Reply to reviewer comment hess-2017-340-RC2

Heiko Apel[1], Zharkinay Abdykerimova[2], Marina Agalhanova[3], Azamat Baimaganbetov[4], Nadejda Gavrilenko[5], Lars Gerlitz[1], Olga Kalashnikova[6], Katy Unger-Shayesteh[1], Sergiy Vorogushyn[1], Abror Gafurov[1]

[1]GFZ German Research Centre for Geoscience, Section 5.4 Hydrology, Potsdam, Germany

[2]Hydro-Meteorological Service of Kyrgyzstan, Bishkek, Kyrgyzstan

[3]Hydro-Meteorological Service of Turkmenistan, Ashgabat, Turkmenistan

[4]Hydro-Meteorological Service of Kazakhstan, Almaty, Kazakhstan

[5]Hydro-Meteorological Service of Uzbekistan, Tashkent, Uzbekistan

[6]CAIAG Central Asian Institute for Applied Geoscience, Bishkek, Kyrgyzstan

Correspondence to: Heiko Apel (heiko.apel@gfz-potsdam.de)

General referee comment:

This paper proposes to use standard multiple linear regression (MLR) to predict season streamflow for 13 catchments in Central Asia. The predictors are antecedent precipitation, streamflow, temperature, and snow depth. The different combinations of predictors are tested using MLR under the framework of leave-one-out cross validation (LOOCV) and using the metric of predicted residual error sum of squares (PRESS). At the end, “the best 20 forecast models” are picked out for the prediction of future streamflow. In general, the paper is well-written and the results are clearly presented. In the meantime, there are comments for further improvements of the paper:

First of all, it is widely known that the predictability of seasonal streamflow is generally from two sources, i.e., catchment storage and future climate [Hamlet and Lettenmaier, 1999; Chiew and MacMahon, 2002; Wood et al., 2002; Schepen et al., 2012; Crochemore et al., 2017]. However, in this paper, the predictors of future climate, which can be atmospheric circulation indices and GCM/RCM outputs, are not considered at all. That is to say, this paper only accounts for the predictability from catchment storage. As a result, the forecasts as are presented in this paper are not deemed “best” and they can be further improved. The authors are encouraged to consider circulation indices in seasonal streamflow forecasting. It is noted that NOAA provides a collection of more than 30 climatic indices (https://www.esrl.noaa.gov/psd/data/climateindices/list/).

We thank the reviewer for the constructive comments. We fully agree that the predictability of seasonal streamflow depends on the information about catchment storage and future climate, particularly rainfall. However, in Central Asia much of the discharge stems from snow melt, i.e. the winter accumulation, resp. the precipitation in winter. In the Altai catchments and along the Northern rim of the Tien Shan some additional precipitation occurs during spring and early summer (March-July).

This precipitation is eventually considered as observations in the late forecasts presented here. However, reliable information about the spring precipitation in advance could possibly improve the early forecasts. We actually studied the seasonal predictability of precipitation in Central Asia using NAO, ENSO and EA indices as well as automatically selected seas surface temperature regions as predictors in a preceding paper (Gerlitz et al., 2016). Although some skillful models were obtained for winter precipitation, the variability of the seasonal precipitation in Central Asia was strongly underestimated. Furthermore, summer precipitation in Central Asia is usually convective, i.e. is triggered by surface heating and associated atmospheric instability. Summer precipitation sums are composed of few single events (occasionally of high intensity) which, however, are rather randomly distributed and non-predictable. Therefor we did not include the seasonal forecasts of precipitation in the presented linear models, because no additional gain in performance can be expected. Another (practical) reason for this decision was the envisaged operational use by the CA hydromet services. Using station data as presented is fairly easy for them to include in the operational routines, while using climate indices could pose a mental as well as technical barrier for the staff in the services.

Regarding the term “best” we of course refer to the best forecast obtained with the presented approach and predictors. We do by no means imply that the presented models are the best forecast obtainable in general and will make this clear in the revised manuscript.

Second, the analysis of predictive uncertainty is too simple to be informative in this paper. It is pointed out that for ensemble and probabilistic forecasts, the attributes of reliability and skill are of key importance [Murphy, 1993, What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting]. Reliability can be diagnosed using the PIT reliability diagram or PIT histogram [e.g., Wang et al., 2009; Crochemore et al., 2017]. Meanwhile, Skill can be measured using the continuous ranked probability score (CRPS), which is for both deterministic and ensemble forecasts and is equivalent to the mean absolute error (MAE) for deterministic forecasts [Hersbach, 2000]. In addition to the illustrative plots of predictive uncertainty, the authors are encouraged to perform a comprehensive examination of forecast reliability and skill.

