Reply to Chris Delaney (reviewer #1)

General Comments:

Thank you for the opportunity to review this manuscript. I found this article of significant interest in that the water municipality for which I am employed is currently investigating the use of ensemble flow forecasts to help inform the operations of the reservoirs under our management. I appreciated that this article considered both short range (continually adjusted) and long range (emergency response) reservoir operations. I feel the subject matter covered in this article is very relevant to current reservoir management challenges, because reservoir operations are becoming increasing constrained by increasing demand, release constraints due to habitat and environmental concerns, and changes in hydrology due to climate change. Incorporating seasonal flow forecasts into a decision support system could provide useful information for operators to help meet these challenges. I found this article very well written and most concepts very well explained with a few minor exceptions as covered below.

We thank the reviewer for the positive comments and practical suggestions for improvement.

Specific Comments:

Section 2.3.2, Page 6, Equation 3: It is unclear if Equations 4 and 5 are the cost functions to be used in the rolling horizon objective function (Equation 3).

Yes, Equation 4 and 5 define the penalty cost used in the rolling horizon objective function (Equation 3). We clarified this in the revised version of the manuscript.

Figure 8: I could not make sense of the results provided in this figure. The scenarios with higher releases have higher storage levels. If both scenarios have the same inflows then this does not make sense. Is it possible the symbology does not match for the storage and release hydrographs?

It is indeed the case that the scenario with the higher releases has higher storage levels. This occurs because the capacity of this reservoir is larger. In our study, reservoirs with higher release requirements are designed with higher storage capacity to maintain a consistent reliability (see Section 2.2.1 and Table 2).

Technical Corrections:

Appendix 1, Page 12, Line 25: Drift equation is not numbered. This should be Equation 8. Coefficient of variation is not defined. This should be defined or a reference provided.

Thanks for picking up this error. We added the equation number and the definition of coefficient of variation (i.e., ratio of the standard deviation to the mean of the annualized inflow).

Figure 6: Panels a, c, e, and g should be labeled Panels a-d. Panels b, d, f, and h should be labeled Panels e-h. It would be useful to define “critical decision periods” in the figure caption.

Figure 7: This figure should be labeled Figure 8, because it is referenced in the report after the current Figure 8.

Figure 8: This figure should be labeled Figure 7. Figure caption should include that the results presented are for the “Serpentine” reservoir.

We implemented all suggested improvements to figures.

Reply to reviewer #2

General Comments:

The paper concerns an interesting and emerging topic: quantifying the actual improvement of seasonal forecasts in water resources management and reservoir operation in particular. I think the experimental setting is valuable and provides interesting results, which are worth to be published, but I disagree about the way the results are presented, which may lead, in my opinion, to misleading interpretations. The paper distinguishes between "continually adjusted operation", when decisions are adjusted at frequent intervals, and "emergency response operation", when key decisions are taken infrequently. As an example of "continually adjusted operation", the paper considers a reservoir operated to track a constant target reservoir storage. As an example of "emergency response operation", the papers consider a reservoir operated to meet a constant supply demand. The paper concludes that there is a clear relationship between forecast skill and value in the case of "continually adjusted operation" while this is not in the case for "emergency response operation". Another conclusion is that, in this second type of operation, it is of fundamental importance considering skillful forecasts at certain, well defined moments when critical decisions are taken. In my opinion, the results should not be commented in light of the type of decision process (in both cases decision are taken with the same frequency, i.e., every month), rather considering the type of reservoir operating objective and its dynamical response to forecasts. When the reservoir is operated for water supply, the release decision does not directly depend on the predicted inflow, because the reservoir storage partially buffers the variability of the inflows. As a consequence, poor inflow forecasts may not directly affect the operating performance (of course, this depends on reservoir capacity and storage-demand ratio). Prolonged poor forecasts (on long lead times, for example) may instead negatively affect the operating performance. When the reservoir is operated to track a certain target storage, the release decision must mimic the inflow much more closely (the reservoir storage does not provide any buffer in this case). For this reason, a skillful forecast implies high operating performance and viceversa, which turns into a clearer relationship between forecast skill and value. On top of this, we should consider that the storage tracking objective depends basically on the forecast of few lead times (if not on the first lead time only), while the supply objective is more dependent on forecasts on longer lead times (which allow for the hedging decisions mentioned in the paper). This means that, in general, the tracking storage operation benefits of more skillful forecasts than the supply objective, because both FoGSS forecasts and synthetic forecasts show a decreasing skill with lead time (Fig. 2 and 3). This may explain the fact that a certain forecast skill may be linked to a wider range of forecast values when considering the water supply objective (Fig. 5). I argue whether a different definition of forecast skill should be used in the two cases (for example, skill computed on different lead times, i.e., the ones relevant to the dynamic of each reservoir objective). I'll provide more comments on these points in the following part.

Summarizing, I think that the paper deserve publication after a major revision. It might bring valuable insights on the topic of seasonal forecast value for reservoir operation, but the results should be discussed highlighting the effect of the different dynamics of the operating objectives (fast response to forecasts for storage tracking and slow response for water supply) on the forecast skill-value relationship, rather than focusing on the duality between "continually adjusted operation" and "emergency response operation", as it is in the current version of the manuscript.

We greatly appreciate this detailed and thoughtful review. In particular we value the reviewer's the suggestion to distinguish the two reservoir types by operation rather than frequency of decision process. This shift in emphasis allowed for a clearer, more coherent interpretation of the results. We made the distinction between the two reservoir objectives in these simple terms: one targets constant storage by varying the release, the other targets constant release by allowing the storage level to vary. This shift changes the role of storage from a target to a buffer, with consequent effects on forecast value that are brought out by our results. Rather than referring to the reservoirs as "continually-adjusted" and "emergency response", we refer to "level objective" and "supply objective". Discussion now places more emphasis on the role of a storage buffer in obviating the need for accurate forecasts during much of the operation.

Specific Comments:

Page 3, Lines 4-8. How is the sampling of the "error injected" parameter performed? Is there one sampling for each forecast? How does the error increase with lead time? The interpretation of the results would benefit by a more comprehensive description.

Something we failed to mention in our paper was that the parameters of synthetic forecast model that define the way error increases with lead-time are defined using the actual FoGSS forecasts available for each catchment. This ensures that even though the forecast is synthetic, its decay with lead time is realistic. We sample the "injected error" once for each generated forecast.

Page 3, Lines 25-26. "In months where forecasts are not informative, FoGSS is designed to return a climatological forecast." How is this performed? If this is the case, why does it happen that in Fig. 3 there are some negative skills (this is particularly evident for the Eppalock reservoir)?

FoGSS applies statistical methods to correct errors in both the precipitation and hydrological components of the forecast, with the aim of ensuring forecasts are at least as skilful as climatology. The details are reasonably complex, but the principles are more straightforward. Essentially, precipitation forecasts can be considered as being corrected with a 'model output statistics' type approach (a term coined by Glahn and Lowry, 1972). (The actual method used, the Bayesian Joint Probability modelling approach, is described in detail by Wang et al. 2009, Wang and Robertson 2011, and as a post-processor of GCM forecasts, by Schepen and Wang 2014.) The principle here is that the relationship between the predictor, x, *(e.g. forecast precip from a GCM for February at a lead-time of 3 months) and* *the predictand, y, (e.g. observation for that month) can be assumed to follow a regression, such that:*

$$y = dx + \mu \qquad (1)$$

where d and μ are parameters. We optimise this relationship on the basis of historical forecasts, and where forecasts are consistently poor, d can approach zero. That is,

$$y \approx \mu . \qquad (2)$$

This means that the forecast essentially returns a constant, μ (i.e., a climatology). In this way, precip forecasts should never be much worse than climatology (they can be slightly worse under cross-validation).

We pursue a similar approach for the hydrological component, with slightly different mechanics. In FoGSS, we completely separate uncertainty due to precipitation forecasts from uncertainty due to hydrological modelling. This means we do not apply a different regression to each lead-time, as we characterise hydrological model errors only from hydrological simulations that are forced by observations. That is, we apply a regression with parameters that vary by month, but these (12) sets of regression parameters are applied to all lead-times. This usually works quite well for forecasts (in concert with the precip corrections described above), but in instances where flow is often zero (as in Eppalock in late summer/early autumn), the hydrological model and the error model can tend to over-predict flows at longer lead-times. In the way that FoGSS is currently configured, the incidence of zero flows is always underestimated where observed flows are zero more than half the time, which results in small but persistent positive biases. While the actual discrepancies in flows in these cases are very small (e.g. perhaps the model predicts a mean flow of 0.6 mm, while climatological mean flow is 0.1 mm and, as already noted, observations are often zero), the skill (expressed as % of climatology flows, as in Figure 2) can appear to be very poor, because it is relative. We are currently working on methods to handle zero flows more proficiently, but this is a significant technical challenge. For the purposes of this paper, because negative skill occurs only when flows are very small, and forecast errors are extremely modest (often fractions of a mm), these issues have virtually no bearing on the utility of the forecasts to reservoir operation. We see this in our results where we demonstrate the strong value of forecasts at Eppalock.

We added to the discussion of these issues in the paper as follows:

"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. Because flows are so low, even small small positive biases result in high relative errors in February and March. However, because inflows are so low during these months, these errors have very little influence on annual (or even seasonal) water balances."

Page 4, Lines 24-26. In the standard operating policy, how is the release modulated when the demand cannot be fully met? I would include the description of this aspect also in Appendix 1.

When there is insufficient water to meet the demand the release is simply constrained to the total volume of water left in storage plus available inflows. In other words, the operator releases as much as possible. We made sure this is clear in our revision.

