Complex relationship between seasonal streamflow forecast skill and value in reservoir operations

Sean W. D. Turner[1,2], James C. Bennett[3,4], David E. Robertson[3], Stefano Galelli[5]

[1]Pacific Northwest National Laboratory, College Park, MD, USA.
[2]SUTD-MIT International Design Centre, Singapore University of Technology and Design, 487372, Singapore.
[3]CSIRO, Melbourne, Clayton, Victoria 3168, Australia.
[4]Institute for Marine and Antarctic Studies, University of Tasmania, Battery Point, Tasmania 7004, Australia.
[4]Pillar of Engineering Systems and Design, Singapore University of Technology and Design, 487372, Singapore.

Correspondence to: Stefano Galelli (stefano_galelli@sutd.edu.sg)

Abstract. Considerable research effort has recently been directed at improving and operationalising ensemble seasonal streamflow forecasts. Whilst this creates new opportunities for improving the performance of water resources systems, there may also be associated risks. Here we explore these potential risks by examining the sensitivity of forecast value (improvement in system performance brought about by adopting forecasts) to changes in the forecast skill for a range of hypothetical reservoir designs with contrasting operating objectives. Forecast-informed operations are simulated using rolling-horizon, adaptive control and then benchmarked against optimised control rules to assess performance improvements. Results show that there exists a strong relationship between forecast skill and value for systems operated to maintain a target water level. But this relationship breaks down when the reservoir is operated to satisfy a target demand for water; good forecast accuracy does not necessarily translate into performance improvement. We show that the primary cause of this behaviour is the buffering role played by storage in water supply reservoirs, which renders the forecast superfluous for long periods of the operation. System performance depends primarily on forecast accuracy when critical decisions are made—namely during severe drought. As it is not possible to know in advance if a forecast will perform well at such moments, we advocate measuring the consistency of forecast performance, through bootstrap resampling, to indicate potential usefulness in storage operations. Our results highlight the need for sensitivity assessment in value-of-forecast studies involving reservoirs with supply objectives.

1 Introduction

Coupled natural-engineered water resources systems provide a multitude of services to society. A properly functioning system can ensure reliable public water supply, support agricultural and industrial activity, produce clean hydroelectricity, provide amenity, sustain ecosystems and protect communities against damaging floods. But these benefits are by no means guaranteed; the performance of a given system depends on the quality of its operating scheme and the intelligence used to support management decisions on the storage, release and transfer of water. Typically, such operating decisions are governed by control rules based on observable system state variables. For example, the operator might select from a predefined lookup table the desired volume of water to release from a reservoir based on the time of year, volume of water held in storage and current catchment conditions (soil moisture, snow pack, etc.). The problem with this approach is that the decisions it recommends are optimal only under the narrow range of historical forcing conditions upon which they are trained. This is a major concern given emerging evidence of sharp trends and abrupt regime shifts in streamflow records and paleo reconstructions (Turner and Galelli 2016a). Flexible, real-time operating schemes that adapt in response to seasonal streamflow forecasts are thus the vanguard of water resources management practice, seen widely as the natural successor to predefined control rules (Rayner et al. 2005, Brown 2010, Gong et al. 2010, Brown et al. 2015).

A move toward schemes informed by seasonal streamflow forecasts would benefit from a wealth of recent science advances, including new ensemble seasonal streamflow forecasting methods, adding to existing ensemble streamflow prediction (ESP) and regression methods (e.g., Wang and Robertson 2011; Olson et al. 2016; Pagano et al. 2014; see review by Yuan et al. 2015). Seasonal streamflow forecast services are becoming available in countries such as the United States, Australia and Sweden. An emerging field of research has begun to demonstrate the value of seasonal streamflow forecasts when applied to real-world water management problems, such as determining the appropriate water release from a reservoir—the focus of the present study. Water release decisions can be improved with seasonal forecasts across a variety of reservoir types, including hydropower dams (Kim and Palmer 1997, Faber and Stedinger 2001, Hamlet et al. 2002, Alemu et al. 2010, Block 2011), water supply reservoirs (Anghileri et al. 2016, Zhao and Zhao 2014, Li et al. 2014) and reservoir systems operated for multiple competing objectives (Graham and Georgakakos 2010, Georgakakos et al. 2012). Operators considering whether to adopt a forecast-informed operating scheme should be encouraged by these outcomes. But they also need to understand the associated risks and uncertainties (Goddard *et al.*, 2010). If the new scheme increases the benefits of a system by, say, 20% in a simulation experiment, then can the operator assume that 20 % will be guaranteed when the scheme is implemented in practice?

To explore uncertainty in the value of seasonal forecasts applied to reservoir operations, we conduct two simulation experiments using reservoir inflow time series recorded at four contrasting catchments located in Australia. Our first experiment uses synthetically generated forecasts of varying skill to test for sensitivity in simulated forecast value across a range of reservoirs. Forecast value is calculated using cumulative penalty costs incurred for deviation from a predefined objective over a 30-year simulation. We define two simple, contrasting objectives: a "supply objective", which aims to maintain a target release by allowing storage to vary; and a "level objective", which aims to maintain a target storage level by varying the release. As we shall see, the contrast in performance between the two operational settings is striking. Our second experiment aims at explaining this outcome by applying an advanced seasonal streamflow forecast system to a range of fabricated reservoirs with deliberately adjusted design parameters. Results provide new insights into the risks operators take when applying seasonal forecasts to critical management decisions in systems dominated by a supply objective.

2 Materials and methods

2.1 Inflow records and forecasts

Our experiments are based on four reservoir inflow records (Table 1), which were selected because they represent a range of hydrological regimes (perennial, ephemeral, intermittent) across different regions of Australia. For each inflow record, we study the period 1982 – 2010 (Figure 1), for which forecasts are available.

2.1.1 Synthetic forecasts: Martingale Model of Forecast Evolution (MMFE)

Our first experiment is a sensitivity test for forecast value as a function of forecast quality. To generate many forecasts of varying quality, we use the Martingale Model of Forecast Evolution (MMFE) (Heath and Jackson, 1994). This model can be considered superior to one that simply imposes random error on observed values, since it captures the way in which forecast error decreases as the forecast horizon shortens and more information becomes available to the forecaster (known as the evolution of forecast error) (Zhao et al. 2011). Here we vary an 'injected error' parameter, which controls the error of the synthetic forecast. The injected error takes values between 0 and 1, where 0 generates a perfect forecast and 1 generates a sufficiently error-laden forecast to ensure that our experiments include a wide range of forecast performance. (Note that an error injected of 1 should not be interpreted as having any physical meaning, such as equivalence to climatology.) Because the model uses probabilistic sampling to generate forecasts for a given error, the deviation of the forecast from the

observation will vary in time, although the temporal average of the error will match the error injected given enough data points. The code for this model is available open source (Turner and Galelli, 2017).

Here the synthetic forecasts are constructed to overlay the four inflow time series described above. For each catchment, we generate 1000, 12-month ahead, monthly-resolution synthetic forecasts. The quality of the forecasts is varied by sampling from a uniform distribution between 0 and 1 to feed the injected error parameter. Each forecast should be considered a separate deterministic forecast rather than a member of a forecast ensemble. Figure 2 displays the goodness-of-fit for these forecasts as a function of the error injected at each forecast lead-time (forecasted against observed values for the period 1982 – 2010). The goodness-of-fit measure is the normalised Root Mean Squared Error (nRMSE), which is the RMSE divided by the standard deviation of observations. Since zero error corresponds to the perfect forecast, all lead times have nRMSE of 0 when no error is injected. As the injected error increases, the performance gap between short and longer lead time forecasts widens, reflecting a deterioration of forecast performance that one would expect with a weaker forecasting system.

2.1.2 Actual forecasts: Forecast Guided Stochastic Scenarios (FoGSS)

In our second experiment, we apply the forecast guided stochastic scenarios (FoGSS) experimental streamflow forecast system (Bennett et al. 2016; Bennett et al. 2017, this issue). FoGSS combines dynamical climate forecasts, statistical post-processsing, rainfall-runoff modelling and statistical error modelling to produce 12-month ensemble streamflow forecasts. The method behind FoGSS is complex, and accordingly we only give an overview here. A full description, including detailed equations, is available in Bennett et al. (2016) and Schepen and Wang (2014). FoGSS makes use of climate forecasts from the Predictive Ocean and Atmosphere Model for Australia (POAMA) (Hudson et al., 2013; Marshall et al., 2014), post-processed with the method of calibration, bridging and merging (CBaM; Schepen and Wang 2014, Schepen et al. 2014, Peng et al. 2014) to produce ensemble precipitation forecasts. CBaM corrects biases, removes noise, downscales forecasts to catchment areas and ensures ensembles are statistically reliable. The precipitation forecasts are then used to force the monthly Water Partitioning and Balance (Wapaba) hydrological model (Wang et al. 2011). Hydrological prediction uncertainty is handled with a 3-stage error model, which reduces bias and errors, propagates uncertainty, and ensures streamflow forecast ensembles are reliable (Wang et al. 2012; Li et al. 2013; Li et al 2015; Li et al 2016). In months where forecasts are not informative, FoGSS is designed to return a climatological forecast. FoGSS produces 1000-member ensemble streamflow forecasts in the form of monthly-resolution time-series with a 12-month forecast horizon.

FoGSS hindcasts are available for selected Australian catchments for the years 1982-2010 (based on the availability of POAMA reforecasts), including the four catchments examined in this study. The hindcasts are generated using a leave-5-years-out cross-validation scheme (Bennett et al. 2016), which ensures that the performance of FoGSS hindcasts are not artificially inflated. We characterise forecast performance with a skill score calculated from a well-known probabilistic error score, the continuous ranked probability score (CRPS; see, e.g., Gneiting and Raftery, 2007). The skill score is calculated by:

$$CRPSS = \frac{CRPS_{Ref} - CRPS}{CRPS_{Ref}} \times 100\% \qquad \text{Equation 1}$$

where $CRPS$ is the error of FoGSS forecasts and $CRPS_{Ref}$ is the error of a reference forecast, in this case a naïve climatology. The climatology reference forecast is generated from a transformed normal distribution (Wang et al. 2012), fitted to streamflow data using the same leave-5-years out cross-validation as applied to the FoGSS forecasts (Bennett et al. 2016).

FoGSS exhibits a range of performance across the catchments used in this study (Figure 3). In the Upper Yarra, Burrinjuck and Eppalock catchments, FoGSS forecasts are generally skilful at lead times of 0-2 months, extending to more than 3 months at certain times of year (in particular for the Upper Yarra and Burrinjuck catchments). Skill is much less evident in

the Serpentine catchment, only appearing evident in a few months of the year (January, August, September, November), even at short lead times. Generally, at longer lead times forecasts are at worst similar to climatology. The only exception is the Eppalock catchment for February and March, where strongly negative skills occur. In the Eppalock catchment, February and March usually experience very low (to zero) inflows. FoGSS forecasts in the Eppalock catchment are slightly positively biased at longer lead times. However, because inflows are so low during these months, these errors have very little influence on annual (or even seasonal) water balances.