Many thanks for the suggestion. Because we are using deterministic models, we have now evaluated the skill in terms of the MAE and plotted it in Figure 4. The MAE is normalized to the mean seasonal discharge for each basin, just as already shown for the RMSE. The MAE skill is very similar to the RMSE, being in the range of 10%-20% for the January forecasts, and below 10% for the most important April forecast:

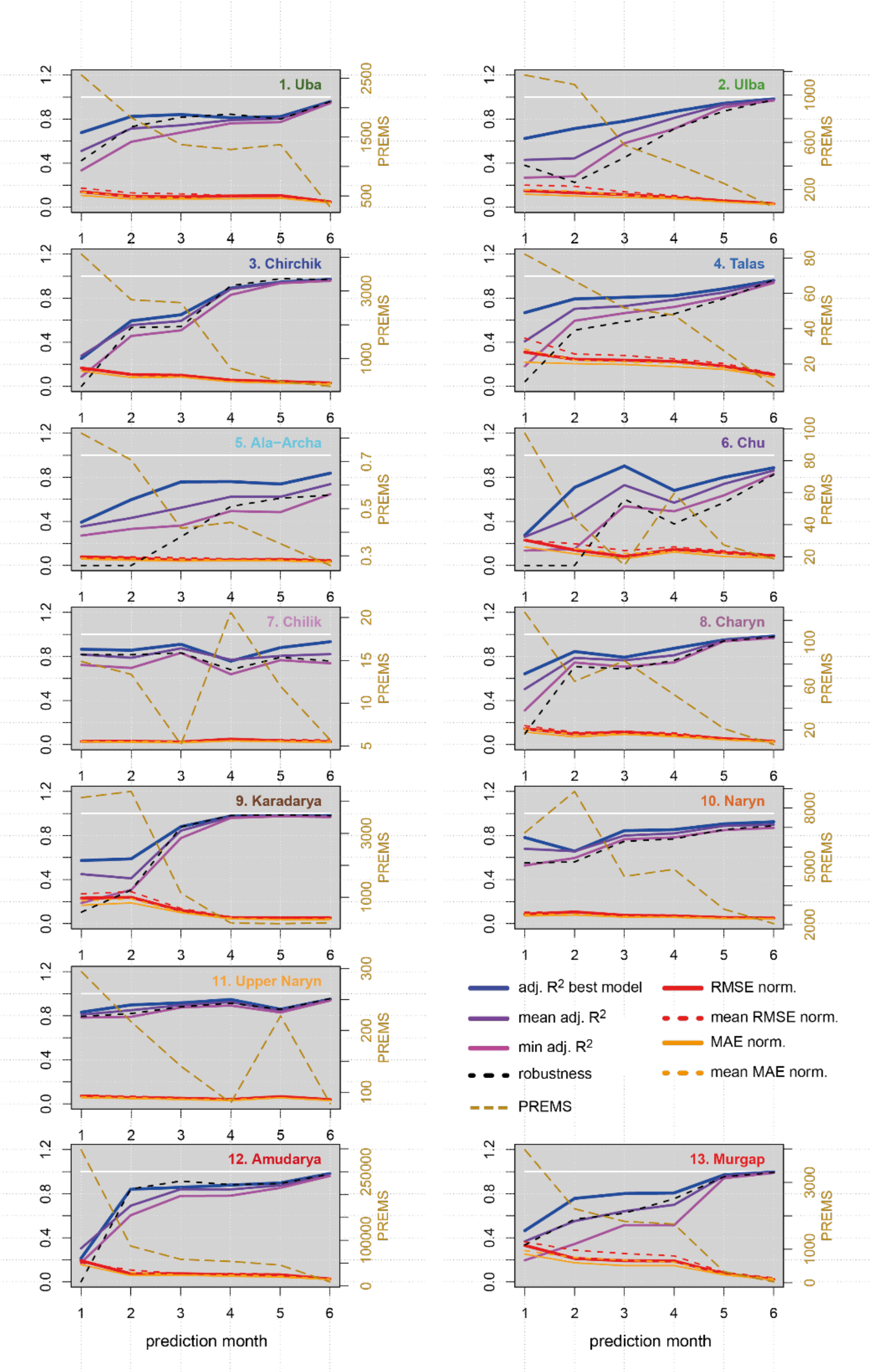


Figure 4: Performance of the prediction models for the different catchments and prediction months. Adj. R^2 best model is the adjusted R^2 of the single best LOOCV model, mean adj. R^2 is the mean adj. R^2 of the best 20 LOOCV

models, min adj. R^2 is minimum adj. R^2 of the best 20 LOOCV models, robustness is mean LOOCV-adj. R^2 of the best 20 models divided by the mean adj. R^2, RMSE/MAE norm. is the root mean squared error/mean absolute error of the single best model normalized to mean multi-annual seasonal discharge, mean RMSE/MAE norm is the mean root mean square error/mean absolute error of the best 20 LOOCV models normalized to the multi-annual seasonal discharge; PREMS is the predictive residual sum of squares (PRESS) of the single best model, divided by the number of prediction months.