Page 4, Lines 26 and followings. How does the maximum release change in the different reservoir configurations? Is it linked to the reservoir capacity? Does the maximum release affect the ability of meeting the operating objectives (for example, should it be designed so to be able to always meet the demand)? The authors mention something in Appendix 2, but I would add a sentence in the main text as well.

We simply set the maximum release to the demand multiplied by two for all systems, which means that larger capacity reservoirs have greater release capability. We think this assumption suffices for our synthetic study, although we concede that the release capability of actual reservoirs may vary widely depending on design.

Table 2. I think the readability of the table would benefit from a description of the meaning of "draft ratio", "drift" etc. in the caption. If I'm not wrong, "storage ratio" is not defined in the text. What does it represent? What are the actual reservoir features? It would be nice to have the figures for better understanding the reservoir settings in' the different experiments. Is the "critical period" computed assuming no inflow to the reservoir or climatology or what else? The numbers seem in some cases very high and I don't fully understand how to interpret these large numbers (see also the following comment).

We added some definitions to the caption, as suggested. The storage ratio (years) is simply mean annual inflow divided by the storage volume. It represents the number of years it would take to fill the reservoir if there was no release—so it provides some indication of reservoir's potential recharge rate. The critical period is computed using the historical inflow record. Since the reservoirs are designed to 0.95 reliability under the historical record, they are guaranteed to fail when operated with standard operating policy. So all simulations contain a period from full to empty that is used to determine the critical period. The critical period can be very long because often the drawdown rate is moderated by inflows. In particularly large systems with low recharge rates, it's often the case that the inflows will partially recover the reservoir before drawdown commences again. This can lead to very a long critical period.

Figure 4. "Emergency response" case: why do the Burrinjuck and Eppalock reservoir empty, if their critical periods are 84 and 102 months respectively? "Continually adjusted" case: why does it happen that the reservoir storage is above the target but the releases are equal to 0? (see for example, Burrinjick and Eppalock reservoirs at the beginning of the time series) This behavior seems to be sub-optimal, because a release greater than zero would contribute in reducing the objective cost.

All reservoirs are designed with a reliability of less than 1, so emptying is guaranteed (no emptying would mean no requirement to cut back demand, and so the reliability would be 100%). In figure 4 we observe that even with an extremely long critical period (102 months in the case of Eppalock) the reservoir does in fact empty toward the end of the drought period (taking precisely 102 months).

The question regarding releases being zero despite high storage is very interesting. It most cases we observe that when the storage is above target the release will be large (because the objective is to bring the storage back down to target). However, there are some instances in which the model appears indifferent to the large storage level (such as the beginning of some of the time series, as identified by the reviewer). The reason is that in these instances the impact of zero release or full release is the exact same. The inflow in these instances is so large that even when maximum release is applied, the reservoir will continue to spill. So deviation from the objective function is the same irrespective of the release decision. Our model takes the lowest release as default in these instances, but because the reservoir would spill anyway the memory of the decision is wiped out and has no bearing on the end cost.

Page 6, Lines 13-14. “We test two operating objectives: one that rewards a judicious response to an emergency (emergency response objective) response and one that rewards judicious continual adjustments (continual adjustment objective)”. As already mentioned, I don’t agree with this terminology because decisions are taken at a constant pace when considering both the objectives. The “emergency response objective” could easily become a “continual adjustment objective” if, for example, the demand was (relatively) large and the reservoir (relatively) small, because the release decision could change much more frequently. Viceversa, the “continual adjustment objective“ in Fig. 4 – Eppalock reservoir behave as an “emergency response objective”, because the release decision is most of the time equal to the maximum release and does not change frequently, just because the storage is for a long time higher than the target. I would suggest changing this terminology because it is misleading in my opinion... Page 7, Lines 1-2. “For the continually adjusted operating setting we find that the release must be adjusted constantly through the operating horizon to keep storage close to the target level of 75%”. As mentioned before, this objective is directly influenced by the variability of the inflow and reservoir release should change frequently to keep a constant storage. An example can be found in Fig.4 – Eppalock reservoir, where the storage is at the target storage, but the release keeps changing because of the inflow to the reservoir. I would include this explanation in the text.

We agree with this critique of the terminology and reconstructed the paper consistent with reviewer's recommendation to compare the reservoir settings by focusing on the objectives and consequent role of storage rather than frequency of decision adjustment.

Page 7, Lines 24-29. I would appreciate a discussion on the reason why the two objectives behave differently. The interpretation I propose in the "General comment" may represent a possible interpretation.

See previous response.

Page 7, Lines 29-32. I am not sure that the two objectives can be compared just on the basis of the value of the "injected forecast error". In fact, as commented in the "General comment", the "continually adjusted objective" is mainly driven by the forecast at one or few lead times, while the "emergency response objective" is driven by forecasts on longer lead times. On these lead times, the forecast skill is poorer by construction (if I understood how the synthetic forecasts were produced). What is the authors' opinion on this? Would it make sense to compute the forecast skill on different lead times, e.g. short lead times for the first objective and long lead times for the other objective? If, for example, it would turn out that the forecast skill in Fig. 5 a-d for injected error equal to 0.2 is comparable to the forecast skill of Fig.5 e-h for injected error equal to 0.6, spread of the forecast value would be similar in the two cases.

Whilst we accept that reliance on forecasts at longer lead times will impair the cost reduction achieved, we expect the influence of lead time on the spread of performance to be modest. The reason is that the spread of forecast skill for a given injected error is similarly tight across all lead times (fig. 2). Nonetheless, we think this point is worth raising in the description, even if only to alert the reader to the point that longer lead-time forecasts are likely to be more important in the supply targeted reservoirs. The use of the longer lead time forecasts for the supply targeted reservoirs (upper panels of Figure 5) explains the better cost improvements with the perfect forecast and the much sharper drop off in performance when error is injected.

Page 7, Lines 33-34. Why do the Burrinjuck and Serpentine reservoirs behave differently from the other two reservoirs? What are the features that they have in common which may justify the observed behavior? I would be nice if this result could be generalized.

This is actually a misstatement—thanks for picking this up. Eppalock is also sensitive to the increase in error, so it is Upper Yarra that is the outlier here. What we should have stated was: "Upper Yarra tolerates greater increases in error injected before forecast-based operations begin to be outperformed by control rules". We suspect the main factor at play here is the very short critical period for Upper Yarra relative to the other reservoirs (when draft ratio is 0.5). This means the storage buffer is less influential and adjustments to the release decision are required more frequently (i.e., it behaves a little more like the "continually adjusted" reservoir), as indicated in Figure 4. We expanded this section to include some reasoning for the behavior observed.

Page 7, Lines 38 and followings. "These results show that the measure of forecast error, quality, skill or goodness-of-fit - if based on the entire forecast period – cannot predict accurately whether that forecast will be valuable in an emergency-type operational setting”. What do you mean when you write: “if based on the entire forecast period”?

Here we're referring to measures of forecast quality that are based on an assessment of the hindcast as compared to observed conditions, over, say, 30 years. When we say "entire forecast period" we mean the skill measured across the hindcast (as opposed to selecting points within the hindcast, e.g., during drought, when the forecast performance may be more important to the performance in operations). We clarified this in the re-write, as it’s a critically important point.

Page 8, Lines 2-4. “This may be because the emergency response objective is constructed to be sensitive to a few serious shortfalls in meeting demand, while the continual adjusted objective rewards consistent performance over all the months assessed”. Both the objectives have a squared term, which should penalize to the same extent big deviations from the target (It could be different if there was an absolute value in Eq. 5, instead of the squared term). I think that the explanation may rely in the buffer effect of the reservoir, as explained in the “General comment”.

We agree that the explanation relies on the buffer effect of the reservoir to the extent that when reservoir is full the forecast is made redundant (the decision to release will be made irrespective of forecast quality). The result of this phenomenon is that the available improvements from forecasts in large supply targeted reservoirs depend solely on the quality during isolated periods when the reservoirs are drawn down. So (referring to the previous comment) measuring forecast performance across the whole hindcast may not indicate the value that will be reaped by that forecast in operation.

Page 8, Lines 15-19. This sentence is not fully clear to me.

We replaced this section with: “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.) While this ignores the spread of the ensemble, the chosen method provides a clear indication of the contribution of the forecast to the performance of the operation. In contrast, methods that use the spread of the ensemble in the decision process are complex, often requiring arbitrary decisions by the user. This makes experimentation laborious and results hard to diagnose…” [example of multi-stage stochastic optimization with recourse].

Page 8, Line 25. “our own prior experiments with this approach”. Please provide a reference, if any.

We don’t have a reference here, so we simply removed the statement.

Page 8, Line 35. I would explicitly state that the second experiment focuses mainly on the “emergency response objective”.

Agreed.

Figure 6. I would include a marker on the left hand side figures. I was surprised to see that there are more critical periods in case of low draft ratios (see fig. a, c, e). I would expect that higher draft ratios drive more critical situations. Could the authors comment on this?

It's important to remember that higher draft ratio reservoirs also coincide with larger storages (since all reservoir storages here are designed for reliability of 0.95). A general pattern that emerges in such circumstances is that reservoirs with larger demand (and storage) recover less readily, leading to a concentration of the 5% of time with failure on a single drought period. Smaller reservoirs can fail but then recover quickly, so the 5% of failure periods tend to occur multiple times over the simulation period. We added a sentence in the paper to notify the reader that this is the expected behavior.

Page 9, Line 10-12. "since the reservoir is drawn down at the end of the simulation, meaning the implications of late, sacrificial decisions are unavailable to quantify overall performance". Does this mean that the penalty on the final storage is not considered in the optimization? If so, why is it not considered?