2.2 Reservoir setup

2.2.1 Reservoir model and design specifications

We use monthly resolution reservoir simulation and operating schemes in both experiments. Each reservoir obeys basic mass balance, meaning volume of water held in storage (S_{t+1}) is equal to the previous month's storage (S_t) plus total inflow to the reservoir (Q_t) minus volume of water released (R_t). (Evaporation and other water losses are ignored for simplicity.) The release R_t is constrained physically to a maximum of the available water in storage plus any incoming inflows during period t (Equation 2).

$$S_{t+1} = S_t + Q_t - R_t - Spill_t \qquad \text{Equation 2}$$

$$Spill_t = \max(S_t + Q_t - R_t - S_{cap}, 0)$$

$$subject\ to \quad 0 \le S \le S_{cap}\ ;\ 0 \le R_t \le \min(S_t + Q_t, R_{max})$$

where S_{cap} is the capacity of the reservoir and R_{max} is the maximum water release, taken in this study as twice the release target to give the operator ample storage level control. All excess water is spilled.

Rather than using the real-world specifications of the four reservoirs corresponding to our inflow records, we vary the size and operation of reservoirs. This approach gives two important advantages. First, it allows us to specify operating objectives relevant to the study question (level objective versus supply objective). Second, it enables us to examine the value of forecasts for reservoirs sensitive to different types of hydrological conditions. Specifically, by changing the design parameters of a reservoir, it becomes sensitive to droughts of different intensity and duration. So a wider range of reservoirs allows us to test reservoir performance across a variety of time periods within the simulation.

To fabricate these reservoirs we begin by assuming a time-based reliability of 0.95 in all instances. Time-based reliability is the ratio of non-failure months—months during which the demand for water is satisfied in full—to the total number of months simulated. This reliability target can be considered a realistic service standard, since in designing these reservoirs we assume a standard operating policy where reservoirs release to meet as much of the demand as possible from the water available in storage and from incoming flow. A constant demand for water is assigned for eight alternative reservoirs by varying the draft ratio (ratio of demand to mean inflow) for values between 0.2 and 0.9 in increments of 0.1. The reservoir capacity required to achieve the target reliability is then determined for each demand using an iterative simulation procedure (storage-yield-reliability analysis). Since the reliability is held constant across all reservoirs, an incremental increase in the draft ratio results in a larger design storage capacity—as shown in Table 2. In other words, when the demand on a reservoir increases, the storage must also be increased so that the required reliability (0.95) is achieved. As demand and storage increase, drift decreases and critical period increases. Critical period gives the time taken for the reservoir to empty under recorded droughts, whilst drift indicates the presence of within-year or over-year behaviour (drift greater than 1 normally suggests that the reservoir will fill and spill each year). The wide variance across these indicators suggests that as demand is adjusted, the storage dynamics are affected and the reservoirs will be sensitive to different hydrological events. For example, a reservoir with large demand and storage will easily tolerate short-duration periods of extremely low inflow, but will be

vulnerable under very long periods of moderately low flows. Conversely, small reservoirs with lower demands will fail easily under short duration droughts, but will usually tolerate moderately low flows for long periods, because the demand will be too small to cause drawdown. Appendix 1 provides more detailed definitions of the parameters and variables discussed above. All computations are executed using R package *reservoir* (Turner and Galelli 2016b) using observed inflows for the period 1982 – 2010.

2.3 Operating schemes

If we allow that the objective of a reservoir can be described adequately by a mathematical function, we can quantify operating performance by imposing penalty costs for deviations from that objective. Then to understand the value of a forecast-informed operating model, we need simply to compare that performance against a benchmark. We therefore apply two operating schemes in this study: a *benchmark scheme* that ignores forecasts and a *forecast-informed scheme* that makes use of forecasts. Since we are primarily interested in the value added by applying the forecasts to the operation, we must ensure that the performance differences between the two models are attributable to the forecast information rather than conceptual differences in the operating schemes applied. We therefore select two schemes that are conceptually similar (see section 2.3.2), whilst recognising standard, common practice. Our benchmark scheme guides the reservoir operation using control rules, which are established by optimising release decisions for historical conditions. Control rules (often termed "release policies", "hedging rules", or "rule curves") are very commonly applied in practice (Loucks et al. 2005), so they provide a realistic benchmark. Our forecast-informed scheme effectively adjusts those control rules in response to new information available through the forecast.

2.3.1 Benchmark scheme: control rules

The control rules we devise can be thought of as a look-up table that specifies reservoir release as a function of two state variables: volume of water held in storage (discretised uniformly into a manageable number of values) and month-of-year. In practice—and in simulation—the operator simply observes the current reservoir level and then implements release for the time of year as specified by these rules. These rules are designed with respect to the operating objectives and constraints of the system, and can be considered risk-based in the sense that they are conceived to minimise the expected cost of release decisions across the distribution of the inflow for each month. Costs are based on penalties associated with failure to meet the objectives of the reservoir (see Section 2.4).

The most rigorous way to design such rules is by optimisation. Here we use *stochastic dynamic programming* (SDP), which offers four significant advantages. First, SDP handles non-linearity in both the operation of the system and the objective functions. Second, SDP accounts for the effect of uncertainties, in this case stemming from inflows, on system dynamics. Third, SDP finds the optimal operation for a given model of the system (as opposed to other approaches that approximate the optimal solution). Fourth, SDP returns a cost associated with each combination of state variables, in this case the volume in storage and the month-of-year, known as Bellman's function. Bellman's function is useful for the forecast-informed operating scheme introduced in the following section. The inputs to our SDP model are the reservoir specifications, reservoir objective function and inflow time series, which provides inflow distributions for each month-of-year. The control rules are optimised by solving a backwards recursive procedure (Bellman 1956, Loucks et al. 2005), which is detailed in Appendix 2. We retrain the control rules for each year of simulation using the same data (1982-2010) and with the same leave-five-years-out cross-validation scheme employed in FoGSS (Section 2.1.2). SDP suffers from two well-known drawbacks: the exponential growth of computation with the number of state variables, and the need for an explicit model representing each component of the water system (Castelletti et al. 2010). These issues limit the application of SDP to relatively-small systems (e.g., maximum three to four reservoirs), but do not represent an obstacle in our study, which focuses on single-reservoir systems.

2.3.2 Forecast-informed scheme: rolling-horizon, adaptive control

To inform operations with forecasts, we adopt a *rolling-horizon, adaptive control* scheme—also known as *model predictive control* (Bertsekas 1976). The idea behind this scheme is that the a deterministic forecast can be used to run short simulations ($t = 1,2,\ldots,H$, where H is the forecast length in months) to evaluate changes in storage that would be experienced under alternative sequences of release decisions. The release decision sequence ($R_1, R_2,\ldots,R_H$) is optimised to minimise the cost over the forecast horizon H plus the cost associated with the resulting storage state:

$$\min_{R_{1,2,\ldots,H}} \left\{ \left[\sum_{t=1}^{H} C_t(R_t, S_t) \right] + X(S_{H+1}) \right\} \quad \text{Equation 3}$$

where C_t is the penalty cost calculated from the reservoir's objective function (see equations 4 and 5, below), and $X(\cdot)$ is a penalty cost function that accounts for the long-term effects of the release decisions being made. The latter helps avoid a short-term, greedy policy that optimises solely for operations in the following H months. We set the function $X(\cdot)$ equal to Bellman's function obtained when designing the control rules, since it contains costs that represent the risk of a given storage level for each month-of-the-year (Appendix 2). By using Bellman's function in this way, we effectively append the forecast-informed scheme to the control rules. In effect, this means that the information contained in the forecast is used to adjust the decisions that would be taken using the benchmark scheme—hence our prior statement that the two schemes are conceptually similar.

The optimisation problem is solved at each time step using deterministic dynamic programming, giving the precise optimal release sequence for the forecast horizon ($R_1, R_2,\ldots,R_H$). The first of these (R_1) is implemented in simulation and the remainder are discarded, since the optimisation is repeated on the next time step as a new forecast is issued (hence the term "rolling-horizon", Mayne et al. 2000). A deterministic approach does not exploit fully the information contained within the ensemble; this is its major drawback as compared to a stochastic approach. Yet, this does not mean While this approach ignores the spread of the ensemble (and therefore a key element of its value—Boucher et al., 2012), it provides a clear indication of the contribution of the forecast to the performance of the operation and is thus still a standard when dealing with seasonal forecasts (e.g., Anghileri et al., 2016). In contrast, methods that use the spread of the ensemble present a number of technical challenges. One cannot simply optimise the release decision by minimising the expected cost across all ensemble members, because this discounts the operator's ability to adjust the release in response to new information, resulting in over-conservative release decisions and thus weak performance (Raso et al., 2014). The established approach to incorporating information from the spread of the ensemble is Multi-Stage Stochastic Optimisation, which applies a reduced form of the ensemble known as a scenario tree to guide corrective decisions as new forecast data are revealed (Shapiro *et al.*, 2014). Whilst this approach has been applied in a handful of water related studies, including short-horizon problems (Raso *et al.*, 2014) as well as using seasonal streamflow forecasts (Housh et al. 2013, Xu et al. 2015), it relies on arbitrary decisions (such as the preferred scenario tree nodal structure), is computationally demanding, and is highly complex, making experimentation laborious and results hard to diagnose. For these reasons, we pursue the deterministic model predictive control method described above.

2.4 Operating objectives

We test two operating objectives: one that rewards meeting target releases (*supply objective*) and one that rewards meeting target storage levels (*level objective)*. The supply objective encourages full release of water to meet target demand except under drought conditions:

$$C^{supply} = \sum_{t=1}^{T} [\max(1 - R_t/D, 0)]^2 \quad \text{Equation 4}$$

where D is the demand and C^{supply} is the penalty cost used in the adaptive control scheme (Equation 3). The squared term creates an impetus to cut back the release to reduce the risk of major shortfalls that would occur if the reservoir failed (i.e., becomes fully depleted). Reservoir failure is often associated with highly damaging consequences, such as large water restrictions imposed on households and businesses. Operators therefore tend to hedge against the risk of failure by cutting back the release in small and frequent increments that are, in the long-run, preferable and ultimately less costly than relatively infrequent major shortfalls that would result from total storage depletion (Draper and Lund 2004).

The level objective encourages controlled releases to maintain a target storage level, which could represent operation for flood control (e.g., maintain sufficient flood buffer storage), amenity (e.g., avoid unsightly drawdown) or hydropower (maintain high hydraulic head). The objective penalises deviations from a target storage S^*, which is set arbitrarily to 75% of total storage capacity in the present study:

$$C^{level} = \sum_{t=1}^{T} (1 - S_t/S^*)^2 \quad \text{Equation 5}$$

where T is the final month of the simulation and C^{level} is the penalty cost used in the adaptive control scheme (Equation 3).

