Responses to comments referee comments on “Improving statistical forecasts of seasonal streamflows using hydrological model output” by D. E. Robertson et al.

Referee comments are in black and author responses are in blue.

Anonymous Referee #1

Received and published: 21 September 2012

This paper presents a comparison of the impact of using a hybrid statistical streamflow forecasting model over the operational statistical streamflow forecasting model currently used by the Australian Bureau of Meteorology. The new predictor that is used is the previous month’s modelled streamflow from the hydrologic model WAPABA along with a prediction about future climate influences (same as for the operational model). The inclusion of the WAPABA streamflow makes it a hybrid model as the streamflow forecasts now incorporate information from a dynamical hydrologic model. The authors find that the hybrid model generally leads to improvements in forecast skill for most catchments. The investigation into the reasons for the improved skill is thorough and should provide a useful contribution to statistical streamflow forecasting literature. The paper is well written although I think it could be expanded in a few sections as outlined below to make the work clearer. I recommend that the paper is accepted subjected to minor revisions.

Page 8707 Lines 8 - 26 - I found the details on the current operational BOM method to be a bit too terse. I can see that the details are in Robertson and Wang (2012) but the current paper required me to reread several times before it all came together.

We have reworded and expanded this paragraph to hopefully make the approach clearer. One of the challenges we face with this section is balancing the description of work which is reported elsewhere with the new work which is the focus of this paper.

Page 8708 Line 27 - 28 - On my initial reading I found this section quite confusing. The explanation on Page 8717 cleared things up for me but you might want to consider rewording this section to make it clearer.

We have reworded this sentence to make it clearer exactly what we have done.

Page 8709 Line 17 - I think it would be useful to explicitly state that since the skill scores measure a reduction in error, that a large positive value is good and a negative value means an increase in error compared to reference case. I think it would also be useful to note that negative skill occurs where the climatological forecasts are better.

A sentence to this effect has been included here.

Figures I’m not really a fan of the presentation of Figures 3, 4 and 9. I think it’s too hard to tell the difference between the different colours especially for the panels further away from the colour scale on the right . I would prefer to see the use of bands of colour to represent specific ranges of values rather than a continuous range. If you feel that you need to maintain the continuous range then I think at least around 0 skill value you need to have a more neutral colour.

We thank the referee for this suggestion and have complied with the advice. Figures 3, 4 and 9 (now Figures 4, 5 and 11) are now presented in color bands rather than continuous color ranges.

Figure 6 - it's difficult to see the light blue uncertainty bars particularly in the lower parts where they are over the grey bands for the climatological values.

Figures 6, 7 and 8 (now Figures 7, 8 and 9) have been adjusted to increase the contrast between the blue uncertainty bars and the grey climatology values.

Anonymous Referee #2

General Comments

This is a well-written, well-organized paper describing an incremental variation on a statistical forecasting scheme for seasonal streamflow. Unlike many papers in this genre, the focus is an approach that is in current operations, as opposed to the more common research setting, and the authors investigate a potential improvement to that approach, the addition of hydrological model based predictors. The results are positive, and reasonably well described, and the authors helpfully attempt to diagnose cases in which skill from the new approach is present. That said, the paper is not overly ambitious, and a separate paper in review at this time (in WRR) explores a different variation on the same scheme, using similar figures and analyses – inviting the question of whether the two could have been combined into a single more comprehensive paper. Nonetheless, the work that has been done constitutes an interesting addition, but I strongly suggest that the authors justify some of their methodological choices with more analysis, and do a more to shed light on the findings. The suggestions detailed below should not be particularly onerous, and I also leave it to the authors to be creative in fully investigating/analyzing the predictor options in front of them.

We thank the referee for their positive comments on the paper. The focus of this paper is improving the representation of initial catchment conditions in the statistical forecasting model, while the focus of the paper submitted to WRR is on extracting the most information on future climate conditions from a range of climate indices. The approach taken in each paper is different, but the results (ie improvements in skill etc) are presented in a similar manner to this paper because they are effective communication methods. We strongly believe that each paper should have a clear hypothesis that it is testing and therefore have submitted the two papers separately.

Specific Comments

8703 ln 12: actually snow water equivalent rather than depth or extent is typically used

Comment noted and snow water equivalent referred to explicitly in the text

ln 16: antecedent conditions …– true in non-snowy locations, certainly

I think that this is also true in snowy locations, where antecedent streamflows and rainfall totals would be very crude indicators of the snow conditions.

ln 22: 'more refined' – perhaps be more specific as to the predictors of interest?

Clarified that more refined predictors better represent dynamic catchment processes

Ln 24 – this paragraph is a good discussion

Thankyou

8704 ln 9: 'implemented' is more accurate than investigated here – or both – could cite traditional ESP references, given that the US NWS has employed ESP operationally

Sentence now reads "... investigate and adopted ..."

ln 21: might also see the papers of Lifeng Luo since 2005

ln 24: Wood & Schaake, 2008 is also addresses this issue

References added.

