Response to Interactive comment on "State Updating and Calibration Period Selection to Improve Dynamic Monthly Streamflow Forecasts for a Wetland Management Application" *by* Matthew S. Gibbs et al.

Anonymous Referee #1

This article investigates 2 scientific issues in the context of rainfall-runoff seasonal forecasting: (a) the advantages of state(s) updating and (b) the sensitivity of the choice of the data used for calibration (calibration period length). These 2 scientific issues have been / are widely discussed in the hydrological community. Authors choose to put them in the context of complex wetland management application. This intention is relevant, since these issues are of crucial importance for operational matters.

First, it is worth noting that the manuscript is most often very clear (in particular, the introduction is efficient). Some suggestions are made below (detailed comments) to make the manuscript clearer (some parts are easier to understand when checked again after a further reading). The methodology is quite well detailed (I reckon that it is sufficient for anyone who wishes replicating the study) and the results are well presented.

However, many (too many ?) details concerning the application context are provided (section 2). I am afraid that I missed understanding how they infer with the scientific issues: results and discussion section do not make clear to me whether and how this particular context has implication on the way these issues are dealt with and on the results of the study. In a similar way, some details are given about the data used in an operational context, but it is not clear how they impact the results of this study. For example, this is the case of the precipitation forecasts (see detailed comments). Indeed all the results are not discussed in depth, with respect to these options (e.g., results obtained with observed rainfall versus results obtained with forecasted rainfall) and with respect to the context and practical purposes of the wetland management (whereas this appears in the submitted title): results are presented in only 2 pages and a half and the discussion is shorter (1 page). Even if all the tested cases are very useful for the specific case study and application, they are then not fundamental for the reader who focuses more on the 'generic' scientific issues than on the specific context of wetland management. The authors should consider removing them in order to make the reading and the analysis easier, rather than providing everything they learnt from their case study (again: even if it is quite interesting per se). They may prefer explaining how their findings are related to their specific case study and practical application.

One of the topics of interest for the sub-seasonal to seasonal hydrological forecasting special issue was user needs for seasonal forecasts. As such, more detail than typical on the case study application to wetland management was included in the manuscript. However, it is agreed that this is a distraction from the more generic scientific issues of interest to most readers. As such, section 2 will be shortened to be more targeted and remove surplus detail, such as that of the water balance model. See responses to detailed comments below for further changes in this regard.

It is believed that all the test cases considered remain relevant to the contribution of the manuscript and general scientific issues, i.e. with and without state updating, two different calibration period lengths, and observed and forecast rainfall. The least contribution may come from the two rainfall sources, but these cases represent two different tests. The observed rainfall allow the effect of the changes on model performance to be isolated, without the effects and errors associated with the forecast rainfall propagating through. However, the forecast rainfall case is necessary for the practical application, and without this case the results cannot be considered realistic.

As mentioned previously, the 2 scientific issues have been explored by many previous studies. That is why this article has to do thorough review of literature in order to emphasize on the novelty of their study or to compare their results to those of other studies:

- The way how non-stationarity is treated is very satisfying. The explicit distinction between physical catchment non stationarity and other model non-stationarity is necessary (while not always made); it is introduced in a very clear manner. I only suggest the authors to give a more explicit definition of the model parameters (the discussion is indeed implicitly present behind), since some previous studies proposed parameter variations to compensate many different non-stationarities (up to model structural deficiencies), as nicely pointed out in the introduction. This issue is particularly relevant in this study because the authors chose to update their model but only selected state updating, while many other approaches exist, one of them being parameter updating: this choice, which is very consistent, may be better explained. While not being a specialist of the choice of data calibration, I found the quoted references relevant. I only wish that these articles (e.g., Luo et al., 2011, which clearly inspired the methodology adopted by Gibbs et al.) would have been quoted not only in a generic way but also in sections 4 (Results) and 5 (Discussion) as benchmarks for the results: the results confirm previous studies in a large part; is there any interesting difference?

A description of the model parameters and structure will be included in the revised manuscript, see response to detailed comments below. The rolling approach used to calibrate the model parameters (both the hydrological model and error model) could be considered a form of parameter updating, where the parameters are updated every year based on the most recent data. This concept of parameter updating, and the rolling approach used, will be made more explicit in the introduction.

As suggested, the Discussion section will be expanded to contrast the findings in this work against the relevant references cited (e.g. Luo et al. 2011, Brigode et al. 2013), to highlight where there are similarities or different findings.

