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

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 has been shortened and more targeted. 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.

2) 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?

The description of the model parameters and structure has been expanded in Section 3.1 and Table 1. 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 has been clarified in Section 3.3. The introduction has also been updated to outline the approach to parameter updating based on analogous periods in the historical record:

The degradation in model predictive performance due to catchment non-stationarity can impact on the decisions informed by these forecasts. To address this concern, a number of studies have calibrated model parameters to subsets of the available data, by attempting to find periods in the historical record that are analogous to conditions expected in the prediction time period, and by tailoring the time period selection to compensate for deficiencies in the model structure or input data (Brigode et al., 2013; de Vos et al., 2010; Luo et al., 2012; Vaze et al., 2010; Wu et al., 2013; Zhang et al., 2011). Often there is a trade-off between the benefits of a longer calibration period which exposes the model to a more diverse range of conditions and tends to improve parameter identifiability, versus the benefits of a

shorter calibration period that exposes the model to the most recent – and hence often the most relevant – dynamics in the catchment. Demonstrating and understanding the impact of this trade-off on model predictive performance is a key research gap pursued in this study.

The Discussion has been updated to contrast to previous studies and the interesting differences found in this work. Most notably, this difference is the benefit of state updating at a lead time longer than a few days. This is outlined in Section 5.1:

Most previous studies have used state updating in a short term flood forecasting context, and found limited effect of the initial conditions after a number of days (e.g. Berthet et al., 2009; Randrianasolo et al., 2014; Sun et al., 2017). However, forecasting of flood peak and timing is a different application to the forecasting of streamflow volumes. A number of data driven modelling studies have demonstrated that monthly streamflow lagged by one (or more) months provided some useful information for forecasting at a one month lead time (e.g. Bennett et al., 2014; Humphrey et al., 2016; Yang et al., 2017). This study demonstrates that these benefits also hold when CRR models, rather than data-driven approaches, are used as the forecasting model.

3) 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.

See the above comment (2) for the changes due to the benefits at longer lead times. Thank you for the very constructive suggestions for relevant literature here and in the detailed comments. The literature review has been updated to include these references and points to improve the representation of the previous literature in the field of data assimilation.

In CRR models, catchment conditions are represented by (usually multiple) model storages, referred to as "state variables". The values of these storages at the start of a forecast period are typically determined using a warm-up period, which allows the internal model states to reach reasonable values. Given the expected influence of the initial conditions on the simulated streamflow, observed data can be assimilated into the model to update the state of the model storages. The most commonly used approaches in hydrological data assimilation include direct updating of storages (for example Demirel et al., 2013), Kalman filtering, particle filtering, and variational data assimilation (see Liu and Gupta, 2007). Berthet (2010) considered a number of tests for different updating approaches for the GRP model, a CRR model commonly used in short term streamflow forecasting applications.

4) 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 following has been added to the manuscript.

The assumptions of the post-processor residual error model used to estimate predictive uncertainty for monthly volumes are different to the assumptions of the residual error model used in the likelihood function for calibrating the daily GR4J model. As outlined in Section 3.2, the GR4J model is calibrated at the daily scale to observed streamflow using the standard least squares likelihood function, because it better captures the high daily flows, important for estimating the monthly volumes. The post-processing error model for the monthly volumes is designed to capture the predictive uncertainty in these monthly volumes, in particular the heteroscedasticity and skew of the residuals (McInerney et al., 2017; Refsgaard, 1997). These choices of residual error models at the daily and monthly time scales contribute to the study objectives of reliable forecasts at the monthly time scale, and are common in forecasting applications (for example, Lerat et al., 2015).

5) 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 has been 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.

6) 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. Clarification of the daily scale GR4J model have been added in Sections 3.1 and 3.3. The following has been added at the start of Section 3.5 to be clear when the monthly aggregation occurs (prior to the application of the post-processor error model):

The monthly streamflow forecasts are obtained by aggregating the daily GR4J simulations. In order to quantify predictive uncertainty using a residual error model, the monthly-aggregated GR4J simulations, $\boldsymbol{Q}^{\theta}$, are compared to observed monthly streamflow volumes, $\widetilde{\boldsymbol{Q}}$. The quantification of error is based on residuals errors, defined by the differences between observed and simulated monthly streamflow. Separate error models are estimated for the GR4J predictions for each catchment and for each type of forcing data (observed or forecast rainfall), as follows:

7) 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

8) 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 has been updated as follows:

The parameters of these models have a limited relationship to measureable catchment attributes (e.g. soil horizon depth) (e.g. Fenicia et al., 2014), and typically require calibration to observed streamflow data (noting that physical models also require some calibration (Mount et al., 2016; Pappenberger and Beven, 2006)).