The penalty of the final period is considered in the optimization undertaken for each forecast available. The issue here is that the impacts of late decisions for the full simulation period can't actually be compared fairly using a final period storage penalty. Clearly, ending with a higher storage would imply less cost using the storage penalty. But if the reservoir would recover anyway (if, for example, the inflows immediately after the simulation period were high) then the smarter action would have been to allow the storage to deplete more and meet the demand in full. In such an instance, the final storage penalty would misrepresent the value of the decisions taken leading up to the end of the simulation. In other words, we can't fairly evaluate the decisions taken within a given drawdown period unless the drought is allowed to play out in full (particularly when we're using lead times of up to one year).

Page 9, Line 28. "the hedge comes too late at the end of 2006 ". This is currently not visible in Fig. 8. I wonder if including the trajectories obtained with the perfect forecast might clarify what is "too late".

We agree that this is perhaps unclear from the figure presented, but we think that including the perfect forecast may make this plot overcomplex. Instead we annotated the figure to highlight the points raised.

Page 9, Line 31-32. I don't understand the comment. The reservoir does not fully recover in both the trajectories. This might be because the inflow is too small in comparison of the reservoir storage.

We clarified this in our re-write. It is true that neither trajectory recovers. What this means is that the impact of earlier release decisions matters for both operating modes. The control rules hedged (correctly), whilst the forecast-based operation didn't—and because neither reservoir recovers, these decisions become important further into the future, resulting in weaker forecast-based operations relative to the control rules.

Page 9, Lines 41-42. “forecast skill must be consistently available”. “Consistently” means “on long lead times”?

See earlier comment for ***Page 7, Lines 38****. This is something we defined more clearly in our revision. By ‘consistently’, we do not mean “on long lead times”. We mean that the skill of the forecasts, measured across, say, 30 years, should be consistently good if improved performance in operation is to be achieved. Forecast skill is usually calculated by averaging error scores over long periods, which can emphasise single forecasts that were either very good or very poor (relative to climatology). For example, a large majority of forecasts may perform slightly worse than climatology, but one or two forecasts perform very well, resulting in a positively skillful forecast (on average). However, in this example, the forecasts are unlikely to be useful to reservoir managers, because skill is not consistently available. This is particularly the case in supply reservoirs with large storage buffering capacity because the forecasts are seldom relied upon.*

Page 10, Lines 14 and following. I would revise the conclusion (as well as the abstract and title) to soften (or remove) the distinction between the “continually adjusted operation” and “emergency response operation”, as already commented several times in this review.

We amended the terminology used to contrast the reservoir operating settings in line with the reviewer’s earlier suggestion (see response to the first reviewer comment).

Page 10, Lines 19-20. It seems that the results of the second experiment are driven mostly by the Millennium Drought. This exceptional sequence of several dry years contributes in enhancing the value of forecasts. Would the results be much different if his extremely exceptional event was not considered? It would be interesting to have a look at the figures computed without the last years, to have an idea of the forecast value in case of less extreme droughts.

In forecast verification, we wish to understand how forecasts are likely to perform in future. This is best achieved by looking at forecast performance for the longest periods possible, making it undesirable to shorten the verification period. In addition, as future observations cannot be known with certainty, it only makes sense to condition verification on forecasts of dry periods (not observations). This is obviously not possible for multi-year droughts, because our forecasts only predict out to 12-months.

Technical corrections

Page 6, Equation 4-5. Should “T” be “H”, given the notation in Eq. 3? Define all the variables, not only “D”.

H is the 12-month forecast horizon, whereas T is the entire simulation period. We clarified this point and define all variables, as suggested.

Page 8, Line 12. “24 reservoirs” should be “32 reservoirs”.

Updated.

Page 9, Line 20. Fig. 8 seems to be cited before Fig. 7 in the text.

Corrected.

Reply to reviewer #3

The paper contributes to a better management of storage reservoirs by the use of (synthetic) seasonal forecasts. A major focus of the work is the assessment of the impact of different operating policies (emergency response versus continually adjusted) in combination with forecast of different forecast skills. The general setup of the experiments addresses the long-term operation of a reservoir system (monthly time steps) in application to a drought management.

General comments: The research topic is highly relevant. The practical value of seasonal forecasts, either by the classical ESP approach or weather models, needs validation in application to the management of water resources. The presented methodology seems to be a suitable tool to address the skill of actual or synthetic seasonal forecasts, furthermore, the authors address approaches to generate synthetic forecasts with defined skills to conduct systematic experiments.

Thanks for the positive feedback.

My main doubts are as follows: The classification of "continually adjusted" and "emergency response" objectives is misleading and gets the paper into a wrong direction. In the way implemented, the "continually adjusted" objective is a constant setpoint (75%, see page 6, line 25) for the reservoir storage. This is a very unlikely parametrization for a storage reservoir with water supply objectives and an annual hydrological cycle. The motivation of such a guide curve is to shift water from the wet to the dry season in order to guarantee a reliable water supply under consideration of an uncertain, variable yield. On the other hand, the "emergency response" objective has the character of a (soft) constraint. Both are incomplete if used exclusively and actual reservoir operation typically include both elements among others for flood control, recreation, hydropower etc.

We agree that the terms "continually adjusted" and "emergency response" are inapt, and in response to this comment (and the comments of Reviewer 2), we changed the terminology we use to describe the operating objectives. Rather than referring to the reservoirs as "continually-adjusted" and "emergency response", we refer to "level objective" and "supply objective". We also agree that reservoirs are typically multi-objective, requiring releases that consider both supply and other objectives. Our aim here is to show how those different objectives affect the value that might be gleaned from a forecast applied to the operation of a reservoir. The cleanest way to do this in an experimental set up is to separate the operating modes into two classes and then compare the performances using identical forecasts.

After the introduction into seasonal forecast, the synthetic forecast used in the experiments are disconnected from the actual products available. You should address the skill of actual seasonal forecast products as a benchmark for the synthetic forecast used.

We failed to mention that the two sets of forecasts are not entirely disconnected. The error structure embedded in the synthetic forecast model is trained using a member of the FoGSS forecasts. This ensures that even though the forecast is synthetic, its decay with lead time is realistic.

The paper may get published after major revisions. My advice is to give up the classification of "continually adjusted" and "emergency response" objectives and focus on the added value of seasonal forecasts of various skills in application to the reservoir management application.

In response to this comment (and those made by reviewer #2), we made the distinction between the two reservoir objectives in these simple terms: one objective targets constant storage by varying the release, the other targets constant release by varying the storage (or by allowing the storage to vary). This shift changes the role of storage from a target to a buffer, with consequent effects on forecast value that are brought out by our results. Rather than referring to the reservoirs as "continually-adjusted" and "emergency response", we used "level objective" and "supply objective". Discussion now places more emphasis on the role of a storage buffer in obviating the need for accurate forecasts during much of the operation.

We appreciate that a study into the added value of forecasts of different skills would constitute an interesting study. However, we feel that the originality of our work stems from the comparison of the operating types and in particular the surprising unpredictability of forecast value when applied to the supply objective.

Detailed comments:

Page 1, line 25: Do not forget the dimensioning of such a system, note that the storage volume of a reservoir is an explicit design decision.

The context here is systems that have been designed and are already in operation. We changed to "the performance of a given system depends on…"

Page 2, lines 7-18: Very clear example of the misleading classification into "continually adjusted" and "emergency response". You address flood control as "continually adjusted", but drought management as "emergency response". You could turn it around with the argument that relevant floods occur only "every 20 years by design". This is misleading, because the typical reservoir operating policy will reserve both a free volume due to flood control, a minimum volume for water supply, both seasonal dependent.

As noted above, we changed the description of the classification to highlight that the distinction is for a storage objective and a supply objective (rather than continually-adjusted and emergency response, which seems to have caused some confusion). Note for the storage objective, the operator is assumed to be unconcerned with maintaining supply, and vice versa. The typical policy of operating for both storage volume and supply is deliberately neglected, because we're trying to get at the influence of the operation type on the predictability of the forecast value in operation.

Page 4, lines 10-16: Revise this paragraph. Spill should be included in Equation 2. Either use inflow or release volume consistently if you like to refer to a volume, or alternatively use inflow and release if this is in flow units, but them introduce a time step in the equations.

We modified this as suggested.

Page 5, lines 22-32: You refer to advantages of the SDP. But against what kind of other technique? Furthermore, this description is biased and it appears that SDP has no disadvantages at all.

SDP is the conventional approach to design reservoir operating rules, so this is why we adopted it. There are multiple variants of SDP (such as approximate dynamic programming or stochastic dual dynamic programming; see Castelletti et al., 2010) as well as other approaches to reservoir operation (such as the parameterization-simulation-optimization framework; see Koutsoyiannis and Economou, 2003), which aim at improving the scalability of SDP to larger water systems. We understand that the paragraph might be biased towards the advantages of SDP, so we discussed briefly about its main disadvantages.

Page 5, line 34 -: This seems to be a deterministic technique only, please clarify.

Yes, this is a deterministic technique that optimizes a sequence of release decisions using a (deterministic) forecast of the inflow process. We clarified this aspect in the revised version of the manuscript. Justification for use of a deterministic approach are included in section 4.1, where state that "operations are 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., we take the median of the ensemble at each lead time). While this ignores the spread of the ensemble, the chosen method provides a clear indication of the contribution of the forecast to the performance of the operation. In contrast, methods that use the spread of the ensemble require in the decision process are complex, often requiring arbitrary decisions by the user. This makes experimentation laborious and results hard to diagnose…"

Page 7, lines 1-3: Results do not belong in here.