8705, ln 15: here might be a good place to introduce a brief nomenclature by which to refer to the two alternatives that are being evaluated, since they are referenced over and over in the paper. E.g., the 'RW12' predictors versus 'model+Qlag1'.

The nomenclature used in the paper has been carefully considered to reduce the chance of confusion about the ideas being expressed. While we appreciate that on occasion the adopted nomenclature can appear cumbersome, the actual number of times any given phrase is used is rather small. Therefore, we believe that an abbreviated nomenclature will not add considerably to the readability of the paper.

8706, ln 4: Even though the model reference no doubt includes illustrations of model performance, it would be useful to include a few time series from the model calibration and validation (possibly also scatter plots, in a multi-part figure) to give the reader a feel for the quality of the model.

Presentation of model calibration and validation results as applied in this paper is not as simple as the reviewer suggests because cross validation simulations are used throughout. As the reviewer suggests the paper which describes the WAPABA model (Wang et al., 2011) contains comprehensive assessment of the performance of the model on a larger range of catchment than those used in this study. We do summarise the major findings on the performance of the model reported by (Wang et al., 2011).

Ln 11 paragraph: A major suggestion. The authors make a curious choice here – rather than use soil moisture values as indicators of catchment wetness, they use model forecasts of streamflow during the forecast period – surely they are almost linearly related, and the SM values would require less work to generate. I'd like to see a plot of SM vs the predictor flow to see what is gained by doing the simulation. In addition, the authors add a lag 1 observed flow as a predictor. Surely this also is related, not only to SM but to the streamflow used in the RW12 predictor set. Could the authors show the cross-correlations (perhaps with a family of two-parameter scatter plots) between all the catchment wetness variable used in the paper? I would guess that there is a large overlap of the signal they provide to the forecasts, and it would be nice to understand this predictor choice process better, given that the major objective of the paper is to contrast the effects of the two predictor sets. The authors should in any case describe why model SM was not considered as a direct predictor.

A section, including two figures (Figures 13 and 14), has been added in the discussion to address this point.

Ln 11: the result that the mean of an ensemble prediction is the same as the prediction based on the ensemble mean (eg climatology) is quite interesting and would surprise many in the field. Could the authors add a plot illustrating the analysis of the variations here (mean, median, ensemble median.

This could be due to the monthly timestep of the model and the lack of snow, which could make the rainfall runoff response more linear.

The simulations from the WABAPA model represent an integration of the conditions of the state variables. No matter what forcing data are used to run the model to produce the simulations, the signal in the simulations is solely due to the initial conditions of the state variables. Therefore, while the magnitude of the simulations produced using different forcing data may differ, the underlying relationship between the simulations and streamflows during the forecast period should remain similar.

8707 ln 16: Could the authors add a high-level schematic to illustrate the predictor choices in the two alternatives being considered. I think a reader should be able to understand 80% of the paper just by scanning the figures and tables – the schematic would help.

We have carefully considered the reviewer's suggestion and believe that the addition of a schematic diagram is unlikely to add value to the manuscript. We believe that a simple high-level schematic would repeat material that is contained in other figures, e.g. Figure 2 (now Figure 3) which clearly identifies that two sets of predictors are being compared.

8708 ln ~1: One main weakness to the paper is that it only really tries one variation on the basic scheme, and the addition of lag1 flows to alternative 2 probably reintroduces some of the signal from antecedent flow in alt. 1. Fig 9 does show it to add skill over WAPABA alone, but I would suggest that any additional analysis on the individual contributions of the catchment wetness predictors, or perhaps in different combinations, would strengthen this paper and yield more insight on possible choices. I leave it to the authors to consider what may be the best additional analyses to show. I also wonder, for alt. 1, which predictor is chosen (rainfall or flow) when the 'best predictor' is applied. Seeing a timeseries of the predictors (rainfall, ant. flow and ant. precip or lag 1 precip, together with the predictand, all normalized) might give more insight to the reader as to their covariation. Obviously this could not be shown for all sites, but perhaps for a few.

A new paragraph has been added to the introduction to describe one of the key motivations of this study was to reduce artificial skill introduced by predictor selection. Investigating alternative predictors or combinations of predictors and their relative contributions to forecast skill essentially becomes a predictor selection exercise, which will introduce artificial skill. We therefore have adopted the approach to fix the predictors for alternative 2, assess the skill of forecasts made using fixed predictors, and investigate the forecast skill changes relative to the existing methods used in the operational systems.

Presenting the predictors used for 'alt 1' is not straightforward. The predictors used for 'alt 1' are cross-validated and therefore it is possible that for a given location and forecast season, the predictors used to produce forecasts vary between years. This occurs because several candidates have very similar performance and omitting 5 years from the performance assessment can influence which predictor is best. Figures summarising of the 'best predictor' representing initial catchment conditions are provided as supplementary material.