9) 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, the manuscript has been updated as follows:

It is generally considered that longer calibration periods produce more robust parameter estimates, as a longer period exposes the model to a more diverse range of catchment conditions and flow events (Wu et al., 2013), however this is not always the case (for example Brigode et al., 2013).

10) 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. The manuscript has been updated as follows:

In this work, the term "non-stationary" is used to refer to situations where physical changes are expected to have occurred in a catchment, and where there is evidence to reject the hypothesis of stationarity. In practice, catchments may have different "degrees" of non-stationarity, depending on the evidence available to reject the hypothesis of stationarity, the degree of change in a catchment, and the time scales over which the changes take place.

- Page 3

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

This has been clarified to refer to groundwater abstractions, as opposed to other potential causes of groundwater depletion:

Examples of non-stationary changes in catchment conditions that could be expected to change the rainfall-runoff relationship include changes in land use or land-cover (e.g., deforestation, urbanization), land drainage, interception (e.g. dams or diversions), groundwater abstractions or responses to changes in climate (Milly et al., 2015).

12) 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, has been included and the paragraph edited as follows:

Predictive uncertainty quantification is another major aspect of practical streamflow prediction. Many approaches are available to quantify predictive uncertainty, from approaches that identify a range of model parameters that represent the behaviour of the catchment using approaches such as generalised likelihood uncertainty estimation (GLUE) (Beven and Binley, 1992), to post-processor approaches (e.g. Krzysztofowicz and Maranzano, 2004), to disaggregation approaches that attempt to characterise each individual source of error explicitly (e.g. Kavetski et al., 2003; Vrugt et al., 2005). In this work, predictive uncertainty is estimated using an aggregated post-processor residual error model. The residual error model represents the differences between the hydrological model predictions and observed data, without trying to identify the contributing sources (Evin et al., 2014). The post-processor approach is chosen because it can lead to more robust estimates of predictive uncertainty compared to joint calibration of all parameters (i.e. estimating CRR model and error model parameters concurrently) (Evin et al., 2014).

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

The terminology has been clarified, see response at 12) above. “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

14) 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 has been changed to:

Much of the skill in seasonal streamflow forecasts over periods following rainy seasons is commonly attributed to accurately representing initial catchment conditions (Koster et al., 2010; Pagano et al., 2004; Wang et al., 2009). In contrast, forecast skill over periods following dry seasons is generally attributed to both initial catchment conditions and meteorological inputs (Maurer and Lettenmaier, 2003; Wood and Lettenmaier, 2008). The impact of the 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), although this time frame is generally catchment dependent.

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

The change has been included throughout.

16) 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 has been changed to:

Studies on observed data assimilation and CRR model state updating have focused primarily on flood forecasting with short lead-times. The benefits at longer lead-times (e.g. seasonal) to forecast water availability have received less attention in the published literature.

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

This is a useful clarification, and “Calibration Period Length (CPL)” is now used throughout.

18) Line 21: to enhance "seasonal" forecasting skill?

The objectives have been updated as:

Evaluate the ability of state updating in a daily CRR model to improve predictive performance when forecasting streamflow volume for the upcoming month.

19) 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 objective has been updated to “assess” rather than “demonstrate”

Assess the degree to which using a shorter calibration period, that is more representative of the forecast period, can improve predictive performance, in particular when there is evidence of catchment non-stationarity.

- Page 5

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

Correct, the typo has been corrected.

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

A boxplot of the monthly flow in Drain M has been included to demonstrate the variability (Figure 2).

- Page 6

22) 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 at 1), we agree that Section 2 includes superfluous detail that will be removed. This section on the separate water balance model in eWater Source has been removed altogether.

23) 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.

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

Correct, the typo will be corrected throughout.

25) 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 (see response 18).

26) 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. This sentence has been removed.

27) 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 has been updated to clarify that there is a rainfall gradient from south to north:

The mean annual rainfall for the region is in the range of 600 mm in the north to 675 mm in the south

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

The acronym has been 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

29) 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?

