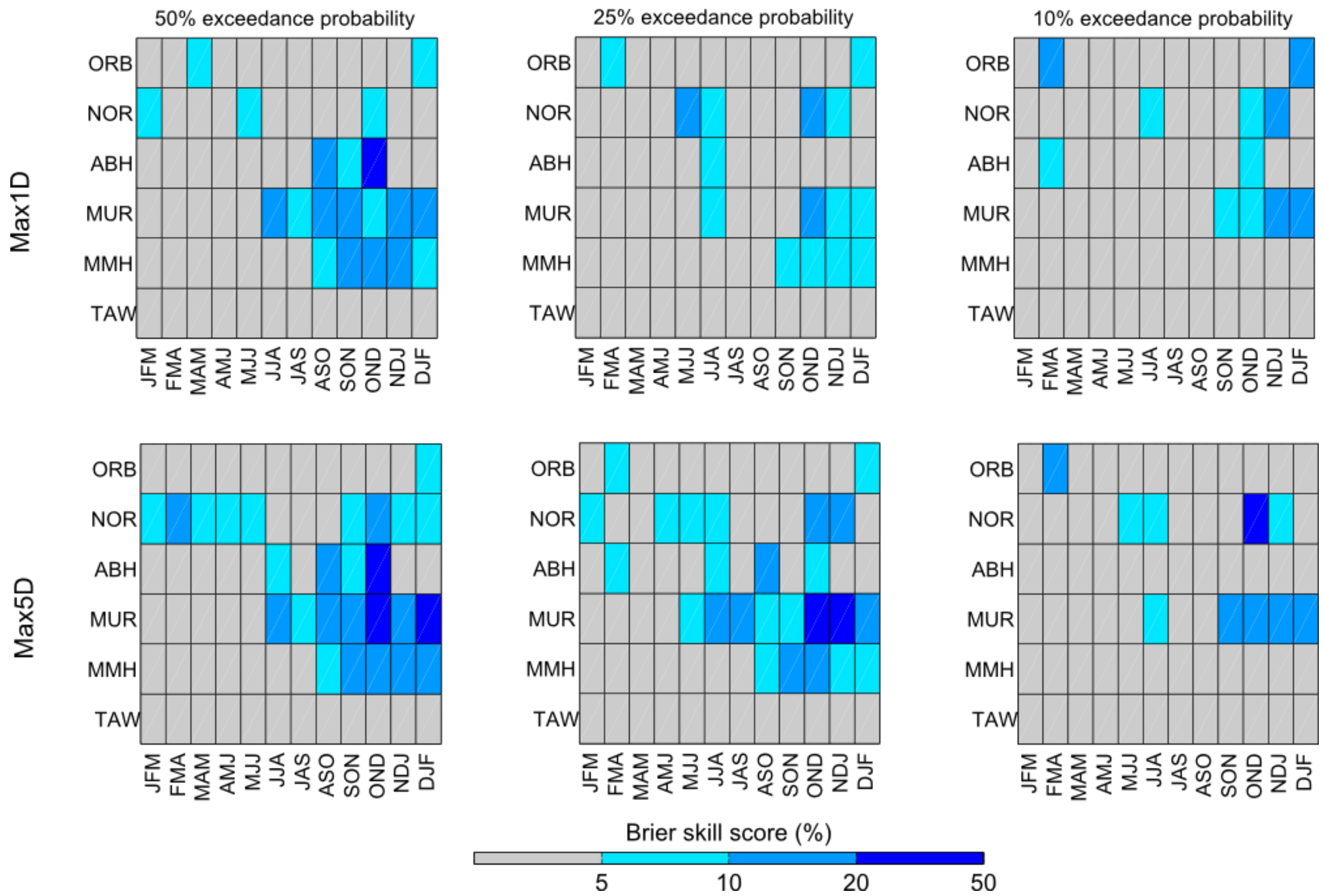

Fig. 12 Evidence of skill from the log-likelihood ratio at three streamflow thresholds for 3-month forecasts. Scores show evidence of skill of BJP-BMA forecasts over climatology forecasts. Categories are taken from Schepen et al. (2012a): little evidence of skill where LLR<2; positive evidence where 2<LLR>4; strong evidence where 4<LLR>6; very strong evidence where LLR>6.