Ln 5: the first part of this paragraph is well known – could be more concise – as with the leave 1+X out part. The concern about autocorrelation could be tested, of course, just by calculating it. I would guess it's not a huge concern in reality.

In Australian conditions there are multiple year episodes of wet and dry conditions that have long lasting effects, for example connections/disconnections between surface and groundwater systems. Simple autocorrelation tests may not necessarily diagnose independence between calibration and verification periods. The effect of low frequency signals relating to multiple year episodes of wet and dry conditions will potentially be overwhelmed in autocorrelation tests by higher frequency inter and intra annual variations.

8709 ln 23: define 'event' if it hasn't been already. Also, isn't this notion 'all events contribute' obvious? Ie, that's typically the way scores are calculated.

We have removed the word 'event' from the manuscript and rephrased this section to improve clarity.

8710 ln 2: note that aside from PIT, CRPS has a reliability component in its decomposition that can be extracted.

Yes the CRPS can be decomposed to assess the forecast reliability, but for the purposes of this paper the PIT provides an adequate tool to demonstrate the point that there is no significant change in reliability.

8711 ln 10: could you analyze these hypotheses about predictor value behavior by plotting the predictor values? One should be able to see 'wetting up' and perhaps plot that against 'increase in skill' to see if this diagnosis is true.

At this point we are trying to provide a general overview and interpretation of the results relating to physical process that are occurring in the diverse range of catchments considered rather than a detailed analyses and hypothesis testing. A figure (Figure 2) illustrating the seasonal cycle of streamflows for all catchments has been included to support this interpretation.

8715 ln 21: can you show this linear relationship in a plot? For a number of the catchments, perhaps?

A figure demonstrating this is included for the catchment and season for which results are presented.

8718 ln 10: I would guess that the difference in skill between prior period simulations and forecast period flow simulations is simply due to lags in runoff generation – ie rainfall prior to the forecast doesn't raise flow in the prior period (always) but may in the forecast period. It should raise SM, however.

The point being made in this paragraph relates to knowledge of climate during the forecast period, rather than the state of the catchment at the forecast time as the reviewer appears to be implying. The sentence referred to now explicitly mentions using forecasts from dynamic climate models to inform seasonal streamflow forecasts to make this clearer.

Ln 20: any statistical forecast approach shouln't be much affected by bias – biases are inherently corrected in the forecast procedure, which often normalizes predictors. Co-variation of predictor and predictand, regardless of bias, is the determining characteristic of predictor quality.

We agree. The point we are making is that it doesn't matter if the simulations are biased.

Fig 6-8: I don’t particularly love these figures – in part because the figure space is not well used by the data (ie half of some plots is blank), but also because the biases shown would be more evident if obs were plotted against fcst, rather than as an additional symbol overlain onto a plot of forecast spread vs forecast median). I suggest that the authors experiment and see if they can more simply convey the relationship between forecast and obs – ie either fcst as f(obs) or forecasts (in horizontal mode) on the x-axis and obs on the y, together with a 1:1 line.

We have investigated several methods to present this information. No matter how these data are plotted the primary criticism, that the space not being well used, is true. We prefer to plot the using the approach presented in the manuscript because the forecast is produced before the observation is available. The key point of these figures and the associated discussion in the text is to assess how the forecasts are different using the alternative sets of predictors and what streamflow is observed given the forecast is made. Therefore, we believe it is most appropriate to plot the forecasts and observations using the approach adopted.

Fig 3: It’s nice to see an improvement, but this again brings us back to the paper’s main weakness, that it presents an incremental variation that yields a small improvement. To their credit the authors attempt to understand the reason for the increase, but given the narrow focus, the paper might do well to explore and more fully report on some variations – ie, catchment predictors with lag1 flow only, ant. rainfall only, ant. flow only, SM, a best predictor approach applied to all of them, and so on, together with some cross-correlation plots to show how much unique info there is in the candidate wetness predictors In that case this figure might have several more bars. If the authors pull a predictor table into R and use the default plot function, I think it will generate just such a family of scatter plots.

This comment is a reiteration of previous comment, which we have addressed in two ways as highlighted earlier.

Firstly we have clarified in the introduction that one of the motivations for this study was to reduce artificial skill in the forecasts. Using a fixed set of predictors achieves this goal and the focus of the paper is on the magnitude of skill improvements and why the skill may increase.

Secondly, we have added a discussion on the benefits of using the simulated streamflow as the predictor rather than the condition of model state variables at the forecast time.

Reference

Wang, Q. J., Pagano, T. C., Zhou, S. L., Hapuarachchi, H. A. P., Zhang, L., and Robertson, D. E.: Monthly versus daily water balance models in simulating monthly runoff, Journal of Hydrology, 404, 166-175, 2011.