- Concerning the data assimilation and model updating issue, the bibliography is poorer (see detailed comment for page 3). Many references could be added. Since the authors chose to use the GR4J model and since the state updating they chose is the same one as the approach adopted for the GRP model ('adaptation' of the GR4J model for forecasting, used by the French flood forecasting centres), it is also worth mentioning this work (see detailed comments below). Beyond the references issue, it may (should ?) be noted that this study explores the benefits of model updating for seasonal forecasting, whereas many, if not most, studies consider shorter lead-times. This aspect has to be mentioned, since it is well known that the effects of model updating most often vanish whe the lead-time increases. In my opinion, keeping benefits at large lead-time is one of the (surprising) key result of this study and may be usefully emphasized.

Thank you for the very constructive suggestions for relevant literature here and in the detailed comments. The literature review will be updated to include these references and points to improve the representation of the previous literature in the field of data assimilation.

See the response to detailed comments on emphasizing one of the key results of this work, that model updating still improved forecast performance at the lead-time of one month. This includes highlighting that this has not been focused on in the literature in the introduction section, and highlighting this in the abstract and conclusions. This change is proposed in the introduction:

The impact of assimilating observed data into CRR model state variables at the start of a probabilistic forecast on ~~probabilistic seasonal forecasts~~ has focused on short lead-times, however any benefits at longer lead-times (e.g. seasonal) has had limited evaluation~~. It could be expected that by assimilating this extra information into the model state variable(s) the precision of the forecast, and potentially its reliability, could be increased.~~

A few methodological choices may deserve a little more discussion or explanation: The calibration algorithm is a rather complex one, but used with assumptions which are known to be not met in most cases (page 9, line 7: independent, homoscedastic residuals). Moreover, these assumptions are not consistent with the choice of the model error post-processor (a Box-Cox transformation is used in order to take into account the heteroscedasticity of these same residuals). Why did the authors pick a complex approach with unverified and inconsistent assumptions rather than a simpler one? It let the reader think that the authors used "components" available on the shelf or a pre-existing tool, which is quite understandable. But then they have to justify these choices (and why a so complex calibration method when much simpler ones are easily available?).

It was considered that a calibration method capable of estimating a posterior distribution of parameter values was required. This allowed the change in the distribution of suitable parameter values to be considered over time (Figure 8). In contrast, a simpler algorithm (gradient or evolutionary method) that provides a point estimate of the calibrated parameter values would not show the parameter identifiability (spread in values), and in turn the trends over time may not have been able to be separated from the variability in estimates. As such, the choice of algorithm is not considered overly complex, and a significantly simpler approach to estimating parameter uncertainty is not known to the authors.

The assumptions of the GR4J model likelihood function and the error model likelihood function are different, as they are applied in different situations:

- The GR4J model was calibrated at the daily scale to observed rainfall. The likelihood function used for the calibration of the GR4J model is arguably the simplest available, the standard least squares. While this approach is simple, it is expected to enable the model to capture high flows, as outlined. While the assumptions made by this likelihood may not met, the function is considered fit for purpose. This discussion in Section 3.2 will be expanded on to make this point clearer.
- The error model is applied based on the outputs from the GR4J model aggregated to the monthly time scale. The Box Cox transformation is used to capture heteroscedasticity and skew in residuals, which allows more reliable forecasts. Both observed and forecast rainfall were used to drive the GR4J model, and the error structures based on the forecast rainfall are very different from observed rainfall. As such, the errors, and error assumptions, would be expected to be different in these two cases.

Given the different objectives of the two models, and the different time scales adopted, this inconsistency is considered to be reasonable, though we agree in general the effects of this type of inconsistency warrant a separate investigation. This point will be included in Section 3.5. The approach used is not without precedent, as other forecasting systems implemented in Australia also adopt different objective functions at different stages of the modelling chain (Lerat et al., 2015).

Furthermore, one point is not discussed but may deserves some attention. Like the hydrological model, the model error post-processor is calibrated (not in a joint manner however). Why does the study on the impact of the calibration data period length on the calibration only focus on hydrological parameters and not on the post-processor parameters as well (mu, sigma)? This can indeed be treated independently (therefore not necessary in this article), but this research issue may be usefully mentioned. Are the post-processor parameters concerned by the rolling calibration?

The error model parameters are also determined using the rolling calibration approach, and this will be clarified in Section 3.5. This way any trends in the error model parameters are captured in the same way as those for the hydrological model. For the sake of brevity, a figure similar to Figure 8 has not been included for the error model parameters.