Figure 4 gives storage behaviour and release decisions implemented for 0.95 reliability reservoirs (draft ratio = 0.5) operated for the supply objective (Equation 4) and level objective (Equation 5), under rolling horizon, adaptive control operation with a perfect 12-month forecast. The figure shows the contrast in the frequency of decision-making for the two operating objectives. For the supply objective we see that the release is adjusted only under drought—predominantly during the Australia's Millennium Drought—and that there are multi-decade periods in which the operator simply releases to meet demand. For the level objective we see that the release must be adjusted constantly through the operating horizon to keep storage close to the target level of 75%. The main aim of the experiments described below is to elucidate how this distinction in operating behaviour affects the usefulness of applying seasonal forecasts in operations.

3 Experiment 1 – Characterising the uncertainty of forecast value in reservoir operations

3.1 Experiment description

The purpose of the first experiment is to examine the nature of uncertainty in forecast performance under two contrasting operating objectives (level objective versus supply objective). For this experiment we hold the reservoir design specifications constant (mid-range draft ratio of 0.5 selected for all four inflow time series). For each of the four reservoirs we follow these steps:

1. A set of control rules is optimised with the SDP approach over the period 1982-2010, where the objective is to minimise the sum of penalty costs over the simulation.
2. The adaptive control, rolling horizon scheme is run for a synthetic forecast generated by MMFE over the 1982-2010 period. The value of the forecast is measured by the percentage reduction in penalty cost relative to the control rules over the entire 1982-2010 period.
3. Step 2 is repeated 1000 times, once for each set of synthetic forecasts generated with the MMFE.

4. Steps 1-3 are executed twice—once for the supply objective and once for the level objective. The exact same set of 1000, monthly resolution, 12-month-ahead MMFE forecasts is applied in each case.

We then assess the performance of the forecast-informed operating scheme against the forecast error injected by the MMFE.

3.2 Results for experiment 1

Figure 5 shows the value of the forecast-informed scheme for each reservoir. The value of forecasts is presented as the reduction in cost relative to control rules (%). A positive cost reduction indicates that the forecast-informed scheme outperforms control rules, and a negative cost reduction indicates that control rules outperform the forecast-informed scheme. Forecasts with zero error (i.e., perfect forecasts) outperform control rules in all cases, regardless of the objective. Interestingly, when operated with a perfect forecast, the reservoirs operated to meet the supply objective enjoy a significantly larger percentage increase in performance (40 – 60%) compared with the reservoirs operated to the level objective (20 – 40%). This occurs because the target in the level objective reservoirs will often be achievable within one or two months of operation, meaning the perfect forecast skill available at longer lead times is surplus to requirement. The supply targeted reservoirs, in contrast, will benefit from the entire forecast as they drawn down during drought.

More striking is the contrast in behaviour between operational objectives as the forecast error is increased. For the supply objective (panels a – d) forecast value declines rapidly, becoming highly unstable with the injection of a moderate error into the forecast. For the level objective (panels e – h), the forecast value decreases relatively slowly, and the points remain tightly grouped for errors up to ~0.4. Taking Burrinjuck (Figure 5a) as an example, we find that an injected forecast error of 0.2 could result in cost reductions anywhere from -5% to +40% for the supply objective (i.e., the forecast-informed operations are outperformed by simple control rules by up to 5% in some instances). The same forecasts applied to the level objective (Figure 5e) result in cost reductions in the narrow region of 24 to 26%. Serpentine reservoir presents even greater sensitivity to injected error. Here an injected error of 0.3 gives cost reductions ranging between -50% and +50% with for the supply objective. The same forecasts appear to guarantee beneficial cost reductions of 5 to 15% when operating with a level objective.

These results show that for the supply objective, the measure of forecast error, quality, skill or goodness-of-fit do not always accurately predict whether that forecast will be valuable. We believe that this unexpected phenomenon relates to the role played by storage. When operated to the level objective, storage plays no role as a buffer. The release is simply adjusted to keep storage a desired level. Because inflows fluctuate constantly, release must be adjusted throughout the operation in response to forecasts issued (recall Figure 4). At moments when forecasts skill is weak, release decisions may underperform relative to control rules. At moments when forecast skill is strong, release decisions will improve on control rules. For the supply objective, however, storage actively buffers inflows. When storage levels are high, the operator can be assured that a short period of low inflows need not threaten the system performance, because the target release can be met by drawing on stored water. In such a case, it does not matter how accurate the next inflows forecast is: the release target will be met regardless. Generally, very large reservoirs may withstand a number of consecutive drought years before storages drop to levels that raise concern. Only then will the option of reducing the release be considered. The value of the forecast will be determined solely by its skill at a small number of periods during which the storages are sufficiently depleted to warrant hedging the release. Forecast skill is often measured by averaging errors over a long period of time (as done in Figure 3). Figure 5 shows that it is possible for a skilful forecast (measured on average) to generate a net reduction in performance if the skill level dips during the critical point in time where the forecast is mobilised. Similarly, it is also possible for a forecasting system that is on average unskilful to generate a net improvement in operating performance if forecasts happen to be accurate during that critical moment.

This ability of storages to buffer inflows also explains why the Upper Yarra Reservoir, under the supply objective, shows a stronger correlation between forecast value and forecast quality than Burrinjuck, Eppalock and Serpentine. Upper Yarra tolerates injected error in the forecast of up to ~0.4 before negative performance gains are observed—compared to ~0.2 injected error for the other three reservoirs. At draft ratio of 0.5, Upper Yarra has the shortest critical period, lowest storage ratio and highest drift value (a result of low variance in inflows for 1982-2010 relative to the other storages). In other words, the storage buffer in Upper Yarra will tend to provide less time between full and empty during drought. In such systems, adjustments to release decisions are required more frequently (as observed in Figure 4).

We now turn to experiment 2, which explores further the behaviour observed with the supply targeted reservoirs. We need to understand whether the same behaviour occurs with an actual forecast service (as opposed to synthetic forecasts). And we wish to explore further the possibility that variance in forecast skill through time is the explanation.

4 Experiment 2 – The importance of critical drought timing on forecast value

4.1 Experiment description

The primary aim of Experiment 2 is to determine whether the periods during which critical decisions are made can explain the wide variation in forecast value for a given forecast skill level when applied to reservoirs with the supply objective. For this experiment we keep the forecast input consistent and instead vary the timing of critical decision points in the simulation. This is achieved by adjusting the reservoir specifications in such a way that they respond to different types of drought (as described in section 2.2.1) so that critical decision periods change. Control rules are designed for all 32 reservoirs (four inflows, eight reservoir set ups) using the SDP approach as above. Operations are then simulated using both the control rules and the deterministic model predictive control model using the median value from the full FoGSS forecast ensemble (i.e., a deterministic forecast is constructed by taking the median of the ensemble at each lead time).

We compute the value of FoGSS forecasts in relation to both an upper benchmark (perfect forecast) and a lower benchmark (control rules):

$$Performance\ Gain = \frac{C^{cntrl} - C^{fcast}}{C^{cntrl} - C^{perfect}} \quad \text{Equation 6}$$

where C^{cntrl}, C^{fcast} and $C^{perfect}$ are the total penalty costs associated with the control rules, forecast-informed operation, and perfect forecast operation respectively. A performance gain of 1 is generally unattainable as it signifies that the forecast is perfect. A performance gain of 0 indicates equal performance with control rules. Negative performance gain suggests that the forecast-based scheme is more costly than control rules (as shown in Figure 5, $C^{perfect}$ is always less than C^{cntrl}, meaning the denominator in Equation 6 is always positive). The use of the upper bound in this performance score ensures that the variance in performance will be a function solely of the critical drought timing.

4.2 Results for experiment 2

The left hand panels in Figure 6 (a, c, e, g) specify times at which operating decisions become critical (herein termed "critical decision periods"). These periods are defined as moments when supply is cut back when operating with perfect forecasts (i.e., moments when the operator should be adjusting the release). A general pattern that emerges when a reservoir's storage capacity and demand are simultaneously increased is that reservoirs with larger demand (and storage) recover less readily, leading to a concentration of the failure on a single drought period. In contrast, smaller reservoirs with relatively low demand often fail but then recover quickly, so the failure periods tend be short, occurring multiple times over the simulation period. Indeed we see here that for smaller reservoirs the operations tend to be sensitive to short dry spells, so hedging decisions are required on a more frequent basis during the simulation. The exception is Eppalock, for which the

critical period of the reservoir is relatively insensitive to changes in the design demand (Table 2). For the reservoirs located in south-eastern Australia (Burrinjuck, Eppalock and Upper Yarra), critical decision periods tend to coincide with the severe Millennium Drought (~2001-2009; van Dijk et al. 2011) occurring towards the end of the simulation period. Critical decision periods for the Serpentine also occur to the end of the record, reflecting the long-term trend of declining inflows since 1975 (Petrone et al. 2008).

The right-hand panels (Figure 6 b, d, f, h) show how operating performance varies with draft ratio. The FoGSS-informed operating model offers performance improvements (i.e., performance gain > 0) in more than four fifths of reservoirs tested. Performance gains are achieved for all reservoirs specified for Eppalock and Upper Yarra, and six of the eight reservoirs specified for Burrinjuck. Performance for Serpentine is relatively poor, with only three of seven reservoirs improving under forecast-informed operation (the 90% draft reservoir is omitted in this case, since the end of the simulation period prevents us from quantifying the implications of a late, sacrificial release decision on overall performance). This is partly the result of the generally low skill of FoGSS forecasts with respect to climatology forecasts in the Serpentine catchment (Figure 3), and is also due to the consistency of FoGSS performance through the validation period (discussed in the ensuing paragraphs). Generally, the forecast-informed schemes improve performance over control rules most in reservoirs that must meet high demand (draft ratio > 0.7). For these reservoirs, critical decisions tend to be concentrated in the Millennium Drought period—during which climatology is a poor predictor of inflows, and thus forecast information offers substantial benefits over control rules.