We think this is matter of stylistic preference, since the graphs presented are intended to justify the chosen objectives and do not show any result of either experiment executed. We included a statement to the effect that the graphs are given merely to highlight the suitability of the operating objectives chosen.

Page 8, line 20: Do you refer to Multi-stage Stochastic Optimization rather than Dynamic Programming? In the following, this paragraph reads more like a methodology section, not a results one.

Thanks for picking this up—we meant Multi-stage Stochastic Optimization. The description of prior results has been removed.

References

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

Koutsoyiannis, D., Economou, A. (2003). Evaluation of the parameterization-simulation-optimization approach for the control of reservoir systems. Water Resources Research, 39(6).

Complex relationship between seasonal streamflow forecast skill and value in reservoir operations~~Value of seasonal streamflow forecasts in emergency response reservoir management~~

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

[1]~~SUTD-MIT International Design Centre, Singapore University of Technology and Design, 487372, Singapore~~Pacific Northwest National Laboratory, College Park, MD, USA.
[2]SUTD-MIT International Design Centre, Singapore University of Technology and Design, 487372, Singapore.
[3~~2~~]CSIRO, Melbourne, Clayton, Victoria 3168, Australia.
[4~~3~~]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 surprising behaviour is the buffering role played by storage in water supply reservoirs, which renders the forecast superfluous for long periods of the operation. The main driver of system performance is the forecast accuracy at the timing during which critical decisions are made—namely during severe drought. Our results emphasise the importance of forecast skill consistency and highlight a need for sensitivity assessment in value-of-forecast studies involving reservoirs with supply objectives

1 Introduction

Coupled natural-engineered water resources systems ~~benefit society in a variety of ways~~provide a multitude of ~~vital~~ 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 ~~largely~~ on the quality of its operating scheme and the intelligence used to support ~~key~~ management decisions ~~pertaining to~~on the storage, release and transfer of water. Typically, 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 ~~optimal~~ 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). ~~Of particular importance is the operator's ability to inform decisions with accurate estimates of near-term water availability.~~ 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)~~. The widely held view is that that real-time (or “online”) operating schemes could supplant conventional, “offline” operating schemes, which specify fixed management decisions, or “control rules,” as a function of system state variables like time of year, snowpack depth, soil moisture, total water held in reservoirs, and so on.t has become apparent that offline control rules can~~ tend to.~~misguide the operator when a system undergoes significant change. A climatic shift, for example, can cause breakdown in established relationships between current system state variables and future water availability, leading to suboptimal operating decisions (Turner and Galelli 2016a).~~ ~~Second~~A move toward schemes~~,~~ informed by seasonal streamflow forecasts would benefit from a wealth of recent ~~the~~ science advances, including ~~of seasonal streamflow forecasting has advanced significantly in recent years. A number of~~ new –ensemble seasonal streamflow forecasting methods~~ have been developed~~, 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). ~~New s~~Seasonal streamflow forecast services are becoming available in countries such as the United States, Australia and Sweden. ~~Simultaneously,~~

~~a~~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 ~~demonstrably~~ improved with seasonal forecasts ~~in a variety of situations~~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 ~~would surely~~should be encouraged by these outcomes. But they would also wish to understand the associated risks and uncertainties. For example, if the new scheme increases the benefits of a ~~given~~ 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? ~~A study into the nature and causes of possible uncertainty of the value of forecast-informed operation would provide valuable new evidence to support this field of research.~~

~~Whilst the studies cited above are highly informative, providing strong motivation for further integration of seasonal forecast services into water resources management practice, they tend to focus on systems for which the water release decision is adjusted at frequent intervals—an operating mode we refer to as *continually adjusted operation*. Generally, reservoirs operated for flood control, hydropower and amenity fall into this category, as do low-reliability water supply reservoirs with short critical drawdown periods. This contrasts with *emergency response operations*, for which key management decisions are made relatively infrequently. The most obvious case is a drought action on a large, urban water supply system, such as water use restrictions imposed once every twenty years by design. Emergency response operations may need to be studied as a separate problem to continually adjusted operations, for the following reasons: the stakes tend to be very high and consequently the decisions are often highly politicized, attracting significant public attention (Porter et al. 2015); and the infrequent nature of the decision means that system performance will be determined solely by decisions made at critical moments, such as at the onset of major drought. The ultimate value of a forecast-based operating scheme will depend on the forecast quality during those critical moments, as well as on the operator’s willingness to act.~~

~~In order t~~To explore ~~the value of applying seasonal forecasts in emergency response operations~~~~possible~~ 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. ~~The purpose of o~~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.~~is to determine whether there is any difference between the two operational settings – continually adjusted and emergency response – in terms of the value they reap from seasonal streamflow forecasts. To address this question we compare the relationship between forecast performance and operational value for each case. This is achieved by fabricating for each catchment a hypothetical reservoir and 1000 synthetic inflow forecasts of varying quality, from near-perfect to low-skilled, and then simulating using a rolling horizon, adaptive control operating model.~~ As we shall see, the contrast in the relationship between forecast skill and forecast value between the two operational settings is striking. ~~The purpose of o~~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. ~~is to understand whether this result can be explained by forecast performance during periods for which operating decisions become critical. For this experiment we apply an advanced seasonal streamflow forecast system. By varying the reservoir design for each catchment we shift the critical decision points onto different periods, allowing us to investigate the importance of isolated moments of forecast performance to the overall value of the forecast in operation.~~ Results provide new insights into the risks operators take when applying ~~a~~ seasonal forecasts to critical management decisions in ~~emergency response systems~~ 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 ~~implement~~ vary an ~~the~~ ‘~~error~~ injected error’ parameter, ~~that~~ which ~~allows us to~~ controls the error of the ~~generated~~ synthetic forecast. ~~This~~ The ~~“error~~ injected error” ~~parameter~~ takes values between 0 and 1, where 0 ~~implies~~ generates a perfect forecast and 1 ~~yields~~ 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 at any given moment 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. ~~We generate four sets of~~For each catchment, we generate 1000, 12-month ahead, monthly-resolution synthetic forecasts. The quality of the forecasts ~~of~~ is varied by~~varying quality 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). 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 ~~post-processes~~makes use of climate forecasts from the Predictive Ocean and Atmosphere Model for Australia (POAMA) (Hudson et al., 2013; Marshall et al., 2014), ~~with~~ 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 ~~hydrological~~ 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~~ 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~~, with these small biases resulting in high relative errors in February and March. However, because inflows are so low during these months, these relative errors have very little influence on annual (or even seasonal)~~

~~water balances~~. 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).

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

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

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

where S_{cap} is the capacity of the reservoir and R_{max} is the maximum water release, ~~volume of water that can be released during any time period,~~ taken in this study as twice the release target to give the operator ample storage level control. All excess water is spilled~~, i.e., *Spill*$_t$ = max ($S_t + Q_t - R_t - S_{cap}$, 0)~~.

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 (~~continuously adjusted operations versus emergency response~~level objective versus supply objective). Second, it enables us to examine the value of forecasts for reservoirs sensitive to different types of ~~drought~~hydrological conditions~~, such that overall forecast value becomes dependent on the quality of the forecast at different time periods in the simulation (necessary for experiment 2)~~.

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 ~~releases~~ water is ~~cannot be met~~satisfied in full—to the total number of months simulated. ~~A target of 95%~~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 the demand in full if water is available to do so (~~this design assumption means~~meaning that the resultant reservoirs will be very unlikely to empty when operated with more advanced techniques). A constant demand for water is assigned for eight alternative reservoirs by varying the~~a~~ 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~~, which summarises the reservoir designs, shows that the storage must be increased in order to meet greater demand at the target reliability~~. 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 ~~indicates within-year reservoirs that~~suggests that the reservoir will ~~re~~fill and spill each year). ~~In other words,~~The wide variance across these indicators suggests that as demand is adjusted, the storage dynamics are affected and the reservoirs ~~becomes~~ 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 ~~long-duration events with~~ 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 – ~~2011~~2010.

2.3 ~~Operating schemes~~Benchmark scheme: control rules

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. ~~But~~ Then to understand the value of a forecast-informed operating model, we ~~also~~ 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~~specifies~~ control rules~~ to govern the operation of a reservoir. Control rules are~~, which are established by optimising ~~reservoir operations~~ release decisions ~~on the basis of~~for historical ~~inflows~~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 Conventional operating 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 (~~described in~~see Section 2.4).

