# Review comments from Mark Thyer

**Comment #1.1.** General Comments This paper presents a modification to existing approaches for handling autoregressive errors in streamflow modelling in a forecasting context. I applaud this paper for undergoing a detailed analysis of the issues that are encountered when endeavouring to deal with both heteroscedasticity and autocorrelation in hydrological modelling errors. Something which we think should be straightforward, but is actually quite challenging to get right. The paper is fairly well written, but needs some improvement (see minor issues). The results presented, while quite promising, are currently not sufficiently convincing to warrant publication. Please see the list of major issues below. These issues need to be addressed prior to publication.

Response: Thank you for the careful and constructive review. We have attempted to address the major issues you have raised, while keeping the paper as brief as possible.

Major Issues

**Comment #1.2.** More metrics are required to verify performance.

Currently the three methods, AR-Norm, AR-Raw and RAR-Norm are evaluated by visual inspection of a few events and using the NSE as an evaluation criteria. A wider range of metrics is needed. In a forecasting context, it is not simply the NSE which is used to evaluate predictions, users are also interested in the statistical properties of the predictive streamflow distribution, such as reliability and precision. It is common for these metrics to also trade-off against one another, so it would be interesting to see if that occurs in this case. Furthermore, the NSE is heavily weighted towards better predictions of high flows. It is recommended that authors use metrics that evaluate the full predictive streamflow distribution and use precision and reliability metrics, such as they have used in past, e.g. Wang and Robertson [2011] or see for example Evin et al. [2014].

Response: Thanks for this suggestion. We have added a number of metrics to bolster our conclusions, including the probabilistic verification scores CRPS (which measures both accuracy and reliability), RMSEP (which measures accuracy of forecast in probability) and PIT-Uniform probability plots to assess reliability. These show that there is little to distinguish between the three models with probabilistic measures; all show similar accuracy and reliability (though again, RAR-Norm tends to produce slightly better CRPS and RMSEP skills scores than the other models.) In addition, we analyse the NSE of forecasts when flows are rising and falling. These analyses confirm the general tendency of the AR-Norm model to perform least well when flows are rising, and the tendency of the AR-Raw model to perform least well when flows are receding. In addition, these analyses show that the RAR-Norm model reflects the best tendencies of the AR-Raw and AR-Norm models.

**Comment #1.3.** Robustness of the results with respect to the hydrological model.

Line 20 page 6044 makes the point that AR-Raw performs better than AR-Norm and state “this suggest that more robust performance can be expected of base hydrological models with AR models are applied to raw errors”. Sectio, 4.2 is devoted to discusses that the AR-Norm model, produce poor performance of the hydrological model. However, this is based only a single hydrological model, GR4J. When Evin et al. [2014] applied an equivalent to the AR-Norm model (but with linear heteroscedatic errors, rather than log-sinh transformed) to the 12 MOPEX catchments they found similar poor model performance for GR4J for some catchments, but this did not occur when the HBV model was applied. This provides strong evidence that the problems with ARNorm is not necessarily generic, but hydrological model specific. It is recommended that the authors trial a different hydrological model, e.g. HBV, and see if the results are similar. If they are, then this provides a greater robustness of the model results, and greater confidence for the hydrological community to adopt this method.

Providing more metrics with a wider range of hydrological models would be better test the extent of the problems with AR-Raw and AR-Norm and the robustness of the results. For example, Figure 3, shows the error over-correction problem with AR-Norm occurs in only 10-20% of cases, which is not very high. Given also that the poor performance of the AR-Norm method is hydrological model specific, further testing and metrics are required to verify the robustness of the proposed approach.

Response: We concede that other rainfall-runoff models may not be as prone to poor base model performance as GR4J. We have stopped short of investigating additional hydrological models however, to keep our paper brief. We address the reviewer's concern as follows:

1) We now explicitly acknowledge that the sometimes poor performance of the base hydrological model may be particular to GR4J
2) Adding some of the MOPEX catchments used by Evin et al. 2014 (see response to Comment # 1.4, below) has allowed us to draw more directly from Evin's work, which suggested that HBV could lead to more robust base model performance. We refer to this study explicitly when we discuss the performance of the base hydrological model
3) Because the RAR-Norm model restricts the magnitude of updates that can be applied by the AR-Norm model, more reliance is placed on the base hydrological model to accurately simulate flows. This will generally encourage the base hydrological model to perform strongly compared with the AR-Norm, irrespective of the hydrological model used. If the base hydrological model is already performing strongly (as might be expected, e.g., of HBV) then the RAR-Norm model is unlikely to undermine this performance. We see evidence of this in our experiments with GR4J (which we know can perform poorly): when the performance of the GR4J base hydrological model is strong relative to the updated forecasts for both AR-Raw and AR-Norm models (e.g. in the Tarwin, Mitta Mitta, or Guadalupe catchments), the RAR-Norm model base hydrological model also performs strongly. In other words, if the problem does not exist in the other models, RAR-Norm does not introduce it.