We also evaluated the reliability by means of PIT diagrams, as suggested. The plot below shows the PIT diagrams for every catchment and all forecast months using the prediction of the selected ensemble models. The PIT diagrams show that the model ensemble predictions are in most cases close to the 1:1 line, i.e. provide reliable forecasts. However, in some cases the predictive uncertainty is under-estimated, i.e. the predictive uncertainty bands presented in Figure 6 are too narrow. We further calculated a PIT score as the area between the PIT curve and the 1:1 line as a summarizing indicator for the reliability. The theoretically least reliable model has a score of 0.5, a perfect model a score of 0. The highest score, i.e. the lowest reliability, of all models is 0.2, with the majority of the models being in the range of 0.07-0.15. Interpreting the scores with the curves in the PIT diagram it can be stated that the reliability of the models is good for PIT scores <= 0.1. For higher scores the predictive uncertainty is likely to be underestimated. We will include this analysis in the revised manuscript as suggested by the reviewer, and provide the PIT scores as guidelines for the interpretation of the predictive uncertainty bounds.

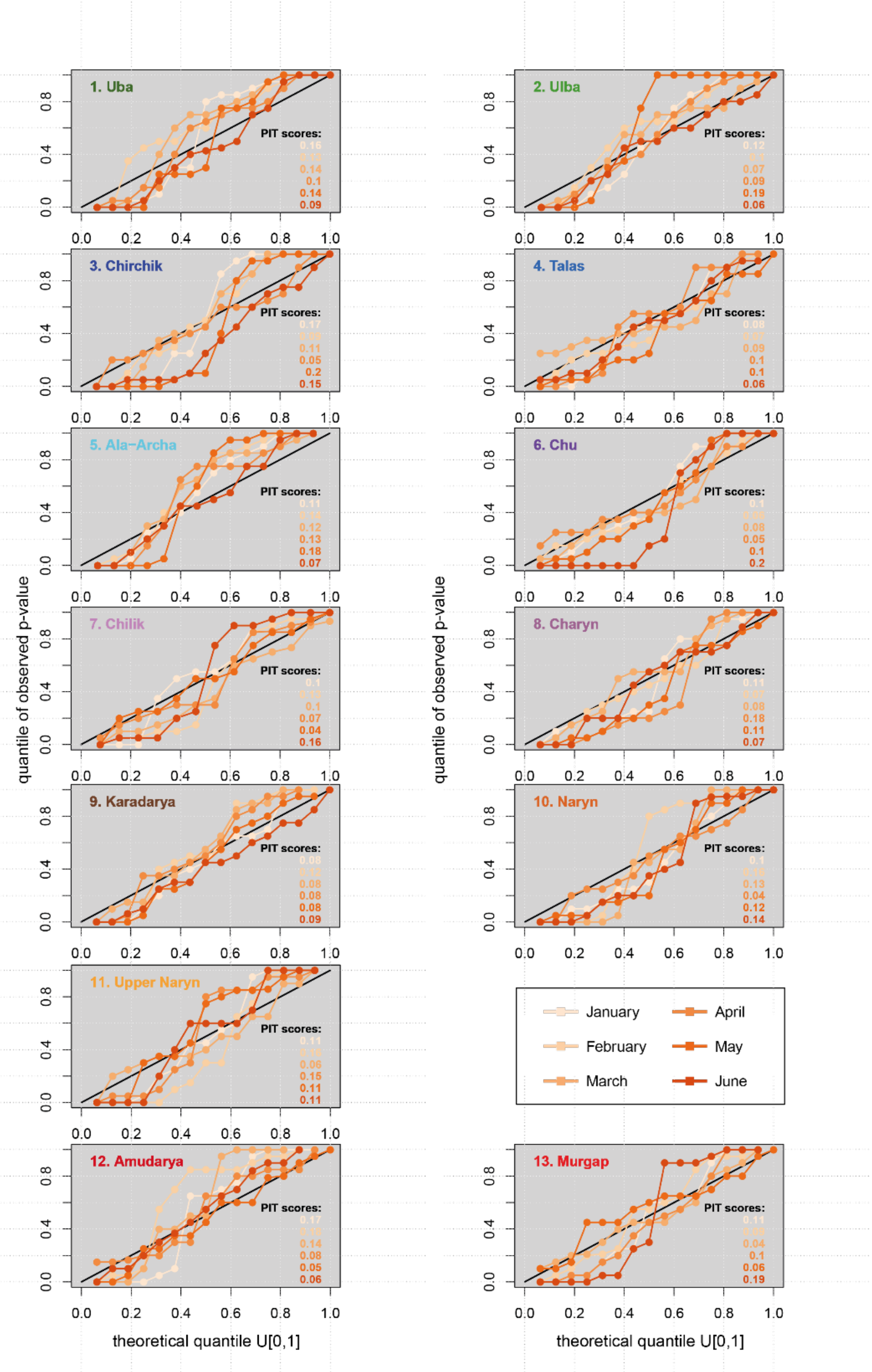
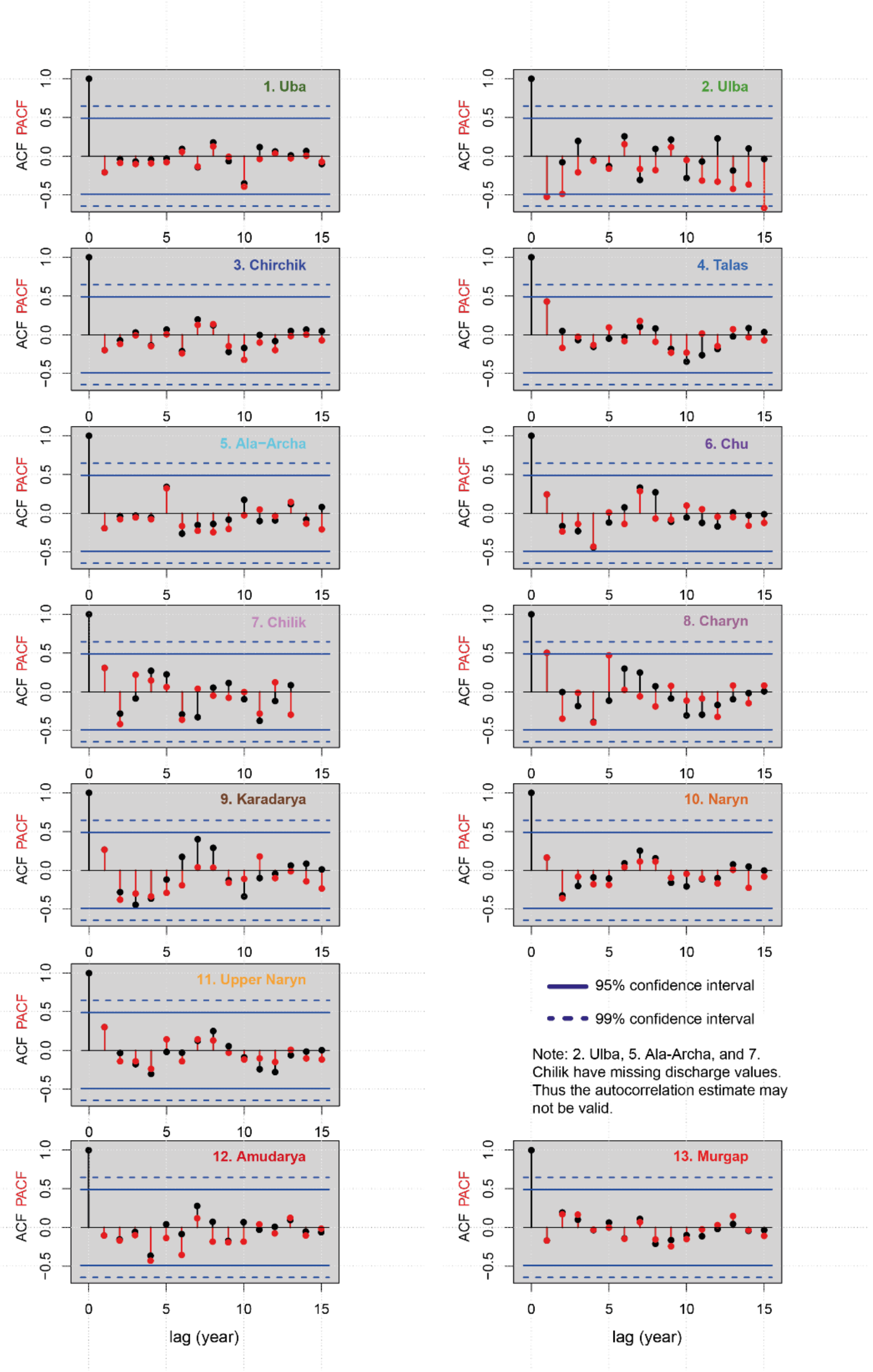