That is correct. 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. The manuscript has been modified as follows:

*POAMA-2 predictions have a coarse spatial resolution (~250 km), which does not capture the spatial variability in catchment-scale rainfall. For the purposes of this application, the POAMA-2 rainfall hindcasts (i.e. forecasts developed by applying the modelling system to the historical period) at the relevant pixel were downscaled to each climate station in the study region (**Error! Reference source not found.**) using the statistical downscaling method detailed in Shao and Li (2013). Further details of the downscaling approach are provided in*

Humphrey et al. (2016).30) 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 Section 2.1:

*Mosquito Creek flows into Bool Lagoon (Catchment C1 in **Error! Reference source not found.**, area 1002 km^2). Drain M commences at the outlet of Bool Lagoon, and a large catchment flows into Drain M between Bool Lagoon and a diversion point at Callendale (Catchment C3 with an area of 2200 km^2). Finally, the Drain M local catchment contributes flow downstream of the Callendale diversion point, flowing into Lake George (Catchment C2, area 383 km^2).*

31) 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. The conflicting water requirements for the case study site have been clarified in Section 2.2 as follows:

Drain M serves multiple competing demands on the water resources available in this catchment system. These demands influence the decision to use the regulators along the system:

a) *Bool Lagoon has water requirements that influence releases from the lagoon into Drain M.*
b) *Lake George has water requirements to maintain the estuarine ecology of the lake, to support its significance as a biological resource, and as a resource for recreational fishing.*
c) *The ocean outlet requires some flow to prevent sediment from entering Lake George and to maintain connectivity to the sea (which allows fish movement and aids fish recruitment). However, high flows may impact on sea grasses, due to their low salinity and high nutrient load.*
d) *The wetlands of the Upper South East to the north typically benefit from as much water as possible from the Drain M system.*

32) 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. The section has been shortened as follows:

The identification of high quality data is important because biases and systematic changes in the measurement of hydrological data can significantly affect model calibration and lead to non-stationarity in the estimated model parameters (Westra et al., 2014). Analysis of the data and monitoring stations suggested that streamflow data uncertainty is expected to be low, given the regular cross sections of the weirs used for monitoring stage and upstream drains, and the high number of gaugings (between 78 and 166 flow gaugings at each flow station) available to develop stage-discharge relationships.

- Page 8

33) 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.

See response to 32)

33) 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 information has been highlighted earlier in section 2.1 as follows:

*In the region where the case study catchments are located, plantation forestry expanded substantially in the late 1990s. Changes in the relationship between rainfall and runoff also occurred during this period, evidenced by the reduced slope in the plot of cumulative runoff against cumulative rainfall (double-mass analysis) in **Error! Reference source not found.** (Searcy et al., 1960; Yihdego and Webb, 2013). The runoff ratio in catchment C1 is approximately 0.045 before year 2000, but reduces by 70% to 0.013 after 2000. The runoff ratio in catchment C2 is around 0.088 before year 2000, but reduces by 30% to 0.061 after 2000. This comparison provides stronger evidence of non-stationarity in catchment C1 than in catchment C2. Other studies have also investigated the link between changes in the hydrology and changes in land use in the region (Avey and Harvey, 2014; Brookes et al., 2017). These changes have implications on the choice of calibration data period, as data from the 1970s may not be representative of hydrological conditions in the 2000s.*

34) 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) has also been 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.

35) 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 has been clarified in the manuscript as follows:

Note that the catchments considered have a relatively slow streamflow response. Consequently, the pre-specified split to the routing store of 0.9 in the original specification of the GR4J model may be too low for these catchments. To mitigate this potential deficiency, we have modified the GR4J model so that the split between the routing store and the direct runoff is included as an explicit calibration parameter termed split.

- Page 9

36) 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 and has been removed.

37) 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

38) 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 trade-off has been highlighted in two places in the revised manuscript:

Section 1:

Often there is a trade-off between the benefits of a longer calibration period which exposes the model to a more diverse range of conditions and tends to improve parameter identifiability, versus the benefits of a shorter calibration period that exposes the model to the most recent – and hence often the most relevant – dynamics in the catchment. Demonstrating and understanding the impact of this trade-off on model predictive performance is a key research gap pursued in this study.

Section 3.3

Calibration period lengths of CPL = 10 years and CPL = 20 years length are considered, to assess the trade-off between using a longer calibration period to expose the model to more diverse catchment conditions and improve parameter identifiability, versus using a shorter calibration period length to expose the model to more recent hydrological dynamics.