The arguments above are now covered in the discussion (Lines 418-437).

As noted in the response to Comment #1.2, we have added more metrics and analysis, as well as three extra catchments, and we hope that these demonstrate that the RAR-Norm model is preferable to both the AR-Norm and AR-Raw models in general. As we show in the proof in the Appendix, and argue in the discussion, the potential of the AR-Norm model to over-correct rising flows is likely to be generic (irrespective of hydrological model or transformation applied). In addition, while you are right in saying that the AR-Norm model is susceptible to over-correction for as little as 10% of flows, it is often these instances – when flows are rising rapidly – that are of most interest to forecasters (e.g., for forecasting floods). We therefore argue that the problem of over-correction by the AR-Norm model is a salient one and that the RAR-Norm model addresses this problem successfully.

**Comment #1.4.** 3. Ability to compare results with previous studies. This is more a general comment of an issue which is a common blight for the progress of the hydrological scientific community. One of the big challenges for reviewers (and readers in general) is the ability to compare results between different studies, due to differences in implementation. As an example, Evin et al. [2013] showed that the equivalent to AR-norm was better than AR-Raw, while Evin et al. [2014] showed that AR-Norm can degrade hydrological model performance for GRJ, but not HBV. While Schaefli et al. [2007] showed that AR on raw errors lead to better inference, while this study showed a AR-hybrid (norm and raw) (see minor comment 3) works better than both AR-Norm and AR-Raw. However in all these studies, there are differences in their approach and case study application. For example, Evin et al. [2013,2014] used a linear heteroscedastic residual error model, Schaefli et al. [2007] used a mixture of Gaussians for their error model, while this study used a log-sinh transformation with modification for zero flow occurrences. Furthermore, each study had a different set of case study catchments. It concerns me that the conclusions of each of these studies could be sensitive to these differences rather than differences in the way the AR is handled, and it makes it very difficult for hydrological science to

move forward. This is the reason why Evin et al. [2014] choose to use the MOPEX dataset, as it least provides a common set of catchments to previous studies. I would suggest to these authors to include the 12 MOPEX catchments as used by Evin et al. [2014] to enable better comparison. This is not an essential criteria, but it would increase the ability to compare the results, and test its compare robustness against previous results.

Response: We agree that comparability of results is highly desirable. To this end, we have included 3 of the catchments used by Evin et al. [2014], and specifically note that these are chosen for the purposes of comparison to that study. In addition, we apply the same cross-validation strategy as Evin et al. 2014 to these catchments, to enable direct comparison to Evin et al.'s findings. We did not use Evin's remaining 9 catchments, for the simple that these are all impacted by snow, and this was not the focus of our study. We discuss the results of the three US catchments with reference to Evin et al. 2014. We find that the additional of the US catchments supports our initial findings, and thank the reviewer very much for this suggestion.

Minor Issues

**Comment #1.5.** Page 6039 Line 20-25. The assertion that these equations represent the median needs further derivation (perhaps in an appendix), as it is not clear to me. For example, the error term e(t) is completely dropped from eqs 4 and 5. This assumes that median of Z-1(et)=0, now median(et)=0, but, I'm not convinced that median of Z-1(et)=0, due to the use of the log sinh transformation which takes into account zero flow occurrences.

Response: Thank you for reading our manuscript so closely. Following your suggestion, we explain why the updated streamflow is the median of the ensemble streamflow forecast in Appendix A.

**Comment #1.6.** Page 6045, Eq(8). It is very confusing using the subscript (R) for both AR-Raw and AR-Norm. Please use a different subscript for RAR-Model

Response: We have carefully and thoroughly updated the notations and avoided the use of the subscript (R).

**Comment #1.7.** RAR model is essentially a hybrid of AR-Norm when it over-corrections, use ARRaw. Suggest to change name of RAR_Norm to AR-hybrid. Also, why did the authors choose not use the phi term, i.e. Q(s,t)+ phi*[Q(t-1) – Q(s,t-1)] in last line of eq 8. Some justification of this is needed.

