Response to interactive comment by Anonymous Referee #1

In this study performance of streamflow forecasts for Kharif Season (April-September) in the Upper Indus Basin of Pakistan is assessed. Streamflow forecasts are generated using the Bayesian joint probability (BJP) approach. Several predictors such as antecedent flow, climate indicators, and ESP based streamflow forecasts are used to test the performance of the streamflow forecasts. The study finds that in general BJP streamflow forecasts based on predictors antecedent flow and climate indicators perform the best. Variation in the skill is found for the focus basins, and for the early and late part of the season. In general, the manuscript is well organized and methods are technically sound. I do have a few comments/suggestions, some of which are moderate to major, which need to be addressed before publication.

Response: Thank you for your encouraging and helpful review.

Major comments:

(1) It would be helpful, mostly for the readers who are not well aware of the seasonal cycle of climate in the region, to add a figure for both basins that show the seasonal cycle of precipitation, temperature, runoff/streamflow. Similar to Fig. 2 of this manuscript http://journals.ametsoc.org/doi/full/10.1175/JHM-D-14-0213.1. Such a figure would provide a needed background to the readers about the region and also help interpret the results of streamflow forecasts evaluation.

Response: We agree and have added Figure 2 to show the seasonal cycle of streamflow, precipitation and calculated potential evapotranspiration for our two study basins.

(2) The authors mention lack of climate forecasts skill in this region. I would encourage them to show a map(s) of the long-term skill of at least rainfall (winter) and temperature (winter and summer) in the region. I think a case for using statistical forecasts such as ones presented in this study can be made better if statistical forecast skill is demonstrated relative to the skill of dynamical forecasts, not just climatological forecasts. As of now, there are several global dynamical forecasts systems that provide operational seasonal forecasts. One of them being the North American Multimodel Ensemble (NMME, http://www.cpc.ncep.noaa.gov/products/NMME/).

Response: We agree with the comment that it can be useful to assess statistical forecast skill together with dynamical forecast skill. The reason that this has not been done in this study has do with the strong practical application focus of the tool development. Because of the need to model glacier and snow processes, sophisticated hydrological models that could ingest climate forecasts are of limited availability. Furthermore, the region does not have ready access to real-time seasonal climate forecasts, although this problem can be overcome. The purpose of this study is therefore to develop tools that can be easily implemented based on infrastructure that exists today. As infrastructure (like hydrological models) improves, an extended study on a fully dynamical forecasting system will be highly appropriate. As dynamical forecasts need to be post-processed for use in hydrological forecasting, the linkage of dynamical climate model forecasts with hydrological models in the region would require substantial additional research.

Additionally, a recently published assessment of forecast skill over Pakistan and Afghanistan for NMME May 1 hindcasts (for May to November) concluded that the MMEM, that generally exceeded the skill of any individual model, provides little benefit over climatology (Cash, B. A., Manganello, J. V., and Kinter, J. L.: Evaluation of NMME temperature and precipitation bias and forecast skill for South Asia, Clim Dyn, 10.1007/s00382-017-3841-4, 2017). Given this assessment, we would not expect NMME forecast precipitation or temperature for our study region to add skill as additional predictors to the BJP. We have thus added the following text to the Introduction:

"*Cash et al. (2017) assessed monthly North American Multi-Model Ensemble (Kirtman et al., 2013) hindcasts initialised May 1 for May to November for South Asia, including the mountainous areas of Afghanistan and Pakistan. They concluded that the multi-model ensemble mean temperature and precipitation forecasts, while generally exceeding the skill of any individual model, provided little benefit over climatology*."

In the Introduction, we have provided more information on the background of this study and clarify the strong practical application focus of this study, including why a fully dynamical forecasting system has not been investigated. We have also added to the discussion, regarding potential further research with dynamical models.

(3) The authors use March streamflow is the only predictors reflecting antecedent conditions, it is not clear why other variables such as snow water equivalent, soil moisture, total water storage were not used. Nowadays observations (through remote sensing) or simulations (e.g. through GLDAS https://ldas.gsfc.nasa.gov/index.php) of those variables are readily available. Especially in a region where snowmelt runoff is dominant, I would think snow and soil moisture would provide some streamflow forecast skill.

Response: We did investigate a MODIS-based snow cover product (post-processed to remove cloud cover effects) and did find a relatively high correlation with streamflow. For example, for the Jhelum catchment the MODIS snow cover at the end of March has a 0.68 correlation with Kharif season streamflow. However, the MODIS data only commenced 2000 and so this limited availability of data resulted in its exclusion from the final set of model evaluations. As March flow correlation with Kharif season streamflow is also 0.68, a snow cover predictor is not expected to be a demonstrably better predictor than March flow in this case.

In response to the Editor's additional recommendations, we undertook analysis of GLDAS-SWE hindcast performance for our two study basins. Firstly there is the issue of record length, as GLDAS-2.0 only covers to 2010 (i.e. it isn't up to date and thus unsuitable for real-time operational forecasting) and GLDAS-2.1 only commences in 2000, i.e. a shorter period than the predictors assessed in the manuscript. Such a short period was also our justification for not using MODIS snow cover as a predictor, which similarly commenced in 2000. Despite this short record length, we have assessed 2000-2015 annual correlation of v2.1 SWEMarch with Kharif (April-September) flow and QMarch for our two study basins.

- For Jhelum at Mangla, SWEMarch and QMarch correlation with Kharif flow are 0.50 and 0.73, respectively. Also, there is a 0.87 correlation between SWEMarch and QMarch, suggesting SWEMarch doesn't provide additional information and hence skill above that provided by QMarch.
- For Indus at Tarbella, SWEMarch and QMarch correlation with Kharif flow are comparable at 0.56 and 0.55, respectively. The correlation between SWEMarch and QMarch is 0.77, suggesting they are not independent predictors.

Therefore, the limitation of short record length, lack of higher correlation with flow than that of the QMarch predictor, and relatively high cross-correlation with QMarch, leads us to conclude that GLDAS-SWE would not provide additional skill as a predictor for the BJP. We have added the following to section 4.1 Skills score to address this:

"MODIS (Hall et al., 2010) snow-cover area and GLDAS-2.1 (Rodell et al., 2004) snow-water equivalent, additional measures of antecedent conditions, were also assessed as potential predictors. A significant limitation is the shorter record lengths for MODIS and GLDAS, as available data for both start in 2000. This is of particular concern for the BJP's leave-one-out cross-validation, as using short records to identify dynamical mechanisms is susceptible to spurious skill. Correlation analysis, cognisant of the short 2000-15 period, found that these snow products have a similar or lower correlation with Kharif flow (QKharif) compared to March flow (QMarch), and are relatively highly correlated with QMarch. Thus the limitation of short record length, lack of higher correlation with flow than that of the QMarch predictor, and relatively high cross-correlation with QMarch, leads us to conclude that they would not be expected to provide additional skill as a predictor for the BJP."

(4) I would also encourage the authors to provide some more details regarding the PIT plots in the method section. To my knowledge PIT is not a typical metric used for forecast evaluation so it would help the readers to get a bit more details on them and also briefly describe what each type of the figures (a through e) highlights regarding the forecast skill.

Response: We agree and have extended the explanation of the use of PIT plots to evaluate probabilistic forecasts as follows (new text in **bold**).

"Reliability refers to the statistical similarity between the forecast probabilities and the relative frequencies of events in the observations, which can be verified using probability integral transforms (PITs). The PIT represents the non-exceedance probability of observed streamflow obtained from the CDF of the ensemble forecast. If the forecast ensemble spread is appropriate and free of bias then observations will be contained within the forecast ensemble spread, with reliable forecasts having PIT values that follow a uniform distribution between 0 and 1 (Laio and Tamea, 2007). ***Thus PIT plots are an efficient diagnostic to visually evaluate whether the forecast probability distributions are too wide or too narrow or are biased (under or over estimating) in their prediction of the observed distribution (Wang and Robertson, 2011). As outlined by (Thyer et al., 2009), PIT plot points falling on the 1:1 line indicate that the predicted distribution is a perfect match to the observed; observed PIT values of 0.0 or 1.0 indicate the corresponding observed data falls outside the predicted range, hence the predictive uncertainty is significantly underestimated; PIT values clustered around the midrange (i.e. a low slope in the 0.4 -0.6 uniform variate range) indicate the predictive uncertainty is overestimated; PIT values clustered around the tails (i.e. a high slope in the 0.4 -0.6 uniform variate range) indicate the predictive uncertainty is underestimated; and if PIT values at the theoretical median are higher than those of the uniform variate the predictions have an underprediction bias, and vice versa if they are lower than the uniform variate then the predictions have an overprediction bias****."*

Minor comments:

(5) P2, L24: Not only P and T but other atmospheric forcings as well.

Response: We have changed the text to remove specific mention of P and T, it now reads: *"Dynamical approaches use hydrological models initialised with observed inputs up to the beginning of the forecast season (to account for antecedent conditions) that can be driven either by historical or modelled climate inputs."*

(6) P2, L29: This statement regarding the skill of dynamical forecast skill should be made more specific, e.g. mention the regions and seasons etc.

Response: We have added a summary of published evaluations of dynamical climate forecast skill for the region, as follows:

"Dynamical (i.e. climate model) forecasts of precipitation and temperature are often not sufficiently skilful in this region. For example, Kim et al. (2012) assessed retrospective seasonal forecasts of the Asian summer monsoon from ECMWF System 4 (Molteni et al., 2011) and NCEP CFSv2 (Saha et al., 2014), finding low skill for precipitation prediction and poor simulations of the Indian summer monsoon circulation. Cash et al. (2017) assessed monthly North American Multi-Model Ensemble (Kirtman et al., 2013) hindcasts initialised May 1 for May to November for South Asia, including the mountainous areas of Afghanistan and Pakistan. They concluded that the multi-model ensemble mean temperature and precipitation forecasts, while generally exceeding the skill of any individual model, provided little benefit over climatology."

(7) P3, L23: Summer streamflow would depend upon winter T too, as winter T would influence snow accumulation. Please revise.

Response: We have revised to include reference to winter T, as follows:

"The predominant source of flow in the UIB is snowmelt, with glacier melt a secondary source, with 80% of flow occurring during the June-September summer period. Interannual flow variability is thus controlled by two processes, snow accumulation as determined by winter precipitation and temperature and meltwater generation as determined by summer temperatures. Hence snowmelt-generated flow is a function of winter precipitation and temperature and also summer temperature, whereas glacier melt is primarily a function of summer temperature, although glacier melt is also influenced by snow cover (Charles, 2016)."

(8) P4, L5-10: These sentences are confusing and hard to understand.

Response: These sentences have been replaced by the following sentence:

"Useful climate indices should relate to the weather prior to the forecast season, providing an indication of snow accumulation, and also to the weather within the forecast season, influencing temperature and hence snow and glacier melt rates."

(9) P5, L21: Please see comment #2.

Response: See response to comment #2 above.

(10) Results in Table 1 and 2: It is not clear if those results are after cross-validation or before? Or are the results presented in Figure 3 onward are cross-validated? I would suggest comparing the cross-validated skill vs the skill calculated using the entire period.

Response: These are cross-validated results, and the Table captions have been updated to reflect this. In the paper we will re-emphasis our case that cross validated results are more representative of the real skill and reduce the chances of overfitting. Hence we do not want to show results without cross-validation, given statistical methods are prone to artificial skill and overfitting.

(11) Section 3.3: Suggest dividing this section into three sub-sections to discuss each of the verification scores separately.

Response: As we rely on cited references for the provision of detailed descriptions of the skill scores, we feel there is not enough stand-alone material for each score to justify three sub-sections.

(12) P8, L2: It is surprising to see that MEI (May-June) from the previous year is a skillful predictor. Could you comment on why that may be? During May-June, ENSO events are in initial development stage and sometimes may change signs in the later part of the year so it is surprising that in this case, you are finding MEI May-June to be a skillful predictor for the streamflow of the following year.

Response: We hypothesised that the MEI (May-June) predictor skill relates to autumn/winter snow accumulation, a lag of 4 months to snow accumulation from October onwards. It is thus not unreasonable that circulation systems bringing moisture into the region during autumn/winter are influenced by the forcing initiated by ENSO processes during the summer. We investigated this further and have added the following text to the discussion:

"*For Mangla, the predictor combination that gave the best Kharif season cross-validated skill scores included an ENSO-based predictor (SSI_{March}) immediately before the season (Table 1), which makes sense intuitively as it represents a climate driver of both the snow accumulation before and precipitation conditions during the Kharif season. In contrast, for Tarbela a much earlier ENSO-based predictor (MEI_{MayJun}, i.e. May-Jun the year before) provides higher skill scores than the equivalent predictor immediately before the season (MEI_{FebMar}) (Table 2). To try to understand the dynamical mechanism by which MEI_{MayJun} is providing skill in forecasting Q_{Kharif}, we compared MEI correlations with GLDAS SWE_{March}, Q_{March} and Q_{Kharif}. Results were inconclusive, and perhaps impeded by the short record lengths given SWE is only available from 2000, as while MEI_{MayJun} has a higher correlation with Q_{Kharif} than MEI_{FebMar} (0.76 versus 0.63, respectively) it has a slightly lower correlation with SWE_{March} (0.48 versus 0.52, respectively). Hence MEI_{MayJun} does not appear to be a long-lead predictor of snow accumulation, and so the differences in skill scores may be due to spurious correlations. Therefore we recommend both this model and the Q_{March} and MEI_{FebMar} model be compared and assessed for future events.*"

and have added the following finding to the conclusions:

"*For Tarbela the Q_{March} and MEI_{MayJun} model gave the best skill, however because we could not determine the dynamical mechanism(s) by which the relatively long lag between MEI_{MayJun} influences snowpack accumulation and flow, we cannot rule out the possibility that the skill is due to spurious correlation. Therefore we recommend both this model and the Q_{March} and MEI_{FebMar} model be compared and assessed for future events and, more generally, that BMA be trialled in future research to combine the skill of multiple BJP models as, for example, undertaken in Australia in (Pokhrel et al., 2013)*"

(13) P8, L14-15. I thought that in some cases March flow was the highest skill predictor and adding any predictor didn't increase the skill so why are you using both March flow and climate predictors here?

Response: For Indus at Tarbela (Table 2), inclusion of a climate predictor in addition to the March flow predictor increased skill in all cases. For Jhelum at Mangla, there were mixed results as adding the climate predictor slightly decreased the skill for the full 1975-2015 period but increased the skill for the 2001-2015 period. As the 2001-2015 period is the comparison period for SRM, we think it is acceptable to use these predictor combinations.

(14) Figure 5 and 6. These figures are used to compare the skill of BJP vs SRM based streamflow forecasts. I think it would be better to combine them both into one figure. Maybe just show SRM forecasts with a different color.

Response: We are concerned that combining multiple forecast evaluations onto single figures would look cluttered in some circumstances, such as Panel (d), hence out preference is to maintain these figures as they are.

(15) Conclusion: The last two bullet points are not really findings. I suggest to separately discuss them after listing the findings. Also please mention here the current state-of-the-practice for generating streamflow forecasts in the region and the value the methods explored in this study will add.
Response: We agree and now present these last two bullet points as a separate paragraph after the list of findings. We have added a description of current methods, including discussion of the value added by our new approach. This modified text is as follows:

"*In future research, BJP forecast models could readily be developed and assessed for other tributaries, e.g. the Chenab and Kabul, subject to availability of flow data. This would allow an overall assessment of UIB flow forecasting for the major contributing basins. The present method used by the Indus River System Authority (IRSA) to forecast UIB Kharif streamflow is based on historical analogues. The IRSA use their database of the previous 60 years of flow to select years where the historic March flows are within 5% of the current March flow and use the corresponding historical Kharif flows (within 5%) for their forecast scenario. The selection of the historical scenario is also informed by forecasts from the Pakistan Meteorological Department, forecasts provided by WAPDA (e.g. SRM forecasts), and present snow conditions in the catchment. The forecasts are continuously revised as the season progresses.*

Sufficiently skilful BJP forecasts could also inform scenario selection, providing for the first time a probabilistic approach to forecasts in contrast to a single forecast as currently used. However probabilistic forecasts (such as a the BJP) can be misinterpreted if they are unfamiliar to the water management professionals using them to inform decisions (Pagano et al., 2002;Ramos et al., 2013;Rayner et al., 2005;Whateley et al., 2015). Hence the successful transfer of BJP forecast tools to operational use within Pakistan would require guidance for building BJP models and generating forecasts, test cases with example results, face to face training, and on-going support."

