

The authors would like to thank the editor and two anonymous referees who kindly reviewed the earlier version of this manuscript and provided valuable suggestions and comments. We present our response to the comments and suggestions from the reviewers.

Referee 1

This research article develops a multi-site seasonal rainfall and streamflow forecast for the Huai River Basin in China. It is refreshing to see Bayesian statistics, as opposed to frequentist statistics, used to explore teleconnections between sea surface temperatures and rainfall and streamflow. The approach appears to tick all boxes required of Bayesian statistics, such as estimating the prior and posterior distribution using well established software and methods. The plot of the posteriors (coefficients) shows nicely the spatial variations between stations and the model covariates and give a sense of parameter uncertainty in the model making this a transparent forecasting method. The m-fold cross validations also builds confidence in the forecast skill to a degree. While the paper is nice and compact a more in-depth discussion of the results is important. The following thoughts may provide a starting point:

As is usual with all forecasting approaches, the extremes are not captured well by this forecast approach. This is particularly evident for streamflow at Bengbu for 1997, 1999, 2003 and 2009 (Figure 5). This warrants more discussion considering that this station is a “best performer” looking at the skill scores (Figure 7). Importantly it is these outliers that are most important to forecast. Why is this method missing these events? Here are some thoughts on what this discussion may contain;

- 1) Is a log transform of the original data appropriate and/or does the back transform need to be bias corrected
- 2) Could a hierarchical generalized linear model assuming gamma distributed error structure be more appropriate considering the original data is skewed

We tested the assumption that the log-transformed data is normally distributed and gamma distributed using the K-S test. The null hypothesis that the log transformed data follows normal distribution or gamma distribution cannot be rejected at the 5% level. Also, the p-value corresponding to the log-transformed flows is lower than the p-value from the gamma distribution. One could conduct a formal hypothesis test to compare the p-values from log-normal and gamma. However, given small data sample this could mean very little differentiability between the two model assumptions. We choose lognormal over gamma here for simplicity. Residual analysis based on the quantile plots and skewness tests showed that the lognormal assumption is valid for much of the stations barring a few. Table 1 here shows the K-S test results for log-normal and gamma distribution assumptions conducted for all the stations. Figure 1 here shows the quantile-quantile plots of the residuals from the partial pooling Hierarchical model.

We suspect that a) the predictors we are using may not capture the full range of the response, b) the marginal distribution of the hydroclimatic variables in the region may correspond to a finite mixture of processes, and hence the right and left tail probabilities may reflect processes that are different, and c) the relationship between the predictors and the predictands may be nonlinear, especially as it relates to extreme values. Given these possibilities, the model presented here is a relatively crude approximation of reality, and yet it is generally successful in terms of the usual skill scores and criteria that are presented. The reviewer correctly notes that this kind of result is not atypical. However, resolving these issues with the short record is not always easy. Our emerging work with Gaussian Process models suggests that there is promise in developing that methodology to address some of these issues, in a Bayesian framework. However, we feel that the current application is of interest in its own right, both for the region in China and for the clear results and application of the method

We have added some of this discussion in the manuscript now.

Table 1: p-values for the K-S test.

Station #	p-value	
	Log Normal	Gamma
1	0.1961	0.004727
2	0.4196	0.952
3	0.1939	0
4	0.66	0.9141
5	0.6863	0.934
6	0.7649	0.81
7	0.7795	0.6969
8	0.8443	0.9914
9	0.4189	0.535
10	0.6494	0.8872
11	0.8187	0.8997
12	0.9429	0.9369
13	0.454	0.6368
14	0.7515	0.9426