Response:

While the RAR-norm uses errors calculated from transformed and untransformed flows, it is not a formal combination of the AR-Norm and the AR-Raw models. This is because we do not apply a rho term to the error in the untransformed domain when we apply the restriction. In addition, the model is conceptually much more similar to AR-Norm, and indeed the model functions as an AR-Norm model for the large majority of the time. Accordingly, we prefer the moniker RAR-Norm.

**Comment #1.8.** Figure 3 – Q(M,t) is used before it is defined. Please define it earlier in the manuscript.

Response:: All notations have been updated for better readability. We use $D_t \le \left| Q_{t-1} - \tilde{Q}_{t-1} \right|$ in the revision. Please refer to Section 2.1 for the definitions of the notations.

**Comment #1.9.** . Agree with B. Schaefli, the superscript notation is hard to read. Please change to increase readability

**Comment #1.10.** Response: We have carefully and thoroughly updated the notations and avoided the superscript in the old version. Agree with B. Schaefli, re structure, the new method RAR should be presented in Section 2. All methods should be in a method section, all results in a results section

Response: We have changed the structure according to comments from B. Schaefli, and hope this is easier to follow.

**Comment #1.11.** Please also provide details on the algorithm used to maximize the likelihood – was it SCE or something else?

Response: The Shuffled Complex Evolution (SCE) algorithm (Duan et al., 1994) is used to minimize the negative log likelihood. (Lines 153-155).

# Review comments from Bettina Schaefli

**Comment #2.1.** This manuscript proposes a new method to correct forecasted streamflows based on the forecast error of the previous time step. The method represents a modification of the commonly applied autoregressive correction. The paper is well written, the method and the results concisely presented and discussed. However, the presented results did not convince me that the new method really outperforms the reference method; this might easily be improved by showing more details of the performed tests.

Response: Thank you very much for the time and effort you have taken to carefully review this manuscript. We feel your comments are very valuable for us to improve the quality of the manuscript considerably. We appreciate your positive feedback. We have paid serious attention to your suggestions and have attempted to address all your concerns, especially on the model performance comparison. Please refer to specific responses to your comments below.

My suggestion for moderate revisions of this paper are:
General comment on used terms

**Comment #2.2.** I would carefully revise the used wording to clearly distinguish between "forecast" (prediction of the system state at a given moment in time) from the more general "prediction". At the moment, the two terms are used interchangeably, which might sometimes be misleading, especially because the discussed streamflow correction only applies to forecasting.

Response: We have now used the word *forecast* throughout the paper, and have removed the word *prediction*. We occasionally use the word 'simulation' to differentiate instances where forecast rainfalls would never be used to force a rainfall runoff model, and make the distinction in the text, as follows:

"Our study is aimed at streamflow forecasting applications, so we preserve the distinction between observed and forecast forcings by referring to streamflows modelled with observed rainfall as *simulations* and those modelled with forecast rainfall as *forecasts*. As the forecast rainfall we use is observed rainfall, the terms *forecast* and *simulation* are interchangeable." Lines 198-203

Intro

**Comment #2.3.** As far as I see, the Kavetski et al. 2003 reference does not discuss forecasting and thus also not updating procedures but parameter estimation. Please give here references for papers that actually use streamflow correction / updating in a forecast setting.

Response: Thanks for the suggestion. We now use Morawietz et al., 2011 in the revision as a reference for updating procedures used in the context of streamflow forecasting.

**Comment #2.4.** In the general discussion of streamflow prediction errors, you might want to add the recent reference by Pianosi, F., and Raso, L.: Dynamic modeling of predictive uncertainty by regression on absolute errors, Water Resources Research, 48, W03516, 10.1029/2011wr010603, 2012.

Response: Thanks. We have added Pianosi and Raso (2012) as a reference for heteroscedastic prediction errors.

Method, section 2 and 5

**Comment #2.5.** Eq. 2 as well as following eqs. does not show the involved parameters

Response: We have carefully revised all equations and notations. The corresponding new equation comes with the definitions of the parameters involved.

**Comment #2.6.** Eq. 4: part of the equation does result in the "median value"? should be corrected;

Response: Thanks for reading the manuscript so closely. We provide the proof in Appendix A to show why the updated streamflow is the median.

**Comment #2.7.** Reference for max. likelihood formulations in eq. 6 , 7?

Response: we add Li et al. (2013) as a reference for the likelihood formulations.