Response to interactive comment by Anonymous Referee #2

The authors have assessed three different methods - Bayesian Joint Probability (BJP), the Snowmelt Runoff Model (SRM) and a hybrid approach (SRM - Ensemble Streamflow Prediction inflow means as additional predictor in BJP approach) for forecasting seasonal streamflow to the two largest dams in the Upper Indus Basin, Pakistan. The authors concluded that BJP approach is simple and it worked well to provide probabilistic seasonal streamflow forecasts. The topic is relevant for publication in HESS. Overall, the paper is well written. I recommend a moderate revision to the manuscript and the following concerns need to be addressed:

Response: Thank you for your encouraging and helpful review.

Major Concerns:

1. Under the BJP approach, was the conditional multivariate normal distributions fit over the entire season or on monthly basis? How many samples were generated through Monte Carlo Simulations under the BJP approach? Provide details.

Response: We have modified the text to include clarifying information, as follows:

"The cross-validated BJP forecast performance was assessed for 1975-2015 (41 seasons), with the BJP models calibrated on a seasonal basis (i.e. 40 data points) using 1000 MCMC samples for each of the leave-one-out calibrations."

2. Page 7, lines 24-27, the skill of using March flow and/or one climate predictor looks very similar to each other. The authors are recommended to use statistical significance test to compare if the skills are significantly different from each other.

Response: We have added uncertainty estimates to the skill scores Table 1 and Table 2, from bootstrapping, to aid interpretation of how different the forecast models are from one another. This leads to a discussion on model selection, with the addition of the following text to section 4.1 Skill scores:

"Table 1 presents cross-validated BJP forecast skill scores using the trialled combinations of antecedent flow and climate predictors for the Kharif season for Jhelum at Mangla, together with bootstrapped 10th to 90th percentile ranges to assess model uncertainty. These ranges were obtained by resampling 1000 random sequences of years of the same length as the observed record, i.e. with replacement, and calculating skill scores for each sample."

"Given there is large uncertainty in skill scores we do not aim to select a 'best' model. However, as there are many models with positive skill (i.e. better than climatology) then using skilful models is plausible. Ideas on how to do this are discussed in section 5."

and the following addition to section 5 Discussion:

"More generally, the skill score uncertainty ranges presented in Table 1 and Table 2 highlight that no 'best' forecast model can be selected for either basin. Attempting to select a best model would ignore model uncertainty and thus not make best use of forecast skill across the range of models trialled. To address this, probabilistic forecasts from multiple BJP models can be combined using Bayesian Model Averaging to produce combined forecasts with higher skill than that obtainable from any individual model (Wang et al., 2012a). Thus trialling a BMA approach is recommended, although it is beyond the scope of this current work."

3. Given that most of the streamflow at Indus River at Tarbela is snowmelt driven, use of a direct or indirect indicator of snow as one of the predictors, along with the projected summer air temperature can improve the forecasting skill. The authors are encouraged to consider global precipitation (for winter) and air temperature forecasts as predictors, which can represent snow as one of the inputs to the model.

Response: We agree that GCM forecasts of precipitation and temperature could potentially be used as predictors, however in this work we are assuming that the water resources practitioners do not readily have access to GCM seasonal climate forecast data (including hindcasts, needed for model establishment). Hence out approach relies on information regarding temperature and precipitation being captured by our selected climate index predictors (statistically). This could be the subject of future research with dynamical models, so we will mention this in our revised discussion.

4. It is not clear why MEI for May and Jun from previous year enhanced the skill score for Indus at Tarbela? Explain.

Response: We hypothesised that the MEI (May-June) predictor skill relates to autumn/winter snow accumulation, a lag of 4 months to snow accumulation from October onwards. It is thus not unreasonable that circulation systems bringing moisture into the region during autumn/winter are influenced by the forcing initiated by ENSO processes during the summer. We investigated this further and have added the following text to the discussion:

"*For Mangla, the predictor combination that gave the best Kharif season cross-validated skill scores included an ENSO-based predictor (SSI_{March}) immediately before the season (Table 1), which makes sense intuitively as it represents a climate driver of both the snow accumulation before and precipitation conditions during the Kharif season. In contrast, for Tarbela a much earlier ENSO-based predictor (MEI_{MayJun}, i.e. May-Jun the year before) provides higher skill scores than the equivalent predictor immediately before the season (MEI_{FebMar}) (Table 2). To try to understand the dynamical mechanism by which MEI_{MayJun} is providing skill in forecasting Q_{Kharif}, we compared MEI correlations with GLDAS SWE_{March}, Q_{March} and Q_{Kharif}. Results were inconclusive, and perhaps impeded by the short record lengths given SWE is only available from 2000, as while MEI_{MayJun} has a higher correlation with Q_{Kharif} than MEI_{FebMar} (0.76 versus 0.63, respectively) it has a slightly lower correlation with SWE_{March} (0.48 versus 0.52, respectively). Hence MEI_{MayJun} does not appear to be a long-lead predictor of snow accumulation, and so the differences in skill scores may be due to spurious correlations. Therefore we recommend both this model and the Q_{March} and MEI_{FebMar} model be compared and assessed for future events.*"

and have added the following finding to the conclusions:

"*For Tarbela the Q_{March} and MEI_{MayJun} model gave the best skill, however because we could not determine the dynamical mechanism(s) by which the relatively long lag between MEI_{MayJun} influences snowpack accumulation and flow, we cannot rule out the possibility that the skill is due to spurious correlation. Therefore we recommend both this model and the Q_{March} and MEI_{FebMar} model be compared and assessed for future events and, more generally, that BMA be trialled in future research to combine the skill of multiple BJP models as, for example, undertaken in Australia in (Pokhrel et al., 2013).*"

5. Page 7, line 1, how good or better the skill enhancement is if SSCRSP (or SSRMSE) changes from 21 to 24.3 (within moderate skill range in Table 1)? Does it reduce uncertainty? Clarify.

Response: As described in the verification methods section, improvements in CRPS reflect improvement in accuracy and/or sharpness and improvements in RMSE reflect improvements in accuracy of the median only. So an inference can be made through comparative analysis of the various skill metrics. We have added uncertainty estimates to the skill scores Table 1 and Table 2, from bootstrapping, to aid interpretation of how different the forecast models are from one another. This leads to a discussion on model selection, with the addition of the following text to section 4.1 Skill scores:

"*Table 1 presents cross-validated BJP forecast skill scores using the trialled combinations of antecedent flow and climate predictors for the Kharif season for Jhelum at Mangla, together with bootstrapped 10th to 90th percentile ranges to assess model uncertainty. These ranges were obtained by resampling 1000 random sequences of years of the same length as the observed record, i.e. with replacement, and calculating skill scores for each sample*."

"*Given there is large uncertainty in skill scores we do not aim to select a 'best' model. However, as there are many models with positive skill (i.e. better than climatology) then using skilful models is plausible. Ideas on how to do this are discussed in section 5*."

and the following addition to section 5 Discussion:

"*More generally, the skill score uncertainty ranges presented in Table 1 and Table 2 highlight that no 'best' forecast model can be selected for either basin. Attempting to select a best model would ignore model uncertainty and thus not make best use of forecast skill across the range of models trialled. To address this, probabilistic forecasts from multiple BJP models can be combined using Bayesian Model Averaging to produce combined forecasts with higher skill than that obtainable from any individual model (Wang et al., 2012a). Thus trialling a BMA approach is recommended, although it is beyond the scope of this current work*."

6. In Table 3, it will be good to know the correlations that are statistically significant (e.g. at 95% confidence interval) based on the sample size.

Response: We have added statistical significance in the table.

7. Page 10, lines 2-7, the hypotheses listed are not clear. As mentioned by the authors earlier, it is already known the snowmelt plays an important role for Indus River at Tarbela. So it not a hypothesis. Also, the results indicated that adding NAO, when used as a predictor, did not improve forecasting skill.

Response: We have re-worded to avoid the confusion caused by the term "hypothesised". Table 2 shows that the NAO predictor did add some skill, however not as much as the selected ENSO based predictor. The text now reads as follows:

"*These higher correlations in the late Kharif (relative to the early Kharif) for Indus at Tarbela would relate to the correspondingly higher relative skill scores shown in Figure 3 for late Kharif, corresponding to late-season glacier melt processes that are a significant component of the inflow to Tarbela but not Mangla (Mukhopadhyay and Khan, 2015).*

It is also interesting to reflect on the relative performance of the NAO climate predictor, which does not provide any skill for inflow to Mangla (Table 1) but offers comparable skill to several of the ENSO indices trialled for Tarbela (Table 2). This indicates NAO may have some skill with regards to late season glacier melt. Overall, these results concur with investigations showing a stronger relationship between ENSO and precipitation and weaker relationship between NAO and precipitation in recent *decades (Yadav et al., 2009a;Yadav et al., 2009b) resulting in the prevalence of ENSO as the better predictor of winter snowpack magnitude*."

Minor Concerns:

8. Did the models use monthly (or daily) data for the model fitting? If so, it needs to be clearly stated.

Response: The BJP is calibrated to seasonal data (i.e. 41 data points 1975 to 2015). This has been clarified in the text as follows:

"*The cross-validated BJP forecast performance was assessed for 1975-2015 (41 seasons), with the BJP models calibrated on a seasonal basis (i.e. 40 data points) using 1000 MCMC samples for each of the leave-one-out calibrations*."

RC2:9. Page 6, lines 27 – 30, RMSEP needs to be used instead of RMSE. Also RMSEP needs to be defined in the text.

Response: We considered RMSEP but concluded similar relative results are obtained from RMSE, based on extensive experience across many previous studies (co-authors Wang, Schepen and Robertson). Hence we have continued with RMSE.

RC2:10. In figures 3a, 4a, 5a and 6a, what are the bounding lines (is it 95% Confidence Interval)?

Response: As stated in the figure caption, these are Kolmogorov 5% significance bands. We have clarified this in our revised text referring to these figures.

Editor Decision: Reconsider after major revisions (further review by editor and referees)
(03 Mar 2018)
by Andy Wood

Comments to the Author:
The authors' responses to reviewer concerns are appreciated, and the author should proceed in making the proposed corrections, with a few exceptions. In a number of key areas noted below, the authors can do more to strengthen the paper – primarily through (1) following reviewer suggestions further toward assessing additional, likely predictors from publically available data sources; and (2) further investigating and justifying the use of such a long-lag climate index (MEI). While it is well known that such indices provide predictability, such demonstrations have typically been at shorter lags (ie, values immediately prior to the prediction period). The authors must do more to defend the idea (through analysis) that predictability pathways exist, and that shorter lag indices (eg, the MEI itself, but in January of the same year as the prediction) cannot provide better skill. The current result is not well supported by existing literature on index-based hydroclimate prediction. Further notes on these points are given below.

Response: Thank you for providing recommendations to help strengthen the paper. We will undertake the proposed corrections and address the identified exceptions.

Regarding the comment "*The authors must do more to defend the idea (through analysis) that predictability pathways exist, and that shorter lag indices (eg, the MEI itself, but in January of the same year as the prediction) cannot provide better skill*."
In Table 2 of the submitted manuscript, skills scores using the MEI for short-lag February-March (i.e. immediately before the forecast season) and long-lag May-June are both shown, both as individual predictors and in combination with Q_{March}. The skills scores show that the short-lag MEI_{FebMar} provides considerably less skill than the selected long-lag MEI_{MayJun}. We acknowledge this is not sufficiently discussed and so have undertaken further analysis, as recommended, and have added the following to the discussion:

> "*For Mangla, the predictor combination that gave the best Kharif season cross-validated skill scores included an ENSO-based predictor (SSI_{March}) immediately before the season (Table 1), which makes sense intuitively as it represents a climate driver of both the snow accumulation before and precipitation conditions during the Kharif season. In contrast, for Tarbela a much earlier ENSO-based predictor (MEI_{MayJun}, i.e. May-Jun the year before) provides higher skill scores than the equivalent predictor immediately before the season (MEI_{FebMar}) (Table 2). To try to understand the dynamical mechanism by which MEI_{MayJun} is providing skill in forecasting Q_{Kharif}, we compared MEI correlations with GLDAS SWE_{March}, Q_{March} and Q_{Kharif}. Results were inconclusive, and perhaps impeded by the short record lengths given SWE is only available from 2000, as while MEI_{MayJun} has a higher correlation with Q_{Kharif} than MEI_{FebMar} (0.76 versus 0.63, respectively) it has a slightly lower correlation with SWE_{March} (0.48 versus 0.52, respectively). Hence MEI_{MayJun} does not appear to be a long-lead predictor of snow accumulation, and so the differences in skill scores may be due to spurious correlations. Therefore we recommend both this model and the Q_{March} and MEI_{FebMar} model be compared and assessed for future events.*"

Response to Rev1-Q2/Rev2-Q3 – It would be worth assessing the skill of the NMME average at least (which is available operationally) over the region to see whether it can be included in a statistical framework as a predictor – even if it is not used in hydrological modeling. Even though the assessment recognizes limitations in the ability of practitioners to use data such as NMME, it should not be hard for the authors to extract the NMME data and evaluate whether it would be worth developing as an additional predictor.

Response: While we agree that assessing readily available dynamical forecasts of precipitation and temperature over the region could aid in identifying additional predictors, a recently published assessment of forecast skill over Pakistan and Afghanistan for NMME May 1 hindcasts (for May to

November) concluded that the MMEM, that generally exceeded the skill of any individual model, provides little benefit over climatology (Cash et al. 2017) [1].

Given this assessment, we would not expect NMME forecast precipitation or temperature for our study region to add skill as additional predictors to the BJP. We hope, however, that the usage of climate model forecasts can be investigated in the future in a more rigorous way and thus have added to the discussion the following:

"Future research could investigate whether dynamical seasonal forecasts of temperature have skill of relevance to forecasting glacier melt, however as noted above such skill has not been determined to date (e.g. Cash et al., 2017), and is beyond the scope of this assessment given our focus of developing practical and easily implementable forecast tools using readily available inputs."

Response to Rev1-Q3/Rev2-Q3 -- Similarly, it would be worth following the reviewer's suggestion to see whether modeled SWE anomalies from GLDAS (also publicly available) would be a similarly useful predictor. Even if they have the same correlation with streamflow, if it is uncorrelated with antecedent flow, it may add skill. The reasons for not using MODIS are fair.

Response: We have investigated GLDAS SWE over our two study basins. Firstly there is the issue of record length, as GLDAS-2.0 only covers to 2010 (i.e. it isn't up to date and thus unsuitable for real-time operational forecasting) and GLDAS-2.1 only commences in 2000, i.e. a shorter period than the predictors assessed in the manuscript. Such a short period was also our justification for not using MODIS snow cover as a predictor, which similarly commenced in 2000.

Despite this short record length, we have assessed 2000-2015 annual correlation of v2.1 SWE_{March} with Kharif (April-September) flow and Q_{March} for our two study basins.

- For Jhelum at Mangla, SWE_{March} and Q_{March} correlation with Kharif flow are 0.50 and 0.73, respectively. Also, there is a 0.87 correlation between SWE_{March} and Q_{March}, suggesting SWE_{March} doesn't provide additional information and hence skill above that provided by Q_{March}.
- For Indus at Tarbella, SWE_{March} and Q_{March} correlation with Kharif flow are comparable at 0.56 and 0.55, respectively. The correlation between SWE_{March} and Q_{March} is 0.77, suggesting they are not independent predictors.