The most rigorous way to design such rules is by optimisation. ~~In this study~~ Here we use *~~Stochastic~~ stochastic ~~Dynamic~~ dynamic ~~Programming~~ 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 ~~non-linear~~ 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 ~~using~~ with the same leave-five-years-out cross-validation scheme employed in FoGSS (Section 2.1.2). SDP suffers from two well-known drawbacks~~ that are worth~~: 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 *m~~M~~odel ~~Predictive~~ predictive ~~Control~~ 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~~= 12~~ 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,…,R_H) is optimised to minimise the cost over the forecast horizon H plus the cost associated with the resulting storage state:

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

where C_t is the penalty cost calculated from the reservoir's objective function (see ~~defined mathematically in the following section~~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 ~~the~~ 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~~ month-~~of~~ of-~~the~~ the-year (Appendix 2). By using Bellman's function in this way, ~~here~~ we effectively append the forecast-informed scheme to the control rules. In ~~essence~~effect, this means that the information contained in the forecast is used ~~effectively~~ 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,…,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). While this approach ignores the spread of the ensemble, it provides a clear indication of the contribution of the forecast to the performance of the operation. In contrast, methods that use the spread of the ensemble present a number of technical challenges. ~~While this ignores the spread of the ensemble, we pursue this method because using an ensemble forecast in a multi-stage optimization scheme requires recourse in the control. We~~ One cannot simply optimis~~z~~e the release decision by minimi~~z~~sing 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 ~~Optimization~~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. ~~Moreover, our own prior experiments with this approach, in which we reduced the FoGSS ensembles to scenario trees using both the information flow modelling approach (Raso *et al.*, 2013) and the neural gas algorithm (Turner and Galelli 2016c), yielded performances no better on average than those obtained using the median of the ensemble in a deterministic, rolling horizon approach.~~

2.4 Operating objectives

We test two operating objectives: one that rewards ~~close tracking of a~~meeting target releases ~~rewards a judicious response to an emergency~~ (*~~emergency response~~ supply objective*) ~~response~~ and one that rewards meeting~~rewards judicious continual adjustments~~~~close tracking of~~ target storage levels (*~~continual adjustment~~level objective*). The ~~emergency response~~supply objective encourages full release of water to meet target demand except under drought conditions:

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

where D is the demand and C^{supply} ~~–~~ is the penalty cost used in the adaptive control scheme (~~e~~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 ~~continual adjustment~~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:

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

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

Figure 4 gives storage behaviour and release decisions implemented for 0.95 reliability reservoirs (draft ratio = 0.5) operated for the ~~water~~ supply objective (~~left-hand panes~~Equation 4)~~)~~ and ~~storage~~ level objective (~~right-hand panes~~Equation 5), under ~~objectives described above (here the operation is~~ rolling horizon, adaptive control operation with a perfect 12-month forecast~~)~~. The figure ~~figure is intended to~~shows the contrast~~s~~ in the frequency of decision-making for the two ~~types of operation~~ operating objectives~~ under study~~. For ~~emergency response operations~~the supply objective we ~~find~~ 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 ~~continually adjusted operating setting~~level objective we ~~find~~ 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.~~This contrast shows that the two chosen objectives are suitable for representing emergency response and continually adjusted operation in this study.~~

3 Experiment 1 – ~~Comparing impact of operating mode on value of forecast-informed operations~~ Characterising the uncertainty of forecast value in reservoir operations

3.1 Experiment description

The purpose of the first experiment is to examine ~~whether~~ the nature of uncertainty in forecast performance under two ~~a change in~~ contrasting operating objective~~s~~s (~~continually adjusted~~level objective versus supply objective~~emergency response~~) ~~affects the relationship between forecast quality and forecast value in operation~~. 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 ~~emergency response~~supply objective and once for the ~~continual adjustment~~ 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~~Figure 4~~ shows the value ~~of forecasts achieved by~~of the forecast-informed scheme for each ~~of the four inflow time series~~reservoir. The ~~V~~value of forecasts is presented as the ~~relative~~ reduction in~~of~~ cost~~s~~ relative to ~~(%) with respect to~~ control rules (%). A positive cost reduction indicates ~~outperformance~~that the forecast-informed scheme outperforms control rules, and ~~and~~ a negative cost reduction indicates ~~underperformance of~~ ~~: positive/negative values indicate the~~that control rules outperform the forecast-informed scheme~~ outperforms/underperforms the~~relative to~~ control rules~~. Forecasts with zero error (i.e., perfect forecasts) ~~uniformly~~ outperform control rules in all cases, regardless of the objective. Interestingly, when operated with a perfect forecast the ~~supply targeted reservoirs~~reservoirs operated to meet the supply objective enjoy a significantly larger percentage increase in performance (40 – 60%) ~~relative~~compared with~~ to~~ the ~~level targeted~~ 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 ~~substantial~~ 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 ~~settings~~objectives as the forecast error is increased. ~~As error increases,~~For the supply objective~~ the value of forecasts behaves differently for the emergency response~~ (panels a – d) ~~and continually adjusted (e – f) objectives. For the emergency response objective (Figure 4a-d),~~ forecast value declines rapidly, becoming ~~becomes relatively unpredictable~~highly unstable with the injection of a moderate error ~~as soon as error is introduced~~ into the forecast. ~~This contrasts markedly with~~ For the ~~continually adjusted~~level objective (panels ~~Figure 4~~e – h), ~~for which~~ the forecast value ~~generally~~ decreases relatively slowly, and the points remain tightly grouped for errors up to~~ as forecast quality is eroded. For errors up to~~ ~0.4~~ the points are tightly grouped and the forecasts are valuable, showing that forecast error correlates strongly with forecast value for these reservoirs~~. Taking Burrinjuck (Figure 5~~Figure 4~~a) 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 ~~emergency response~~supply- ~~targeted~~ ~~operations~~objective (i.e., the forecast-informed operations ~~could be~~are outperformed by simple control rules by up to 5% in some instances). The same forecasts applied to the ~~continually adjusted~~level~~-targeted operations~~ objective (~~Figure 4~~Figure 5e) result in cost reductions in the narrow region of 24 to 26%.

These results show that for the supply objective, the measure of forecast error, quality, skill or goodness-of-fit ~~cannot~~do not always accurately predict ~~accurately~~ whether that forecast will be valuable~~ in a supply-targeted reservoir~~. We hypothesise that this unexpected phenomenon relates to the role played by storage. When operated ~~for the level target~~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. ~~At moments when forecasts skill is weak, release decisions may underperform relative to control rules.~~ ~~But because the forecast is mobilised constantly throughout the operation, the net will always be an improvement (assuming the forecast is, on average, skilful relative to climatology).~~ For the supply objective, however, storage ~~plays a vital role in maintaining the desired performanc~~actively buffers inflowse. When storage levels are high~~s are sufficiently healthy~~, 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. In very large reservoirs, it may take a number of consecutive drought years for storages to drop to levels that raise concern. Only then will the option of reducing the release be considered. ~~This is crucially important, because it means that forecasts issued out to twelve months will have absolutely no prospect of changing the release decision (meet demand in full) for long periods of operation during which storages are near capacity.~~ 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. ~~And because each synthetic forecast of given overall error level will vary in its quality through time, the additional value of the forecast measured over a 30-year operating period will vary as a function of the forecast set applied~~ Forecast skill is often measured by averaging errors over a long period of time (as done in Figure 3). ~~. It would in theory be possible~~ Figure 5 shows that it is possible for a '~~an overall~~ skilful' forecast (measured on average) ~~set of forecasts~~ to generate a net reduction in performance if the skill level dips during the critical point in time where the forecast is mobilised.

This ~~mechanism~~ability of storages to buffer inflows ~~would~~ also explains why~~, under the supply objective,~~ the Upper Yarra Reservoir, under the supply objective, ~~appears to be less sensitive to change in forecast value~~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.

~~For the emergency response objective, the Burrinjuck and Serpentine reservoirs are particularly sensitive to forecast errors, as errors <0.2 can result in forecast-informed operations being outperformed by control rules. We have numerous forecasts of low error that deliver consistently strong performance in the continually adjusted operational setting, but which can cause a reduction in performance in the emergency response setting.~~ At draft ratio of 0.5, Upper Yarra has the shortest critical period, lowest storage ratio and highest drift value ~~(the only reservoir of the four with drift > 1, indicating within-year storage behaviour)~~(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 ~~the~~ release decisions are required ~~at more periods throughout the operation~~more frequently (as observed~~, in fact,~~ in Figure 4). ~~Perhaps surprisingly, we observe several forecasts with high error that deliver consistent reduction in performance in the continually adjusted operational setting, but which can cause improved performance in an emergency operational setting.~~

~~These results show that the measure of forecast error, quality, skill or goodness-of-fit – if based on the entire forecast period – cannot predict accurately whether that forecast will be valuable in an emergency-type operational setting. This may be because the emergency response objective is constructed to be sensitive to a few serious shortfalls in meeting demand, while the continual adjusted objective rewards consistent performance over all the months assessed.~~ We now turn to experiment 2, which ~~aims at explaining the phenomenon observed here~~at exploringes 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 aim of the second experiment~~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 ~~performance~~ skill level when applied to reservoirs with the supply objective~~in emergency response operations~~. 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 ~~24~~ 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.)~~Operations are then simulated using both the control rules and the forecast-informed model using the median of the value from the full FoGSS forecast ensemble (i.e., the median of the ensemble is taken as the expected inflow at each lead time)~~. ~~While this ignores the spread of the ensemble, we pursue this method because using an ensemble forecast in a multi-stage optimization scheme requires recourse in the control. We cannot simply optimize the release decision by minimizing 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 Optimization, 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 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. Moreover, our own prior experiments with this approach, in which we reduced the FoGSS ensembles to scenario trees using both the information flow modelling approach (Raso *et al.*, 2013) and the neural gas algorithm (Turner and Galelli 2016c), yielded performances no better on average than those obtained using the median of the ensemble in a deterministic, rolling horizon approach.~~

We compute the ~~performance attained with~~value of FoGSS forecasts in relation to both an upper benchmark (perfect forecast) and a lower benchmark (control rules) ~~the performance attained using a perfect forecast (i.e., a 12 month ‘forecast’ of the observed inflow)~~:

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

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 ~~we observe~~ 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~~–d~~) specify times at which ~~operations~~ operations decisions become critical (herein termed “critical decision periods”. These periods are defined as moments when ~~(defined as points in which perfect forecast operations implement (i.e.~~ supply is cut back~~s~~) when operating with ~~forecasts are~~ perfect forecasts (i.e., moments when the operator should be adjusting the release). A general pattern that emerges ~~in~~ 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 5% of time with failure on a single drought period. In contrast, smaller reservoirs with relatively low demand often fail but then recover quickly, so the 5% of failure periods tend to occur multiple times over the simulation period. Here we see that f~~F~~or 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), ~~these~~ ~~ciriical~~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 ~~horizon~~period. Critical decision periods for the Serpentine also occur to the end of the record, reflecting the long-term trend of ~~declines~~ declining ~~in~~ inflows ~~over this period~~since 1975 (Petrone et al. 2008). ~~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).~~

The right-~~-~~hand panels (Figure 6 b, d, f, h~~e–h~~) show how operating performance varies with draft ratio. The FoGSS-informed operating model offers performance improvements (i.e., ~~Performance~~ performance ~~Gain~~ 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 reservoir is drawn down at the end of the simulation, meaning the implications of late, sacrificial decisions are unavailable to quantify 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~~d~~ 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~~Figure 8 (a, b)~~ gives storage and release time series (2005–2011) for the Serpentine reservoir with 50% (where performance gain is positive) and 80% (negative performance gain) draft requirements. Whilst the differences in release between control rules and forecast-informed operations appear modest in these time series, the practical implications of these differences c~~ould~~an 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), ~~During this period~~ the storage depletes and recovers (fully) a number of times.~~,~~ Within the sequence shown ~~and~~ there is a two-year period beginning~~from~~ mid-2007 ~~to mid-2009~~ 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 ~~2011~~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

Field Code Changed

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 7~~Figure 8~~c–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 forecast error in ~~emergency~~ supply-targeted~~response~~ operations, demonstrated with the synthetic forecasts in Section 3.2 (~~this~~ also ~~holds~~ true for ~~the~~ Burrinjuck Reservoir).