There are certain cases for which seemingly minor changes in the critical decision periods result in large differences in performance gain. To understand this behaviour, we can examine specific cases. Figure 7 gives storage and release time series (2005–2011) for the Serpentine reservoir with the 50% draft requirement (where performance gain is positive) and with the 80% draft requirement (negative performance gain). Whilst the differences between control rules and forecast-informed operations appear modest in the release time series, the practical implications of these differences can be substantial (e.g., a public supply system that runs dry for an entire month, versus one that supplies sufficient water for basic household activities). For the 50% draft reservoir (panels a, b), the storage depletes and recovers (fully) a number of times. Within the sequence shown there is a two-year period beginning mid-2007 during which storage and inflows are sufficiently healthy that no hedging is required. Performance gain is effectively determined by the differences between control rules and forecast-informed operations during just two periods: the first half of 2007 and the period from mid-2009 to December 2010. Overall, the forecast-informed operation improves performance in this reservoir because it instructs the operator to hedge significantly from mid-2009, thus avoiding total reservoir depletion and 100% release shortfall incurred by the control rules. The information provided by FoGSS for this specific time period suffices to avoid reservoir failure and thus reduce the penalty cost by enough to overcome an earlier mistake (the hedge comes too late at the end of 2006). This contrasts with the Serpentine reservoir with 80% draft requirement, for which the forecast causes reduced performance relative to control rules (Figure 7c–d). The storage dynamics brought about by the larger storage capacity and draft ratio mean that the 80% reservoir is heavily depleted during the entirely of chosen sequence. This means that more points along the sequence become important for decision making (refer back to Figure 6g). As for the 50% draft reservoir, we observe an intelligent decision from mid-2009, and the same misstep at the end of 2006. But the 80% draft reservoir never fully recovers after 2006, so all release decisions during this period become locked into memory and contribute to future performance. There appears to be a period in late 2008 during which the forecast performance dips and the operator is instructed to meet the full target release, resulting in costly reservoir failure a few months later. Moreover, the year 2005 also becomes important for this reservoir, and it appears that FoGSS underestimates future flow since an unnecessary and costly hedge is implemented. This simple example demonstrates that a simple shift of emphasis onto some different periods can make the difference between a forecast that outperforms control rules in operation and one that does not. This example is consistent with the high sensitivity of the

Serpentine Reservoir to injected forecast error in supply-targeted operations, demonstrated with the synthetic forecasts in Section 3.2 (also true for Burrinjuck Reservoir).

We have shown earlier that FoGSS forecasts are skilful, on average, for the 1982-2010 period (Figure 3). Yet this masks the degree to which skill varies over shorter periods. As we have seen, the supply objective can result in a situation where the forecast is mobilised in only a few, crucial periods, meaning that forecast skill may need to be consistently available to warrant its use in supply-targeted operations. To demonstrate the consistency of FoGSS forecast skill, we calculate CRPS skill (equation 1) of lead-0 forecasts for a block of 12 consecutive months, randomly selected from the 1982-2010 validation period. This calculation is repeated by bootstrapping with 5000 repeats. We repeat this procedure for blocks of 2, 3, 4, 5 and 6 years. Figure 8 shows the ranges of skill from the bootstraps as box and whisker plots. The probability that any given 1-year period will have positively skilful forecasts is not statistically significant ($p>0.05$) for all reservoirs. As the blocks get larger, the probability of finding instances of negative skill reduces. For 3-year blocks, forecasts are significantly skilful ($p<0.05$) for both the Eppalock and Upper Yarra reservoirs. However, for Serpentine and Burrinjuck reservoirs, forecasts are not significantly skilful until we test skill for 5-year blocks. That is, FoGSS forecasts are less consistently skilful for the Serpentine and Burrinjuck reservoirs than for the Eppalock and Upper Yarra reservoirs. Less consistent forecast skill helps explain why the forecast-informed scheme does not always outperform control rules in the Serpentine and Burrinjuck reservoirs. An important practical implication of measuring the consistency of skill in this way is that it does not require knowledge of future conditions. This measure can be used to predict the ability of future forecasts to help meet supply objectives.

5 Discussion and conclusions

Our findings have general relevance for an increasingly water constrained world, where the demand for water and variability of climate are, in many regions, intensifying simultaneously. Intelligent use of skilful forecasts has the potential to reduce the instances of supply failure, and to extend the life of existing infrastructure at very little cost; forecast systems are very cheap compared to developing new supply infrastructure. But this potential can only be realised if the limitations of forecasts are acknowledged and their utility to specific systems and operating objectives is understood.

Our analysis shows that the benefit to reservoir operators offered by forecasts varies considerably with the objective of the reservoir. For operations that target a constant storage level, there is a clear relationship between forecast accuracy and benefit: as forecasts become more accurate, operational performance improves. This relationship is much less clear in supply-targeted reservoirs, where synthetic experiments showed that even reasonably accurate forecasts may offer little improvement over conventional control rules. This arises because reservoirs operated to the supply objective can buffer variability in inflows to a greater extent than reservoirs operated to the level objective. We conclude more generally that seasonal forecasts are more likely to raise performance in instances when reservoirs are less able to buffer variability in inflows or demand. This has important implications for older reservoirs. In our experiments, we have fabricated our reservoirs to specific draft ratios and reliabilities with recent inflows records. In practice, reservoirs have long service lives (typically decades), leaving them vulnerable to possible changes to the inflow regime beyond their construction (e.g., Bennett et al. 2012). In severe cases, an older reservoir may no longer be able to buffer inflows as effectively as when it was constructed, even if demand is static. Our findings imply that skilful seasonal streamflow forecasting systems may be able to compensate for some of the losses in performance in such instances.

While the value of forecasts was strongest for the level objective, we have shown that forecasts can also offer value to reservoirs operated to a supply objective. The real-world example of the FoGSS forecast system showed that skilful forecasts improve supply-targeted operations in the majority of reservoirs used in this study. Meeting the supply objective essentially

requires effective action in only a few crucial instances. Accordingly, we contend that if forecast skill is consistently available, forecasts will improve the operator's ability to manage a system to meet a supply objective. We therefore recommend measuring the consistency of forecast skill as a useful predictor of the value of forecasts to supply objectives.

It appears that the operator of a supply-targeted system will need to accept greater risk than the operator of a level-targeted system when adopting a given seasonal forecast service. This may explain the reluctance of operators of large urban water supply systems to adopt seasonal forecasts—an inaccurate forecast at the critical moment may humiliate managers if the implications of missteps are felt by the public. Slow response to an oncoming drought resulting from overestimation of water availability could result in grave consequences in an urban system. For example, the severe rota cuts imposed on millions of people in São Paulo have been attributed to tardy management decisions at the onset of a major drought (although in this case the failed management actions were attributed to political factors rather than a weak operating scheme) (Meganck et al. 2015). On the other hand, an underestimate of water availability can lead to over-hasty and ultimately unnecessary supply restrictions that may weaken the operator's ability to act decisively the next time a drought emerges. Whilst a skilful forecast service would actually improve these decisions on average over a very long period of time (given enough decision points), managers of such systems may experience only a few such episodes in their entire careers. By adopting a new operating scheme they expose themselves to criticism in the event that the scheme fails to work at the time that matters most. This is particularly true for emergencies, which attract significant public attention and political interest (Porter et al. 2015). It is worth emphasising that the vast majority of dams and reservoirs are operated at least partially for sustaining a target release; the practitioner community's reluctance to adopt a forecast-informed operating scheme is understandable in this light.

Our results also carry implications for future study into the value of forecasts in reservoir operations. The high variability of the performance of supply-targeted systems presents potential pitfalls for case studies assessing the value of forecasts. The unstable relationship between forecast accuracy and operating performance means that even good forecasts may result in poor operational performance. Or perhaps worse, mediocre forecasts may show strong performance for supply objectives, giving potential users false confidence in the forecast-informed operating scheme. When assessing the value of forecasts in any system with a supply target, then, we offer three recommendations:

1. That sensitivity of a given system to forecast performance be assessed, with appropriate operating objectives, perhaps with synthetic forecasts as in our study;
2. That long records and a large number of reforecasts are used to assess performance, and if these are not available that the conclusions of the study be moderated accordingly;
3. That the consistency of forecast skill be established, over the longest period possible, under stringent cross-validation.

The onus is on the analyst to determine whether the forecast service is sufficiently and consistently skilful to satisfy the operator's averseness to adopting a management system that might cause more harm than good during his or her short career.

6 Summary

The increasing improvement and availability of seasonal streamflow forecasts opens new opportunities for the adoption of adaptive operating schemes to inform water resources management. Consequently, research is needed to determine the value of forecasts for a range of design and operating settings. This can be done by measuring improvement in system performance as defined by the operating objectives. We use a rolling-horizon, adaptive control approach to demonstrate that the relation between forecast performance and operational value varies significantly when comparing level-targeted and supply-targeted operations. We demonstrate a clear and strong relation between forecast skill and value for reservoirs operated to meet target levels (*level objective*)—operational value increases as the accuracy of the forecast improves. In contrast, good forecast

accuracy across the simulation period does not necessarily translate into performance improvement for reservoirs operated to meet supply targets (*supply objective*). This is because reservoirs are able to better buffer variability in inflows when operated to meet the supply objective. We demonstrate with an experimental forecast system, FoGSS, that forecasts add value to 25 of the 32 reservoirs tested, when they are operated to meet the supply objective. For reservoirs operated to a supply objective, the driver of operating performance is the forecast accuracy during a small number of periods where critical decisions are made. We conclude that for forecasts to complement operations without imposing downside risks, forecast skill has to be consistently available.

APPENDIX 1 – Definitions of reservoir parameters and analysis techniques

All reservoir analyses executed in this study comply with standard, common techniques outlined in mainstream literature (e.g., Loucks et al. 2005, McMahon and Adeloye 2005).

Time based reliability:

For a monthly time series, the time-based reliability considers the proportion of months during the simulation period that the target demand is met in full, namely

$$Reliability = \frac{N_s}{Total\ number\ of\ months} \qquad \text{Equation 7}$$

$$0 \leq Reliability \leq 1$$

where N_s is the number of months that the target demand is met in full. Whilst the time-based reliability chosen in this study is 0.95, this does not necessarily mean that reservoir will fail as frequently as once every twenty months. This is because a fail period typically lasts more than a single month. For this reason the time-based reliability is often close to the annual reliability (years in which failure does not occur over total number of years simulated).

Standard Operating Policy (SOP)

Standard operating policy (SOP) is a default mode of operation in water supply reservoirs. SOP assumes that the operator releases to meet demand in full if there is sufficient water in storage and inflow. If available water (i.e., stored water plus inflow) is insufficient to meet demand then all available water will be released.

Draft ratio:

The ratio of demand, or target release, to the mean inflow over the period of record.

Storage-yield-reliability analysis

Storage-yield-reliability analysis refers to the procedure used to determine the storage capacity required to meet a demand (or yield) at a specified time-based reliability. This is done using an iterative simulation procedure. First, the demand and a trial storage capacity are implemented in the reservoir model. The reservoir is then simulated assuming standard operating policy. The resulting release time series is analysed to determine the time-based reliability of the trial reservoir. The storage capacity is iterated (bi-section method) according to whether the target is missed or exceeded. After a number of iterations, an optimal storage capacity is attained.

Critical period

The critical period is defined as the number of months taken for the reservoir to deplete from full to empty (also known as *critical drawdown period*), assuming standard operating policy. The critical period is a function of the demand, storage capacity, and inflow rate during drought. Some reservoirs experience more than one critical period during a simulation. In such cases we take the average of all critical periods.

Drift

Drift (*m*)—also known as *standardised net inflow*—indicates the resilience of a reservoir as well as its tendency for within-year behaviour (i.e., tendency to spill at least once each year).

$$m = \frac{1 - DR}{Cv} \qquad \text{Equation 8}$$

where *DR* is the draft ratio of the reservoir (demand over mean inflow) and *Cv* is the coefficient of variation of the annualized inflow time series, defined as the ratio of the standard deviation to the mean of the annualized inflow.