Therefore, the limitation of short record length, lack of higher correlation with flow than that of the Q_{March} predictor, and relatively high cross-correlation with Q_{March}, leads us to conclude that GLDAS SWE would not provide additional skill as a predictor for the BJP. We've added text to reflect this to section 4.1 Skill scores, as follows:

"*MODIS (Hall et al., 2010) snow-cover area and GLDAS-2.1 (Rodell et al., 2004) snow-water equivalent, additional measures of antecedent conditions, were also assessed as potential predictors. A significant limitation is the shorter record lengths for MODIS and GLDAS, as available data for both start in 2000. This is of particular concern for the BJP's leave-one-out cross-validation, as using short records to identify dynamical mechanisms is susceptible to spurious skill. Correlation analysis, cognisant of the short 2000-15 period, found that these snow products have a similar or lower correlation with Kharif flow (QKharif) compared to March flow (QMarch), and are relatively highly correlated with QMarch. Thus the limitation of short record length, lack of higher correlation with flow than that of the QMarch predictor, and relatively high cross-correlation with QMarch, leads us to conclude that they would not be expected to provide additional skill as a predictor for the BJP.*"

Response to Rev1-Q12/Rev2-Q4 – The significance of MEI as a predictor at a lag of nearly a year to the predictand still needs further justification. The Marriotti text does not adequately characterize the processes and potential predictability at such lags explicitly. Although one possible route toward this predictability could be, as suggested, the MEI prediction of fall winter climate variability, hence winter snow accumulation, presumably this could be more directly captured

[1] Cash, B. A., Manganello, J. V., and Kinter, J. L.: Evaluation of NMME temperature and precipitation bias and forecast skill for South Asia, Clim Dyn, 10.1007/s00382-017-3841-4, 2017.

using the actual variables (ie, winter snow peak, accumulated winter precip) than a long-lag teleconnection. Even the use of MEI at a shorter lag from the predictand would intuitively carry more information. This result suggests a spurious set of correlations without greater analysis of the underlying physical mechanisms. Could the paper establish whether the MEI is skillfully related to the proposed intermediate dynamics though additional analyses? Would using predictors such as winter precipitation anomalies provide similar predictability? Further analysis would result in a broader and more insightful study to benefit potential implementation.

Response:
Please see our first response above, addressing these points and outlining the additional analysis we have undertaken and the resulting discussion we've added to the paper.

~~Assessment of methods for s~~Seasonal streamflow forecasting in the Upper Indus Basin of Pakistan: an assessment of methods

Stephen P. Charles[1], Quan J. Wang[2], Mobin-ud-Din Ahmad[3], Danial Hashmi[4], Andrew Schepen[5] ~~and~~, Geoff Podger[3] and David E. Robertson[6]

[1]CSIRO Land and Water, Floreat, 6014, Australia
[2]The University of Melbourne, Parkville, 3010, Australia
[3]CSIRO Land and Water, Canberra, 2601, Australia
[4]Water and Power Development Authority, Lahore, Pakistan
[5]CSIRO Land and Water, Dutton Park, 4102, Australia
[6]CSIRO Land and Water, Clayton, 3168, Australia

Correspondence to: Stephen P. Charles (Steve.Charles@csiro.au)

Abstract. Timely and skilful seasonal streamflow forecasts are used by water managers in many regions of the world for seasonal water allocation outlooks for irrigators, reservoir operations, environmental flow management, water markets and drought response strategies. In Australia, the Bayesian joint probability (BJP) statistical approach has been deployed by the Australian Bureau of Meteorology to provide seasonal streamflow forecasts across the country since 2010. Here we assess the BJP approach, using antecedent conditions and climate indices as predictors, to produce Kharif season (April-September) streamflow forecasts for inflow to Pakistan's two largest Upper Indus Basin (UIB) water supply dams, Tarbela (on the Indus) and Mangla (on the Jhelum). For Mangla, we compare these BJP forecasts to (i) ensemble streamflow predictions (ESP) from the snowmelt runoff model (SRM) and (ii) a hybrid approach using the BJP with SRM-ESP forecast means as an additional predictor. For Tarbela, we only assess BJP forecasts using antecedent and climate predictors as we did not have access to SRM for this location. Cross validation of the streamflow forecasts show that the BJP approach using two predictors (March flow and an ENSO climate index) provides skilful probabilistic forecasts that are reliable in uncertainty spread for both Mangla and Tarbela. For Mangla, the SRM approach leads to forecasts that exhibit some bias and are unreliable in uncertainty spread, and the hybrid approach does not result in better forecast skill. Skill levels for Kharif (April-September), early Kharif (April-June) and late Kharif (July-September) BJP forecasts vary between the two locations. Forecasts for Mangla show high skill for early Kharif and moderate skill for all Kharif and late Kharif, whereas forecasts for Tarbela also show moderate skill for all Kharif and late Kharif, but low skill for early Kharif. The BJP approach is simple to apply, with small input data requirements and automated calibration and forecast generation. It offers a tool for rapid deployment at many locations across the UIB to provide probabilistic seasonal streamflow forecasts that can inform Pakistan's basin water management.

1 Introduction

The Asian Development Bank rates water security in Pakistan as ‘hazardous’ (the lowest of five classes), ranking it 46th out of 48 countries in the Asia-Pacific region, with only Kiribati and Afghanistan ranked lower (Asian Development Bank, 2016). Other studies confirm Pakistan’s relatively high levels of exploitation of river flows and groundwater, associated water stress and resultant exposure to climate change (Döll et al., 2009;Wada et al., 2011;Schlosser et al., 2014;Kirby et al., 2017). Given the high demands on the main water source, the Indus River, its year to year flow variability has a significant impact on security of supply in the Indus Basin Irrigation System (IBIS) of Pakistan. Better management outcomes could be achieved if a reliable understanding of Kharif (summer, April-September) water availability at the beginning of the season were available. This would improve IBIS water allocation planning, a critical need given the highly seasonal flows (~80% annual flow occurs in the Kharif season), limited storage capacity (~~30 days of supply~~10% of inflows) and increasing water demand for agriculture and energy production. Thus we assess methods for providing seasonal streamflow forecasts for the two largest water supply dams, Tarbela (on the Indus) and Mangla (on the Jhelum), in the Upper Indus Basin (UIB) of Pakistan.

Seasonal streamflow forecasts can be a valuable source of information for water resource managers (Chiew et al., 2003;Anghileri et al., 2016), with both statistical and dynamical forecasting approaches developed and implemented internationally (Yuan et al., 2015). Sources of seasonal streamflow predictability come from initial hydrological or antecedent conditions (e.g. water held in storage in a catchment, in the soil, as ground water, in surface stores, or as snow/ice) and also from the skill of seasonal climate forecasts (Bennett et al., 2016;Doblas-Reyes et al., 2013;Li et al., 2009;Shukla and Lettenmaier, 2011;Shukla et al., 2013;van Dijk et al., 2013;Koster et al., 2010;Wood et al., 2015;Yossef et al., 2013). Statistical approaches relate antecedent catchment conditions and/or climate indices to streamflow using techniques such as multiple linear regression (Maurer and Lettenmaier, 2003). Statistical approaches require predictor-predictand records of sufficient length to determine robust relationships, stationarity in the relationships, and rigorous cross-validation to avoid over-fitting or an inflated skill assessment (Robertson and Wang, 2012;Schepen et al., 2012). Dynamical approaches use hydrological models initialised with observed inputs up to the beginning of the forecast season (to account for antecedent conditions) that can be driven either by historical or ~~climate~~ modelled ~~precipitation and temperature~~climate ~~forecasts~~ inputs (Yuan et al., 2015;Zheng et al., 2013). For example, in Ensemble Streamflow Prediction (ESP), hydrological models are driven by each historical season’s precipitation and temperature series to produce an ensemble of flow forecasts, with this ensemble providing a distribution of plausible flows for the forecast period (Wood and Lettenmaier, 2008). ESP forecasts can also be used as an input predictor to statistical techniques (Robertson et al., 2013). Dynamical (i.e. climate model) forecasts of precipitation and temperature are often not sufficiently skilful in this region. For example, Kim et al. (2012) assessed retrospective seasonal forecasts of the Asian summer monsoon from ECMWF System 4 (Molteni et al., 2011) and NCEP CFSv2 (Saha et al., 2014), finding low skill for precipitation prediction and poor simulations of the Indian summer monsoon circulation. Cash et al. (2017) assessed monthly North American Multi-Model Ensemble (Kirtman et al., 2013) hindcasts initialised May 1 for May to November for South Asia, including the mountainous areas of Afghanistan and Pakistan. They concluded that the multi-model ensemble mean temperature and precipitation forecasts, while generally exceeding the skill of any individual model, provided little benefit over climatology. Given these findings, we have not investigated the use of dynamical forecasts in this assessment. Alternatively, statistically-based forecast methods using robust relationships between climate drivers, antecedent catchment conditions and resultant streamflow can be valuable research and management tools when properly implemented (Plummer et al., 2009;Schepen et al., 2016). Thus in this assessment, we assess three statistically-based forecast options for their practical feasibility in developing seasonal streamflow forecasting models for the study region:

1. A statistical approach using the Bayesian joint probability (BJP) model with predictors accounting for antecedent basin conditions and climate drivers;
2. An ESP approach using the snowmelt runoff model (SRM); and,
3. A hybrid approach using option (1) with an additional predictor – the mean ESP forecasts from (2).

The study is reported as follows: Section 2 outlines the study area, details of the case studies and data used, and climate influences; Section 3 presents the BJP statistical approach, the SRM-ESP approach, and the verification metrics used to assess forecast skill, bias, reliability and robustness; Section 4 presents the results of the BJP and SRM skill scores and performance diagnostics. Section 5 discusses the performance of the forecast approaches, and Section 6 concludes with the main findings and recommendations.

2 Case Study and Data

2.1 Upper Indus Basin

Pakistan's water supply, crucial for its extensive irrigated agriculture industry, hydropower generation, and industrial and municipal water supply, is predominantly sourced from Indus river flow, with groundwater a secondary although important contributor to most demands (with the exception of hydropower). The glaciated and snow covered sub-basins of the UIB, encompassing glaciated headwater catchments within the northern Hindu-Kush, Karakoram and western Himalayan mountain ranges, dominate water generation within the Indus Basin (Alford et al., 2014). The UIB's tributaries include the Indus at Kharmong: Shigar, Shyok and Astore in the Karakoram Himalaya, the Jhelum, Chenab, Ravi and Sutlej in the western Himalaya, and the Hunza, Gilgit, Kabul, Swat and Chitral in the Hindu Kush mountains (Figure 1). These basins can be classified as having a flow regime that is either glacier-melt dominated (Hunza, Shigar and Shyok) or snow-melt dominated (Jhelum, Kabul, Gilgit, Astore and Swat) (Hasson et al., 2014). The predominant source of flow in the UIB is snowmelt, ~~followed by~~with glacier melt ~~as~~ a secondary source, with 80% of flow occurring during the June-September summer period. Interannual flow variability is thus controlled by two processes, snow accumulation as determined by ~~the amount of~~ winter precipitation and temperature ~~(snowfall) and~~and ~~the~~ meltwater generation as determined by summer temperatures. ~~Whilst~~ Hence snowmelt-generated flow is a function of ~~both~~ winter precipitation and temperature ~~and~~ and also summer temperature~~s~~, whereas glacier melt is primarily a function of summer temperature, although glacier melt is also influenced by snow cover (Charles, 2016).

Inflows to two major reservoirs, Tarbela Dam on the Upper Indus and Mangla Dam on the Jhelum River, a major tributary to the Indus system, are investigated (Figure 1). Daily inflow data from 1975 to 2015 was obtained from Pakistan Water and Power Development Authority (WAPDA). Figure 2 presents the seasonal hydroclimatic cycle for these two basins, showing double-peaked (winter and summer) precipitation with inflow peaking in May for Mangla and July for Tarbela. The Tarbela Dam on the main stem of the Indus is one of the largest individual storage in the UIB, crucial for hydropower generation and irrigation supply (Ahmed and Sanchez, 2011). Annual inflows to Tarbela constitute 70% melt water, of which snowmelt contributes 44% and glacial melts contribute 26% (Mukhopadhyay and Khan, 2015). The Mangla Dam on the Jhelum River is (since enlargement) a similar size storage to Tarbela and one of the most important resources in Pakistan for electricity generation and water supply for irrigation (Mahmood et al., 2015). For the Jhelum, the area upstream of Mangla is reported as 33,500 km^2 with an elevation range from 300 m to 6285 m and mean of nearly 2,400 m, the relatively low altitude ensures that there is only 0.7 % coverage by glaciers or perennial snow according to GLIMS glacier database as cited by Bogacki and Ismail (2016). In contrast, the Indus upstream of Tarbela is over five times larger (173,345 km^2) with higher elevation (to 8,238 m, as reported by Immerzeel et al. (2009)) and 11.5% covered by perennial glaciers (Ismail and Bogacki, 2018), such that (as noted above and evident in Figure 2, showing inflow exceeding precipitation) glacier ice melt is a significant contribution to annual flow.

2.2 Climate influences

~~Climate~~ Useful climate indices should ~~provide information on the state of the atmosphere~~relate to the weather ~~preceding~~ prior to the ~~the forecast season and, potentially, the weather conditions to be experienced during the~~ forecast season ~~ahead. The state of the atmosphere relates to the weather experienced by the study region prior to the forecast season and hence the precipitation received and the resultant~~, providing an indication of ~~-~~snow accumulation, and~~pack magnitude. As the state of the atmosphere also relates to~~ also to the ~~trajectory of~~ weather ~~that may be experienced during~~within the forecast season, ~~it additionally may relate to~~influencing temperature ~~in the forecast season and thus~~and hence snow~~melt~~ and glacier melt ~~extent~~rates. A literature review identified ~~that~~ indices related to the North Atlantic Oscillation (NAO) and El Niño Southern Oscillation (ENSO) ~~are~~ as the most likely to provide skill for the UIB (Charles, 2016). These both influence the direction of prevailing winds bringing moisture into the region and thus determine ~~the magnitude of~~ precipitation and temperature conditions ~~and resultant~~influencing the depth and areal extent of ~~snowpack~~ snow accumulation ~~created~~ in the winter and early spring preceding the Kharif (April-September) high-flow season.

The NAO is a measure of the strength of the pressure gradient between the subtropics and polar regions in the north Atlantic, representing a dominant source of variability in circulation and winds influencing the region (Hurrell, 1995;Bierkens and van Beek, 2009). It has a direct influence on the interannual variability of the westerly winds (westerly disturbances) and their water content traversing Europe, the Mediterranean and the Middle East region into the mountains of the UIB (Yadav et al., 2009a;Syed et al., 2010;Filippi et al., 2014). Indices of the NAO have been related to: UIB station winter precipitation (Archer and Fowler, 2004;Afzal et al., 2013;Filippi et al., 2014); western Indus basin's winter snow cover and station precipitation

(Hasson et al., 2014); Pakistan station temperature (del Río et al., 2013); and winter precipitation in northwest India (Kar and Rana, 2014).

ENSO is a dominant pattern of multi-year variability driven by ocean-atmosphere interactions in the tropical Pacific (Wolter and Timlin, 2011), influencing climate globally including the variability of both western disturbances and monsoon processes experienced by the region. The commonly used SOI (Southern Oscillation Index) has been related to: winter Hindu Kush Himalayan region precipitation (Afzal et al., 2013); Indian Summer Monsoon Precipitation (Ashok et al., 2004;Ashok and Saji, 2007); central southwest Asian winter precipitation (Syed et al., 2006); Pakistan station temperature (del Río et al., 2013); and northwest India winter precipitation (Kar and Rana, 2014). Stronger links have been reported between ENSO, western disturbances and interannual winter precipitation variability for recent decades (Yadav et al., 2009a;Yadav et al., 2009b).

3 Methods

3.1 BJP forecasting models

The statistical seasonal forecasting model used is the Bayesian joint probability (BJP) approach of Wang et al. (2009). The BJP offers state-of-the-art capabilities for developing seasonal forecast models that optimally utilise information available on antecedent catchment conditions, large-scale climate forcing (through climate indices) and flow forecast scenarios from hydrological models (Robertson et al., 2013;Robertson and Wang, 2012;Schepen et al., 2012;Wang and Robertson, 2011). The BJP models simulate predictor-predictand relationships using conditional multivariate normal distributions, with predictor and predictand data transformed to normal using either a log-sinh (Wang et al., 2012b) or Yeo-Johnson (Yeo and Johnson, 2000) transformation. BJP parameters are inferred using Markov Chain Monte Carlo methods (MCMC) to account for parameter uncertainty, which can be due to factors such as short data records. Probabilistic (ensemble) forecasts are produced by generating samples from the estimated conditional multivariate normal distributions. When predictor-predictand relationships are weak, the BJP produces reliable forecasts that approximate climatology. The full technical details of the BJP modelling approach are presented in Wang et al. (2009) and Wang and Robertson (2011).