39) 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 has been updated to include these references, in particular Lui and Gupta (2007) as a review of data assimilation approaches:

In CRR models, catchment conditions are represented by (usually multiple) model storages, referred to as "state variables". The values of these storages at the start of a forecast period are typically determined using a warm-up period, which allows the internal model states to reach reasonable values. Given the expected influence of the initial conditions on the simulated streamflow, observed data can be assimilated into the model to update the state of the model storages. The most commonly used approaches in hydrological data assimilation include direct updating of storages (for example Demirel et al., 2013), Kalman filtering, particle filtering, and variational data assimilation (see Liu and Gupta, 2007). Berthet (2010) considered a number of tests for different updating approaches for the GRP model, a CRR model commonly used in short term streamflow forecasting applications.

40) 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 has been made clearer in the revised manuscript.

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

This change has been made.

- Page 11

42) 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 and has been removed.

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

"Case" and "model configuration" has been clarified and made consistent throughout the manuscript, and any forward referencing removed. The clarification is as follows:

Two options for state updating (with versus without) and two options for calibration period length (CPL = 10 years versus CPL = 20 years) are considered. The combination of these options leads to four model configurations. Four different cases are considered for each model configuration, given by the combinations of two catchments (C1 and C2) and two sources of climate data (observed and forecast). This results in a total of 16 scenarios considered.

- Page 12

44) 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.

45) 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 so that the worst-performing case for each metric had some area shown in the plot. However, it is agreed that this was confusing, and a 0-1 normalisation is now used in Figure 4.

- Page 13

46) 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 reference to an unfair advantage has been removed, as 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 has been changed to:

The reference distribution for each month is calculated as the empirical distribution of all observed data in that month, using the entire set of observed data (including data from the prediction period). This approach provides a stringent baseline for the CRPS normalization in Eq. (13).

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

This typo has been corrected.

- Page 14

48) 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 has been clarified in Section 3.5:

When forecast rainfall is used as input to GR4J, an ensemble of daily streamflow forecasts is produced (with a single GR4J streamflow time series per rainfall forecast time series). Each such "individual" daily GR4J time series is then aggregated to a monthly time step. The time series $\boldsymbol{Q}^{\theta}$ is constructed from the time series of medians of the individual monthly streamflow time series. This use of the median streamflow forecasts from the multiple ensembles of the meteorological ensemble forecasts may result in some information loss, but aggregation approach of streamflow forecast ensembles is commonly used in operational applications (e.g. Lerat et al., 2015; Matte et al., 2017; Schepen et al., 2017; Wani et al., 2017). Further work is needed to more fully utilise the information from ensemble forecasts when developing post processing models.

49) 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 (see response 42).

- Page 15

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

The sentence has been removed.

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

This is a useful clarification, and “Calibration Period Length (CPL)” is now used throughout.

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

This section has been largely rewritten in the revised manuscript. Significance tests have not been calculated due to the lack of replicates.

53) 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.

On review of this conclusion, it was found that this was in part due to the different time periods used for the observed and forecast rainfall cases (as outlined in Section 3.7 previously). The observed rainfall results have been recalculated over only the period used for the forecast rainfall case, and now the results are consistent between the two rainfall cases.

54) 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, this point has been removed.

55) 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:

*The differences between the streamflow predictions obtained in the two catchments C1 and C2 (for the case of GR4J forced with observed rainfall) are illustrated in Figure 7 for the most recent period 2009-2011. In catchment C1, using a longer calibration period length tends to yield wider prediction limits and an overestimation of the observed flow in 2009 and 2010, whereas using the shorter calibration length provides a better capture of the catchment response in these two years. In contrast, in catchment C2, which has less evidence of non-stationarity (Section **Error! Reference source not found.**), the calibration period length makes very little difference on the resulting streamflow predictions.*

- Page 16

56) 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. The manuscript has been changed as follows:

An exception to the pattern of the median parameter values being insensitive to calibration period lengths can be seen in 1999, where the use of the 10 year calibration period length produces higher values of X4 and lower values of X2 and the split introduced in this study (Section 3.1). This exception could represent a model fitting anomaly resulting from a shorter calibration period length.

57) 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. This point has been added to section 5.1 (see response 2)

58) 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” has been changed to “sharpness” throughout.

59) 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” has been added to this sentence.