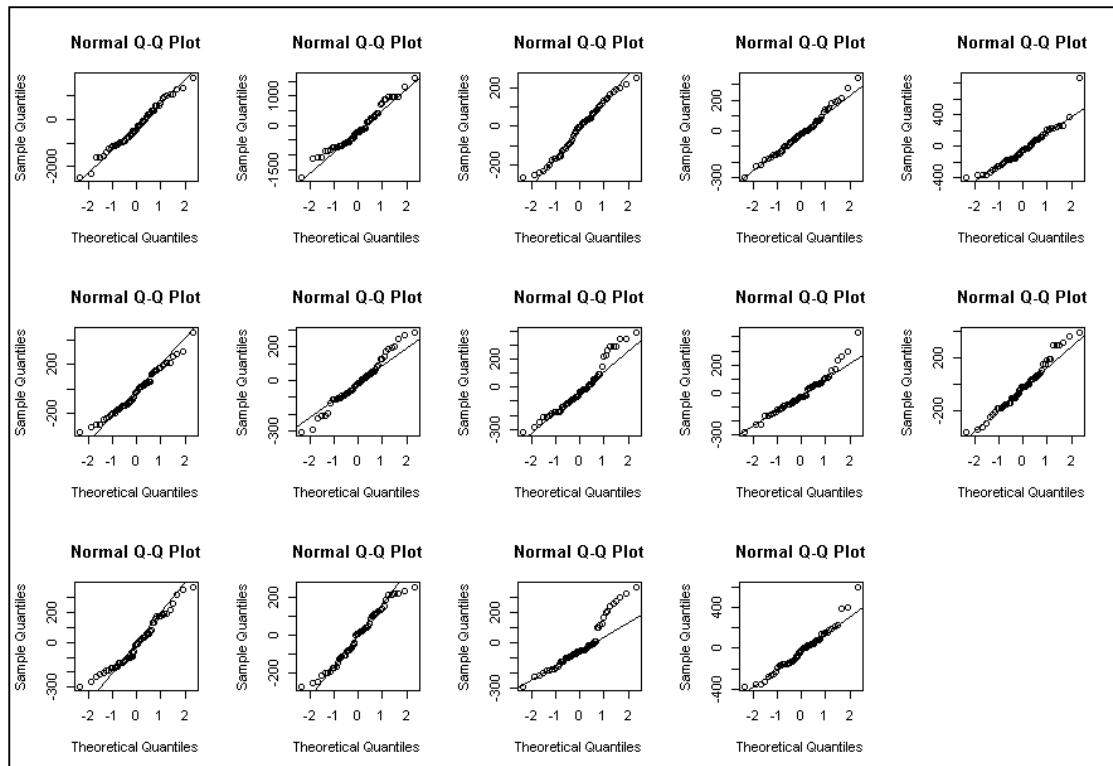


Figure 1: qqplots for the residuals.

- 3) Is this approach confirming the milt of the predictability of rainfall and stream using SSTs

We included this discussion on the manuscript now under the summary and discussion section.

- 4) Is this shortfall a major hurdle for incorporating forecasts in dam management?

Possibly. Decision makers are often influenced by the ability to correctly indicate extreme conditions since the losses from their operations are most sensitive to such states. In our interactions with corporate and public sector decision makers, we have noticed both skepticism induced by failure to predict extremes, even if all performance measures are good, and conversely enthusiasm for the model on noting that the directional (high, average, low) forecasts are quite good. This dichotomy seems to reflect the risk aversion of the decision maker in an interesting way – the first example shows risk aversion to using a new product in lieu of standard operating procedure, where nature can be conveniently blamed. The second example shows a risk aversion to loss of reputation and economic loss incurred by not using a forecast product in a conservative way. We find the latter case to be a bit more common in the private sector. Given that our experience with these cases is based on anecdotal evidence rather than a formal study, we are not sure if we can include it in this paper. Perhaps developing a formal analysis into a paper is a better choice, and to leave the focus of the current paper on the

development and performance of a partial pooling Bayesian approach in this context.

- 5) Further discussion is also warranted on why there is an inconsistency in the skill scores (station 14 is pointed out). Though this confirms the need for multiple skill scores, why not dig out why there is variation in the skill scores. Could this reveal more on forecast performance?