One element may also be better detailed: the results are given at a monthly scale (time step), whereas the GR4J model is a daily one: the way the GR4J model is run has to be precised. This is important, since the model is updated and effects of model updating decrease when the lead-time increases. However, it often does not only depend on the lead-time 'absolute' value but also on the number of time steps to reach this lead-time.

Good point. Details of the time aggregation of the GR4J model results and warm-up period will be added to the model description in Section 3.1 and 3.4.

In a nutshell, this article brings some interesting results, even in a field explored by many previous studies, and deserves publication. The suggestions made in order to improve the manuscript lead me to propose a moderate to major revision (however another round of submission afterwards does not seem necessary).

Thank you for the constructive comments that will help to improve the manuscript substantially.

DETAILED COMMENTS

- Page 2

Line 22 ("As these models are conceptual, they require calibration [...]"): they are not the only models that do so. Even the (so-called) physically-based models which could theoretically not need calibration, are most often calibrated, for various practical reasons.

Agreed, the manuscript will be updated to reflect that in fact most if not all environmental models require some level of calibration.

Lines 22 - 23: Brigode et al. (2013) show that this general a priori (longer calibration periods produce more robust parameters estimates) is not always verified. Therefore if this article is quoted (and it should be, in my opinion), it would be fair to indicate their results.

Agreed, and explanation of results of Brigode et al. (2013) will be included in the manuscript.

Lines 30 - 31: the definition of catchment non-stationarity which is proposed, is very interesting since it 'focuses' (restrains to) the physical object non-stationarity. However, it seems to be a binary state: the catchment is or is not stationary. Have the authors considered the notion of a "degree" of non-stationarity? Indeed, all the listed factors of non-stationarity are not expected to have the same consequences over the catchment behaviour. Might the rolling calibration approach be a tool to assess the relationship between "degrees" of non-stationarity and parameters evolution? (see also detailed comment on Fig. 3)

This is a good point, indeed natural systems are likely to be in varying degrees of non-stationarity, as nothing in nature is entirely static. This point will be included in the manuscript.

- Page 3

Line 1: is groundwater depletion a physical change (of the catchment) or the consequences of some of the listed catchment changes?

This clarification will be included, land use change has been shown to be a contributing factor to the observed groundwater depletion (Avery and Harvey, 2014, Brookes et al. 2017).

Lines 15 - 20: the bibliography review is rather poor: it gives some "extreme approaches" between the (too) simple GLUE and the very detailed BATEA (or similar approaches). Furthermore, GLUE is a quite old approach, giving a reference of 2008 is a bit strange (unfair?), as it appears more recent than much more advanced and sophisticated approaches, as those developed by Kavetski, Vrugt and others. Since the chosen approach is a model error post-processor, I suggest Krzysztofowicz and Maranzano (2004).

Thank you for the reference, it will be included. The reference for GLUE will be changed to Beven and Binley (1992); the reference to Krzysztofowicz and Maranzano (2004) will also be included.

Line 17: "using a model error post-processor" rather than "using a post-processor error model" ?

The terminology will be clarified. "post-processor error model" is preferred by the authors, as the error model is developed and applied after the hydrological model has been calibrated, rather than calibrated in conjunction with the hydrological model (i.e. post-processor), and "error model" emphasises that an error model has been used.

- Page 4

Line 3 ("up to one month"): I understood this paragraph as a bibliography review giving general results (not specific to some catchments). However, it gives some values of the "influence duration" of the initial state, which strongly depends on the catchment characteristics. I am pretty confident in the fact that it easy to find catchments where the impact of the initial conditions is important during several months (even years).

This paragraph is indeed intended to review general results, and the "up to one month" statement is from the cited references (Li et al., 2009; Wang et al., 2011). However, it is agreed that this is not a universal rule, and will change depending on the catchment. The sentence will be changed to:

The impact of initial catchment condition is particularly pronounced when forecasting over short lead times, *typically* up to one month (Li et al., 2009; Wang et al., 2011), *however this time frame is catchment dependent.*

Line 5: "warm-up" rather than "warmup"?

The change will be included throughout.

Lines 14-16: it may be specified that this impact has been deeply evaluated for shorter lead-times (this emphasizes the character of novelty of the study). Furthermore, I disagree with the second sentence as it has been shown that the impact decreases quite fast (for

many not too slow catchments) and is almost negligible at a seasonal scale (see e.g. Berthet et al. 2009 that the authors quote elsewhere). That is one very interesting aspect of the results of the submitted study.