**Comment #2.8.** In general, I think the superscript notation is not nice to read, why not use two different variable names and subscripts for the parameters?
Response: Thanks for the suggestion. We have carefully and thoroughly revised the notations to increase the readability. For example, we don't use complex superscripts any more.

**Comment #2.9.** I am not convinced by the current structure with section 5 presenting the new approach; instead of having an "idea-flow" paper structure (method - result 1 - new method - result 2), I would introduce the new method in section 3.
Response: This suggestion is really valuable to improve the presentation of this manuscript. We have followed the suggestion to change the structure of the manuscript.

**Comment #2.10.** Eq. 8: the same variable name is used for something new, to avoid, what is QM?
Response: We have updated the notations completely and avoided the duplication of variable names.

**Comment #2.11.** P. 6045 last line: word missing
Response: In the revision, the estimation of all three models is described in Section 2.2. We don't need this sentence any more.

**Comment #2.12.** P. 6046 first paragraph: would be useful before eq. 8
Response: We have re-worded the motivation/idea behind the RAR-Norm model and placed this paragraph before the definition as suggested by you.

**Comment #2.13.** Likelihood formulation of the new approach?
Response: We have added the likelihood formulation.

Case study
**Comment #2.14.** The GR4J model: do any specificities of the model influence the obtained results? (to be mentioned in results section?)
Response: There may be. GR4J may be more prone to fluctuations in base hydrological model performance than other models, as pointed out by our other reviewer. We have added the following discussion of this matter:

"We note that the poor performance of the hydrological model may be specific to the GR4J model, and many not occur in other hydrological models. Evin et al. (2014) estimated hydrological model and error model parameters jointly using GR4J and another hydrological model, HBV, for the three US catchments tested here. While they did not assess the performance of the base hydrological models, they found that HBV tended to perform more robustly when combined with different error models. It is possible that we may have achieved more stable base model performance had we used HBV or another hydrological model. We note, however, that our conclusions can probably be generalised to other hydrological models that do not offer robust base model performance under joint parameter estimation (e.g. GR4J). Because the RAR-Norm model essentially limits the range of updating that can be applied through the AR-Norm model, it will tend to rely more heavily on the base hydrological model, and therefore will tend to favour parameter sets that encourage good stand-alone performance of the base model. For those hydrological models that already produce robust base model performance under joint parameter estimation (perhaps HBV), RAR-Norm is unlikely to undermine this performance for the same reasons. We see some evidence of this in our experiments with GR4J:

when the performance of the base hydrological model is already strong relative to the updated forecasts for the AR-Norm and AR-Raw models (e.g. the Tarwin, Mitta Mitta, or Guadalupe catchments), the RAR-Norm model base hydrological model also performs strongly." (Lines 418-437)

**Comment #2.15.** P. 6041, line 21: "we then predict streamflow": not clear here whether in prediction or in forecast mode

Response: We agree – see responses to comments #2.2 and 2.16. This sentence now reads:

"We then generate streamflow forecasts in that year (1999) with model parameters estimated from the remaining data."

**Comment #2.16.** The use of "simulation" and "prediction" is confusing; I recommend using the term "forecast" for simulations with forecasted rainfall and the term "simulation" in the other case

Response: Thanks for this – we agree this makes things clearer. We have now changed the terms we use, as described in response to comment #2.2.

Results

**Comment #2.17.** I suggest a new results section presenting all the results

Response: Thank you for your valuable recommendation. We have followed your suggestion and made all results into two sections: Section 4). All methods are now described before results are documents.

**Comment #2.18.** In any case, the results of the new method require a separate section (now part of section 5 presenting also the method)

Response: We have followed your suggestion and presented all results in Section 4 (See Comment #2.17)

