Replies to Referee #1

"An intercomparison of approaches for improving predictability in operational seasonal streamflow forecasting"

Pablo A. Mendoza, Andrew W. Wood, Elizabeth Clark, Eric Rothwell, Martyn P. Clark, Bart Nijssen, Levi D. Brekke, and Jeffrey R. Arnold

We thank this reviewer for his time in commenting on our paper. We provide responses to each individual point below. For clarity, comments are given in italics, and our responses are given in plain text.

This is an interesting and well written comprehensive evaluation of over a dozen statistical, dynamical and hybrid seasonal streamflow forecasting techniques. The evaluation is done for about 20 years of 5 reservoirs in mostly snow-dominated climates of the Pacific Northwest US. My suggestions for changes are minor at best, with detailed comments below.

We are very pleased that this reviewer appreciates the contributions of this study.

Title: I don't think "predictability" is the right word for the title. It implies something that's immutable and intrinsic, in the sense of theoretical maximum predictability, which is not something that could be "improved". Predictive skill of certain techniques or a forecasting enterprise can be improved, however.

The reviewer makes a good point. To avoid confusion on the concept of "predictability", we have modified the title to "An intercomparison of approaches for improving operational seasonal streamflow forecasts".

Line 57 "current operational practice in the US still takes little to no advantage of largescale climate information for realtime seasonal streamflow forecasting" and later line 64-65 "these [operational] approaches rely solely on the predictability of [initial hydrologic conditions] and do not leverage any type of large-scale current or future climate information". From my experience as a forecaster, there were only very limited locations and leadtimes where the climate information provided substantial benefits. Things like El Nino indices were used in pacific northwest and southwest US for early (i.e. January) and pre-season (i.e. October-December) forecasts. I think it's strong to say that there was no use at all of climate information.

We have modified the text to reflect the reviewer's experience on this topic, though in truth the vast majority of statistical forecasting locations in the western US do not use climate indices to our knowledge – even in the Pacific Northwest (PNW). The paragraph now reads:

"Despite generally promising findings from this body of work and from a number of agency development efforts (Weber et al. 2012; Demargne et al. 2014), the use of large-scale climate information for real-time seasonal streamflow forecasting in the US remains rare. In the western United States, where snowmelt commonly dominates the annual cycle of runoff, official WSFs are produced via two main approaches: (i) statistical models leveraging in situ watershed moisture measurements such as snow water equivalent (SWE), accumulated precipitation and streamflow (Garen 1992; Pagano et al. 2004); and (ii) outputs from the National Weather Service (NWS) Ensemble Streamflow Prediction method (ESP; Day 1985), which is based on watershed modeling. For the overwhelming majority of forecast locations, these approaches rely solely on the predictability from IHCs (measured or modelled). A small number of locations can be found, however, where climate indices also serve as predictors in the statistical framework, and the NWS has recently implemented techniques through which climate model forecasts may eventually be applied to ESP (Demargne et al. 2014)."

Line 89 Following the list of statistical water supply forecasting techniques. It may be useful to include in that list http://onlinelibrary.wiley.com/doi/10.1111/j.1752- 1688.2009.00321.x/abstract because it also includes z-score regression and describes operational products.

Certainly! This was an oversight as we are aware of that work, thus we have included the aforementioned reference, following the reviewer´s suggestion.

Line 172 The universal use of the log transform on all the predictands. Operationally, forecasters use linear, square root, cube root and log transform statistical models, with log being the most extreme. The use of log everywhere wouldn't have been my first choice, and is probably responsible for "forecast blowouts" like 1993 in the Apr-1 / e panel on figure 11 (far lower right corner, only the lower whisker is visible on the chart). But since it's applied the same everywhere, it means that the intercomparison is valid in a relative rather than absolute sense. You might reassure the reader that you tried other transforms and the results were insensitive.

We regret that we did not try other transformations as we were focused on relative outcomes, though this would have been a reasonable thing to do. In truth, we did place a great deal of importance on the transformation when the work was done, though since then our interactions with CSIRO has opened our eyes to the variation in the effectiveness of different transformations (including, for instance, the log-sinh). We do not have the bandwidth to go back and explore this issue, but for now we will highlight it for the readers based of the text of the comment. Hence, we have added the following sentences:

"In practice, forecasters use a variety of transforms such as linear, square root, cube root, log and log-sinh (Wang et al. 2012). We did not explore alternative transforms, using the log consistently throughout, but recognize that the choice of transform can affect the quality of the forecast."

Line 261 The use of stepwise approach to model building. I think what you're describing here is the case where El Nino is predicting fall precipitation, and by the time January 1 comes around the precipitation is "in the bank" and so continued use of El Nino as a seasonal streamflow predictor after January 1 is redundant, if the equation also includes IHC variables. This was a common operational challenge in the US and a frustration to forecasters.