APPENDIX 2 – Reservoir optimization model details

Control rules (the benchmark scheme) and the rolling horizon, adaptive control (forecast informed scheme) are trained and simulated using the R package *reservoir* (Turner and Galelli 2016b). To develop control rules, the following objective is minimised using a backwards recursive procedure:

$$f_t(S_t) = \min_{R_t} \underset{Q_t}{E}\{C_t(S_t, Q_t, R_t) + f_{t+1}(S_{t+1})\} \qquad \forall S_t, t \in \{1, \dots, T\} \qquad \text{Equation 9}$$

where *f* is the optimal cost-to-go function (which gives the cost of the optimal decision at time step *t*+1), *C* is the penalty cost based on deviation from target operation, S is the volume of water in storage, *R* is the release from storage and *Q* is the inflow. Storage is discretized into 500 uniform values, meaning the resulting look-up table comprises a 500 × 12 (months) matrix of releases. Release is discretized into 40 uniform values between 0 and R_{max}, where R_{max} is twice the demand. Inflow is discretized according to the bounding quantiles of 1.00, 0.95, 0.7125, 0.4750, 0.2375, and 0.00 (as adopted by Stedinger et al. 1984) and the likelihood of each flow class is computed for each month using observed inflow data.

For the rolling horizon, adaptive control (or Model Predictive Control) model, the penalty cost given in Equation 3 is minimised at each time step using deterministic dynamic programming.

Table 1 – Reservoir inflow data; μ and Cv are the mean and coefficient of variation of the annual flow totals respectively.

Inflow site	Regime	μ (Mm^3)	Cv	Area (km^2)	Record	Lat.	Long.	State
Burrinjuck	Perennial	1252.1	0.90	1631	1900 – 2014	-35.00	148.58	NSW
Lake Eppalock	Ephemeral	166.8	0.82	1749	1900 – 2014	-36.88	144.56	VIC
Serpentine	Intermittent	58.4	0.69	664	1912 – 2014	-32.40	116.10	WA
Upper Yarra	Perennial	153.3	0.43	337	1913 – 2014	-37.68	145.92	VIC

Table 2 – Reservoir design specifications and characteristics for 0.95 reliability reservoirs. Drift indicates the reservoir time to recovery from full as well as tendency for within year behaviour. Storage ratio represents the time (mean) to fill the reservoir assuming no outflows (i.e., capacity to mean annual inflow ratio). Critical period is the time period taken to empty the reservoir assuming recorded drought conditions. (Full definitions in Appendix 2.)

	Draft ratio	Design Demand [Mm^3/month]	Drift [-]	Design storage [Mm^3]	Storage ratio [years]	Crit. Period [months]
BURRINJUCK	0.2	18.2	1.14	57	0.05	8
	0.3	27.2	1.00	144	0.13	11
	0.4	36.3	0.85	404	0.37	32
	0.5	45.4	0.71	830	0.76	84
	0.6	54.5	0.57	1685	1.55	104
	0.7	63.5	0.43	2570	2.36	104
	0.8	72.6	0.28	3539	3.25	128
	0.9	81.7	0.14	4699	4.31	152
EPPALCOK	0.2	2.4	0.88	58	0.40	102
	0.3	3.6	0.77	175	1.22	102
	0.4	4.8	0.66	289	2.01	102
	0.5	6.0	0.55	409	2.84	102
	0.6	7.2	0.44	535	3.72	146
	0.7	8.4	0.33	710	4.94	146
	0.8	9.6	0.22	885	6.15	147
	0.9	10.8	0.11	1061	7.37	147
SERPENTINE	0.2	0.51	1.50	2	0.07	6
	0.3	0.76	1.32	4	0.13	11
	0.4	1.0	1.13	7	0.24	15
	0.5	1.3	0.94	11	0.37	15
	0.6	1.5	0.75	15	0.48	15
	0.7	1.8	0.56	27	0.89	93
	0.8	2.0	0.38	53	1.75	100
	0.9	2.3	0.19	88	2.90	112
UPPER YARRA	0.2	2.1	1.91	2	0.02	3
	0.3	3.2	1.67	7	0.06	6
	0.4	4.2	1.43	14	0.11	9
	0.5	5.3	1.19	26	0.20	13
	0.6	6.4	0.96	39	0.31	15
	0.7	7.4	0.72	64	0.50	24
	0.8	8.5	0.48	139	1.09	142
	0.9	9.5	0.24	323	2.54	147

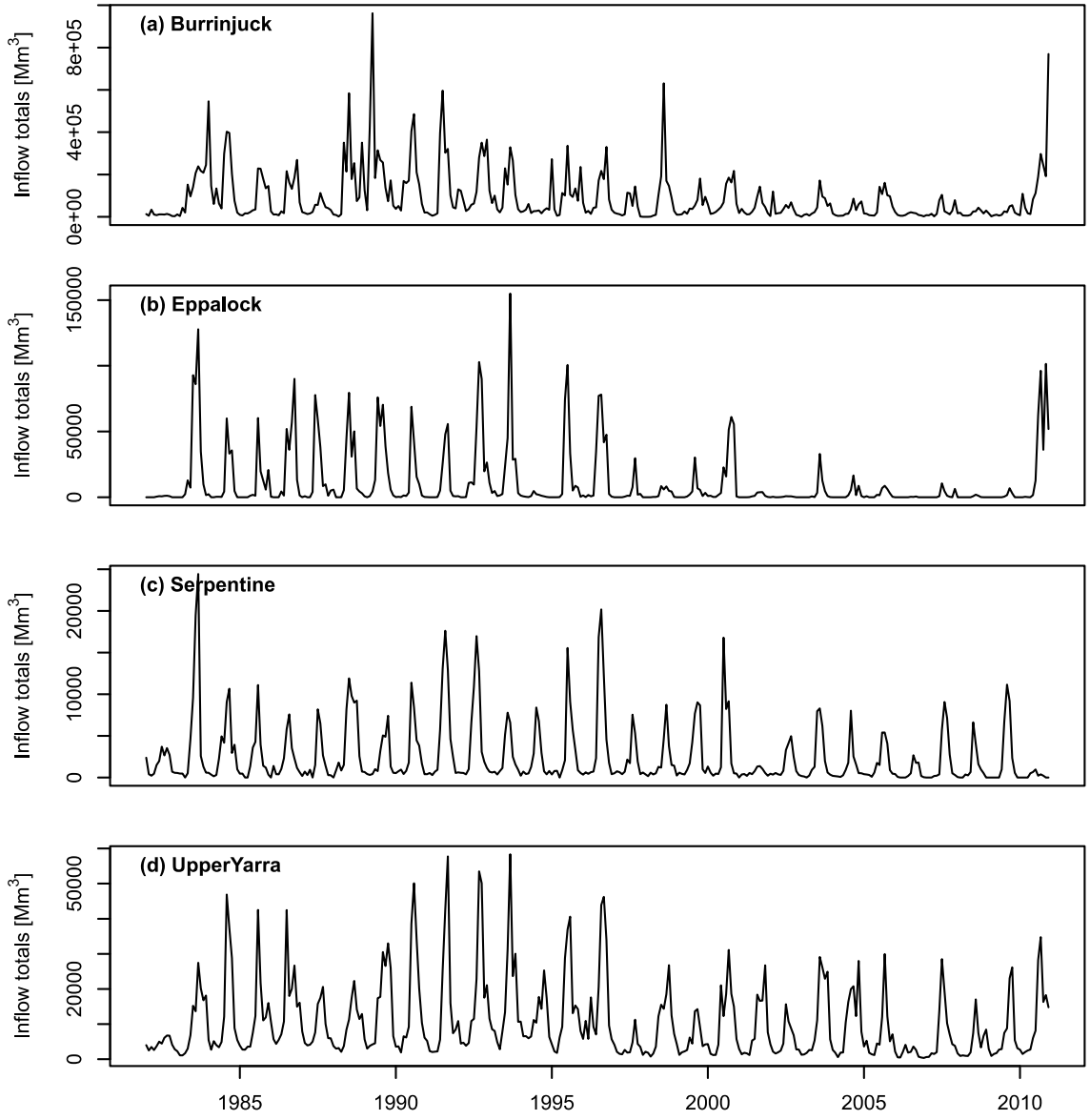

Figure 1 - Reservoir inflow records for (a) Burrinjuck Dam, (b) Lake Eppalock, (c) Serpentine Reservoir and (d) Upper Yarra Reservoir during the 29-year study period Jan 1982 – Dec 2010.

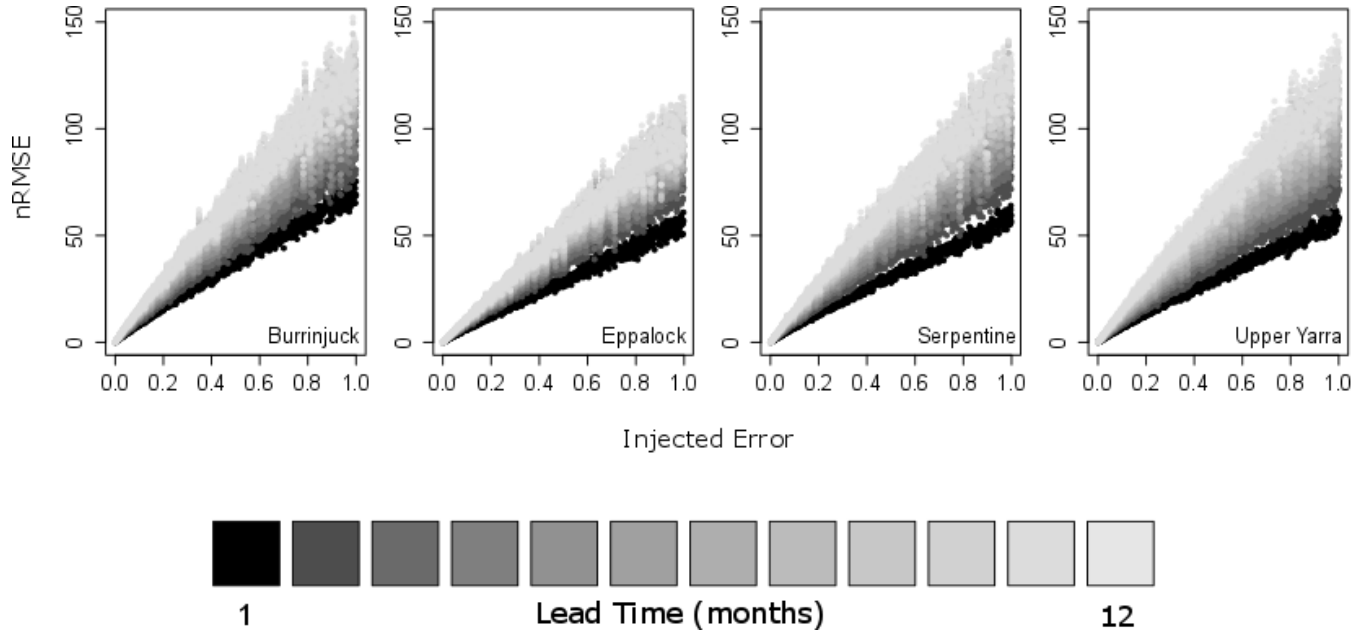

Figure 2 – Normalised Root Mean Squared Error (nRMSE) for varying error injected into synthetic forecasts generated using the Martingale Model of Forecast Evolution (1000 forecasts, monthly resolution, 12 months ahead, giving 12,000 points on each pane).