~~We have shown that~~~~The~~We have shown earlier that ~~results show that~~ FoGSS forecasts are skilful, on average, for the 1982-2010 period (Figure 3). ~~but~~Yet this masks the degree to which skill varies over shorter periods. As we have seen, ~~emergency response~~the supply objective~~s~~ ~~can fail~~can result in a situation where the forecast is mobilised in only a few, ~~short~~ crucial periods, meaning that ~~for~~ forecast skill ~~must be~~may need to be consistently available to ~~aid emergency response objectives~~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~~Figure 7~~ 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 ~~emergency response~~supply objectives.

5 Discussion and conclusions

~~We have shown~~

Our findings have general relevance for an increasingly water constrained world, where the demand for water and variability of climate are, in many regions, increasing 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/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 ~~continually adjusted operations~~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~~for emergency response~~ supply-targeted reservoirs~~ objectives~~, 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 ~~be of greater value~~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 fabricateded~~'designed'~~ our reservoirs to specific draft ratios and reliabilities with recent inflows records~~ and then been able to 'construct' the reservoirs instantly thereafter~~. In practice, reservoirs have long service lives (typically decades), leaving them vulnerable to possible ~~increases~~changes to the~~in the variability of~~ inflow regime~~s~~ beyond~~since~~ their construction (~~perhaps due to climate-driven changes; ~~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 ~~claw back~~compensate for some of the losses in ~~storage reliability~~performance in such instances.

While the value of forecasts was strongest for the level objective, we have shown that ~~Despite this~~forecasts can also offer value to reservoirs operated to a supply objective.~~, w~~ T~~e have shown with t~~he real-world example of the FoGSS forecast system showed that skilful forecasts improve ~~emergency response~~supply-targeted operations in the majority of reservoirs used in this study, relative to conventional control rules. Meeting the ~~emergency response~~supply objective essentially requires effective action in only a few ~~critical ~~crucial instances. Accordingly, we contend that if forecast skill is consistently available, forecasts will better ~~enable emergency response operations objectives to be met~~improve the operator's ability to manage a system to meet a supply ~~-targeted system~~objective. We therefore recommend measuring the consistency of forecast skill as a useful predictor of the value of forecasts to ~~emergency response~~supply objectives.

~~Our results from Experiment 1 show that~~It appears that the operator of ~~an ~~a ~~emergency response~~supply-targeted system will need to accept greater risk than the operator of a ~~continuously adjusted~~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 Sao 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 ~~bad operating decisions~~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 ~~attack ~~criticism in the event that the scheme fails to work at the time that matters most. This is particularly true for ~~in situations of ~~emergenc~~y~~ies, which attract significant public attention and political interest (Porter et al. 2015)~~—~~. ~~and i~~It~~'s~~ 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~~Their~~ 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. ~~Lastly, the~~The high variability of the performance of ~~emergency response~~supply-targeted systems presents potential pitfalls for case studies assessing the value of forecasts. The ~~sometimes weak~~unstable relationship between forecast accuracy and operating performance means that even good forecasts may result in poor operational performance. ~~The converse may be even worse for operators~~Or perhaps worse,~~:~~ ~~by chance,~~ mediocre forecasts may show strong performance for ~~emergency response~~supply objectives, giving potential users~~operators~~ false confidence in the forecast-informed operating scheme~~s~~. ~~For emergency response operations~~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 ~~must be~~is on the analyst to determine whether the forecast service is sufficiently and consistently ~~reliable~~ 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 ~~investigations~~ needed to determine the ~~potential~~ value of forecasts for a range of design and operating settings ~~to which those forecast might be applied~~. This can be done by measuring improvement in system performance ~~improvements~~ as defined by the operating objectives. We use~~d~~ a rolling-horizon, adaptive control approach to demonstrate that the relation between forecast performance and operational value varies significantly when comparing ~~continuously adjusted and emergency response operational settings~~level-targeted and supply-targeted operations. We demonstrate a clear and strong relation between forecast skill and value for ~~continuously adjusted~~~~level-taragteted operation~~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 ~~emergency response~~supply- targets~~ed systems~~ (*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 ~~can benefit~~add value to 25 of the 32 ~~emergency response~~reservoirs tested, when they are operated to meet the supply objective ~~operations in a number of settings, with several notable exceptions~~. ~~In all systems~~For reservoirs operated to a supply objective, the driver of operating performance is the forecast accuracy ~~at the timing~~ during ~~which~~ a small number of periods where critical decisions are made. We conclude that for forecasts to add value, 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 ~~occurs~~ does not occur over total number of years simulated).

Standard Operating Policy (SOP)~~:~~

Standard operating policy (SOP) is ~~A~~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) ~~with~~ at a specified ~~target,~~ 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, ~~the design converges on the target reliability and~~an optimal storage capacity is attained.

Critical period~~:~~

The critical period is ~~taken here as~~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 ~~optimisation problem~~objective is ~~solved~~ minimised using a backwards recursive procedure:

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

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 ~~optimisation problem~~penalty cost given ~~by~~ in Equation 3 is ~~solved~~ 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 (BUI)	Perennial	1252.1	0.90	1631	1900 – 2014	-35.00	148.58	NSW
Lake Eppalock (EPI)	Ephemeral	166.8	0.82	1749	1900 – 2014	-36.88	144.56	VIC
Serpentine (SEI)	Intermittent	58.4	0.69	664	1912 – 2014	-32.40	116.10	WA
Upper Yarra (UYI)	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. 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

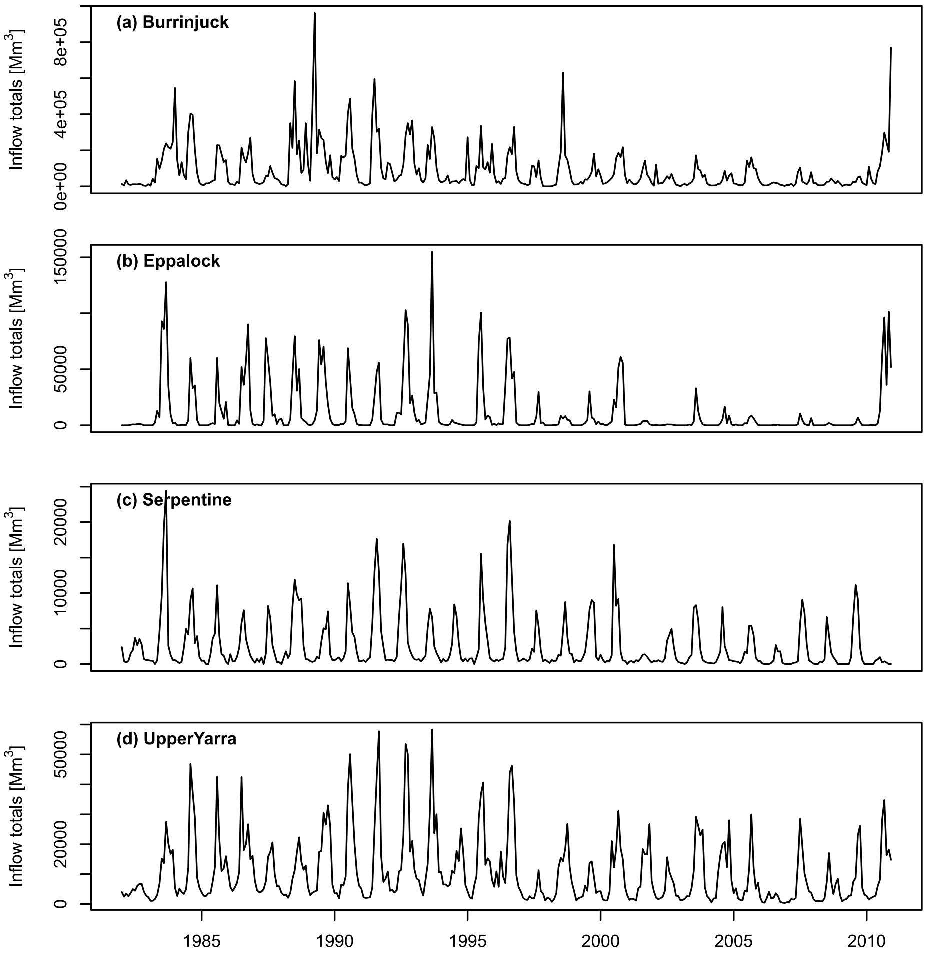

Figure 1 - Reservoir inflow records for (a) Burrinjuck Dam (BUI), (b) Lake Eppalock (EPI), (c) Serpentine Reservoir (SEI) and (d) Upper Yarra Reservoir (UYI) during the 29-year study period Jan 1982 – Dec 2010.