3.2 SRM forecasting model

The Snowmelt Runoff Model (SRM) of Martinec et al. (2008) has been used in several studies in the basin (Butt and Bilal, 2011;Romshoo et al., 2015;Tahir et al., 2011;Bogacki and Ismail, 2016;Ismail and Bogacki, 2017, 2018). WAPDA has procured a version of SRM implemented in MS-Excel©, 'ExcelSRM', and for their 2012 case study for Jhelum inflow into the Mangla Reservoir (NESPAK et al., 2012), ExcelSRM was calibrated using data for 2003 to 2010 and subsequently validated against inflows for 2000 to 2002 and 2007 and 2011(Bogacki and Ismail, 2016). In contrast to the probabilistic forecasts produced by the BJP, the SRM is a deterministic model and so produces a single forecast for a given set of inputs.

Given the inadequacy of seasonal meteorological forecasts for the region (Bogacki and Ismail, 2016;Cash et al., 2017;Ismail and Bogacki, 2018), an ESP approach is used to forecast a range of possible Kharif season inflows. That is, SRM is ~~run with~~

~~observed precipitation (P), temperature (T) and snow cover area inputs up to the end of March~~initiated with end of March observed snow cover, and then ~~simulations are produced for the next~~run to produce six-month Kharif-season scenarios using the P and T inputs from each year in the available historical record, together with the Modified Depletion Curve approach (Rango and Martinec, 1982) to simulate snow cover depletion as a function of that scenario-year's degree-days series. This approach results in an ensemble of simulated inflows and, as well as assessing the SRM-ESP forecasts themselves (from a research version of ExcelSRM we obtained in 2015; i.e. not the current version used operationally), this study has used the mean of the ensemble members for each year as an additional predictor series for input to the BJP (option 3 as outlined in the Introduction).

3.3 Verification

BJP forecast model performance is verified using leave-one-out cross-validated results (Wang and Robertson, 2011). That is, to avoid artificially inflating the skill, for each Kharif season the calibration and assessment does not use that season's data for BJP parameter estimation. The cross-validated BJP forecast performance was assessed for 1975-2015 (41 seasons), with the BJP models calibrated on a seasonal basis (i.e. 40 data points) using 1000 MCMC samples for each of the leave-one-out calibrations.

As noted, the BJP can also use hydrological model simulated flow as an additional predictor (Robertson et al., 2013) and in this case we have SRM simulations for the Mangla inflow for a subset of the investigation period (2001-2015). The SRM is an exception to the leave-one-out cross validation as it is calibrated using all data for the 2003 to 2010 period, with parameters manually tuned "… *in order to keep parameters at smooth values and to maintain a reasonable trend in time*." (NESPAK et al., 2012). The mean flow simulation obtained from driving each year's SRM with all year's available precipitation and temperature are therefore not independent forecasts and so it is not surprising that for 2003 to 2010 period the SRM forecasts can be closer to the observed flows than the median cross-validated BJP estimates. When used as a predictor to the BJP, SRM forecasts are applied with leave-one-out cross validation i.e. for each year in the 2001-2015 period all of the other 14 year's simulations are used to produce an ESP with the resulting mean used as a predictor for that year. Note the BJP is able to extract skill from biased dynamic hydrological model forecasts, as long as the hydrological model simulation bias is systematic and stationary (i.e. not random or with a trend).

Verification assesses the overall skill and the bias, reliability and robustness of the forecasts. This includes assessing whether the bias and reliability of the forecasts varies for different periods of the record (temporal stability) or for different event sizes, e.g. whether there is a limitation in forecasting high- or low-flow seasons. Skill scores, quantifying the skill of the forecasts, allow the direct comparison of the performance of forecasting models that use different sets of predictors. Two common skill scores used here are the root mean squared error (RMSE) that assesses the forecast median and the continuous ranked probability score (CRPS) that assesses the reduction in error of the whole forecast probability distribution (Robertson and Wang, 2013). The skill scores are reported as percentage reductions in error scores of the forecasts relative to the observed historical (climatological) median, for RMSE, and relative to the full distribution of the observed historical (climatological) events, for CRPS. The ‘sharpness’ of a probabilistic forecast distribution (i.e. a narrower peaked distribution rather than a wide, flat distribution) is also a characteristic relevant to forecast skill (Gneiting et al., 2007). Sharp forecasts with narrow forecast intervals reduce the range of possible outcomes that are anticipated, increasing their usefulness for decision makers (Li et al., 2016). This skill can be quantified, for example, as the percentage reduction in the inter-quartile range (IQR) between the forecast’s distribution and the observed historical (climatological) distribution (Crochemore et al., 2017). RMSE. CRPS and IQR skill scores are interpreted as[1]:

- 0 is considered to be a forecast with no skill (equivalent skill to predicting using historical averages or historical reference);
- less than 5 is considered to be a forecast with very low skill;
- 5-15 is considered to be a forecast with low skill;
- 15-30 is considered to be a forecast with moderate skill; and,
- greater than 30 is considered to be a forecast with high skill.

Reliability refers to the statistical similarity between the forecast probabilities and the relative frequencies of events in the observations, which can be verified using probability integral transforms (PITs). The PIT represents the non-exceedance probability of observed streamflow obtained from the CDF of the ensemble forecast. If the forecast ensemble spread is appropriate and free of bias then observations will be contained within the forecast ensemble spread, with reliable forecasts having PIT values that follow a uniform distribution between 0 and 1 (Laio and Tamea, 2007). Thus PIT plots are an efficient diagnostic to visually evaluate whether the forecast probability distributions are too wide or too narrow or are biased (under or over estimating) in their prediction of the observed distribution (Wang and Robertson, 2011). As outlined by (Thyer et al., 2009), PIT plot points falling on the 1:1 line indicate that the predicted distribution is a perfect match to the observed; observed PIT values of 0.0 or 1.0 indicate the corresponding observed data falls outside the predicted range, hence the predictive uncertainty is significantly underestimated; PIT values clustered around the midrange (i.e. a low slope in the 0.4 -0.6 uniform variate range) indicate the predictive uncertainty is overestimated; PIT values clustered around the tails (i.e. a high slope in the 0.4 -0.6 uniform variate range) indicate the predictive uncertainty is underestimated; and if PIT values at the theoretical median are higher than those of the uniform variate the predictions have an underprediction bias, and vice versa if they are lower than the uniform variate then the predictions have an overprediction bias.

4 Results

4.1 Skill scores

BJP models were trialled with combinations of predictors accounting for antecedent flow (flow immediately preceding the forecast season, i.e. March flow for the Kharif forecast) and NAO- or ENSO-based climate indices identified from the literature, as introduced in section 2.2 (Charles, 2016). MODIS (Hall et al., 2010) snow-cover area and GLDAS-2.1 (Rodell

[1] ‘*How are the skill score categories defined?*’ from http://www.bom.gov.au/water/ssf/faq.shtml

et al., 2004) snow-water equivalent, additional measures of antecedent conditions, were also assessed as potential predictors. A significant limitation is the shorter record lengths for MODIS and GLDAS, as available data for both start in 2000. This is of particular concern for the BJP's leave-one-out cross-validation, as using short records to identify dynamical mechanisms is susceptible to spurious skill. Correlation analysis, cognisant of the short 2000-15 period, found that these snow products have a similar or lower correlation with Kharif flow (Q_{Kharif}) compared to March flow (Q_{March}), and are relatively highly correlated with Q_{March}. Thus the limitation of short record length, lack of higher correlation with flow than that of the Q_{March} predictor, and relatively high cross-correlation with Q_{March}, leads us to conclude that they would not be expected to provide additional skill as a predictor for the BJP.

~~-~~SRM-ESP scenario-mean forecasts were an additional predictor trialled for Jhelum at Mangla. ~~The best performing~~Higher skill was generally obtained for predictor combinations ~~used~~ using a flow predictor (March flow) together with either the Multivariate ENSO index (MEI; http://www.esrl.noaa.gov/psd/enso/mei/index.html) (Wolter and Timlin, 1998) or the Southern Oscillation signal index (SSI; http://www.cgd.ucar.edu/cas/catalog/climind/soiAnnual.html) (Trenberth, 1984) as a climate predictor. The seasons of the trialled climate predictors (i.e. their time-lag preceding the forecast season) were selected based on their highest linear correlations with flow (not shown). Table 1 presents cross-validated BJP forecast skill scores using the trialled combinations of antecedent flow and climate predictors for the Kharif season for Jhelum at Mangla, together with bootstrapped 10th to 90th percentile ranges to assess model uncertainty. These ranges were obtained by resampling 1000 random sequences of years of the same length as the observed record, i.e. with replacement, and calculating skill scores for each sample. Combinations including the SRM forecasts (ESP mean) as a predictor are included, however because the SRM results are only available for the 15 year period 2001-2015, they are only providing skill during the 2001-2015 period when included as a BJP predictor for the full 41 year period (1975-2015). These results show:

- The antecedent predictor (March flow, Q_{March}) provides greater skill than any of the individual climate predictors used.
- Two-predictor models using Q_{March} and the SRM_{Kharif} predictor give poorer skill scores compared to using Q_{March} alone.
- Two-predictor models using Q_{March} and one climate predictor slightly improve (in most cases) the skill scores compared to using Q_{March} alone.
- Given there is large uncertainty in skill scores we do not aim to select a 'best' model. However, as there are many models with positive skill (i.e. better than climatology) then using skilful models is plausible. Ideas on how to do this are discussed in section 5.

Addition of the SRM_{Kharif} predictor to the two-predictor models using Q_{March} and one climate predictor does not improve skill scores. Table 2 presents the skill scores for the Kharif season forecasts for the Indus at Tarbela, also using the antecedent flow and climate predictors but without SRM forecasts in this case (as SRM was not available for Indus at Tarbela for this study). Similarly to the results for the Jhelum at Mangla, for the Indus at Tarbela BJP forecasts:

- The antecedent predictor (March flow, Q_{March}) provides greater skill than any of the climate predictors used.

- On the whole, a single climate predictor produces low skill compared to that obtained using Q_{March}, with a notable exception that the MEI_{MayJun} (i.e. the year before) predictor produces skill scores comparable to those obtained using Q_{March}. The selection of MEI_{MayJun} as a predictor is discussed further in Section 5.
- Two-predictor models using Q_{March} and one climate predictor improve the skill scores compared to using Q_{March} alone.
- Given there is large uncertainty in skill scores we do not aim to select a ‘best’ model. However, as there are many models with positive skill (i.e. better than climatology) then using skilful models is plausible. Ideas on how to do this are discussed in section 5.

In addition to calibration for the full Kharif season, BJP calibrations were also undertaken for the early Kharif (April-June) and the late Kharif (July-September) using the relevant flow and ENSO-based predictors (e.g. for late Kharif the June flow was used as an antecedent predictor). A comparison of the resulting skill scores are shown in Figure 3 for the BJP models that gave the highest skill gain relative to climatology for the Kharif, early Kharif and late Kharif periods. It is interesting to contrast the performance for the two locations, with Kharif and late Kharif giving similar results across the two locations whereas for early Kharif a marked difference is seen, with high skill for Jhelum at Mangla contrasting the low skill for Indus at Tarbela. The physical reasons for this contrast would require further investigation, with possible causes discussed in Section 5.

4.2 Performance diagnostics

Here we assess the cross-validated performance of forecasts from BJP models using ~~the~~ an antecedent and climate ~~predictors~~ predictor combination (option 1) ~~that were~~ selected on the basis of skill scores and, for Mangla, compare with the SRM-ESP forecasts (option 2). We do not compare results for the BJP models using the SRM-ESP mean as an additional predictor (option 3), as the addition of this predictor ~~was shown to add~~added little or no skill to BJP forecasts (Table 1).

We use PIT plots for verification of the reliability and robustness of the forecast probability distributions, to assess whether there are biases in the forecasts, or whether the forecast probability distributions are too wide or too narrow (Laio and Tamea, 2007). For reliable forecasts, the PIT values should follow a uniform distribution and hence follow a 1:1 line when plotted against a standard uniform variate. For the BJP forecasts for the full 1975-2015 period, both Jhelum at Mangla (Figure 4a) and Indus at Tarbela (Figure 5a) show reliability (i.e. forecast probability distributions are unbiased and of appropriate spread), ~~with~~ evidenced by the forecast’s PIT values plotting close to the 1:1 lines and within the Kolmogorov 5% significance band. Comparison of Jhelum at Mangla BJP and SRM-ESP forecasts, for the shorter 2001-2015 period for which SRM results are available, show a contrast between reliable BJP forecasts (Figure 6a) and biased SRM forecasts, including five values of 0 or 1 indicating that the SRM forecast distribution is too narrow (Figure 7a).

A feature of robust forecasts is their stability across the full period of record and range of flow magnitudes. Figure 4b and Figure 5b show a uniform spread of PIT values and hence stability across the full period of record, for Jhelum at Mangla and Indus at Tarbela BJP forecasts, respectively. Similarly, forecast stability across the range of flow magnitudes is verified by the uniformity of PIT values against forecast median, as shown in Figure 4c and Figure 5c for Jhelum at Mangla and Indus at Tarbela, respectively. For the Jhelum at Mangla BJP and SRM-ESP comparison, stability across time and flow magnitude are harder to assess given the short 15 year sample size. Figure 6b and c show reasonable stability of the BJP forecasts, although there is a trend over time for this part of the record. The equivalent SRM plots (Figure 7b and c) are not as robust, with (as noted previously) five values at 0 or 1 (from the chronological plot: 2001 is at 0, and 2010, 2012, 2014 and 2015 are at 1).

Robustness is also assessed by plotting forecast quantile ranges and observed flows against the forecast median (Figure 4d and Figure 5d) and chronologically (Figure 4e and Figure 5e, for Jhelum at Mangla and Indus at Tarbela respectively). These show that BJP forecasts reasonably account for the range of observed variability for both locations. The relatively less robust SRM-ESP forecasts are shown in Figure 7d and e, again highlighting the overly narrow forecast distribution range for this version of SRM.

Overall, the performance shown in these figures highlight the reliability and robustness of the BJP forecasts for the Kharif season for Jhelum at Mangla and Indus at Tarbela. For the Jhelum at Mangla BJP and SRM-ESP forecasts for the shorter 2001-2015 period, the results contrast the reliable and robust BJP against some limitations from the SRM, which will be discussed further in the next section.

5 Discussion

The SRM forecasts are examples of the commonly applied ESP approach (Shi et al., 2008;Shukla and Lettenmaier, 2011;Wood et al., 2005). As such, Wood and Schaake (2008) note "*One strength of the ESP approach is that it accounts for uncertainty in future climate, which in some seasons is the major component of forecast uncertainty, by assuming that historical climate variability is a good estimate of current climate uncertainty. A weakness of the approach, however, is that when the uncertainty of the current ("initial") hydrologic state is a significant component of the overall forecast uncertainty …, the deterministic estimate of the forecast ensemble's initial hydrologic state leads to an overconfident forecast—that is, one having a spread that is narrower than the total forecast uncertainties warrant.*"

This can be seen in the poor verification performance of SRM-ESP shown in Section 4.2, with the SRM-ESP from the 15 year sample unable to account for the observed range of flows, i.e. the ESP range is too narrow, even though in terms of overall RMSE and MAPE statistics the SRM-ESP-mean and BJP-median are comparable (18.7% RMSE and 14.3% MAPE for BJP-median; 18.4% RMSE and 12.7% MAPE for SRM-ESP-mean). Given that the observed climate of each individual year would, in most cases, be within the range of the ensemble of climate inputs used to produce the ESP, this indicates SRM formulation could be too strongly reliant on antecedent conditions at the beginning of the forecast season. An additional source of bias, as evidenced by the years of poorest performance being outside the years used for parameter estimation, could be over fitting with the model parameters tied too closely to the range of observed predictor-predictand relationships in the 2003-2010 calibration period.