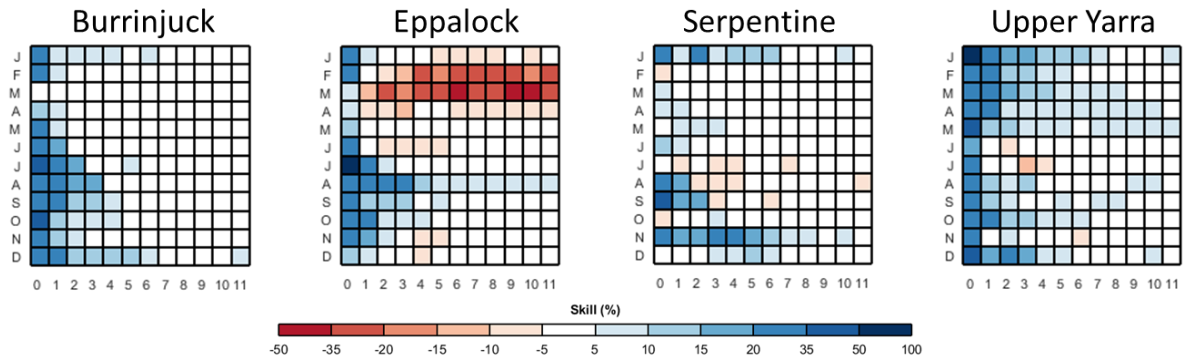

Figure 3 – FoGSS forecast skill measured by the continuous ranked probability score (CRPSS) with respect to climatology forecasts. Rows show target months, columns show lead-time in months.

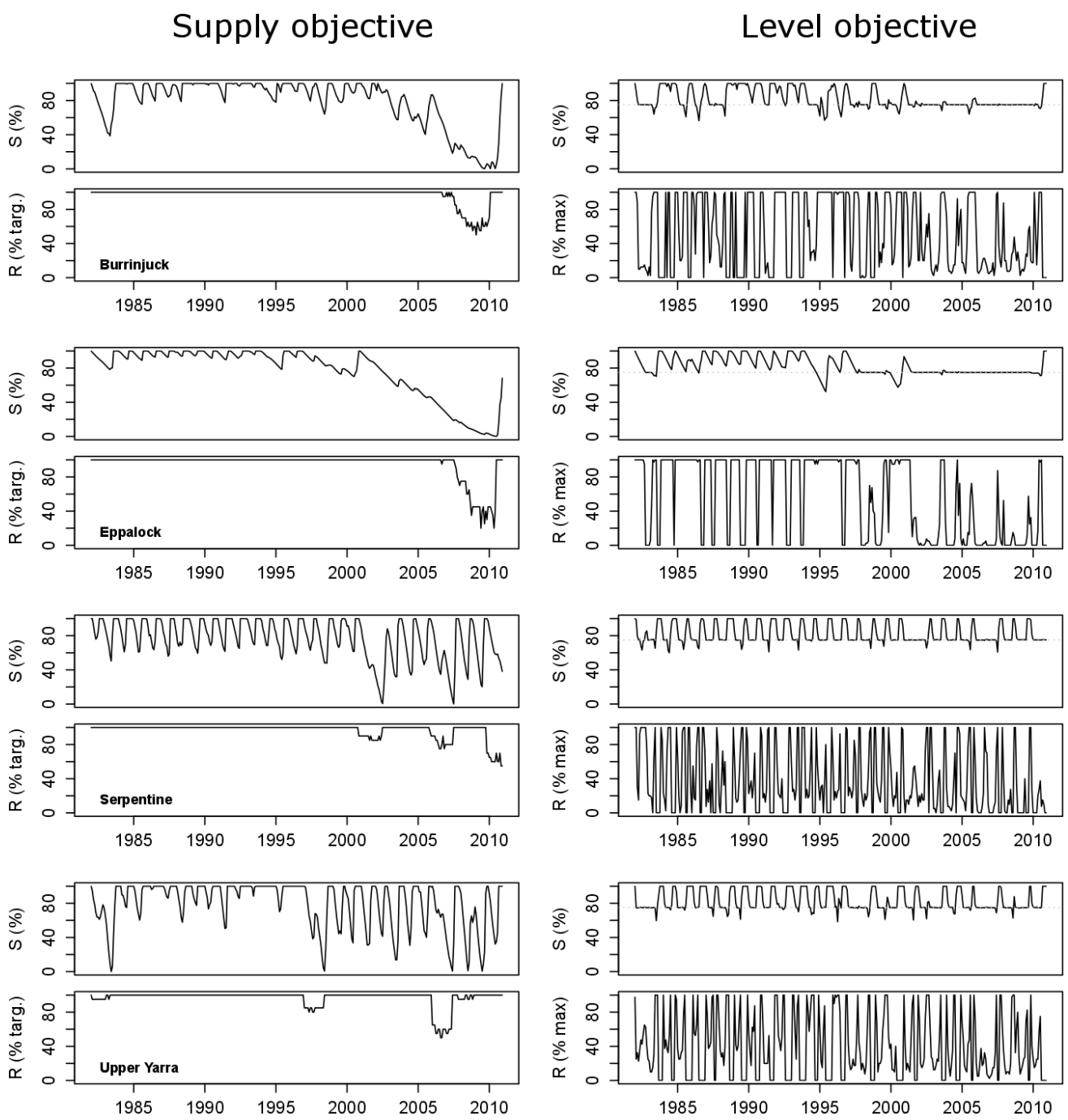

Figure 4 – Behaviour of reservoirs operated to meet the supply objective (left column) and level objective (right column). Simulations use the rolling horizon model with a perfect 12-month (observed) inflow forecast, applied to 95% reliability reservoirs with draft ratio of 0.5. *S* is the storage (as % of capacity) and *R* is the release (given as % of target for emergency response reservoirs and % of maximum possible release for the continually adjusted setting).

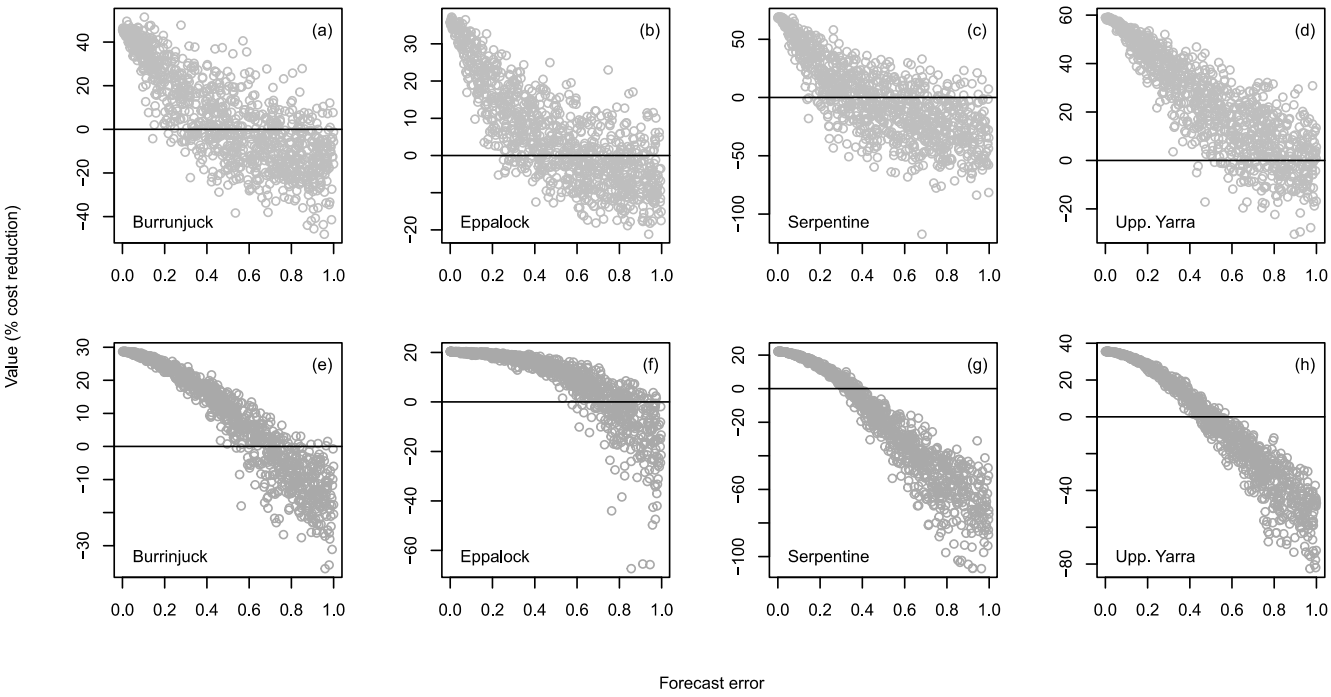

Figure 5 – Value of the forecast-informed scheme over control rules as a function of forecast error for supply objective (a – d) and level objective (e – h) operational settings.