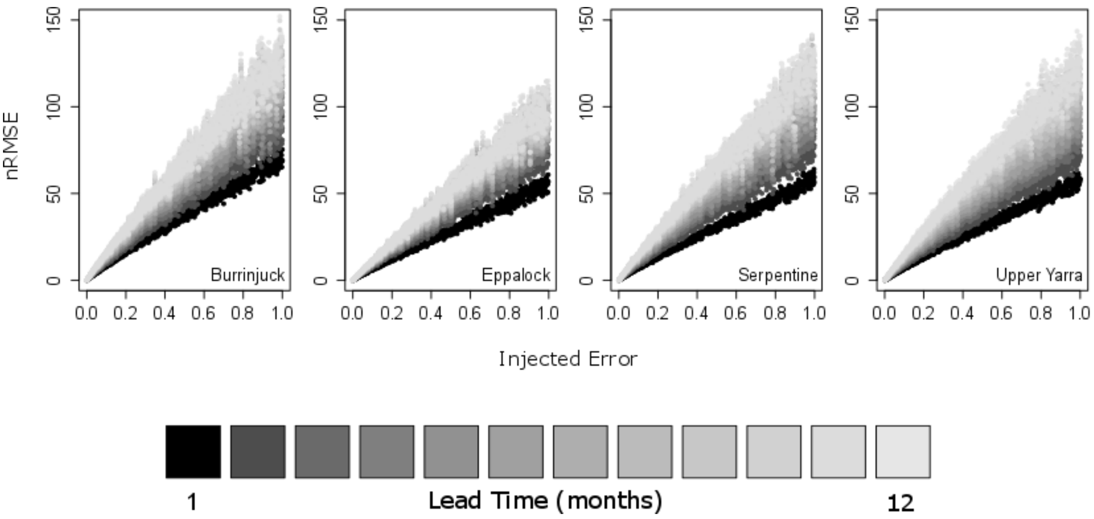

Figure 2 – Normalised Root Mean Squared Error (nRMSE) for varying error injected into synthetic forecasts generated using the Martingale Model of Forecast Evolution (1000 forecasts, monthly resolution, 12 months ahead, giving 12,000 points on each pane).

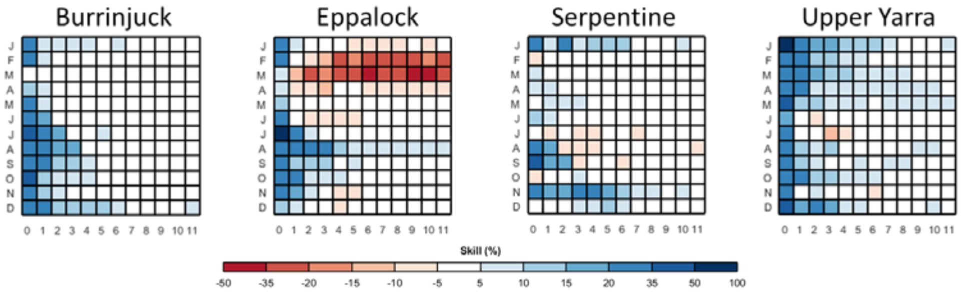

Figure 3 – FoGSS ~~Forecast~~ forecast skill measured by the continuous ranked probability score (CRPSS) with respect to climatology forecasts. Rows show target months, columns show lead-time in months.

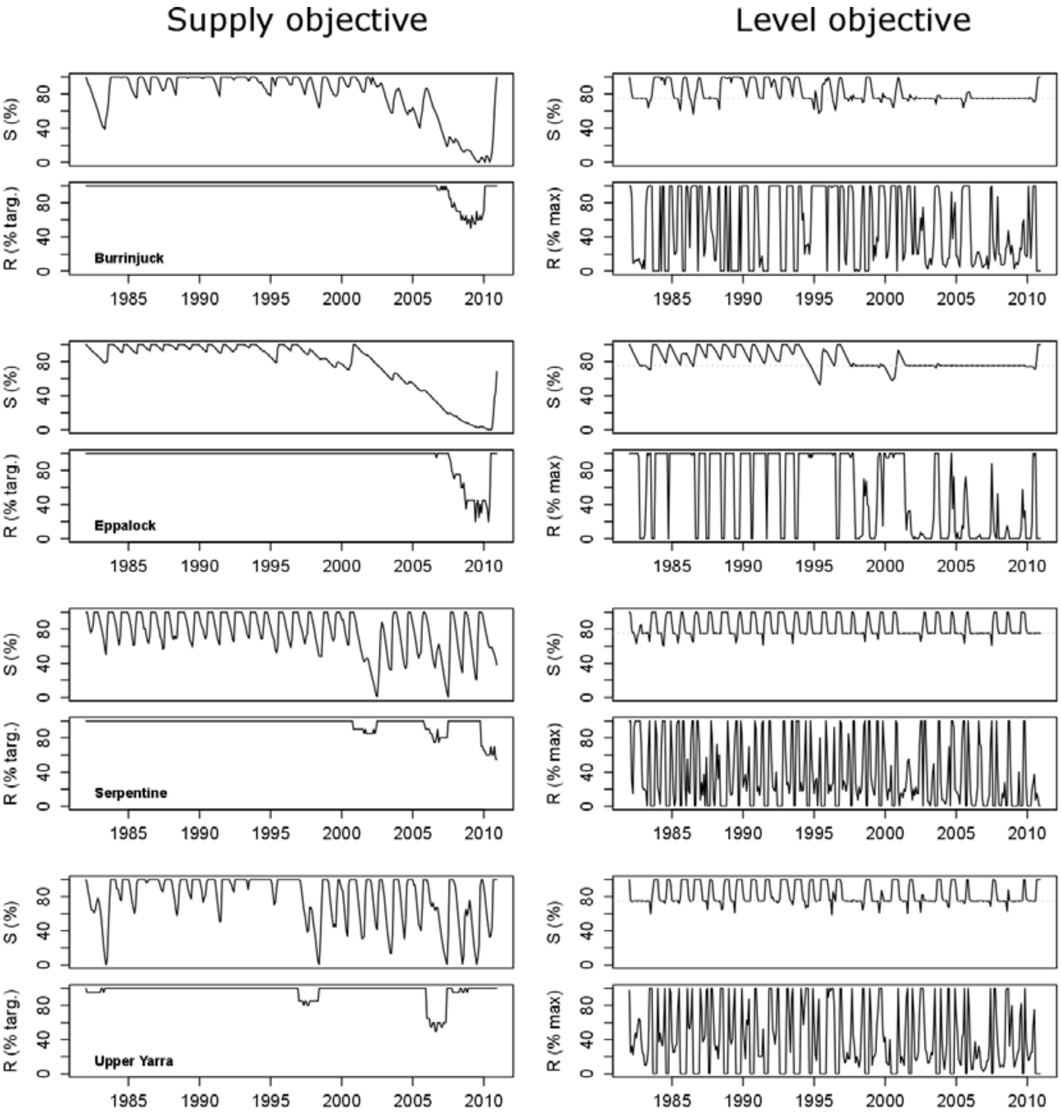

Figure 4 – Behaviour of reservoirs operated to meet th~~undere~~e ~~emergency response~~supply~~-targeted~~ objective (left column) and ~~continually adjusted~~level~~-targeted~~ objective (right column)~~operations~~. Simulations use the rolling horizon model with a perfect 12-month (observed) inflow 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).

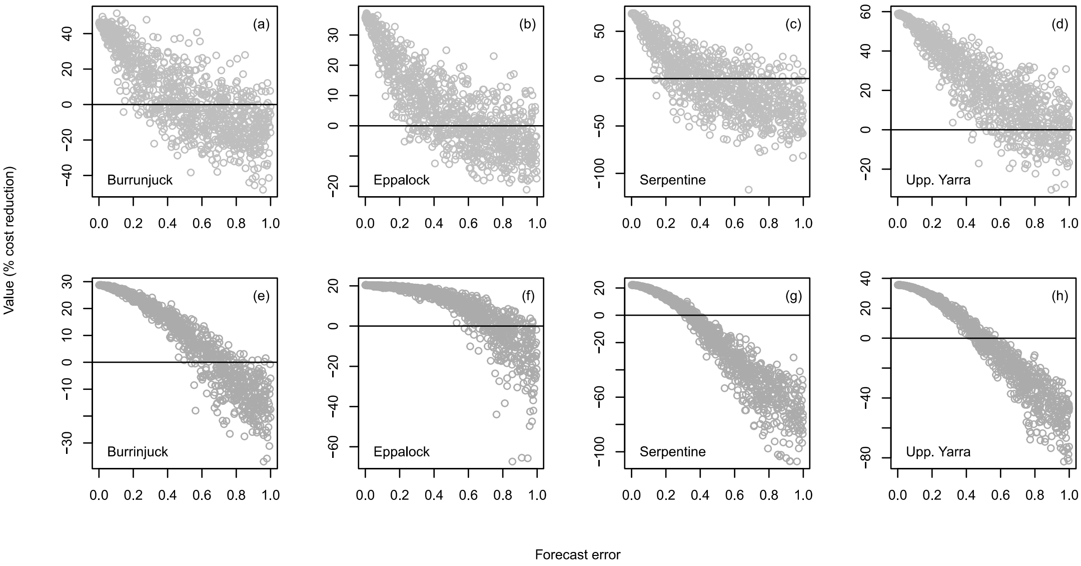

Figure 5 – Value of the forecast-informed scheme over control rules as a function of forecast error for ~~emergency response~~supply objective (a – d) and ~~continually adjusted~~level objective (e – h) operational settings.