- Page 17

60) 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.

This point has been removed, as on review it was considered not directly relevant to the main results presented.

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

See response to 60)

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

See response to 60)

63) 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?

See response to 53)

64) 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.

See response to 60)

- Page 19

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

Parameter meanings and units have been included. A GR4J flow chart was not included for brevity, with the reference to Perrin et al. (2003) use for this purpose.

- Page 20

66) 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 has been adjusted accordingly.

67) 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?

This figure has now been removed as it did not directly contribute to the results presented.

- Page 21 (Fig. 3)

68) 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 (Avey 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, this has been added to Section 3.3:

This methodology allows the identification of changes in parameter distributions over time, without the need to identify specific periods when changes in the rainfall-runoff response may have occurred.

- Page 22 (Fig. 4)

69) 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 following has been added to Section 3.7:

To ensure a consistent comparison of multiple model scenarios, the metrics are computed as follows:

- *the same period is used to calculate the metrics in all cases. This period was determined by the availability of the forecast rainfall, from May 2001 to April 2011.*
- *the performance metrics are normalized by linearly scaling the worst value to a value of 0 and the best value to 1,*

$$M_r = \frac{M - M_w}{M_b - M_w} \qquad (1)$$

where the worst and best values for each metric, M_w and M_b, respectively, are listed in ***Error! Reference source not found.****. The remainder of the presentation, in particular* ***Error! Reference source not found.****, reports the normalized metrics computed using Eq. (15).*

- Page 23 (Fig. 5)

70) 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. The full plot has been provided as supplementary material to show this information.

Page 25 (Fig. 7)

71) Why are the results plotted only from 2008 to 2010?

See response to 70).

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 #2

General:

The manuscript of Gibbs et al. evaluates the effect of calibration setup on the GR4J model performance within 2 Australian basins on 1-month lead-time hydrologic forecast. Authors draw mainly following conclusions based on 2-basin analysis using 5 indicators: (I) the length of calibration period does not necessarily should be as long as possible, in particular when changes in flow regimes are observed. Additionally, (II) the authors state that a simple model state updating improves hydrological forecast at 1-month lead time.

Based on my review, I consider the overall topic to be relevant for HESS, however, some parts of the manuscript need to improved and clarified, as further suggested below.

Thank you for the constructive comments that have improved the clarity and contribution of the manuscript.

Major comments:

1. In general, it is not surprising that updating model initial conditions has benefits on hydrologic forecast. Additionally, it is not surprising that changes of physiographic conditions may change the catchment's response, indicating different information content/validity of observed discharge data on model parameters. This only confirms observations of previous studies, which some of them are cited. Would be nice to more clearly demonstrate benefits over existing/operational approaches (in terms of costs etc). In particular, when the title includes words like "Management Application".

Commentary on the benefits over the existing approach has been added in Section 5.3:

The forecasting approaches developed in this work can support improved water management in the drainage system considered. The approach currently used by the management authority is very conservative: streamflow forecasts are not attempted, and changes in water management are made only once downstream requirements have been met. With the forecasting models and methods developed in this work, it becomes possible to produce streamflow forecasts with a high reliability, improved sharpness and reduced bias. Thus it becomes possible to provide useful probabilistic estimates of how likely it is that the downstream flow requirements will be met in the next month. With this information, managers can more confidently consider increasing the frequency and duration of inundation for many of the wetlands in the region, and can make decisions on management changes much earlier in the season.

The Discussion section has also been expanded to include Section 5.1 to highlight the key findings over previous approaches:

Most previous studies have used state updating in a short term flood forecasting context, and found limited effect of the initial conditions after a number of days (e.g. Berthet et al., 2009; Randrianasolo et al., 2014; Sun et al., 2017). However, forecasting of flood peak and timing is a different application to the forecasting of streamflow volumes. A number of data driven modelling studies have demonstrated that monthly streamflow lagged by one (or more) months provided some useful information for forecasting at a one month lead time (e.g. Bennett et al., 2014; Humphrey et al., 2016; Yang et al., 2017). This study demonstrates that these benefits also hold when CRR models, rather than data-driven approaches, are used as the forecasting model.

2. Unfortunately, the analysis is limited to two basins, which really can't be used to draw any conclusions (as authors also recognise in the end). I would strongly encourage authors to enlarge the number of basins and events. Another two basins may yield completely different results; therefore, generality should be avoided.