This was the intent of this paragraph, to highlight that the impact at seasonal scales has not had extensive evaluation. The paragraph will be changed to:

The impact of assimilating observed data into CRR model state variables at the start of a ~~forecast on~~ probabilistic ~~seasonal~~ forecast has focused on short lead-times, however any benefits at longer lead-times (e.g. seasonal) has had limited evaluation~~. It could be expected that by assimilating this extra information into the model state variable(s) the precision of the forecast, and potentially its reliability, could be increased.~~

Line 18: I suggest to precise "calibration periods choice" or "calibration periods length" rather than only "calibration periods".

The change will be included throughout.

Line 21: to enhance "seasonal" forecasting skill?

Yes, this will be added.

Line 22: Does the article "demonstrate" that calibration period choice can affect forecast skill (that is quite known) or does it assess how much it does so?

The aim will be updated to focus on the assessment, rather than demonstrating a known influence.

- Page 5

Line 11: is it the gauge "A2390514" rather than "A21390514"?

Correct, the typo will be corrected.

Lines 26 - 27: an hydrograph may be useful to support this information.

A hydrograph of flow in Drain M will be included to demonstrate the variability.

- Page 6

Lines 3-13: is the description of the model developed by eWater Source useful for the reader. If I understood correctly, it is not directly related to the model used in this study. If so, this might confuse a bit the reader. E.g., I am not sure that the assumption of a constant inflow of salinity (which is not discussed) is needed by the reader to understand how the authors worked to answer to the scientific questions (which are the core of the article). The multi-objective nature of the calibration is also of no use for the rest of the study. If the fact that this model is used in practice had consequences on the methodological choices for this study, then the authors may consider explaining it (and discuss results with respect to it and to the specific context of wetland management).

As pointed out earlier, we agree that Section 2 includes superfluous detail that will be removed. This section on the separate water balance model in eWater Source will be either substantially shorted or removed altogether.

Line 14 ("To use this model for to inform operations"): it is always tricky for a non native English speaker to ask so to native ones, but may the authors check English here?

Correct, the typo will be corrected.

Line 15: "lead-time" rather than "leadtime"?

Correct, the typo will be corrected throughout.

Line 15: to fully understand the implication of the choice of the 1-month lead-time, it is necessary to know that the CRR model is a daily one (not only because the model is updated). However this information is given at subsection 3.1 (and not very explicitly: the reader has to know that GR4J is a daily model)

Good point, this information will be included as part of the aims and objectives in section 1.3.

Line 19 ("reasonable forecast skill is expected to be possible compared to longer forecast horizons"): may the authors provide some references? How much are the performances expected to decrease for longer lead-times? Furthermore, why did the authors choose to focus on a single lead-time? The evolution of the benefits of the model updating, with respect to the lead-time, in a context of seasonal forecast, would be a very interesting result.

We agree that this comparison of longer lead times, and how they influence the state updating performance, would be a very interesting study. However, it was considered beyond the scope of this work. The statement is based on the fairly obvious observation that the skill of rainfall forecasts is expected to reduce the longer the forecast horizon, and as such the skill of the streamflow forecasts will also reduce. The sentence will be changed to:

and 3) the skill of rainfall, and hence streamflow, forecasts is expected to decrease as the lead-time increases.

Line 20 ("The mean annual rainfall for the region is in the range 600-675 mm"): page 5, lines 3 and 4 suggest some spatial variability. How strong is it? (600 to 675 mm is not very strong difference, compared to some other climates around the globe).

As pointed out, ~10% range is not very strong, but it is spatially consistent, as opposed to representing annual variability. The manuscript will be updated to clarify that there is a rainfall gradient from south to north.

Line 20: is it useful to precise what "FAO56" stands for?

The acronym will be expanded and a reference included (Allen et al., 1998). It stands for Food and Agriculture Organization of the United Nations, and paper 56 relates to guidelines for computing crop water evapotranspiration.

- Page 7

Lines 3 - 4 ("2 rainfall hindcasts [...] were downscaled to the single rainfall gauge scale"): just to be sure, does it mean to the pixel where the gauge is?

This will be expanded on. The pixel size is ~250 km, which tends to smooth out the rainfall events. The downscaling process, mapping the pixel where the rainfall gauge is to the gauge data, is used to restore more representative rainfall events.

Lines 15 - 25: the authors may consider whether this paragraph would not be better written earlier (e.g. among the first paragraphs of section 2).

Agreed, the paragraph on the catchments and where they flow will be moved to the start of section 2.