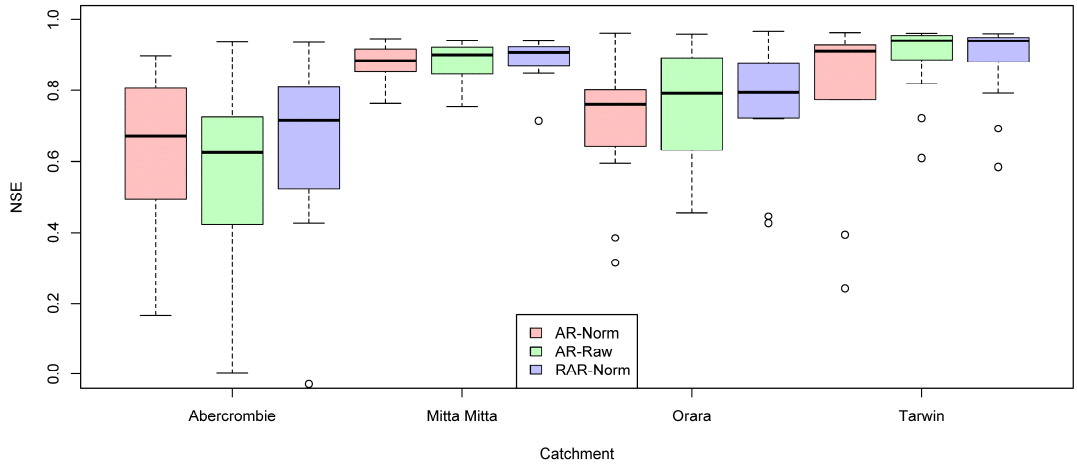
**Comment #2.19.** P. 6047, line 13: "notable better performance": as far as I see, only 2 out of 4 cases show a slightly better performance; from fig. 7, the improvement of RAR over AR-raw is not evident

Response: We have made several additions to demonstrate the performance of our RAR-Norm model:

1) We now assess our model on three additional US catchments, and these confirm our results
2) We use a greater range of metrics and analyses, including assessing the performance of AR models on all instances where streams are i) rising and ii) receding. This demonstrates the general tendency of AR-Norm models to perform least well when flows are rising, as well as the general tendency of the AR-Raw model to perform least well when streamflows are receding. The RAR-Norm model tends to combine the best aspects of the AR-Norm and AR-Raw models.
3) We have chosen a different example that more clearly shows the problem of over-correction of receding flows by the AR-Raw model.
4) We have changed the structure of the paper, following your suggestions, to group all results together. In doing this, we are better able to present that the RAR-Norm model outperforms both other models (see also response to Comment #2.21)

**Comment #2.20.** The NSE results are aggregated, how do they look like for individual cross-validation experiments? Are the NSE samples really significantly better with RAR (different distributions with higher mean) or is this pure chance?

Response: We show box-plots of the NSE values from each cross-validation period for the four Australian catchments below (note that for display purposes we have limited the vertical axis to [0, 1] and this means some of the outliers are not displayed). .These support our contention that the RAR-Norm model generally leads to better performing forecasts in two ways: 1) the means of the box plots are similar or higher than the next best performing model and 2) The distributions NSE scores of RAR-norm tend to be as narrow or narrower than the next best performing model, indicating that the performance of RAR-norm model is more robust under cross-validation.




Box-plots of NSE values for each cross-validation period for Australian catchments. Dark lines, mean values; boxes, interquartile range; whiskers, [0.1, 0.9] intervals; points, outliers.

Discussion

**Comment #2.21.** I am not convinced that the paper shows that the new method leads to a more robust performance of the base hydrological model. This should be shown in a more convincing way by presenting some more detailed results of all the simulations.

Response: We believe the additional metrics, analyses and catchments we have added (see response to comment #2.19) add weight to our conclusion that the RAR-Norm model is an improvement over the conventional AR-Raw and AR-Norm models we test. We summarise this in the conclusion as follows:

"The RAR-Norm model is a modification of the AR-Norm: in most instances it operates as the AR-Norm model, but in instances of possible over-correction it relies on the error in untransformed streamflows at the previous time step. That is, RAR-Norm is essentially a more conservative error model than AR-Norm: in situations where streamflows change rapidly, it opts to update with whichever error (transformed or untransformed) is smaller. This forces greater reliance on the base hydrological model to simulate streamflows accurately, leading to more robust performance in the base hydrological model. The RAR-Norm model clearly outperforms the AR-Norm model in both the updated and base model forecasts, as well as ameliorating the problem of over-correcting rising streamflows. The RAR-Norm model's advantage over the AR-Raw model is less clear: both the base hydrological model and the updated forecasts produced by the AR-Raw model perform similarly to (or sometimes slightly better than) the RAR-Norm model. However, the RAR-Norm model clearly addresses the problem of over-correcting receding streamflows that occurs in the AR-Raw model. As we show, this type of over-correction can seriously distort event hydrographs, and cause forecasts of near zero flows reasonably substantial flows are observed. While these instances are not very common, the failure in the forecast is a serious one. As we note earlier, the over-correction of receding flows is likely to be exacerbated when producing forecasts at lead times of more than one time step. Accordingly, we contend that the RAR-Norm model is preferable to both AR-Norm and AR-Raw models for streamflow forecasting applications." (Lines 458-479)