It is agreed that general results cannot be drawn from the results based on two basins. The drainage system considered was based on a user need for seasonal forecasts, which resulted in the limited application to two basins. The Section 5.4 has been included to outline this important limitation:

The enhancements to predictive performance of streamflow forecasts from state updating and a shorter calibration period have been demonstrated on two catchments. These catchment were selected based on an established user need for seasonal forecasts to improve the water management of a channel drainage system with multiple competing demands. Importantly, the case study catchments in this work are ephemeral and dry, with low runoff ratios. These types of catchments are known to be challenging to model (McInerney et al., 2017; Ye et al., 1997). Future work will evaluate the proposed seasonal streamflow forecasting techniques over a wider range of catchments and environmental conditions.

3. Would be nice to relate your results with another study, which details CRR state updating for 1-month forecast. However, I wonder, whether it is really the effect of state updating here. It is well recognized that the effect of initial conditions (based on discharge observations), in such small basins, diminishes after a couple of days. Please, comment on this.

The result that model updating still improved forecast performance at the lead-time of one month is considered one of the key results of this work. The controlled study for testing the model with and without state updating for all the cases considered (observed and forecast rainfall, two lengths for the calibration period and two basins), demonstrates that for the scenarios considered the effect can be attributed to state updating. The point at Response 2, on the potentially limited generality of this result, is relevant and has been included.

It is agreed that the majority of the literature focused on short term flood forecasting has found limited effect of initial conditions after a few days (e.g. Berthet et al., 2009). However, forecasting of flood peak and timing is a different application to that considered here, forecasting water volume available for environmental use. See response 1. for the inclusion of this point in the manuscript.

4. The description of basin and data is way too long and detailed (3 pages), in particular when the discussion and conclusion do not come to those details at all.

Reviewer 1 raised a similar issue, and as such Section 2 has been shortened to be more targeted and remove surplus detail 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 outlined at Comment 1, commentary on the benefits over the existing approach has been added at Section 5.3.

5. Please, place error bars into figure 4, in the same way as the uncertainty is presented in following figures and provide discussion.

The uncertainty in the streamflow time series had been produced and assessed (e.g Figures 5 & 6) because it is a direct of the output of the modelling setup. However, all of the streamflow replicates are used in the calculation of the performance metrics (reliability, precision and CRPS in particular), and as such the uncertainty in the values of the metrics is not easily derived. To the authors' knowledge, it is no common to provide the uncertainty in performance metrics related to streamflow forecast uncertainty (e.g. Alfieri et al., 2014; McInerney et al., 2017; Pappenberger et al., 2015; Renard et al., 2010).

We do acknowledge that there is uncertainty in the values for the metrics, due to finite number of replicates and the finite length of the observed data (McInerney et al., 2017) and that in the presence of strong autocorrelation in the streamflow error time series this uncertainty may be significant. However, to develop an approach to quantify the metric uncertainties is beyond the scope of this study. Instead, we follow the guidance provided by McInerney et al. (2017) who suggested that these uncertainties are quite small and are unlikely to impact on the conclusions of the study. We do acknowledge that further work is needed in this area.

6. Discussion about alternative types of observations (besides Q) may be provided in the manuscript.

The literature review has been expanded as follows to refer to previous studies that have used other types of observations that could be used for state updating (e.g. either remotely sensed or ground based soil moisture), and why streamflow was focused on in this study.

Updating the states of conceptual rainfall-runoff models is not straightforward, as any environmental model is at best an approximate representations of the real catchment (Berthet et al., 2009). A number of observed data sources can be used to update model storages, including observed streamflow and in-situ or remotely sensed soil moisture. From these options, Li et al. (2015b) suggests that gauged discharge data assimilation is a more effective way to improve short-term forecasts and is still preferred for operational streamflow forecasting purposes.

7. What is the main applicability of your findings? Are they going to be used operational, if yes, what are the benefits over existing forecast method? Please, clarify.

See response to Comment 1.

Minor:

- Third sentence from the Introduction regarding Drain M catchment should be moved somewhere towards the end of Introduction.

This change has been made.

- P 6, L.5: evapoconcentration => evapotranspiration?

This typo has been corrected

- Section name 2.2 "Streamflow and streamflow data": sounds a bit repetitive

The section has been changed to "Streamflow data"

- P 9, L.21 "burn-in" into quotes

This change has been made.