The poorer BJP performance for Indus at Tarbela, as seen in the skill scores relative to Jhelum at Mangla (Figure 3), could be related to the differences in flow generation mechanisms. As the predictors are the source of skill in the statistical BJP approach, examination of correlations between the predictors and flow for the individual months within the Kharif season is insightful. Table 3 shows the (intuitively) expected pattern for Jhelum at Mangla of Q_{March} having a maximum correlation with April flow (0.84) and then maintaining a relatively high correlation until July (0.62) before dropping off for August (0.12). A different process appears to be influencing Indus at Tarbela, as the initial highest Q_{March} correlation with April flow (0.66) drops immediately to 0.18 for May before oscillating between 0.25 (August) to 0.55 (September) for subsequent Kharif months. Similarly, the climate predictor's correlations with the individual month's flows show more of a gradual reduction for Jhelum at Mangla (high for the first four months), whereas for Indus at Tarbela again an oscillatory relationship is seen. These higher correlations in the late Kharif (relative to the early Kharif) for Indus at Tarbela would relate to the correspondingly higher relative skill scores shown in Figure 3 for late Kharif, ~~. It is hypothesised this relates~~corresponding to late-season glacier melt processes that are a significant component of the inflow to Tarbela but not Mangla (Mukhopadhyay and Khan, 2015). Future research could investigate whether dynamical seasonal forecasts of temperature have skill of relevance to forecasting glacier melt, however as noted above such skill has not been determined to date (e.g. Cash et al., 2017), and is beyond the scope of this assessment given our focus of developing practical and easily implementable forecast tools using readily available inputs.

It is also interesting to reflect on the relative performance of the NAO climate predictor, which does not provide any skill for inflow to Mangla (Table 1) but offers comparable skill to several of the ENSO indices trialled for Tarbela (Table 2). This ~~result is also hypothesised to relate to~~indicates NAO ~~being informative of~~may have some skill with regards to late season glacier melt~~,~~. ~~and also corresponds to~~Overall, these results concur with investigations showing a stronger relationship between ENSO and precipitation and weaker relationship between NAO and precipitation in recent decades (Yadav et al., 2009a;Yadav et al., 2009b) resulting in the prevalence of ENSO as the better predictor of winter snowpack magnitude.

For Mangla, the predictor combination that gave the best Kharif season cross-validated skill scores included an ENSO-based predictor (SSI_{March}) immediately before the season (Table 1), which makes sense intuitively as it represents a climate driver of both the snow accumulation before and precipitation conditions during the Kharif season. In contrast, for Tarbela a much earlier ENSO-based predictor (MEI_{MayJun}, i.e. May-Jun the year before) provides higher skill scores than the equivalent predictor immediately before the season (MEI_{FebMar}) (Table 2). To try to understand the dynamical mechanism by which MEI_{MayJun} is providing skill in forecasting Q_{Kharif}, we compared MEI correlations with GLDAS SWE_{March}, Q_{March} and Q_{Kharif}. Results were inconclusive, and perhaps impeded by the short record lengths given SWE is only available from 2000, as while MEI_{MayJun} has a higher correlation with Q_{Kharif} than MEI_{FebMar} (0.76 versus 0.63, respectively) it has a slightly lower correlation with SWE_{March} (0.48 versus 0.52, respectively). Hence MEI_{MayJun} does not appear to be a long-lead predictor of snow accumulation, and so the differences in skill scores may be due to spurious correlations. Therefore we recommend both this model and the Q_{March} and MEI_{FebMar} model be compared and assessed for future events. More generally, the skill score uncertainty ranges presented in Table 1 and Table 2 highlight that no 'best' forecast model can be selected for either basin. Attempting to select a best model would ignore model uncertainty and thus not make best use of forecast skill across the range of models trialled. To address this, probabilistic forecasts from multiple BJP models can be combined using Bayesian Model

Averaging to produce combined forecasts with higher skill than that obtainable from any individual model (Wang et al., 2012a). Thus trialling a BMA approach is recommended, although it is beyond the scope of this current work.

6 Conclusion

This study has assessed the performance and practical feasibility of three options for producing Kharif (April-September) seasonal streamflow forecasts for the Jhelum River inflows to the Mangla Dam in the UIB of Pakistan: option 1, the BJP statistical forecasting technique; option 2, the SRM physically-based model run in ESP mode; and option 3, a hybrid of option 1 with the mean ESP forecasts from option 2 used as an additional predictor for input to the BJP. The option 1 BJP forecast model used antecedent catchment and climatic predictors, with the predictors selected based on BJP skill score performance. The selected predictors represent hydrological conditions immediately preceding the forecast season (i.e., flow of the preceding month – March in this case) and ENSO-based climate indices related to drivers of winter snow accumulation. For an additional comparison, the option 1 BJP approach was also undertaken for the Indus River inflows to the Tarbela Dam.

Overall findings were:

- The best performing BJP models for Tarbela and Mangla inflows are consistent in that both used an antecedent flow predictor and a climate predictor representing ENSO.
- For Tarbela the Q_{March} and MEI_{MayJun} model gave the best skill, however because we could not determine the dynamical mechanism(s) by which the relatively long lag between MEI_{MayJun} influences snowpack accumulation and flow, we cannot rule out the possibility that the skill is due to spurious correlation. Therefore we recommend both this model and the Q_{March} and MEI_{FebMar} model be compared and assessed for future events and, more generally, that BMA be trialled in future research to combine the skill of multiple BJP models as, for example, undertaken in Australia in (Pokhrel et al., 2013).
- There are pragmatic benefits to selecting a BJP model using only antecedent and climate predictors, rather than including SRM mean ESP as an additional predictor even in cases when SRM does provide skill, given that flow and climate predictors are readily available and thus BJP forecasts are easily and quickly produced. The SRM, being a deterministic model, is a much more technically involved and data intensive approach to forecast generation (Bogacki and Ismail, 2016;Ismail and Bogacki, 2018).
- Cross-validated performance of the BJP seasonal forecasts for the 1975 to 2015 Kharif seasons, as shown in the diagnostic and verification statistics presented, highlight that the BJP produces forecasts that are statistically unbiased, robust and reliable. In contrast, the SRM-ESP forecasts show bias particularly for the most recent years outside the SRM calibration period, potentially indicating limitations with the SRM due to lack of cross-validated calibration and resultant over-fitting. Thus SRM-ESP forecasts are overly confident, underestimating the full uncertainty that is captured by the BJP approach.

- High skill was obtained for BJP forecasts of early Kharif flow for Jhelum at Mangla. Moderate skill was obtained for the full and late Kharif season forecasts for both Jhelum at Mangla and Indus at Tarbela. Lower skill was seen for early Kharif for Indus at Tarbela.

In future research, BJP forecast models could readily be developed and assessed for other tributaries, e.g. the Chenab and Kabul, subject to availability of flow data. This would allow an overall assessment of UIB flow forecasting for the major contributing basins. The present method used by the Indus River System Authority (IRSA) to forecast UIB Kharif streamflow is based on historical analogues. The IRSA use their database of the previous 60 years of flow to select years where the historic March flows are within 5% of the current March flow and use the corresponding historical Kharif flows (within 5%) for their forecast scenario. The selection of the historical scenario is also informed by forecasts from the Pakistan Meteorological Department, forecasts provided by WAPDA (e.g. SRM forecasts), and present snow conditions in the catchment. The forecasts are continuously revised as the season progresses.

Sufficiently skilful BJP forecasts could also inform scenario selection, providing for the first time a probabilistic approach to forecasts in contrast to a single forecast as currently used. However ~~Probabilistic~~ probabilistic forecasts (such as a the BJP) can be misinterpreted if they are unfamiliar to the water management professionals using them to inform decisions (Pagano et al., 2002;Ramos et al., 2013;Rayner et al., 2005;Whateley et al., 2015). ~~To enable~~Hence the successful transfer of BJP forecast tools to operational use within ~~IBIS management in~~ Pakistan would require guidance for building BJP models and generating forecasts, test cases with example results, face to face training, and on-going support.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The Pakistan Mission of the Australian Department of Foreign Affairs and Trade provided funds for this study through the Sustainable Development Investment Portfolio Indus project. CSIRO provided additional support through its Land and Water portfolio. We thank the Pakistan Meteorological Department (PMD) for sharing datasets and the Pakistan Water and Power Development Authority Glacier Monitoring Research Centre (WAPDA-GMRC) for sharing datasets and the SRM used in this collaborative work. Thanks also to Luis Neumann~~, David Robertson~~ and Tony Zhao, CSIRO Land and Water, for help and advice. We express our gratitude to the two anonymous reviewers and the editor, Andy Wood, for their inputs which helped improve the paper. We also acknowledge helpful comments received from Muhammad Fraz Ismail regarding the SRM.

References

Afzal, M., Haroon, M. A., Rana, A. S., and Imran, A.: Influence of North Atlantic Oscillations and Southern Oscillations on winter precipitation of northern Pakistan, Pakistan Journal of Meteorology, 9, 1-8, 2013.

Ahmed, K., and Sanchez, M.: A study of the factors and processes involved in the sedimentation of Tarbela reservoir, Pakistan, Environ Earth Sci, 62, 927-933, 10.1007/s12665-010-0578-3, 2011.

Alford, D., Archer, D., Bookhagen, B., Grabs, W., Halvorson, S., Hewitt, K., Immerzeel, W., Kamp, U., and Krumwiede, B.: Monitoring of glaciers, climate and runoff in the Hindu Kush-Himalaya mountains, South Asia Water Initiative Report No. 67668-SAS, Washington DC: World Bank, 2014.

Anghileri, D., Voisin, N., Castelletti, A., Pianosi, F., Nijssen, B., and Lettenmaier, D. P.: Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments, Water Resources Research, 52, 4209-4225, 10.1002/2015WR017864, 2016.

Archer, D. R., and Fowler, H. J.: Spatial and temporal variations in precipitation in the Upper Indus Basin, global teleconnections and hydrological implications, Hydrol. Earth Syst. Sci., 8, 47-61, 10.5194/hess-8-47-2004, 2004.

Ashok, K., Guan, Z., Saji, N. H., and Yamagata, T.: Individual and combined influences of ENSO and the Indian Ocean Dipole on the Indian Summer Monsoon, Journal of Climate, 17, 3141-3155, 10.1175/1520-0442(2004)017<3141:IACIOE>2.0.CO;2, 2004.

Ashok, K., and Saji, N. H.: On the impacts of ENSO and Indian Ocean dipole events on sub-regional Indian summer monsoon rainfall, Nat Hazards, 42, 273-285, 10.1007/s11069-006-9091-0, 2007.

Asian Development Bank: Asian water development outlook 2016: Strengthening water security in Asia and the Pacific, Mandaluyong City, Philippines: Asian Development Bank, 2016.

Bennett, J. C., Wang, Q. J., Li, M., Robertson, D. E., and Schepen, A.: Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model, Water Resources Research, 52, 8238-8259, 10.1002/2016WR019193, 2016.

Bierkens, M. F. P., and van Beek, L. P. H.: Seasonal Predictability of European Discharge: NAO and Hydrological Response Time, Journal of Hydrometeorology, 10, 953-968, 10.1175/2009JHM1034.1, 2009.

Bogacki, W., and Ismail, M. F.: Seasonal forecast of Kharif flows from Upper Jhelum catchment, Proc. IAHS, 374, 137-142, 10.5194/piahs-374-137-2016, 2016.

Butt, M. J., and Bilal, M.: Application of snowmelt runoff model for water resource management, Hydrological Processes, 25, 3735-3747, 10.1002/hyp.8099, 2011.

Cash, B. A., Manganello, J. V., and Kinter, J. L.: Evaluation of NMME temperature and precipitation bias and forecast skill for South Asia, Clim Dyn, 10.1007/s00382-017-3841-4, 2017.

Charles, S. P.: Hydroclimate of the Indus - synthesis of the literature relevant to Indus basin hydroclimate processes, trends, seasonal forecasting and climate change, CSIRO Sustainable Development Investment Portfolio project. CSIRO Land and Water, Australia. 48pp., 2016.

Chiew, F. H. S., Zhou, S. L., and McMahon, T. A.: Use of seasonal streamflow forecasts in water resources management, Journal of Hydrology, 270, 135-144, dx.doi.org/10.1016/S0022-1694(02)00292-5, 2003.

Crochemore, L., Ramos, M. H., Pappenberger, F., and Perrin, C.: Seasonal streamflow forecasting by conditioning climatology with precipitation indices, Hydrol. Earth Syst. Sci., 21, 1573-1591, 10.5194/hess-21-1573-2017, 2017.

del Río, S., Anjum Iqbal, M., Cano-Ortiz, A., Herrero, L., Hassan, A., and Penas, A.: Recent mean temperature trends in Pakistan and links with teleconnection patterns, International Journal of Climatology, 33, 277-290, 10.1002/joc.3423, 2013.

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R. L.: Seasonal climate predictability and forecasting: status and prospects, Wiley Interdisciplinary Reviews: Climate Change, 4, 245-268, 10.1002/wcc.217, 2013.

Döll, P., Fiedler, K., and Zhang, J.: Global-scale analysis of river flow alterations due to water withdrawals and reservoirs, Hydrol. Earth Syst. Sci., 13, 2413-2432, 10.5194/hess-13-2413-2009, 2009.

Filippi, L., Palazzi, E., von Hardenberg, J., and Provenzale, A.: Multidecadal Variations in the Relationship between the NAO and Winter Precipitation in the Hindu Kush–Karakoram, Journal of Climate, 27, 7890-7902, 10.1175/JCLI-D-14-00286.1, 2014.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69, 243-268, 10.1111/j.1467-9868.2007.00587.x, 2007.

Hall, D. K., Riggs, G. A., Foster, J. L., and Kumar, S. V.: Development and evaluation of a cloud-gap-filled MODIS daily snow-cover product, Remote Sensing of Environment, 114, 496-503, http://dx.doi.org/10.1016/j.rse.2009.10.007, 2010.

Hasson, S., Lucarini, V., Khan, M. R., Petitta, M., Bolch, T., and Gioli, G.: Early 21st century snow cover state over the western river basins of the Indus River system, Hydrol. Earth Syst. Sci., 18, 4077-4100, 10.5194/hess-18-4077-2014, 2014.

Hurrell, J. W.: Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation, Science, 269, 676-679, 1995.

Immerzeel, W. W., Droogers, P., de Jong, S. M., and Bierkens, M. F. P.: Large-scale monitoring of snow cover and runoff simulation in Himalayan river basins using remote sensing, Remote Sensing of Environment, 113, 40-49, dx.doi.org/10.1016/j.rse.2008.08.010, 2009.

Ismail, M. F., and Bogacki, W.: Scenario approach for the seasonal forecast of Kharif flows from Upper Indus Basin, Hydrol. Earth Syst. Sci. Discuss., 2017, 1-15, 10.5194/hess-2017-182, 2017.

Ismail, M. F., and Bogacki, W.: Scenario approach for the seasonal forecast of Kharif flows from the Upper Indus Basin, Hydrol. Earth Syst. Sci., 22, 1391-1409, 10.5194/hess-22-1391-2018, 2018.

Kar, S., and Rana, S.: Interannual variability of winter precipitation over northwest India and adjoining region: impact of global forcings, Theor Appl Climatol, 116, 609-623, 10.1007/s00704-013-0968-z, 2014.

Kim, H.-M., Webster, P., Curry, J., and Toma, V.: Asian summer monsoon prediction in ECMWF System 4 and NCEP CFSv2 retrospective seasonal forecasts, Clim Dyn, 39, 2975-2991, 10.1007/s00382-012-1470-5, 2012.

Kirby, M., Ahmad, M.-u.-D., Mainuddin, M., Khaliq, T., and Cheema, M. J. M.: Agricultural production, water use and food availability in Pakistan: Historical trends, and projections to 2050, Agricultural Water Management, 179, 34-46, doi.org/10.1016/j.agwat.2016.06.001, 2017.

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., van den Dool, H., Saha, S., Mendez, M. P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D. G., Tippett, M. K., Barnston, A. G., Li, S., Rosati, A., Schubert, S. D., Rienecker, M., Suarez, M., Li, Z. E., Marshak, J., Lim, Y.-K., Tribbia, J., Pegion, K., Merryfield, W. J., Denis, B., and Wood, E. F.: The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction, Bulletin of the American Meteorological Society, 95, 585-601, 10.1175/BAMS-D-12-00050.1, 2013.