The reviewer correctly identifies the motivation for the technique, in the sense that HESP intends to handle seasonally varying sources of predictability separately, applying the climate predictors only to the portion of the flow variation that has not already been explained by the IHCs, if possible. If the signal from watershed moisture conditions becomes strong and is redundant, $\varepsilon_{climate} = Q - f(IHC)$ (i.e., the residual from that relationship) cannot be explained robustly by climate information and HESP just defaults to Stat-IHC.

Line 379 I think you find that El Nino provides a small amount of predictability in October-December and by 1 January comes, initial hydrologic conditions are comparable to El Nino skill, but then by 1 February and later, IHC are heavily dominant. This is consistent with http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.177.3158&rep=rep1&type=pdf https://scholar.google.com.au/citations?view_op=view_citation&hl=en&user=5hdY14AAAAAJ&ci

tation_for_view=5hdY14AAAAAJ:YsMSGLbcyi4C For many years 1 February was the start of the operational forecasting season and so there is little surprise that hydrologists were underwhelmed with what El Nino had to offer them. It wasn't until leadtimes were pushed back to 1 January, and then back to 1 October that hydrologists became more operationally interested.

Indeed, our skill plots (Figures 5 and 8) align with the findings by Pagano and Garen (2006) and other researchers as to this point, specifically with the progression of seasonal streamflow forecast skill provided in their Figure 1. We thank the reviewer for this observation about the original initialization dates in February, which is encouraging if it indicates a trend toward even earlier start times where the climate information is relatively more important. We add the following sentence:

"This progression of relative predictabilities from climate and watershed moisture conditions (Figures 5 and 8) is consistent with previous findings for the region (e.g., Pagano and Garen 2006)."

Line 410 On explanations of why ESP is under-dispersive- The common way of explaining this is that NWS-style ESP does not consider parameter, data or model uncertainty, only uncertainty of future forcings.

This is a good point. Indeed, ESPs are particularly under-dispersive at late forecast initializations, when uncertainty in IHCs dominates the total streamflow forecast uncertainty (Wood and Schaake 2008). We thank the reviewer for this observation, and have modified the text accordingly:

"For such lead times, the uncertainty in ESP streamflow forecasts is underestimated due to reliance on a single modeled IHC that does not account for modeling errors (Wood and Schaake 2008), such that forecast spread derives only from uncertainty represented by the ensemble of future forcings."

Line 477 Generating custom climate indices beyond El Nino, creates useful information. I feel like this contradicts the statements on lines 383-385 where you say that this technique was the worst performer.

The reviewer refers to a comparison between the three hybrid regression techniques (Stat-Ind-IHC, Stat-CFSR-IHC, and HESP) in terms of probabilistic skill. We did find that for some basins (e.g., Dworshak and Hungry Horse) Stat-CFSR provides higher skill than using custom climate indices (Stat-Ind), outperforming also benchmark techniques at early initializations. Nevertheless, our results also show that when custom-indices are used in combination with stronger predictors, attempting to explain smaller amounts of variance, they are not as robust as using standard climate indices. We add a sentence to help explain this context:

"When used in combination with other, stronger predictors, the parameter estimation cost of the CFSR-PLSR relative to an off-the-shelf index may be more exposed (leading to greater shrinkage of skill after cross-validation)."

References

Day, G. N., 1985: Extended Streamflow Forecasting Using NWSRFS. *J. Water Resour. Plan. Manag.*, **111**, 157–170, doi:10.1061/(ASCE)0733-9496(1985)111:2(157).

Demargne, J., and Coauthors, 2014: The science of NOAA's operational hydrologic ensemble forecast service. *Bull. Am. Meteorol. Soc.*, **95**, 79–98, doi:10.1175/BAMS-D-12-00081.1.

Garen, D. C., 1992: Improved Techniques in Regression-Based Streamflow Volume Forecasting. *J. Water Resour. Plan. Manag.*, **118**, 654–670, doi:10.1061/(ASCE)0733-9496(1992)118:6(654).

Pagano, T., D. Garen, and S. Sorooshian, 2004: Evaluation of Official Western U.S. Seasonal Water Supply Outlooks, 1922–2002. *J. Hydrometeorol.*, **5**, 896–909, doi:10.1175/1525-7541(2004)005<0896:EOOWUS>2.0.CO;2.

Pagano, T. C., and D. C. Garen, 2006: Integration of Climate Information and Forecasts into Western US Water Supply Forecasts. *Clim. Var. Clim. Chang. water Resour. Eng.*, 86–103.

Wang, Q. J., D. L. Shrestha, D. E. Robertson, and P. Pokhrel, 2012: A log-sinh transformation for data normalization and variance stabilization. *Water Resour. Res.*, **48**, 1–7, doi:10.1029/2011WR010973.

Weber, F., D. Garen, and A. Gobena, 2012: Invited commentary: themes and issues from the workshop "Operational river flow and water supply forecasting." *Can. Water Resour. J.*, **37**, 151–161, doi:10.4296/cwrj2012-953.

Wood, A. W., and J. C. Schaake, 2008: Correcting Errors in Streamflow Forecast Ensemble Mean and Spread. *J. Hydrometeorol.*, **9**, 132–148, doi:10.1175/2007JHM862.1.