- Eq. 12 is wrong, sum in the denominator is missing

Correct, the equation has been updated.

- Caption of figure 2: "POAMA" => "POAMA-2"

This change has been made.

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 #3

This manuscript presents an original study focused on two important aspect of monthly streamflow forecasting: state updating and the selection of an appropriate calibration period. The watershed on which the methods are implemented and tested is a very interesting case study. This semi-arid watershed is extensively impacted by human intervention, as it includes diversions and "Drain M", so runoff water can be directed either toward the north to benefit wetlands or toward the south to maintain fish ecosystems and comply with other constraints. Adequate management of this watershed appears very difficult and it is evident from the authors' description of the situation that monthly hydrological forecasts are essential. The authors show that selecting a calibration period during which the conditions are close to the expected forecast conditions can improve forecasts performance substantially.

In my opinion, this is a well-written paper (very clear!) that brings interesting novel knowledge in the field of ensemble monthly streamflow forecasting. I also find it completely appropriate for publication in HESS, especially since the case study raises issues regarding water management and tradeoffs between ecosystem services and human activities.

I appreciate that the authors included a short discussion about the uncertainty related to the gauging measurements (page 8 first paragraph). I also like the idea of adding the "split" parameter to GR4J in the calibration process.

The conclusions drawn by the authors regarding the tradeoff between the length of the calibration database versus its "representativeness" is interesting. I appreciate that they recognize that said conclusions might be limited to their specific case study and that further investigation for a wider range of hydro-climatic regimes would be needed. I only have very few minor comments and suggestions that I think could improve the paper. I strongly recommend that it be published in HESS.

Thank you for the constructive comments that have improved the manuscript.

Minor comments:

1. Why did you consider only the median of the hydrological model predictions rather than the whole ensemble?

This question kept bugging me all along while I was reading. You have access to meteorological ensemble forecasts (page 6 line 30 to page 7 line 5). They are expected to account for the uncertainty related to the meteorological conditions (or at least a part of this uncertainty). Why then did you not keep all the scenarios after passing everything through the hydrological model? Is it to make it more comparable to the case where the hydrological model is forced by (deterministic) observations? If so, I personally don't see why this would be necessary. And then after, you use a statistical method to dress the median back to an ensemble. Is it because you found out that hydrological ensemble forecasts built only from meteorological ensembles were underdispersed and would have needed post-processing?

In my opinion, those choices (i.e. using the median, thus ignoring the other ensemble members, and then dressing the median into an ensemble) really need further explanations/justifications.

The focus of this work was to investigate the impact of the state updating and calibration periods on streamflow predictive uncertainty, and a current best practice post-processing approach was adopted to estimate this predictive uncertainty. It is common practice to use deterministic inputs, often based on summarising an ensemble of forecasts using summary metrics (e.g. mean or median) as the input to post-processing methods (e.g. Lerat et al., 2015; Matte et al., 2017; Schepen et al., 2017; Wani et al., 2017). As such, it was considered beyond the scope of this work to also consider improving methods for applying post-processing models.

However, it is acknowledged that that by only considering the median streamflow simulated across the forecast rainfall ensemble that some information is lost. The paper has been modified as follows:

When forecast rainfall is used as input to GR4J, an ensemble of daily streamflow forecasts is produced (with a single GR4J streamflow time series per rainfall forecast time series). Each such "individual" daily GR4J time series is then aggregated to a monthly time step. The time series $\boldsymbol{Q}^{\theta}$ *is constructed from the time series of medians of the individual monthly streamflow time series. This use of the median streamflow forecasts from the multiple ensembles of the meteorological ensemble forecasts may result in some information loss, but aggregation approach of streamflow forecast ensembles is commonly used in operational applications (e.g. Lerat et al., 2015; Matte et al., 2017; Schepen et al., 2017; Wani et al., 2017). Further work is needed to more fully utilise the information from ensemble forecasts when developing post processing models.*

2. There are a couple of (very minor) elements that could be clearer

- Page 5 line 30: please add a reference for the Ramsar list. I didn't know this list before reading the manuscript so I looked it up on the web. I think that a reference would be helpful to be sure that other readers like me know what you are talking about.

The reference Matthews (1993) has been added.

- Page 9 line 21: What is a "burn-in" period? At first I thought it was a synonym of "warm-up period" in the context of the DREAM algorithm, but I am really not sure.