Koster, R. D., Mahanama, S. P. P., Livneh, B., Lettenmaier, D. P., and Reichle, R. H.: Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow, Nature Geosci, 3, 613-616, 2010.

Laio, F., and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sci., 11, 1267-1277, 10.5194/hess-11-1267-2007, 2007.

Li, H., Luo, L., Wood, E. F., and Schaake, J.: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting, Journal of Geophysical Research: Atmospheres, 114, D04114, 10.1029/2008JD010969, 2009.

Li, M., Wang, Q. J., Bennett, J. C., and Robertson, D. E.: Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting, Hydrol. Earth Syst. Sci., 20, 3561-3579, 10.5194/hess-20-3561-2016, 2016.

Mahmood, R., Babel, M. S., and Jia, S.: Assessment of temporal and spatial changes of future climate in the Jhelum river basin, Pakistan and India, Weather and Climate Extremes, 10, Part B, 40-55, dx.doi.org/10.1016/j.wace.2015.07.002, 2015.

Martinec, J., Rango, A., and Roberts, R.: Snowmelt Runoff Model (SRM) User's Manual, New Mexico State University, Las Cruces, New Mexico., 175, 2008.

Maurer, E. P., and Lettenmaier, D. P.: Predictability of seasonal runoff in the Mississippi River basin, Journal of Geophysical Research: Atmospheres, 108, 8607, 10.1029/2002JD002555, 2003.

Molteni, F., Stockdale, T. M., Balmaseda, A., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T., and Vitart, F.: The new ECMWF seasonal forecast system (System 4), ECMWF Technical Memorandum 656, 2011.

Mukhopadhyay, B., and Khan, A.: A reevaluation of the snowmelt and glacial melt in river flows within Upper Indus Basin and its significance in a changing climate, Journal of Hydrology, 527, 119-132, dx.doi.org/10.1016/j.jhydrol.2015.04.045, 2015.

NESPAK, AHT, and DELTARES: Hydrological Flow Forecast Model for Mangla Catchment, Upgrading of Tools, Water Resources Database, Management Systems and Models Under Sub Component “B1” of WCAP, Final Report, 2012.

Pagano, T. C., Hartmann, H. C., and Sorooshian, S.: Factors affecting seasonal forecast use in Arizona water management: a case study of the 1997-98 El NiÃƒÂ±o, Climate Research, 21, 259-269, 2002.

Plummer, N., Tuteja, N., Wang, Q. J., Wang, E., Robertson, D., Zhou, S., Schepen, A., Alves, O., Timbal, B., and Puri, K.: A seasonal water availability prediction service: opportunities and challenges, 2009.

Pokhrel, P., Wang, Q. J., and Robertson, D. E.: The value of model averaging and dynamical climate model predictions for improving statistical seasonal streamflow forecasts over Australia, Water Resources Research, 49, 6671-6687, 10.1002/wrcr.20449, 2013.

Ramos, M. H., van Andel, S. J., and Pappenberger, F.: Do probabilistic forecasts lead to better decisions?, Hydrol. Earth Syst. Sci., 17, 2219-2232, 10.5194/hess-17-2219-2013, 2013.

Rango, A., and Martinec, J.: Snow accumulation derived from modified depletion curves of snow coverage, Hydrological Aspects of Alpine and High Mountain Areas (Proceedings of the Exeter Symposium, July 1982). IAHS Publ. no. 138., 83-90, 1982.

Rayner, S., Lach, D., and Ingram, H.: Weather Forecasts are for Wimps: Why Water Resource Managers Do Not Use Climate Forecasts, Climatic Change, 69, 197-227, 10.1007/s10584-005-3148-z, 2005.

Robertson, D., and Wang, Q. J.: Seasonal Forecasts of Unregulated Inflows into the Murray River, Australia, Water Resour Manage, 27, 2747-2769, 10.1007/s11269-013-0313-4, 2013.

Robertson, D. E., and Wang, Q. J.: A Bayesian approach to predictor selection for seasonal streamflow forecasting, Journal of Hydrometeorology, 13, 155-171, 10.1175/JHM-D-10-05009.1, 2012.

Robertson, D. E., Pokhrel, P., and Wang, Q. J.: Improving statistical forecasts of seasonal streamflows using hydrological model output, Hydrol. Earth Syst. Sci., 17, 579-593, 10.5194/hess-17-579-2013, 2013.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C. J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D.: The Global Land Data Assimilation System, Bulletin of the American Meteorological Society, 85, 381-394, 10.1175/BAMS-85-3-381, 2004.

Romshoo, S. A., Dar, R. A., Rashid, I., Marazi, A., Ali, N., and Zaz, S. N.: Implications of Shrinking Cryosphere Under Changing Climate on the Streamflows in the Lidder Catchment in the Upper Indus Basin, India, Arctic, Antarctic, and Alpine Research, 47, 627-644, 10.1657/AAAR0014-088, 2015.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H.-y., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M., and Becker, E.: The NCEP Climate Forecast System Version 2, Journal of Climate, 27, 2185-2208, 10.1175/JCLI-D-12-00823.1, 2014.

Schepen, A., Wang, Q. J., and Robertson, D.: Evidence for Using Lagged Climate Indices to Forecast Australian Seasonal Rainfall, Journal of Climate, 25, 1230-1246, 10.1175/JCLI-D-11-00156.1, 2012.

Schepen, A., Zhao, T., Wang, Q. J., Zhou, S., and Feikema, P.: Optimising seasonal streamflow forecast lead time for operational decision making in Australia, Hydrol. Earth Syst. Sci., 20, 4117-4128, 10.5194/hess-20-4117-2016, 2016.

Schlosser, C. A., Strzepek, K., Gao, X., Fant, C., Blanc, É., Paltsev, S., Jacoby, H., Reilly, J., and Gueneau, A.: The future of global water stress: An integrated assessment, Earth's Future, 2, 341-361, 10.1002/2014EF000238, 2014.

Shi, X., Wood, A. W., and Lettenmaier, D. P.: How essential is hydrologic model calibration to seasonal streamflow forecasting?, Journal of Hydrometeorology, 9, 1350-1363, 10.1175/2008JHM1001.1, 2008.

Shukla, S., and Lettenmaier, D. P.: Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill, Hydrol. Earth Syst. Sci., 15, 3529-3538, 10.5194/hess-15-3529-2011, 2011.

Shukla, S., Sheffield, J., Wood, E. F., and Lettenmaier, D. P.: On the sources of global land surface hydrologic predictability, Hydrol. Earth Syst. Sci., 17, 2781-2796, 10.5194/hess-17-2781-2013, 2013.

Syed, F. S., Giorgi, F., Pal, J. S., and King, M. P.: Effect of remote forcings on the winter precipitation of central southwest Asia part 1: observations, Theor Appl Climatol, 86, 147-160, 10.1007/s00704-005-0217-1, 2006.

Syed, F. S., Giorgi, F., Pal, J. S., and Keay, K.: Regional climate model simulation of winter climate over Central–Southwest Asia, with emphasis on NAO and ENSO effects, International Journal of Climatology, 30, 220-235, 10.1002/joc.1887, 2010.

Tahir, A. A., Chevallier, P., Arnaud, Y., Neppel, L., and Ahmad, B.: Modeling snowmelt-runoff under climate scenarios in the Hunza River basin, Karakoram Range, Northern Pakistan, Journal of Hydrology, 409, 104-117, dx.doi.org/10.1016/j.jhydrol.2011.08.035, 2011.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., and Srikanthan, S.: Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, Water Resources Research, 45, n/a-n/a, 10.1029/2008WR006825, 2009.

Trenberth, K. E.: Signal Versus Noise in the Southern Oscillation, Monthly Weather Review, 112, 326-332, 10.1175/1520-0493(1984)112<0326:SVNITS>2.0.CO;2, 1984.

van Dijk, A. I. J. M., Peña-Arancibia, J. L., Wood, E. F., Sheffield, J., and Beck, H. E.: Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide, Water Resources Research, 49, 2729-2746, 10.1002/wrcr.20251, 2013.

Wada, Y., van Beek, L. P. H., and Bierkens, M. F. P.: Modelling global water stress of the recent past: on the relative importance of trends in water demand and climate variability, Hydrol. Earth Syst. Sci., 15, 3785-3808, 10.5194/hess-15-3785-2011, 2011.

Wang, Q. J., Robertson, D. E., and Chiew, F. H. S.: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites, Water Resources Research, 45, n/a-n/a, 10.1029/2008WR007355, 2009.

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

Wang, Q. J., Schepen, A., and Robertson, D. E.: Merging Seasonal Rainfall Forecasts from Multiple Statistical Models through Bayesian Model Averaging, Journal of Climate, 25, 5524-5537, 10.1175/JCLI-D-11-00386.1, 2012a.

Wang, Q. J., Shrestha, D. L., Robertson, D. E., and Pokhrel, P.: A log-sinh transformation for data normalization and variance stabilization, Water Resources Research, 48, n/a-n/a, 10.1029/2011WR010973, 2012b.

Whateley, S., Palmer, R. N., and Brown, C.: Seasonal Hydroclimatic Forecasts as Innovations and the Challenges of Adoption by Water Managers, Journal of Water Resources Planning and Management, 141, doi:10.1061/(ASCE)WR.1943-5452.0000466, 2015.

Wolter, K., and Timlin, M. S.: Measuring the strength of ENSO events: How does 1997/98 rank?, Weather, 53, 315-324, 10.1002/j.1477-8696.1998.tb06408.x, 1998.

Wolter, K., and Timlin, M. S.: El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext), International Journal of Climatology, 31, 1074-1087, 10.1002/joc.2336, 2011.

Wood, A. W., Kumar, A., and Lettenmaier, D. P.: A retrospective assessment of National Centers for Environmental Prediction climate model–based ensemble hydrologic forecasting in the western United States, Journal of Geophysical Research: Atmospheres, 110, n/a-n/a, 10.1029/2004JD004508, 2005.

Wood, A. W., and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, Geophysical Research Letters, 35, n/a-n/a, 10.1029/2008GL034648, 2008.

Wood, A. W., and Schaake, J. C.: Correcting Errors in Streamflow Forecast Ensemble Mean and Spread, Journal of Hydrometeorology, 9, 132-148, 10.1175/2007JHM862.1, 2008.

Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J., and Clark, M.: Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill, Journal of Hydrometeorology, 17, 651-668, 10.1175/JHM-D-14-0213.1, 2015.

Yadav, R. K., Rupa Kumar, K., and Rajeevan, M.: Increasing influence of ENSO and decreasing influence of AO/NAO in the recent decades over northwest India winter precipitation, Journal of Geophysical Research: Atmospheres, 114, D12112, 10.1029/2008JD011318, 2009a.

Yadav, R. K., Yoo, J. H., Kucharski, F., and Abid, M. A.: Why is ENSO influencing northwest India winter precipitation in recent decades?, Journal of Climate, 23, 1979-1993, 10.1175/2009JCLI3202.1, 2009b.

Yeo, I.-K., and Johnson, R. A.: A New Family of Power Transformations to Improve Normality or Symmetry, Biometrika, 87, 954-959, 2000.

Yossef, N. C., Winsemius, H., Weerts, A., van Beek, R., and Bierkens, M. F. P.: Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing, Water Resources Research, 49, 4687-4699, 10.1002/wrcr.20350, 2013.

Yuan, X., Wood, E. F., and Ma, Z.: A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, Wiley Interdisciplinary Reviews: Water, 2, 523-536, 10.1002/wat2.1088, 2015.

Zheng, H. X., Wang, Q. J., Aynul, K., Shao, Q. X., Shin, D., and Tuteja, N.: Evaluation of Downscaled POAMA M24 for Monthly and 3-Monthly Streamflow Forecasts, 20th International Congress on Modelling and Simulation (Modsim2013), 2799-2805, 2013.

Tables

Table 1: ~~Skills~~ Cross-validated skill scores of BJP forecasts for Kharif season for Jhelum at Mangla. Bootstrap 10th to 90th percentile resampling ranges are shown in brackets.

Predictor combination			Skill scores 2001-2015				Skill scores 1975-2015			
Flow[a]	Climate[b]	Model[c]	SSCRPS		SSRMSE		SSCRPS		SSRMSE	
Q_{March}	–	–	21.0	(-9.5 – 41.3)	17.5	(-13.8 – 39.6)	26.9	(15.7 – 35.8)	26.0	(14.5 – 35.3)
Q_{March}	–	SRM_{Kharif}	15.8	(-12.6 – 36.0)	13.2	(-13.2 – 32.4)	25.6	(15.2 – 34.9)	25.4	(15.0 – 34.2)
–	$NAO_{SepOctNov}$	–	-3.2	(-19.5 – 10.5)	-7.1	(-22.6 – 7.3)	-0.6	(-4.8 – 2.9)	-2.3	(-7.3 – 1.8)
–	MEI_{FebMar}	–	14.7	(-1.3 – 28.2)	10.3	(-5.8 – 26.2)	14.2	(5.8 – 21.8)	11.7	(3.1 – 19.9)
Q_{March}	MEI_{FebMar}	–	22.6	(-5.0 – 42.9)	18.5	(-9.0 – 40.7)	25.5	(13.8 – 35.0)	24.7	(13.0 – 33.9)
–	MEI_{FebMar}	SRM_{Kharif}	24.1	(5.2 – 38.2)	21.2	(1.6 – 36.2)	17.7	(7.6 – 26.5)	15.4	(4.5 – 24.4)
Q_{March}	MEI_{FebMar}	SRM_{Kharif}	18.0	(-6.7 – 38.0)	14.8	(-10.0 – 34.8)	23.9	(12.5 – 34.0)	23.2	(11.5 – 32.6)
	SSI_{March}	–	11.4	(-5.3 – 27.5)	7.0	(-9.2 – 27.9)	7.9	(0.6 – 14.8)	6.0	(-1.7 – 13.6)
Q_{March}	SSI_{March}	–	24.3	(-2.5 – 44.9)	20.2	(-7.2 – 42.5)	26.2	(15.7 – 35.8)	25.1	(14.4 – 34.5)
	SSI_{March}	SRM_{Kharif}	24.2	(2.2 – 41.3)	20.5	(-3.2 – 39.5)	12.9	(3.0 – 21.7)	10.3	(-0.4 – 19.6)
Q_{March}	SSI_{March}	SRM_{Kharif}	22.5	(-5.2 – 42.5)	19.1	(-8.0 – 39.0)	24.9	(13.3 – 34.8)	24.4	(13.4 – 33.7)
SRM Scenarios[d]			25.3		20.4		–		–	

[a] 1976-2015

[b] 1975-2015

[c] 2001-2015 SRM-ESP mean, with L1OCV

[d] 2001-2015 SRM-ESP with L1OCV

Table 2: ~~Skills~~ Cross-validated skill scores of BJP forecasts for Kharif season for Indus at Tarbela. Bootstrapped 10th to 90th percentile resampling ranges are shown in brackets.