Line 20 ("It should also be noted that releases from Bool Lagoon [...]"): why is it important to understand this scientific study? (I worry about missing something useful for the interpretation of the results)

This will be clarified. The point being made was it is rare that the upstream catchment contributes to the downstream catchment, as releases from Bool Lagoon are rare.

Lines 30 and following: are the details about the streamflow measurements devices useful?

These details help establish that any perceived non-stationary trends are unlikely to be due to streamflow instrumental measurements. We will clarify this issue and shorten this section to avoid superfluous details.

- Page 8

Line 3: I agree with the fact that indicating the data are of good quality and too often not done, but if the authors want to demonstrate the quality of the rating curves, they may add some information about the number of years during which the 78 and 166 gaugings have been achieved and how much often the rating curves have been modified.

It is agreed that this is an important component of streamflow data quality. Information on rating curve modifications will be included in the reworking of this section (see previous comment).

Lines 9 - 20: since catchment non-stationarity is an important issue for this study, I suggest to make this paragraph a subsection dedicated to this topic (here).

This is a good suggestion which will be adopted.

Lines 23 - 24 ("GR4J [...] explicitly accounts for non-conservative (or 'leaky') catchments"): I agree. However, it should be kept in mind that GR4J has not been designed nor is known to achieve good performances for karstified catchments (mentioned page 7, line 13). Moreover, I am not convinced it is quite appropriate for ephemeral catchments (as suggested by line 21, page 11).

It is agreed that GR4J may not be ideal in ephemeral catchments, as the exponential decay relationship used in the storage reservoirs cannot completely dry out. However, this is a relatively theoretical consideration, for example, if any simulated flow below that which could be adequately measured is considered to be zero, ephemeral behaviour can be represented. As outlined, previous studies have demonstrated good performance for Australian conditions, including ephemeral catchments (Coron et al., 2012; Guo et al., 2017). Westra et al. (2014) will also be added to this list, who applied GR4J in a similar location in southern Australia. Considering alternate model structures was beyond the scope of the study, however, it is possible (likely?) that more appropriate model structures could be identified in future work.

Lines 28 - 31: may the authors explain what motivates their choice of adding a 5th free parameter to calibration? Is it important for their particular catchments or for their methodology in this study?

As noted in the previous comment, GR4J was not designed for this application. The catchments considered have a relatively slow response, and it was considered that the pre-specified split to the routing store of 0.9 may be too low for these catchments. As can be seen in Figure 8, this turned out to be the case for the calibrated parameter values, where higher values of the split parameter were found, in particular for C2. This point will be clarified in the manuscript.

- Page 9

Lines 10 - 12 ("this function [RMSE] provides a focus on the highest flow in the time series, where the majority of the runoff occurs"): it is not necessary. It provides a focus on the largest absolute errors, which indeed most often occur for the largest flows. However, consider a hypothetical model whose errors would be only on low flows.

We agree, the original statement was not strictly correct. The statement will be updated accordingly.

Line 26 ("External influences include model structural limitations [...]"): this confused me, after reading the catchment non-stationarity given on pages 2 & 3. Does it suggest that parameters variation due to structural deficiencies would be considered here?

We agree this sentence does not add value to the discussion in Section 3.3. It will be removed to avoid confusion.

- Page 10

Lines 3-5: Would not it be useful to emphasize the trade-off between a longer calibration period to reduce the parameter uncertainty and a shorter calibration period to mainly take into account the most recent dynamics in the introduction section?

This tradeoff is indeed a key aspects of the study; we agree with the reviewer that it warrants stronger emphasis in the introduction.

Line 13-14: the literature review is also poor about data assimilation and model updating. Generic references may be Refsgaard (1997) and Liu and Gupta (2007). Since the chosen updating approach is the same as the one used for the GRP model (which is a mere adaptation of GR4J for forecasting purposes), I suggest to refer to the work of the team which developed these models. The authors may pick Tangara (2005) and Berthet (2010), both in French, which described the numerous tests of different updating approaches made by the GR4J research team (some of them discussed in section 5! See comment below) and detailed the resulting GRP model. They may prefer Berthet et al. (2010), which provides a much shorter description of the model and the updating techniques but also discusses the impact of the largest errors on the RMSE-based criteria values (see discussion page 9). For a detailed description of the GRP model, the authors may also consult: https://webgr.irstea.fr/en/modeles/modelede-prevision-grp/fonctionnement-grp/. Moreover, since sequential approaches such as ensemble Kalman filter and particle filters are mentioned, I suggest also to add references to Moradkhani et al. (2005, 5005b) and Weerts and El Serafy (2006).