1. Uba
PIT scores:
0.16
0.13
0.14
0.1
0.14
0.09
2. Ulba
PIT scores:
0.12
0.1
0.07
0.09
0.19
0.06
3. Chirchik
PIT scores:
0.17
0.09
0.11
0.05
0.2
0.15
4. Talas
PIT scores:
0.08
0.07
0.09
0.1
0.1
0.06
5. Ala−Archa
PIT scores:
0.11
0.14
0.12
0.13
0.18
0.07
6. Chu
PIT scores:
0.1
0.06
0.08
0.05
0.1
0.2
7. Chilik
PIT scores:
0.1
0.13
0.1
0.07
0.04
0.16
8. Charyn
PIT scores:
0.11
0.07
0.08
0.18
0.11
0.07
9. Karadarya
PIT scores:
0.08
0.12
0.08
0.08
0.08
0.09
10. Naryn
PIT scores:
0.1
0.16
0.13
0.04
0.12
0.14
11. Upper Naryn
PIT scores:
0.11
0.16
0.06
0.15
0.11
0.11
12. Amudarya
PIT scores:
0.17
0.18
0.14
0.08
0.05
0.06
13. Murgap
PIT scores:
0.11
0.09
0.04
0.1
0.06
0.19
January
February
March
April
May
June
quantile of observed p-value
theoretical quantile U[0,1]
0.0
0.2
0.4
0.6
0.8
1.0

Figure: PIT reliability diagrams for every catchment and forecast month. The PIT score is calculated as the area between the reliability plots and the 1:1 line as suggested in Renard et al. (2010). The lower the PIT score, the higher the reliability. The least reliability score is 0.5, the best 0.

There are also some minor comments:

1. As for LOOCV, it can lead to artificial over-estimation of forecast skill if the streamflow series exhibit strong auto-correlation. It is worthwhile to check the serial autocorrelation of streamflow. Or, a more rigorous leave-five-years-out cross validation (L5OCV) ought to be applied.

We checked the autocorrelation and partial autocorrelation of the streamflow time series and plotted it in the figure below. Hardly any autocorrelation at p = 0.05 could be detected. Only for 2. Ulba the partial autocorrelation shows some autocorrelation for lag 1 and 2 just above p = 0.05. But in summary for all catchments, it can be stated that autocorrelation does not exist in the discharge time series, and thus the proposed LOOCV is an appropriate validation method. We propose to include the figure below in the appendix of the revised manuscript and include the statement above in the text.



Note: 2. Ulba, 5. Ala-Archa, and 7. Chilik have missing discharge values. Thus the autocorrelation estimate may not be valid.

2. In terms of predictors of catchment storage, the use of multi-monthly means as the predictor values is sensible.

Thanks for the supporting comment.

3. The paper suggests to use the “the best 20 forecast models”. This setting is empirical and it is rare in peer studies. Please clarify why.

We commented on this already in the reply to reviewer 1, thus we quote the reply here:

The number of models for the ensemble is set subjectively to 20. This selection is aiming at obtaining a sufficient number of models for an ensemble evaluation of the forecasts. With the newly set restriction on model selection (only models with significant predictors), a few ensembles, particularly for the January prediction have less than 20 models, because not enough models fulfilling the new selection criteria could be identified. There is actually no rule for the number of ensembles members applied. We left sufficient amount of freedom for this, in order to enable an expert selection of models by the forecasters of the Central Asian hydromet services. The forecasters have a lot of experience with their catchments, and can decide better which forecasts are valuable for them. The forecasters check every model retained for their performances (quantitatively and qualitatively), and select the models accordingly. This means that in practice fewer models than the 20 presented in the manuscript might be selected, or even more.

Another possible rule for ensemble model selection could be a defined threshold of R^2 for the model. However, due to the high explained variances, the threshold must be very high in order to reduce the number of ensemble members. A fixed R^2 threshold would more likely increase the ensemble members in most cases. The selection of the threshold level would also be subjective.

References:

Gerlitz, L., Vorogushyn, S., Apel, H., Gafurov, A., Unger-Shayesteh, K., and Merz, B.: A statistically based seasonal precipitation forecast model with automatic predictor selection and its application to central and south Asia, Hydrol. Earth Syst. Sci., 20, 4605-4623, 10.5194/hess-20-4605-2016, 2016.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resources Research, 46, 10.1029/2009WR008328, 2010.