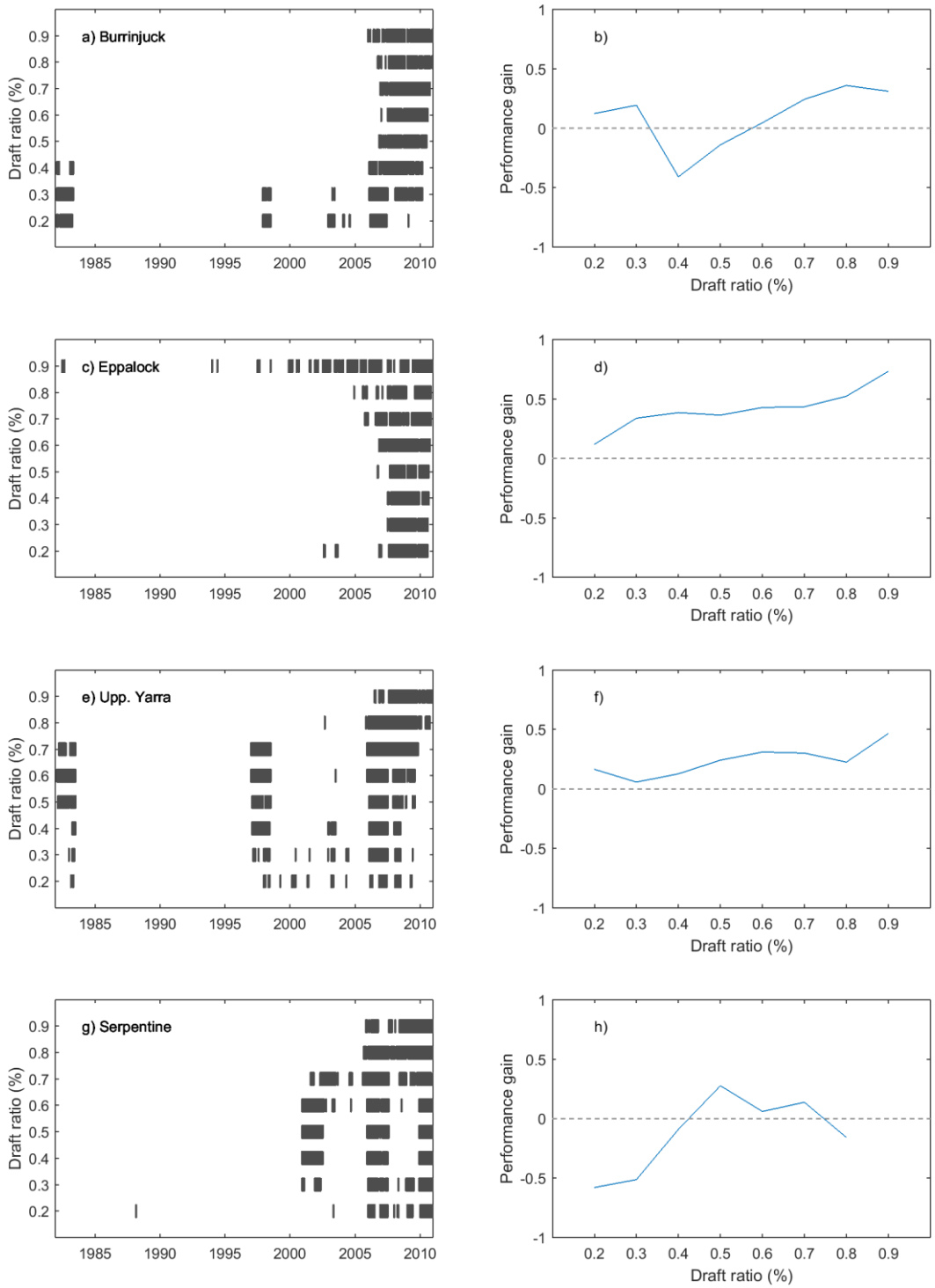

Figure 6 – Panels a, c, e, g give critical decision periods for each reservoir design (draft ratio 0.2, 0.3, …, 0.9). Panels b, d, f, h give performance gain plotted against draft ratio. Critical decision periods are moments during which perfect forecast operations implement supply cutbacks.

5

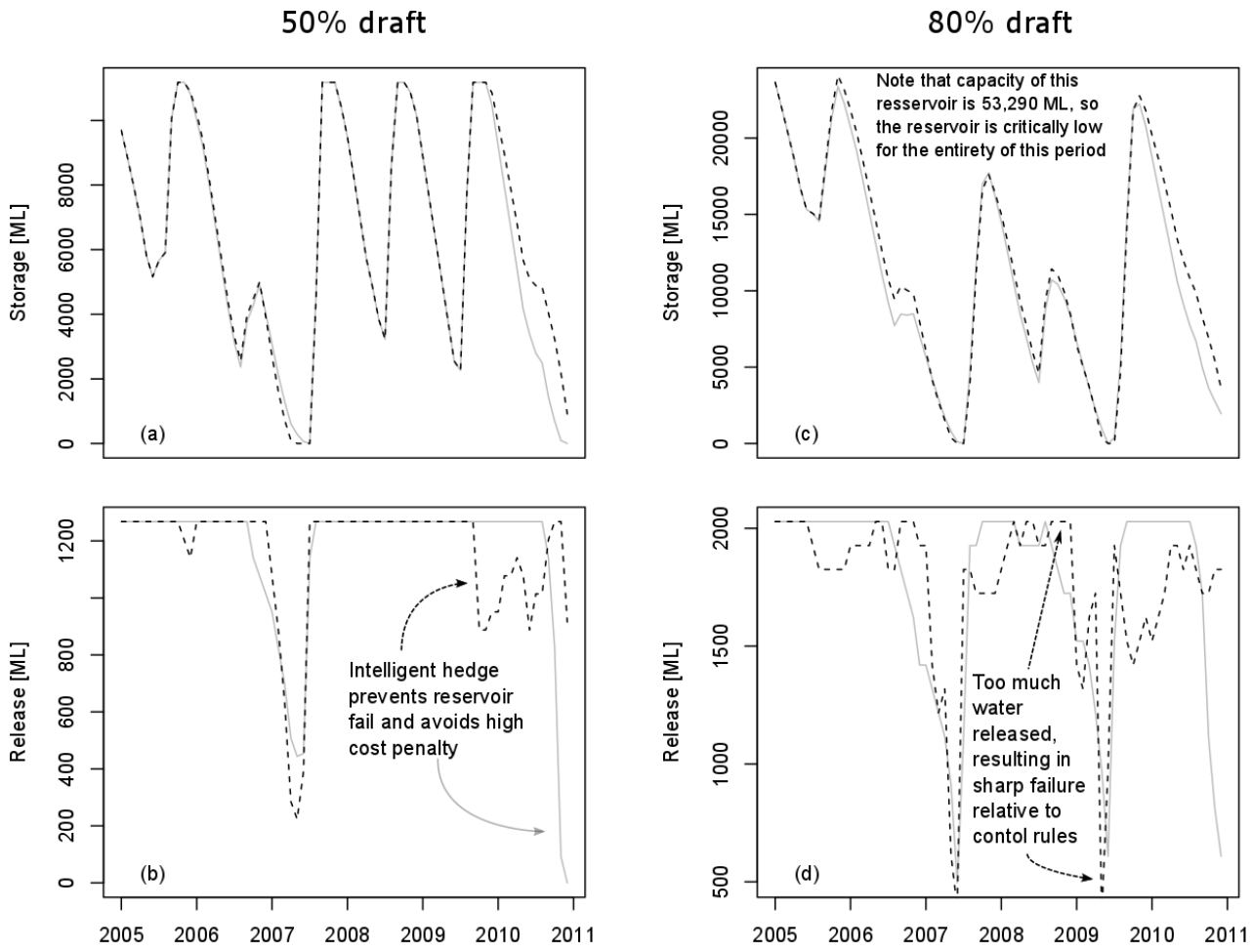

Figure 7 – Storage and release time series for reservoirs in the Serpentine catchment with 50% (a, b) and 80% (c, d) draft ratios. The solid grey line gives operation under control rules whilst the dotted black line gives operation with the FoGSS forecast (median of ensemble).

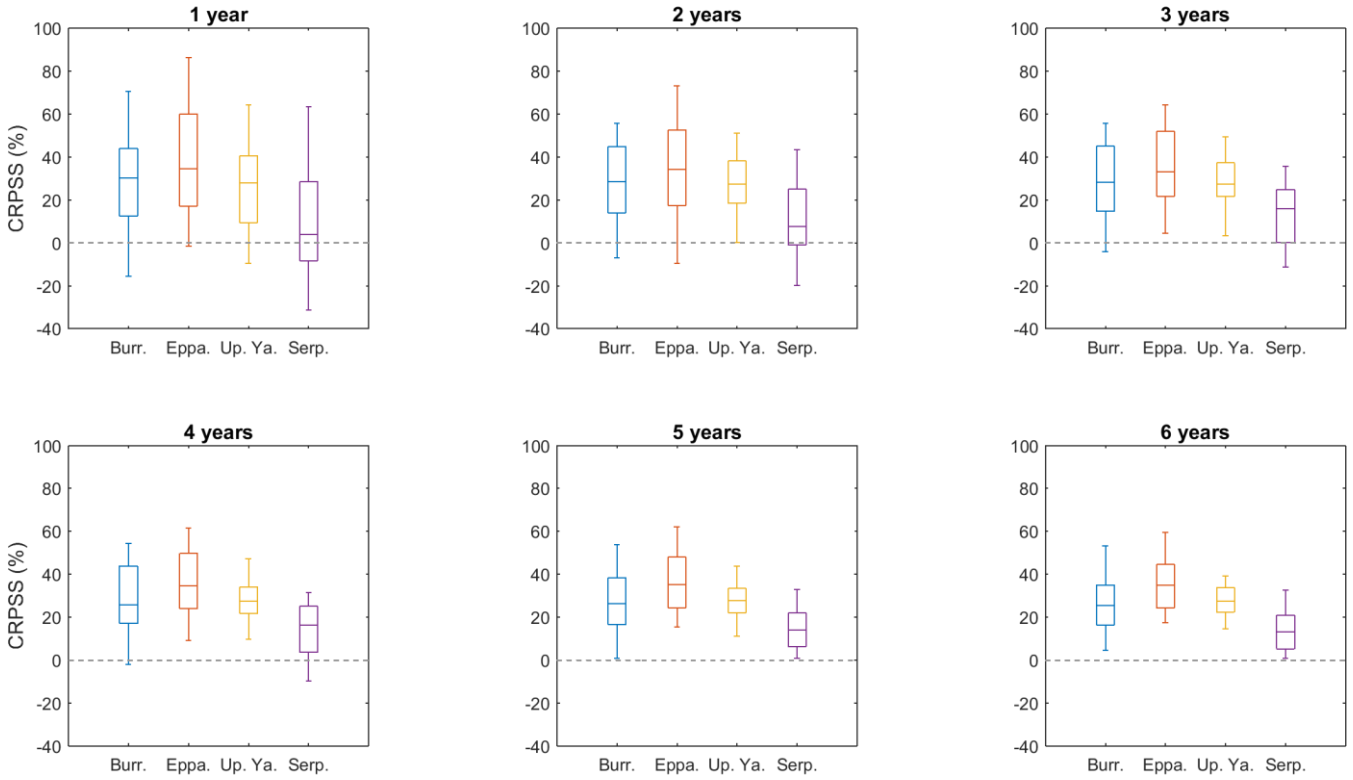

Figure 8 – Variation in skill of lead-0 FoGSS forecasts for blocks of consecutive months. Skill for consecutive months for blocks of 1-6 years is bootstrapped to create the box and whisker plots. Boxes give interquartile range, whiskers give 90% intervals, lines show median values.

References

Alemu, E.T., R.N. Palmer, A. Polebitski, and B. Meaker (2010), Decision support system for optimizing reservoir operations using ensemble streamflow predictions, *Journal of Water Resources Planning and Management*, 137(1), 72-82.

Anghileri, D., N. Voisin, A. Castelletti, F. Pianosi, B. Nijssen, and D.P. Lettenmaier (2016), Value of long- term streamflow forecasts to reservoir operations for water supply in snow- dominated river catchments, *Water Resources Research*, 52(6), 4209-4225.