Thank you for the very constructive suggestions. The literature review will be updated to include these references and points.

Lines 17 - 27: a flowchart would greatly help the reader.

It is considered that the equations provided are the clearest approach to explain the exactly methodology used to update the routing store to simulate the observed flow. Further details on this approach are provided in the original work, Demirel et al. (2013). This link to the earlier work will be made clearer in the revised manuscript.

Line 27: "where X3 is the estimated runoff model parameter" rather than "where X3 is an estimated runoff model parameter"?

This change will be made.

- Page 11

Lines 3 - 4 ("particularly when used to update both model state variables and model parameters"): I agree with the authors, but is it relevant here? (since parameters are not updated here).

It is agreed that this is out of place. This paragraph will be removed and integrated with the revision of the literature review, along with the comment on page 10 Line 13-14.

Line 12 ("Depending on the case"): this is not clear, until the reader reaches section 3.7.

This sentence will be reworded as follows:

As outlined in Section 2.1, both observed and forecast rainfall has been considered. When observed rainfall was used, the predictions used were those from the daily hydrological model, aggregated to the monthly time step. When the ensemble of forecast rainfall was considered, the hydrological model predictions were the median across the ensemble of the hydrological model predictions each day, and then aggregated to the monthly time step.

- Page 12

Line 5: why do the authors prefer to sample the (normalized) residuals rather than picking a number of calculated quantiles (from the Gaussian distribution)?

A direct selection of quantiles would provide an analytical approach to represent the distribution. However, this can be difficult, as the calculated quantiles in normalised space do not correspond to the same quantiles in un-normalised space, due to the combination of the transformation, the truncation of very low flows and the parameter uncertainty. The Monte Carlo Simulation approach used is more computationally intensive, but the most robust. This approach is also more generalised, and would be required is more complex error models were used. For these reasons the approach used has been retained.

Line 18: may the authors explain the choice of the 0.05 and 0.95 as normalized extrema values?

This range was adopted for data visualisation purposes only. If the 0 – 1 range were used, the worst-performing case for each metric would correspond to zero area in Figure 4 and would not appear in the plot area. We decided to avoid this to avoid the impression that some information is missing.

- Page 13

Lines 17 - 21: as pointed out by the authors, the reference distribution has an 'unfair' advantage. Then may the authors explain this choice? Why have they not chosen a simple naive forecast model?

The reference distribution of the monthly streamflow is considered a simple naïve forecast model. Other approaches could be used, autocorrelation with last month's streamflow, for example. The unfair advantage referred to is that the reference distribution has been

calculated using all data, which can be from the "future" for earlier forecast periods. However, it is expected that the forecast models tested should be able to perform better than an uninformed climatology, and as such is considered a useful baseline for the calculation of forecast skill. The paragraph will be changed to:

The reference distribution for each month is calculated as the empirical distribution of all observed data in that month. Note that all of the observed data are used to estimate the reference distribution (including data in the forecast verification period), however, this is still considered a suitable simple naïve forecast model that is useful as a baseline for the assessment of forecast skill. $CRPS_{SS}$ values less than zero indicate the reference distribution produces better performance than the forecast model.

Line 23: check the formula. There is missing sum for the denominator.

This typo will be corrected.

- Page 14

Line 6: is the rainfall forecast used here the ensemble forecasts described in subsection 2.1? I don't think it obvious. If not, why were the ensemble described?

Yes, the ensemble forecasts were used in all cases. This will be clarified here.

Line 6: I found only 2^4 = 16 cases (model with or without updating; 2 calibration period lengths; 2 catchments and 2 rainfall forcings). What do I miss?

Well-spotted - this is an error and will be corrected to 16.

- Page 15

Line 6 ("Any detrimental impacts"): check English.

The sentence will be reworded as follows:

The state updating could be seen to reduce the accuracy of the forecasts occasionally when the model storage was updated to represent a zero flow, and then a low flow that did occur in the following month was underestimated, e.g. in January 2002 for C2 in **Error! Reference source not found.**.

Line 8 (and followings): I suggest to precise "calibration period length" rather than only "calibration period"

This will be corrected throughout.

Line 16: the differences are not much smaller for catchment C1. How much are they significant?

The original statement in the manuscript is correct, the difference between values was smaller for every metric and case when state updating was used. The anomaly is the precision metric for C1 with observed rainfall, which increased for the longer calibration period where all the others decreased. Nonetheless, the differences were still smaller for the "with state updating" case compared to "without". Significance tests have not been calculated due to the lack of replicates. No change is proposed.