Two metrics (Reduction in Error (RE) and Coefficient of Efficiency (CE)) we used in this study measures the goodness of fit of the model; i.e. it compares forecasted streamflow (left out sample) with the actual streamflow data. These metrics are used as an expression of the true R^2 of the regression equation when applied to new data. By assessing the RE and CE under cross-validation, we are essentially providing a measure of the variance explained under validation dataset.

We also used Rank Probability Skill Score (RPSS) to quantify the error in estimating the entire conditional distribution. Whereas the RE and CE measure the forecast skill in conditional mean, the RPSS measures the skill in the probability distribution each year. RPSS is methodologically different from RE and CE, and we agree with the reviewer that one needs multiple verification techniques in evaluating the forecast. In this context, RE and CE impose greater penalty on the few outliers that cause the skew in some stations while RPSS imposes less penalty by evaluating the entire categorical forecast.

- 6) Is there persistence in streamflow from one wet season to another and was this considered?

The following figure 2 shows the autocorrelation function of the flows and rainfall at the two stations. There is no year to year persistence in the flows or rainfall.

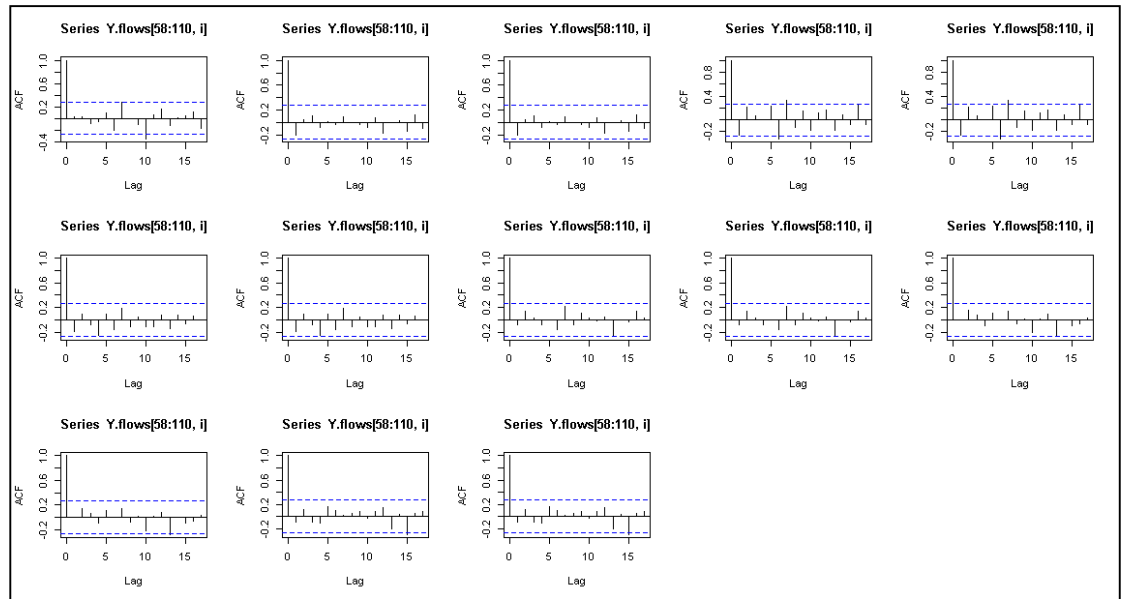


Figure 2: ACF plots for the flows and rainfall.

- 7) Is the streamflow unregulated and if not, how was this accounted for?
Bengbu station is regulated to some extent upstream, but the Lataizu station is not. Given the difficulty is obtaining data from this part of the region, we did not proceed toward developing naturalized flows ourselves. Upstream regulations could affect the daily to monthly flows more so than the seasonal flows that we are using to build the model.