Bellman, R. (1956), Dynamic programming and Lagrange multipliers, *Proceedings of the National Academy of Sciences of the United States of America*, 42(10), 767-769.

Bennett, J. C., Q. J. Wang, D. E. Robertson, and A. Schepen, M. Li, K. Michael (2017) Assessment of an ensemble seasonal streamflow forecasting system for Australia, Hydrology and Earth System Sciences Discussions, 1-36.

Bennett, J. C., Q. J. Wang, M. Li, D. E. Robertson, and A. Schepen (2016), Reliable long-range ensemble streamflow forecasts by combining dynamical climate forecasts, a conceptual runoff model and a staged error model, *Water Resources Research*, 52(10), 8238-8259.

Bennett J. C., F. L. N. Ling, D. A. Post, M. R. Grose, S. P. Corney, B. Graham B, G. K. Holz, J. J. Katzfey, N. L. Bindoff (2012), High-resolution projections of surface water availability for Tasmania, Australia, *Hydrology and Earth System Sciences*, 16, 1287-1303. DOI: 10.5194/hess-16-1287-2012.

Bertsekas, D. (1976), Dynamic programming and stochastic control, Academic Press, New York.

Block, P. (2011) Tailoring seasonal climate forecasts for hydropower operations, *Hydrology and Earth System Sciences*, 15, 1355-1368.

Boucher, M.A., Tremblay, D., Delorme, L., Perreault, L. and Anctil, F., 2012. Hydro-economic assessment of hydrological forecasting systems. *Journal of Hydrology*, 416, pp.133-144.

Brown, C. (2010), The end of reliability, *Journal of Water Resources Planning and Management*, 136(2), 143-145.

Brown, C.M., J.R. Lund, X. Cai, P.M. Reed, E.A. Zagona, A. Ostfeld, J. Hall, G.W. Characklis, W. Yu, and L. Brekke (2015), The future of water resources systems analysis: Toward a scientific framework for sustainable water management, *Water Resources Research*, 51(8), 6110-6124.

Castelletti, A., S. Galelli, M. Restelli, and R. Soncini-Sessa (2010), Tree-based reinforcement learning for optimal water reservoir operation, *Water Resources Research*, 46(9).

Côté, P., and L. Robert (2016), Comparison of Stochastic Optimization Algorithms for Hydropower Reservoir Operation with Ensemble Streamflow Prediction, *Journal of Water Resources Planning and Management*, 142(2), 04015046.

Draper, A.J., and J.R. Lund (2004), Optimal hedging and carryover storage value, *Journal of Water Resources Planning and Management*, 130(1), 83-87.

Faber, B.A., and J.R. Stedinger (2001), Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts, *Journal of Hydrology*, 249(1), 113-133.

Georgakakos, A.P., H. Yao, M. Kistenmacher, K.P. Georgakakos, N.E. Graham, F.Y. Cheng, C. Spencer, and E. Shamir (2012), Value of adaptive water resources management in Northern California under climatic variability and change: Reservoir management, *Journal of Hydrology*, 412, 34-46.

Gneiting T, Raftery AE. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102** 359-378.

Goddard, L., Aitchellouche, Y., Baethgen, W., Dettinger, M., Graham, R., Hayman, P., Kadi, M., Martínez, R. and Meinke, H., 2010. Providing seasonal-to-interannual climate information for risk management and decision-making. *Procedia Environmental Sciences*, 1, pp.81-101.

Gong, G., L. Wang, L. Condon, A. Shearman, and U. Lall (2010), A simple framework for incorporating seasonal streamflow forecasts into existing water resource management practices, *Journal of the American Water Resources Association*, 46(3), 574-585.

Graham, N.E., and K.P. Georgakakos (2010), Toward understanding the value of climate information for multiobjective reservoir management under present and future climate and demand scenarios, *Journal of Applied Meteorology and Climatology*, 49(4), 557-573.

Hamlet, A.F., D. Huppert, and D.P. Lettenmaier (2002), Economic value of long-lead streamflow forecasts for Columbia River hydropower, *Journal of Water Resources Planning and Management*, 128(2), 91-101.

Heath, D.C. and Jackson, P.L., (1994), Modeling the evolution of demand forecasts ITH application to safety stock analysis in production/distribution systems, *IIE Transactions*, 26(3), 17-30.

Housh, M., A. Ostfeld, U. Shamir (2013), Limited multi-stage stochastic programming for managing water supply systems, *Environmental Modelling & Software*, 41, 53-64.

Hudson, D., A. G. Marshall, Y. Yin, O. Alves, and H. H. Hendon (2013), Improving intraseasonal prediction with a new ensemble generation strategy, *Monthly Weather Reviews*, 141(12), 4429–4449.

Kim, Y.O., and R.N. Palmer (1997), Value of seasonal flow forecasts in Bayesian stochastic programming, *Journal of Water Resources Planning and Management*, 123(6), 327-335.

Li, M., Q. J. Wang, and J. Bennett (2013), Accounting for seasonal dependence in hydrological model errors and prediction uncertainty, *Water Resources Research*, 49, 5913–5929.

Li, W., A. Sankarasubramanian, R.S. Ranjithan, and E.D. Brill (2014), Improved regional water management utilizing climate forecasts: An interbasin transfer model with a risk management framework, *Water Resources Research*, 50(8), 6810-6827.

Li, M., Q. J. Wang, J. C. Bennett, and D. E. Robertson (2015), A strategy to overcome adverse effects of autoregressive updating of streamflow forecasts, *Hydrology and Earth System Sciences*, 19(1), 1-15.

Li, M., Q. J. Wang, J. C. Bennett, and D. E. Robertson (2016), Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting, *Hydrology and Earth System Sciences*, 20, 3561-3579.

Loucks, D.P., E. Van Beek, J.R. Stedinger, J.P.M. Dijkman, and M.T. Villars (2005), Water resources systems planning and management: an introduction to methods, models and applications, Paris: UNESCO.

Marshall, A. G., D. Hudson, M. C. Wheeler, O. Alves, H. H. Hendon, M. J. Pook, and J. S. Risbey (2014), Intra-seasonal drivers of extreme heat over Australia in observations and POAMA-2, *Climate Dynamics*, 43(7), 1915–1937.

Mayne, D., R. Rawlings, C. Rao, and P. Scokaert (2000), Constrained model predictive control: stability and optimality, *Automatica*, 36(6), 789-814.

McMahon, T.A. and A.J. Adeloye (2005), Water resources yield, Water Resources Publications, LLC, Colorado.

Meganck, R., K. Havens, and R. M. Pinto-Coelho (2015), Water: Megacities running dry in Brazil, *Nature*, 521(7552), 289-289.

Olsson, J., C. B. Uvo, K. Foster, and W. Yang (2016), Technical Note: Initial assessment of a multi-method approach to spring-flood forecasting in Sweden, *Hydrology and Earth System Sciences*, 20(2), 659-667.

Pagano, T., A. Wood, K. Werner, and R. Tama-Sweet (2014), Western U.S. Water Supply Forecasting: A Tradition Evolves, *Eos, Transactions American Geophysical Union*, 95(3), 28-29, doi: 10.1002/2014eo030007.

Peng, Z., Q. J. Wang, J. C. Bennett, A. Schepen, F. Pappenberger, P. Pokhrel, and Z. Wang (2014), Statistical Calibration and Bridging of ECMWF System4 Outputs for Forecasting Seasonal Precipitation over China, *Journal of Geophysical Research (Atmospheres)*, 119, 7116–7135.

Petrone KC, Hughes JD, Van Niel TG, and Silberstein RP. (2010), Streamflow decline in southwestern Australia, 1950–2008. *Geophysical Research Letters*, 37(11), doi: 10.1029/2010gl043102.

Porter, M.G., D. Downie, H. Scarborough, O. Sahin, and R.A. Stewart (2015), Drought and Desalination: Melbourne water supply and development choices in the twenty-first century, *Desalination and Water Treatment*, 55(9), 2278-2295.

Raso, L., Giesen, N., Stive, P., Schwanenberg, D., and Overloop, P.J. (2013). Tree structure generation from ensemble forecasts for real time control, *Hydrological Processes*, 27(1), 75-82.

Raso, L., D. Schwanenberg, N.C. van de Giesen, and P.J. van Overloop (2014), Short-term optimal operation of water systems using ensemble forecasts, *Advances in Water Resources*, 71, 200-208.

Rayner, S., D. Lach, and H. Ingram (2005), Weather forecasts are for wimps: why water resource managers do not use climate forecasts, *Climate Change*, 69, 197-227.

Schepen, A., Q. J. Wang, and D. E. Robertson (2014), Seasonal Forecasts of Australian Rainfall through Calibration and Bridging of Coupled GCM Outputs, *Monthly Weather Review*, 142(5), 1758-1770, doi: 10.1175/mwr-d-13-00248.1.

Schepen, A., and Q. Wang (2014), Ensemble forecasts of monthly catchment rainfall out to long lead times by post-processing coupled general circulation model output, *Journal of Hydrology*, 519, 2920–2931.

Shapiro, A., D. Dentcheva, and A. Ruszczynski (2014), Lectures on Stochastic Programming: Modelling and Theory, Vol. 16, SIAM.

Stedinger, J.R., B.F. Sule, and D.P. Loucks (1984), Stochastic dynamic programming models for reservoir operation optimization, *Water resources research*, 20(11), 1499-1505.

Turner, S.W.D., and S. Galelli (2016a), Regime-shifting streamflow processes: Implications for water supply reservoir operations, *Water Resources Research*, 52(5), 3984-4002.

Turner, S.W.D., and S. Galelli (2016b), Water supply sensitivity to climate change: An R package for implementing reservoir storage analysis in global and regional impact studies, *Environmental Modelling & Software*, 76, 13-19.

Turner, S.W.D., and S. Galelli (2017): http://github.com/swd-turner/MMFE/, last access: 18 June 2017.

van Dijk, A.I.J.M., H.E. Beck, R.S. Crosbie, R.A.M. Jeu, Y.Y. Liu, G.M. Podger, B. Timbal, and N.R. Viney (2013), The Millennium Drought in southeast Australia (2001-2009): Natural and human causes and implications for water resources, ecosystems, economy, and society, *Water Resources Research*, 49(2), 1040-1057.

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

Xu, B., P.A. Zhong, R.C. Zambon, Y. Zhao, and W.W.G. Yeh (2015), Scenario tree reduction in stochastic programming with recourse for hydropower operations, *Water Resources Research*, 51(8), 6359-6380.

Yuan, X., E. F. Wood, and Z. Ma (2015), A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, *Wiley Interdisciplinary Reviews: Water*, 2(5), 523-536.

Zhao, T., and J. Zhao (2014), Joint and respective effects of long-and short-term forecast uncertainties on reservoir operations, *Journal of Hydrology*, 517, 83-94.

Zhao, T., X. Cai, and D. Yang (2011), Effect of streamflow forecast uncertainty on real-time reservoir operation, *Advances in Water Resources*, 34(4), 495-504.