Line 19 ("the differences were more pronounced for the most practically relevant cases with forecast rainfall [...]"): this is of particular interest for operational purposes (e.g., forecasts) and is worth being emphasized.

This is a good suggestion and will be emphasized in the revision.

Line 20: I don't understand how the model error post-processor compensates the introduced errors. I thought that it only assesses them.

The normalised residual can have a non-zero mean, which can compensate for biases in the hydrological model predictions. But the main purpose of the error model is to quantify the uncertainty in forecasts, not to compensate for errors. For clarity, the paragraph will be changed to remove this reference:

The differences were more pronounced in the more practically relevant case with forecast rainfall, which introduced further errors to be compensated by the state updating approach~~, and the postprocessor error model~~.

Lines 23-24 ("catchment C1 had been identified to have a substantial reduction in the rainfall-runoff relationship over time"). As discussed below (comments on Fig. 3), the catchment appears as rather stationary up to 1990 (approximately) and then also more or less stationary from 1990 to 2010. If it is so, how can the difference in calibrated parameters obtained with the 2 different calibration period lengths for years 2009 2010 be explained by this change around 1990?

The 20-year calibration period was 1989-2008. The 10-year calibration period was 1999-2008. As such, the early 1990s are part of the 20 year calibration period, but not of the 10 year calibration period. The results indicate that the 20 year calibration period resulted in model parameters that produce more flow the 2009 (and 2010) validation year than the 10 year calibration period. This suggests that streamflow data from the 1990s are substantially different from streamflow data from the 2000s. The manuscript will be changed as follows:

From **Error! Reference source not found.**, it can be seen that the model calibrated to the longer period *(e.g. 1989-2008 for the 2009 forecast period)* overestimated the observed flow in 2009 and 2010 in catchment C1, whereas the model calibrated to the shorter period *(e.g. 1999-2008 for the 2009 forecast period)* provided a better reflection of the catchment response.

- Page 16

Line 16 ("A model fitting anomaly resulting from a shorter calibration period"): did the author investigate this "anomaly"? How can it be explained?

This was not investigated further. It is assumed to be due to reduced parameter identifiability. A different combination of parameters (e.g. higher X4 and lower X2 and *split* compared to the periods before and after) resulted in similar values for the objective function for this particular period. This discussion will be added to the manuscript.

Line 26: this is interesting at a seasonal scale, since it has been shown that hydrological models are "stable", i.e. the updating effect vanishes after a number of time steps (e.g. Berthet et al. 2009 at a hourly time step: then after a few days at most for a large majority of the tested watersheds).

Thank you.

Line 29 ("As the range in model predictions should be reduced by forcing the model to simulate the observed streamflow at the start of the forecast period"). Is there any confusion between precision and sharpness? Model updating increase sharpness and precision (at least for the shortest lead-times).

The term “precision” was used in this work as a synonym for “sharpness”, as outlined on Page 12 line 24. However, as “sharpness” is a term more generally used in the foreasting community, “precision” will be changed to “sharpness” throughout.

Line 30 ("the trade-off for an increase in precision would typically be a reduction in the reliability of the predictive uncertainty"): again I assumed that the authors meant "sharpness". I suggest to write that this trade-off for an increase in sharpness ** may ** result in the reduction of the reliability, if the authors do not provide a general (theoretical) explanation. As much as I know, this is a common feature, but exceptions should exist, and the last sentence of the paragraph (page 17, lines 1 - 2) says so.

See above, “precision” will be changed to “sharpness” throughout. Nonetheless, this is a good point, as the result cannot be proven to be generic. The qualification “may” will be added to this sentence.

- Page 17

Line 4 ("update the GR4J production store along with the routing store"): this has been tested by the GR4J team (Berthet, 2010), with no significant improvement in a forecasting context at a hourly time step.

The manuscript will be updated to refer to this study and highlight the potentially limited benefit.

Lines 4 - 5 ("This could be expected"): may the authors give some explanation to found this idea?

The manuscript will be modified to outline this hypothesis. At some times, the routing store was updated to be empty, but the model still overstated the observed flow. Also reducing the size of the production store could further reduce the simulated flow, to match the observed flow in these cases. It should be noted this approach has not been tested to date.

Lines 3 - 9: this paragraph is very interesting, but how is it related to the scientific issues developed in this article?

This paragraph was intended to point toward further work to improve the approaches used. It will be reworded to make this clearer.

