Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2017-381-RC1, 2017

Hydrology and Earth System Sciences
Discussions

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

Received and published: 10 August 2017

August 2017, the 9th

A review of the article entitled "State Updating and Calibration Period Selection to Improve Dynamic Monthly Streamflow Forecasts for a Wetland Management Application", submitted to HESS by Gibbs et al.

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.

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?

- 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.

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?). - 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?

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.

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).

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.

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.

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)

- Page 3

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

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).

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

- 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).

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

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.

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

Line 21: to enhance "seasonal" forecasting skill?

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?

- Page 5

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

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

- 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).

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?

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

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)

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.

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).

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

- 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?

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).

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)

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

- 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.

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).

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).

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?

- 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.

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?

- 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?

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/modele-de-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).

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

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

- 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).

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

- 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)?

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

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

- 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?

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

- 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?

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?

- Page 15

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

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

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

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

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

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?

- 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?

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).

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).

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.

- 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.

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

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

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?

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.

- Page 19

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

- 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.

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?

- 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?

- 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.

- Page 23 (Fig. 5)

Why are the results plotted only up to 2005?

- Page 25 (Fig. 7)

Why are the results plotted only from 2008 to 2010?

REFERENCES

Berthet, L. (2010). Prévision des crues au pas de temps horaire : pour une meilleure assimilation de l'information de débit dans un modèle hydrologique. Ph. D. Thesis, IRSTEA (Antony), AgroParisTech (Paris), https://pastel.archives-ouvertes.fr/pastel-00529652v1

Berthet, L., Andréassian, V., Perrin, C. and Loumagne, C. (2010). How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value of a quadratic criterion. Hydrological Sciences Journal, 55(6): 1063-1073. DOI: 10.1080/02626667.2010.505891

Krzysztofowicz, R. and Maranzano, C. (2004). Hydrologic uncertainty processor for probabilistic stage transition forecasting Journal of Hydrology, 293, 57-73

Liu, Y. and Gupta, H. (2007) Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. Water Resources Research, 43, W07401

Moradkhani, H., Sorooshian, S., Gupta, H. and Houser, P. (2005). Dual state-parameter estimation of hydrological models using ensemble Kalman filter. Advances in Water Resources, 28, 135-147

Moradkhani, H., Hsu, K.-L., Gupta, H. and Sorooshian, S. (2005b) Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the Particle Filter. Water Resources Research, 41, 1-17

Refsgaard, J. C. (1997). Validation and Intercomparison of Different Updating Procedures for Real-Time Forecasting Nordic Hydrology, 28, 65 - 84

Tangara, M. (2005). Nouvelle méthode de prévision de crue utilisant un modèle pluie-débit global. Ph. D. Thesis, IRSTEA (Antony) and École pratique des hautes études de

Paris, https://webgr.irstea.fr/wp-content/uploads/2012/07/2005-TANGARA-THESE.pdf

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2017-381, 2017.