Referee 2

This paper presents a two-level Bayesian model aimed at forecasting multisite rainfall and stream based on exogenous climatic conditions. The methodology and its application to the Huai River catchment are well introduced. A suite of interesting results are presented and the main finding is that “the seasonal forecasts developed using climate precursors contain useful information”. I agree on the general merits of the method which was already thoroughly discussed in a peer reviewed publication. However, I am not convinced that the application of this approach to the Huai river catchment leads to robust findings that are valuable enough to inform the research community or regional forecasting practices. This is mainly because of the lack of an in-depth discussion in which the authors are expected to explain the identified uncertainties and the station similarity using their best understanding of the underlying physical processes. The structure of this manuscript is also confusing at several places. I therefore suggest a major revision.

Thanks for the detailed comments. The manuscript is now revised with the suggestions from both the reviewers and we hope that the discussion section that is more appropriate.

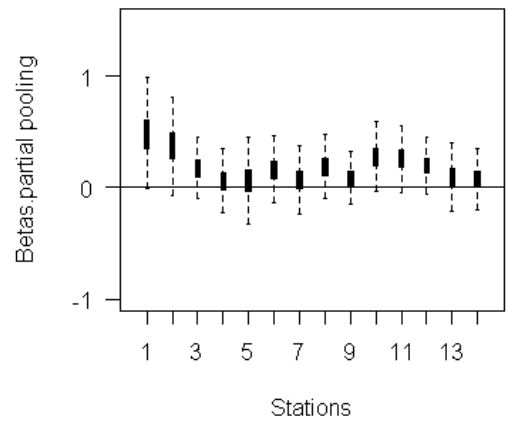
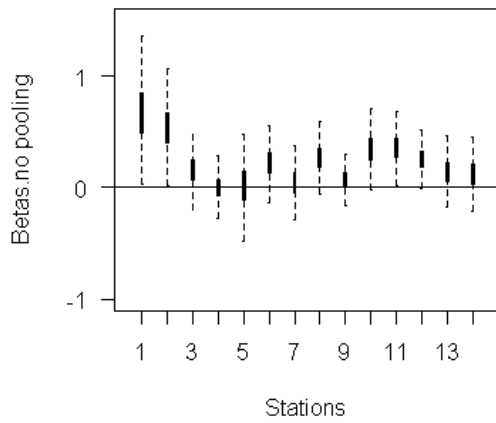
- 1) Figure 7 indicates some consistent patterns across different metrics. Could this be related to the geographic locations of these sites? To facilitate a spatial assessment, I suggest using the same site labels (either numbers or abbreviations) for all figures.

The figures and the tables now have consistent numbers across. The similarity in the skill scores across different stations are a manifestation of the similar stations. For instance, station 1 and station 2 are streamflow stations at the mouth of the river. Their correlation with predictors, and their skill scores are similar, further, supporting our rationale for spatial pooling in the hierarchical model. Similar neighbor stations exhibit similarity in the response and the corresponding model fits. Noting that the correlation of a given predictor is very similar across the stations, the partial-pooling model allows for appropriate information sharing across the stations to reduce the associated uncertainty. This discussion is now improved in the manuscript.

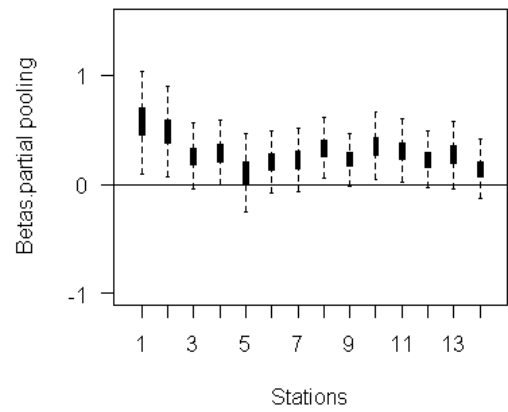
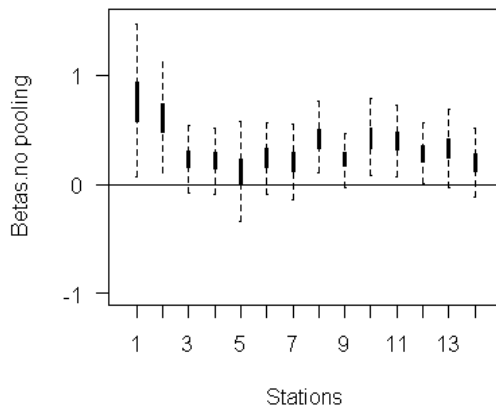
- 2) Compared with the other forecasting studies that involve the same performance metrics or similar climate variables used in this study, how is the forecasting performance of this study?