Predictor combination			Skill scores 1975-2015			
Flow[a]	Climate[b]	Model[c]	SSCRPS		SSRMSE	
Q_{March}	–	–	16.6	(9.0 – 24.0)	18.9	(11.8 – 26.0)
–	$\text{NAO}_{\text{SepOctNov}}$[d]	–	6.1	(1.1 – 11.1)	8.2	(2.8 – 13.7)
Q_{March}	$\text{NAO}_{\text{SepOctNov}}$[d]	–	18.8	(11.9 – 27.0)	21.0	(14.2 – 28.4)
–	$\text{MEI}_{\text{FebMar}}$	–	7.6	(2.5 – 12.7)	10.3	(4.2 – 15.9)
Q_{March}	$\text{MEI}_{\text{FebMar}}$	–	18.6	(9.9 – 25.9)	20.8	(12.0 – 28.1)
–	$\text{MEI}_{\text{MayJun}}$[d]	–	15.6	(7.2 – 23.2)	17.1	(7.9 – 25.8)
Q_{March}	$\text{MEI}_{\text{MayJun}}$[d]	–	25.0	(15.1 – 33.6)	25.0	(14.2 – 34.6)
–	$\text{SSI}_{\text{March}}$	–	1.4	(-1.6 – 4.6)	4.9	(0.3 – 9.0)
Q_{March}	$\text{SSI}_{\text{March}}$	–	16.9	(9.1 – 24.6)	19.3	(11.3 – 26.6)

[a] 1976-2015

[b] 1975-2015

[c] No SRM for Indus

[d] noting lag to calendar-year before flow season

Table 3: Correlation between the flow of the individual months of the Kharif season and the predictors used by BJP (those in bold significant at $p < 0.05$)

	Jhelum at Mangla		Indus at Tarbela	
Q_{Month}	Q_{March}	SSI_{March}	Q_{March}	MEI_{MayJun}[a]
Apr	**0.84**	**-0.50**	**0.66**	**0.41**
May	**0.77**	**-0.41**	0.18	0.22
Jun	**0.62**	**-0.32**	**0.44**	0.28
Jul	**0.62**	**-0.38**	**0.34**	**0.42**
Aug	0.12	-0.03	0.25	**0.37**
Sep	0.18	-0.27	**0.55**	0.30

[a]Year before

Figures

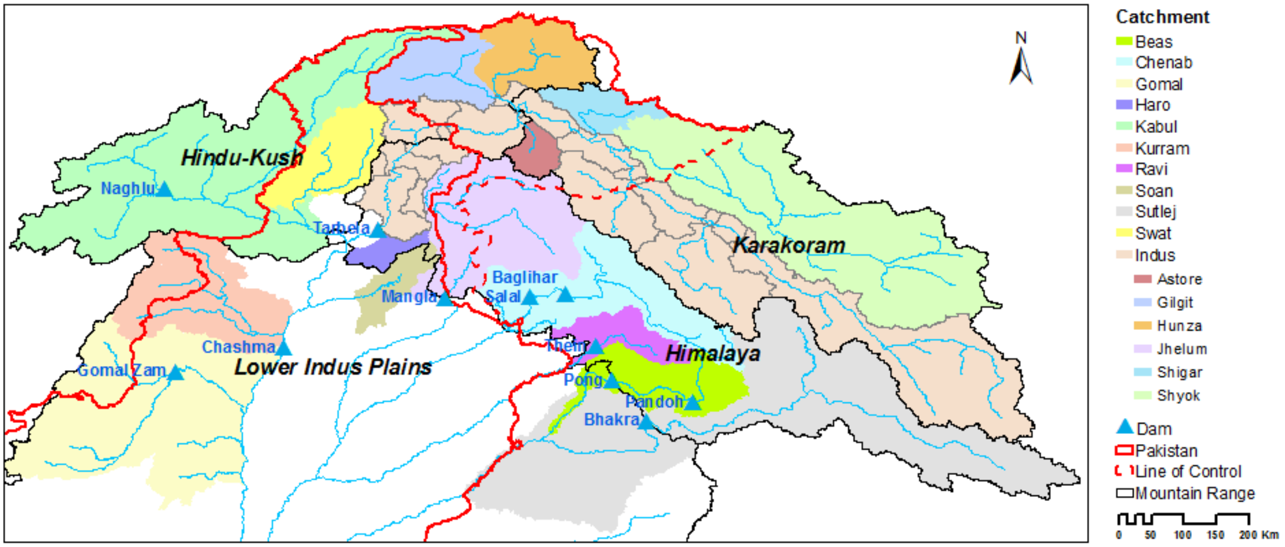


Figure 1: Map of UIB showing sub-basins and location of major dams

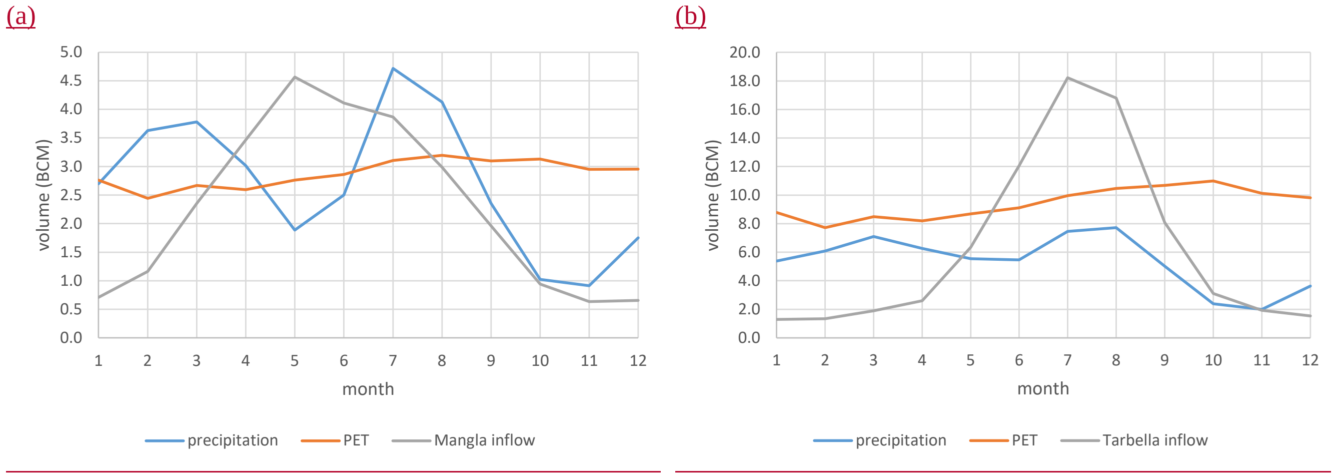


Figure 2: Annual cycle of mean precipitation, PET and inflow for (a) Jhelum at Mangla and (b) Indus at Tarbella (note different volume scales).

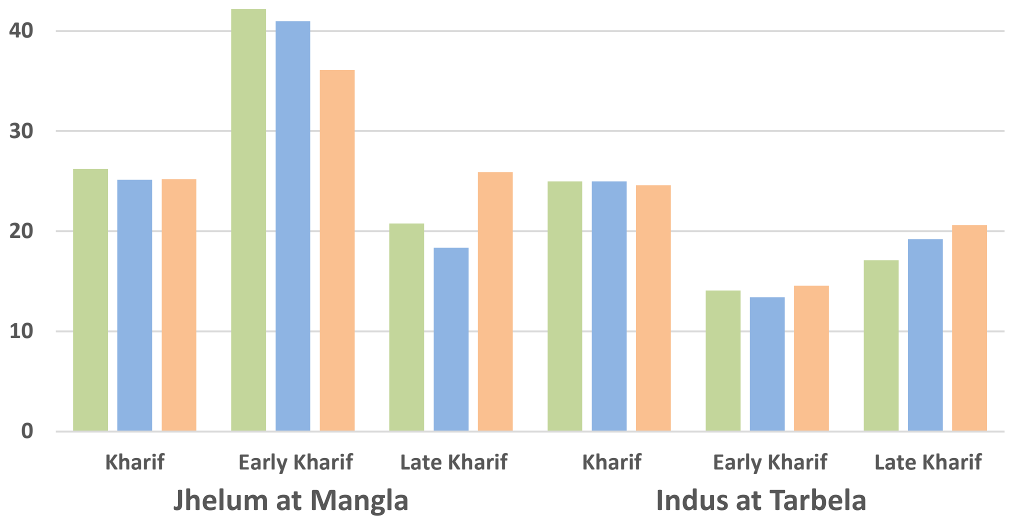
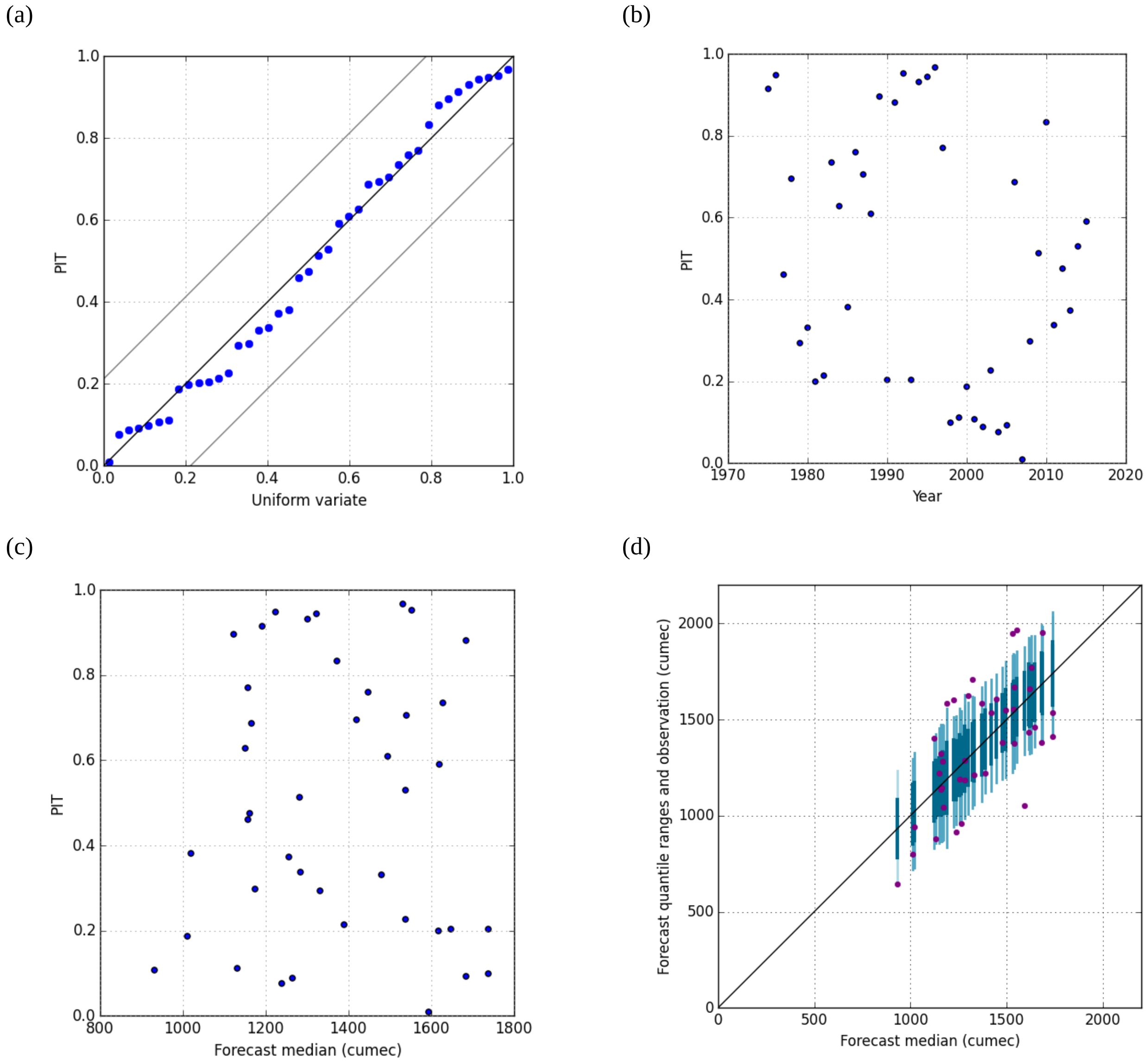


Figure 3: BJP cross-validated skill scores, % skill gain relative to climatology, for CRPS (green), RMSE (blue) and IQR (orange) skill scores. Less than 5 is considered to be a forecast with very low skill. Between 5 and 15 is considered low skill. Between 15 and 30 is considered moderate skill, and higher than 30 is considered to be a forecast with high skill.


(a)
(b)
(c)
(d)
PIT
Uniform variate
0.0
0.2
0.4
0.6
0.8
1.0
Year
1970
1980
1990
2000
2010
2020
Forecast median (cumec)
800
1000
1200
1400
1600
1800
Forecast quantile ranges and observation (cumec)
0
500
1000
1500
2000

(e)

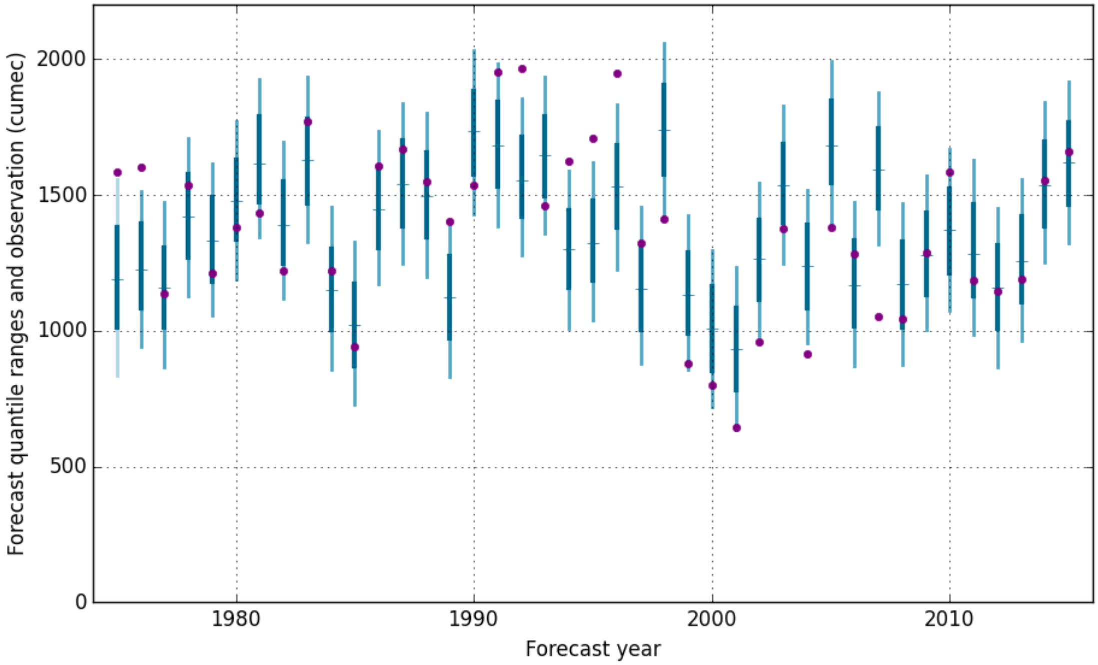
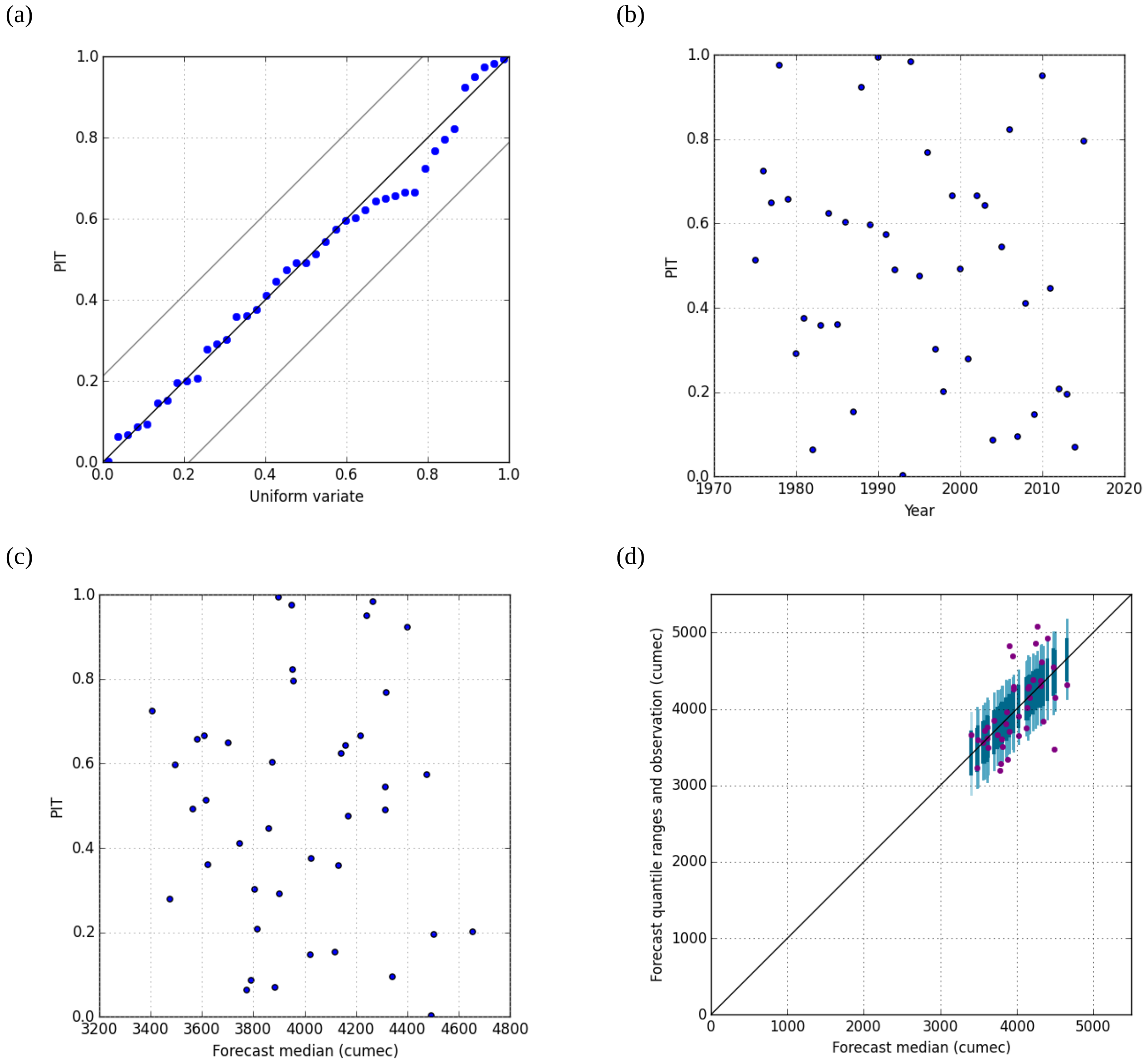


Figure 4: BJP cross-validated forecasts for Jhelum at Mangla Kharif for 1975-2015; (a) PIT uniform probability plot (1:1 black line, theoretical uniform distribution; grey lines, Kolmogorov 5% significance bands; blue points, PIT values of forecast streamflow); (b) chronological PIT plot; (c) median PIT plot; (d) forecast quantiles and observed plotted according to forecast median (1:1 line, forecast median; dark vertical line, forecast [0.25, 0.75] quantile range; light and dark vertical line, forecast [0.10, 0.90] quantile range; dots, observed inflow); (e) chronological forecast quantile range and observations (dark blue, forecast [0.25, 0.75]; light and dark blue, forecast [0.10, 0.90]; crosses, forecast median; dots, observed).