Lines 16 - 19: in my opinion, this is the (or one of the) key findings. How may the authors emphasize it, rather than putting it at the very end of the article?

Agreed. This finding will be added to the abstract and conclusions, and the second aim of the study will be framed to emphasize this aspect of the investigation.

Line 30 ("in most cases"): the study was driven only on 2 catchments... It should be pointed out that a work over a (much) larger number of catchments is needed to ensure the generality of these interesting results.

We agree - our intent was to say that this finding held in most cases considered in the study, i.e. the different metrics with and without state updating. The sentence will be reworded to avoid confusion.

- Page 19

Tab. 1: the authors may usefully add the parameters meanings and their units (and a GR4J flowchart aside).

Good point - this information will be added to the manuscript.

- Page 20

Fig. 1: the drain M is not given in the legend and has the same color as catchment boundaries, which makes it difficult to identify.

Good point - the map will be adjusted accordingly.

Fig. 2: what are the upper and lower bounds? Minimum and maximum of the ensemble? Some predictive quantiles such as 0.05 and 0.95? If the latter, is there any information about the reliability?

That is correct, the upper and lower bounds on the figure legend refer to the minimum and maximum of the ensemble. This will be clarified in the figure caption.

- Page 21 (Fig. 3)

I wonder if there is not a "sudden" change around 1990: catchments C1 and C2 look quite stationary up to 1990, and also after 1990. If it is so, can it really be explained by plantation forestry expansion (which is more a "continuous" factor of non-stationarity). Furthermore, can this question the relevancy of the rolling calibration which might be better adapted for smooth non-stationarity?

A step change is arguably more likely to have taken place over the mid-late 1990s. The early 1990s are also quite wet, which may result in steeper slopes for these years. A cause for a sudden change in the streamflow response around 1990 is not known to the authors. Reductions in groundwater levels due to the forestry expansion has been attributed in other studies (Avery and Harvey, 2014, Brookes et al. 2017).

In any case, identifying such a distinct step change, or even a continuous change, is very difficult to do with certainty. The difficulty in identifying such step changes, and the idea of using the rolling calibration approach as a means to overcome this, will be added to Section 3.3. The discussion on Page 17, lines 16-19, will also be amended to more clearly address this point.

- Page 22 (Fig. 4)

It is very important to insist on the fact that the plot gives relative values of the metrics (higher is better) which are then not consistent with the formulas and details given in subsection 3.6. I was first confused becaused I did not notice (at first) the mention in the y-label (even if I admit that it is written in (sufficiently) large characters). I suggest to change the criteria described in section 3.6 to give only their relative values.

The criteria description in Section 3.6 will be updated to match the values used in Figure 4, and this information will also be added to the caption.

- Page 23 (Fig. 5)

Why are the results plotted only up to 2005?

This was done for illustrative purposes only. Some of the desirable detail in the hydrographs is difficult to see when the whole time series is plotted (~20 years). As such, a subset of the data was shown, to provide a trade-off between showing all of the results, and making a clear point. This point will be stated in the paper, and the full plot provided as supplementary material.

Page 25 (Fig. 7)

Why are the results plotted only from 2008 to 2010?

This was done for illustrative purposes only – same reason as in the point above. Again, this point will be stated in the paper, and the full plot provided as supplementary material.

REFERENCES

Allen, R.G., Pereira, L.S., Raes, D. and Smith M. (1998). Crop evapotranspiration - Guidelines for computing crop water requirements. *FAO Irrigation and drainage Paper 56*. Food and Agriculture Organization of the United Nations, p.300.

Avery, S. and Harvey, D. (2014) How water scientists and lawyers can work together: A 'down under' solution to a water resource management problem. Journal of Water Law 24(2):45-61

Brookes J.D., Aldridge, K., Dalby, P., Oemcke, D., Cooling, M., Daniel, T., Deane, D., Johnson, A., Harding, C., Gibbs, M.S., Ganf, G., Simonic, M. and Wood, C. (2017). Integrated science informs forest and water allocation policies in the South East of Australia. Inland Waters *(In Press).*

Lerat, J., Pickett-Heaps, C.A., Shin, D., Zhou, S., Feikema, P., Khan, U., Laugesen, R., Tuteja, N.K., Kuczera, G., Thyer, M., and Kavetski, D. (2015) Dynamic streamflow forecasts within an uncertainty framework for 100 catchments in Australia [online]. In: 36th Hydrology and Water Resources Symposium: The art and science of water. Barton, ACT: Engineers Australia, 2015: 1396-1403.