For a like comparison of the model performance, we show the improvements of the partial-pooling hierarchical model over the no-pooling model here. The no-pooling model is the traditional way of implementing the regression analysis in forecasts, but estimated using a Bayesian scheme. There is clear benefit in implementing the partial-pooling over the traditional no-pooling regression in reducing the uncertainties in the model parameters (the figure 3 and 4 below shows the comparison of regression coefficients of three predictors from both the model. We can see that the partial pooling has reduced uncertainty in estimating the model parameters) and the posterior distribution (see figure on the 50 years of forecast distribution compared between the two models. In general the prediction uncertainty is reduced, and the RPSS (measure of entire conditional distribution) is greater for the partial pooling model). The no-pooling regression model can be compared to the one used in Kwon et al (2009). However, we are not aware of any group that has conducted seasonal forecasting for the entire Huai river catchment; hence a basin to basin comparison is not possible here.

Predictor: SST1



Predictor: SST2



Predictor: AMO

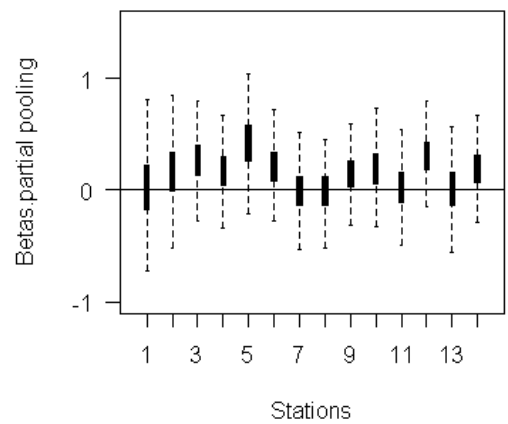
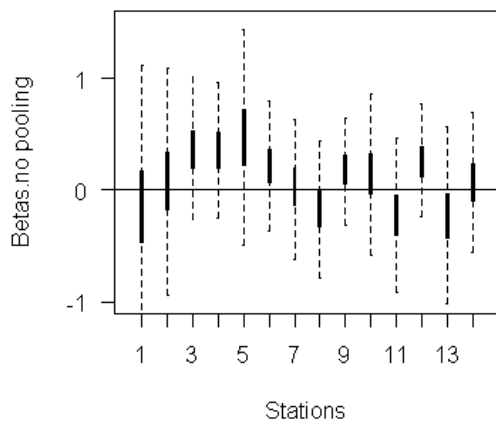


Figure 3: Comparing the partial pooling model to traditional no pooling method for regression coefficients.