(a)
PIT
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Uniform variate
(b)
PIT
1.0
0.8
0.6
0.4
0.2
0.0
1970
1980
1990
2000
2010
2020
Year
(c)
PIT
1.0
0.8
0.6
0.4
0.2
0.0
3200
3400
3600
3800
4000
4200
4400
4600
4800
Forecast median (cumec)
(d)
Forecast quantile ranges and observation (cumec)
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
Forecast median (cumec)

(e)

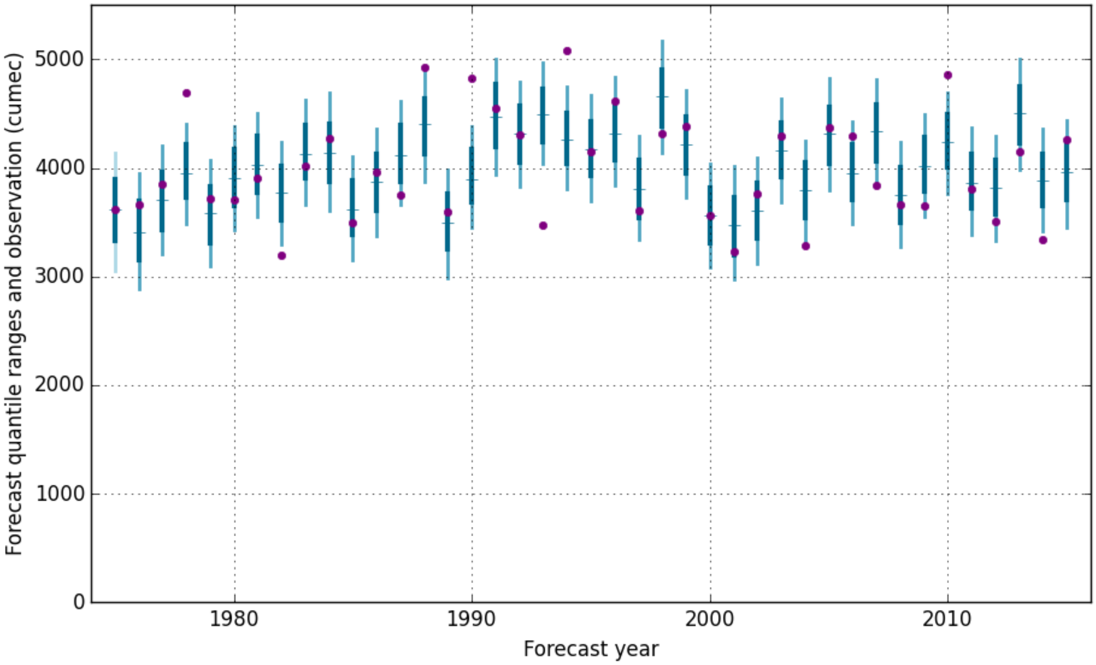


Figure 5: as in Figure 4 for BJP cross-validated forecasts for Indus at Tarbela Kharif for 1975-2015

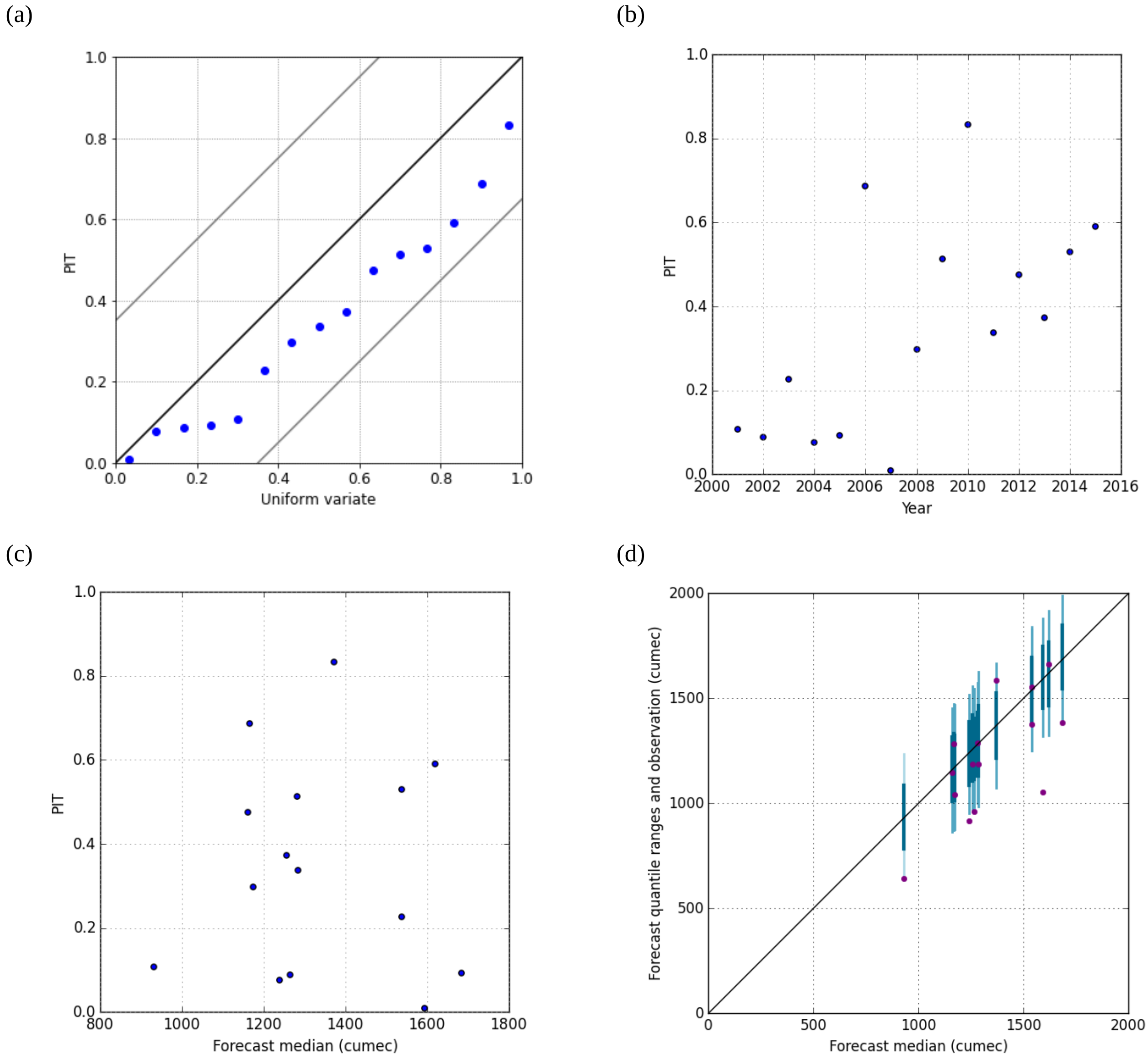
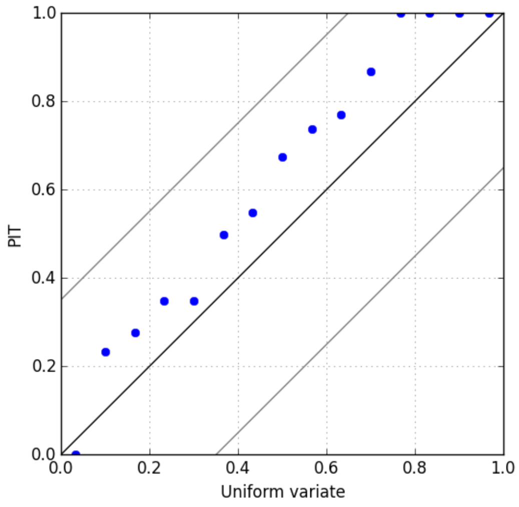
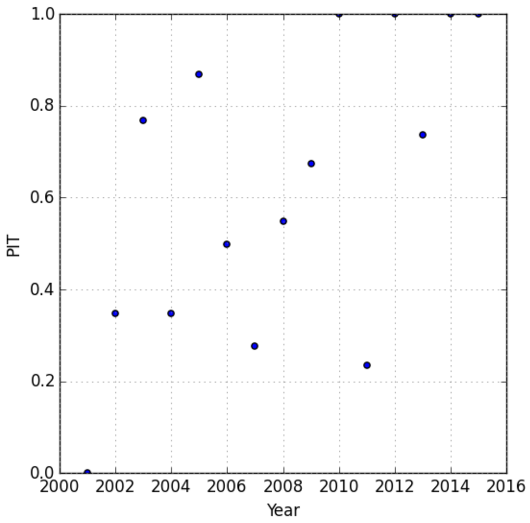


Figure 6: as in Figure 4 for BJP cross-validated forecasts for Jhelum at Mangla Kharif for SRM period of 2001-2015, except for (e) see 2001-2015 period of Figure 4 (e).

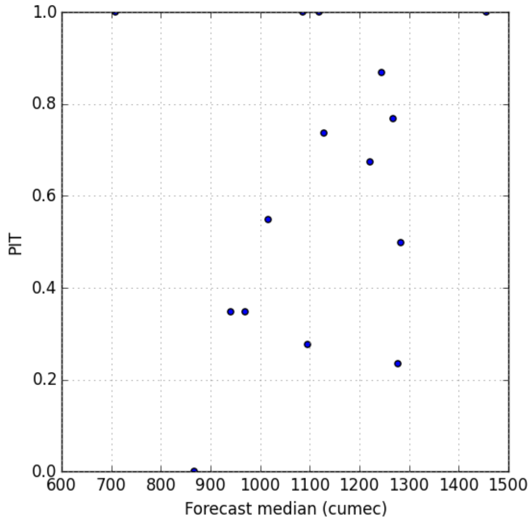
(a)


1.0
0.8
0.6
0.4
0.2
0.0
PIT
0.0
0.2
0.4
0.6
0.8
1.0
Uniform variate

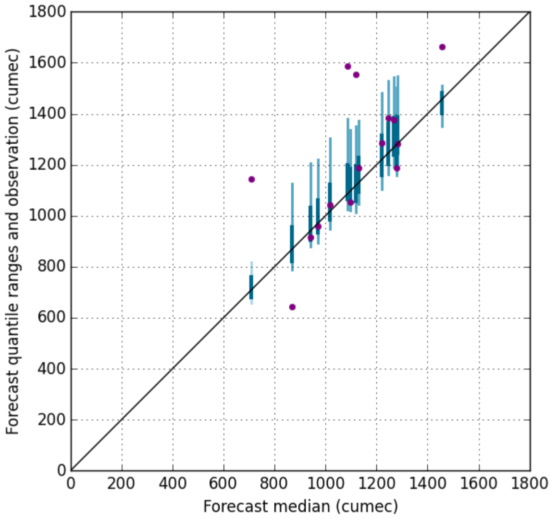
(b)


1.0
0.8
0.6
0.4
0.2
0.0
PIT
2000
2002
2004
2006
2008
2010
2012
2014
2016
Year

(c)


1.0
0.8
0.6
0.4
0.2
0.0
PIT
600
700
800
900
1000
1100
1200
1300
1400
1500
Forecast median (cumec)

(d)


1800
1600
1400
1200
1000
800
600
400
200
0
Forecast quantile ranges and observation (cumec)
0
200
400
600
800
1000
1200
1400
1600
1800
Forecast median (cumec)

(e)

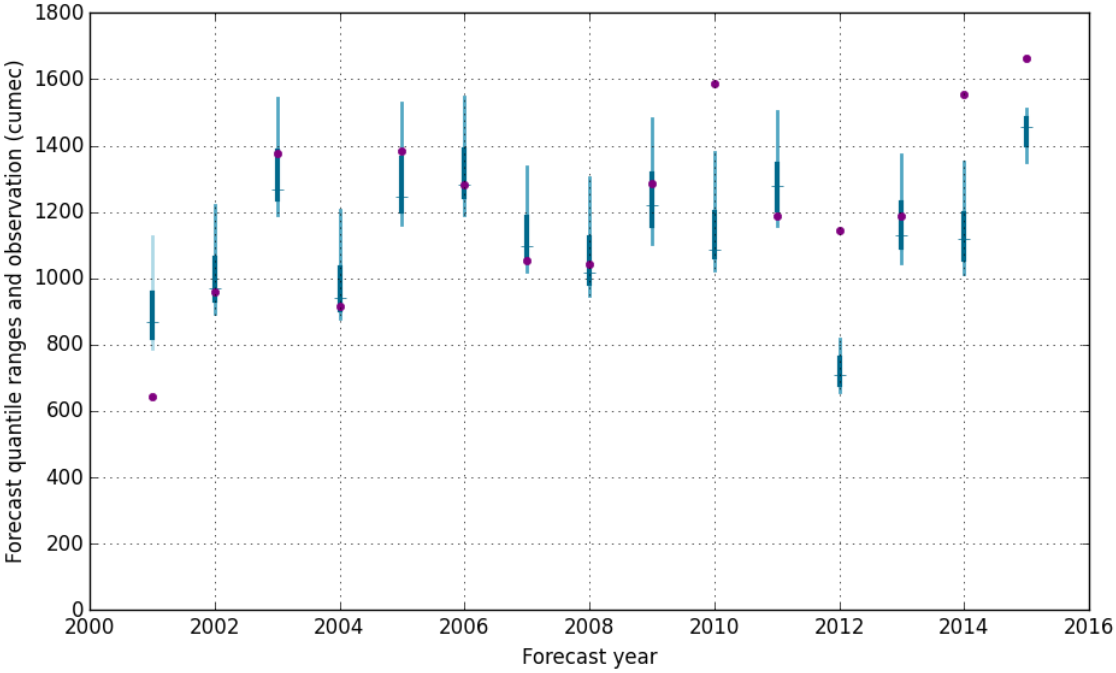


Figure 7: as in Figure 4 for SRM ESP forecasts for Jhelum at Mangla Kharif for SRM period of 2001-2015