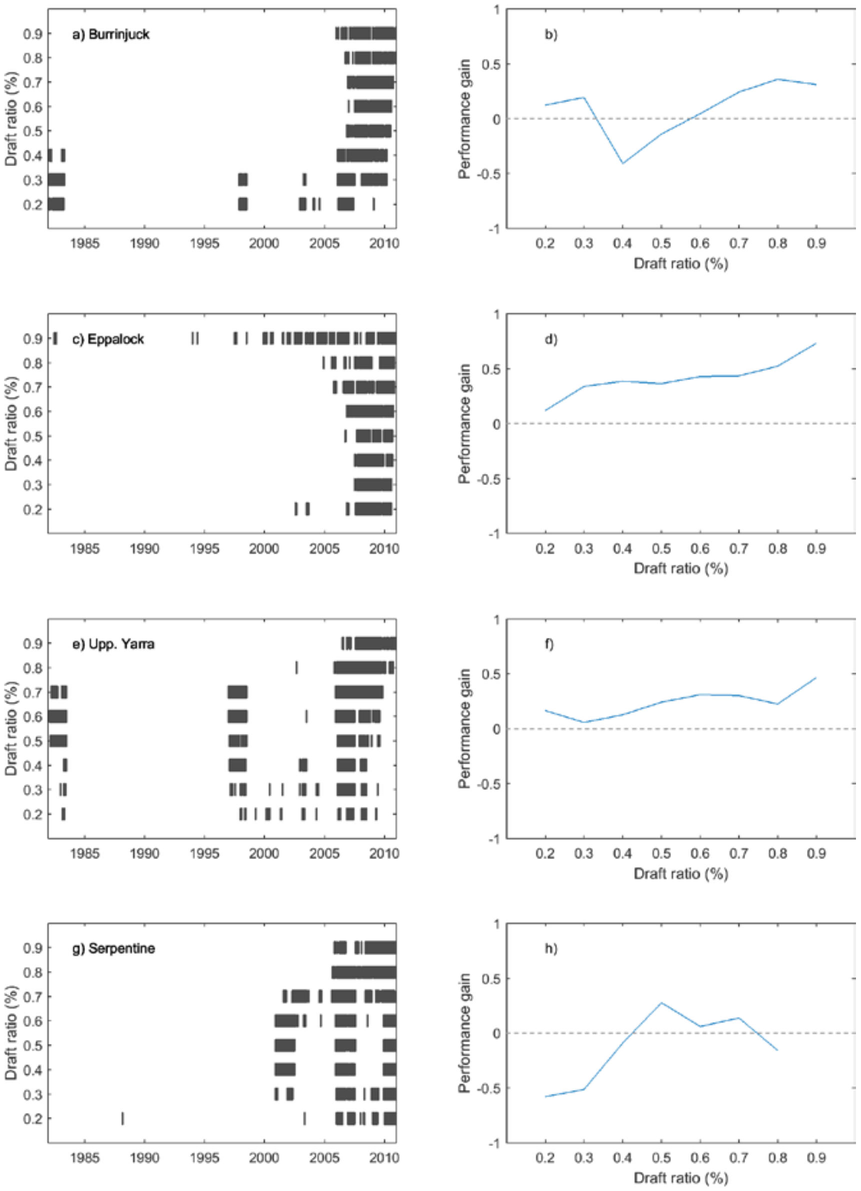

Figure 6 – Panels a, c, e, g~~–d~~ give critical decision periods for each reservoir design (draft ratio 0.2, 0.3, …, 0.9). Panels ~~e–~~b, d, f, h give performance gain plotted against draft ratio. Critical decision periods are moments during which perfect forecast operations implement supply cutbacks.

Formatted: English (Singapore)

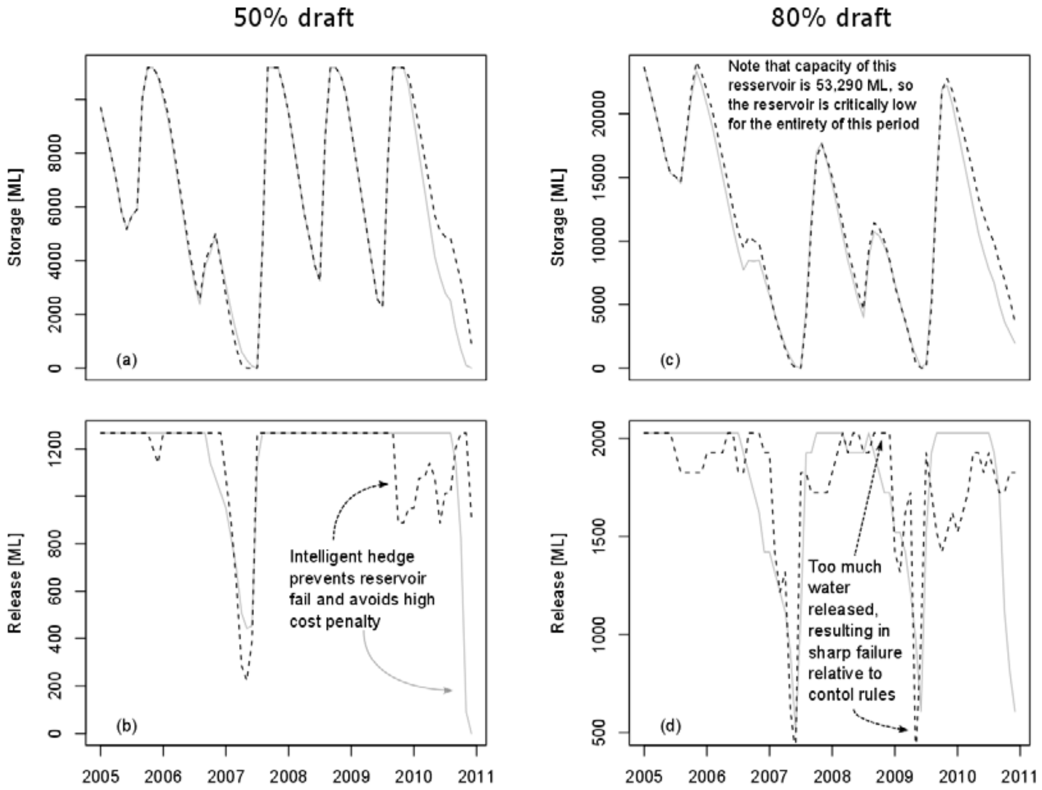

Figure 7 – Storage and release time series for reservoirs in the Serpentine catchment with 50% (a, b) and 80% (c, d) draft ratios. The solid grey line gives operation under control rules whilst the dotted black line gives operation with the FoGSS forecast (median of ensemble).

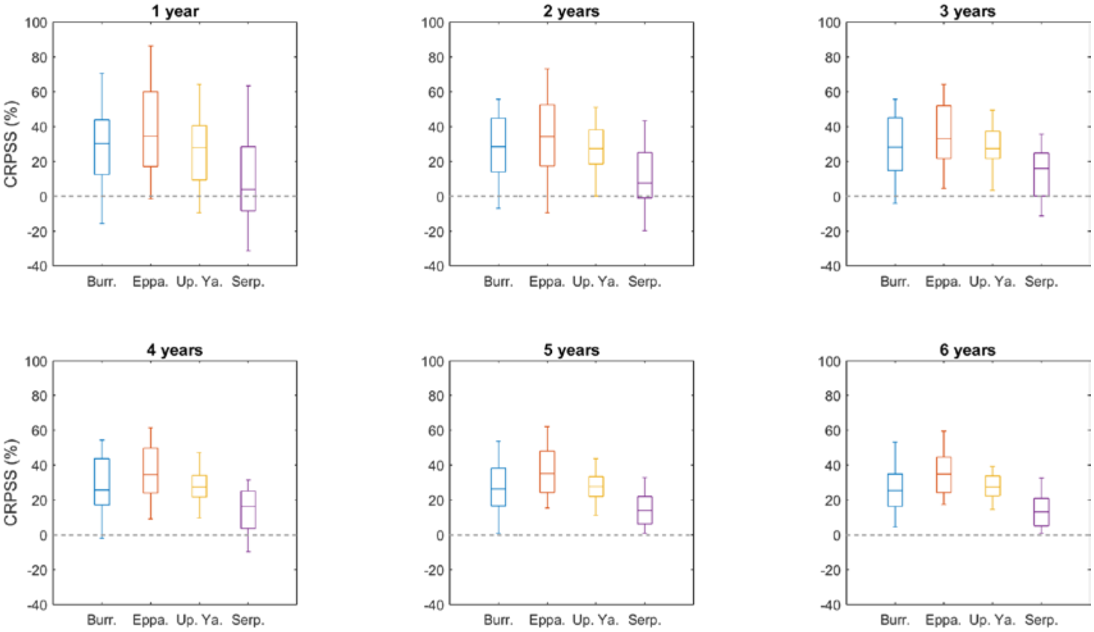

Figure 8 – Variation in skill of lead-0 FoGSS forecasts for blocks of consecutive months. Skill for consecutive months for blocks of 1-6 years is bootstrapped to create the box and whisker plots. Boxes give interquartile range, whiskers give 90% intervals, lines show median values.

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, 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.

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.

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 (2016c), scenario: Construct reduced trees with predefined nodal structures, R package version 1.0.0, Comprehensive R Archive Network, Vienna, Austria.

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.