Review of: Improving Statistical Forecasts of seasonal streamflows using hydrological model output, by Robertson, Pokhrel & Wang,
Recommendation: Accept pending major revisions

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.

Specific Comments

8704 ln 12: actually snow water equivalent rather than depth or extent is typically used
ln 16: antecedent conditions …– true in non-snowy locations, certainly
ln 22: 'more refined' – perhaps be more specific as to the predictors of interest?
Ln 24 – this paragraph is a good discussion
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
ln 21: might also see the papers of Lifeng Luo since 2005
ln 24: Wood & Schaake, 2008 is also addresses this issue
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'.
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.
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.
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.
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.
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.
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.
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.
8710 ln 2: note that aside from PIT, CRPS has a reliability component in its decomposition that can be extracted.
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.
8715 ln 21: can you show this linear relationship in a plot? For a number of the catchments, perhaps?
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.
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.
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.

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.