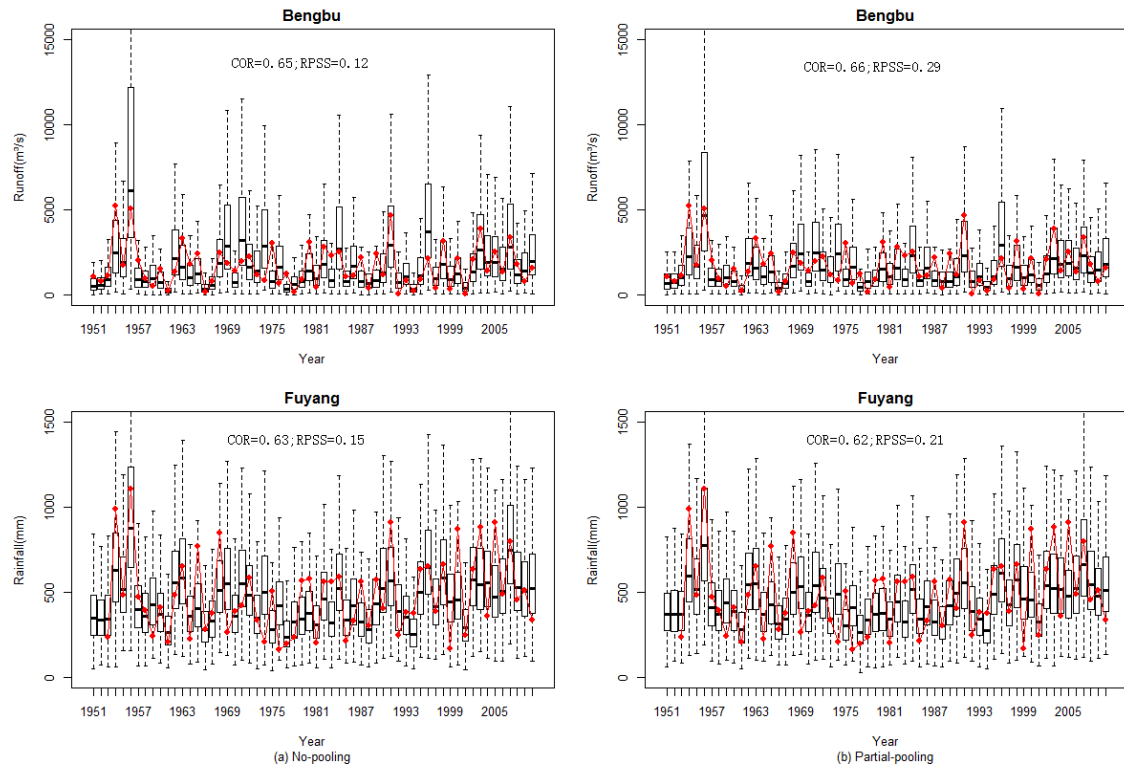


Figure 4: The posterior probability distributions for JJA averaged streamflow and JJA total rainfall (one specific streamflow at Bengbu station and one specific rainfall at Fuyang station) from the no-pooling (a) and the partial-pooling (b) models with observed values in the period of 1951-2010 (red line with dot stand for observed values, correlation (COR) and RPSS calculated by median estimates and observed values)

- 3) What is the most important information that this study can offer for real-world forecasting practice?

Hierarchical Bayesian models and multilevel models are now popular in the computational statistics field. Applications include causal inference, prediction, and comparison for multivariate problems where capturing the group or spatial behavior is warranted. They directly generalize the traditional regression approaches. As mentioned in section 4, the ability to model individual rainfall stations with a few aggregate streamflow stations opens up the discussion for spatial scaling and regional regularization and its utility for basin wide forecast with reduced uncertainty. There are very few groups that are working on developing unified statistical model forecasts that are space and time consistent for larger spatial domains. The methodology can be readily applied in many parts of the developing world where the data availability is sparse or incomplete. Moreover, for any forecast, the residual risk or the forecast uncertainty is represented in the decision making process. A model with reduced posterior uncertainty reduces the false alarm rate and increases the confidence in using the forecasts.

- 4) The description of three performance metrics accounts for a large portion of

section 4. This content should be moved to the methodology section.

As per the suggestion, we moved the description of the performance metrics to a new sub-section under methodology.

- 5) Section 5 is just a short summary with brief thoughts on future work. Nothing is really discussed here.

The manuscript is now improved with discussion.

- 6) Figure 1: No need to show the entire Huai River Basin, as the study area is the Huai river catchment only.

We modified this according to the suggestion.