"burn-in" is the term commonly used in Markov Chain Monte Carlo modelling and is preferred to be maintained. Both "warm-up" and "burn-in" refer to the period at the start of the analysis that is discarded and not considered further, but the applications are different and as such the different terminology considered necessary.

3. There are a few typos in the manuscript and I am unsure of the spelling for 1-2 words:

- Page 3 line 27-28: parenthesis typo, replace "(McInerney et al, 2017)" by "McInerney et al. (2017)"

The typo has been corrected.

- Page 11 equation (5): Something seems wrong with the curly brace

The curly brace is commonly used notation for conditional expressions, in this case to avoid dividing by 0 when λ=0.

- Page 11 line 1: I think that the sentence "(…) are available, for example, the ensemble Kalman filter (…)" should be split, as in: "(…) are available. For example, the ensemble Kalman filter (…)"

The sentence has been largely reworded.

- Page 11 line 12: day "n" and month "t" should be in italics.

The change has been made.

- Page 12 line 17: I think that "straightforward" should be written in one word.

The change has been made.

- Page 13 equation 12: A summation seems to be missing at the denominator. Also, I don't think there should be a "t" index for the average streamflow, since by definition it is independent from time (it is the average of the time series).

The summation typo has been corrected. It is agreed that the index t for the average is also incorrect. Thank you for picking these up.

- Page 14 line 6 (and several other places in the manuscript): Why do you write "observ***ed*** rainfall" but not "forecast***ed*** rainfall" I am not a native English speaker so perhaps I am completely wrong, but I could not help but finding this strange.

This is a strange nuance of the English language. Both "forecast" and "forecasted" can be used as past tense for the verb forecast. It was considered that "forecast" is more commonly used in the literature and is proposed to be maintained throughout.

- Page 14 line 16: Is there a "to" missing in "The changes due adopting (…)"?

The typo has been corrected.

References

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., and Salamon, P.: Evaluation of ensemble streamflow predictions in Europe, Journal of Hydrology, 517, 913-922, 2014.

Avey, S. and Harvey, D.: How water scientists and lawyers can work together: A 'down under' solution to a water resource management problem, Journal of Water Law, 24, 25-61, 2014.

Bennett, J. C., Wang, Q. J., Pokhrel, P., and Robertson, D. E.: The challenge of forecasting high streamflows 1 & 3 months in advance with lagged climate indices in southeast Australia, Nat. Hazards Earth Syst. Sci., 14, 219-233, 2014.

Berthet, L.: Prévision des crues au pas de temps horaire : pour une meilleure assimilation de l'information de débit dans un modèle hydrologique, 2010. AgroParisTech, 2010.

Berthet, L., Andreassian, V., Perrin, C., and Javelle, P.: How crucial is it to account for the antecedent moisture conditions in flood forecasting? Comparison of event-based and continuous approaches on 178 catchments, Hydrology and Earth System Sciences, 13, 819-831, 2009.

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrological Processes, 6, 279-298, 1992.

Brigode, P., Oudin, L., and Perrin, C.: Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change?, Journal of Hydrology, 476, 410-425, 2013.

Brookes, J. D., Aldridge, K., Dalby, P., Oemcke, D., Cooling, M., Daniel, T., Deane, D., Johnson, A., Harding, C., Gibbs, M., Ganf, G., Simonic, M., and Wood, C.: Integrated science informs forest and water allocation policies in the South East of Australia, Inland Waters, 7, 358-371, 2017.

Coron, L., Andreassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F.: Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, Water Resources Research, 48, W05552, 2012.

de Vos, N. J., Rientjes, T. H. M., and Gupta, H. V.: Diagnostic evaluation of conceptual rainfall–runoff models using temporal clustering, Hydrological Processes, 24, 2840-2850, 2010.

Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models, Water Resources Research, 49, 4035-4053, 2013.

Evin, G., Thyer, M., Kavetski, D., McInerney, D., and Kuczera, G.: Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, Water Resources Research, 50, 2350-2375, 2014.

Guo, D., Westra, S., and Maier, H. R.: Impact of evapotranspiration process representation on runoff projections from conceptual rainfall-runoff models, Water Resources Research, doi: 10.1002/2016WR019627, 2017. 2017.

Humphrey, G. B., Gibbs, M. S., Dandy, G. C., and Maier, H. R.: A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network, Journal of Hydrology, 540, 623-640, 2016.

Kavetski, D., Franks, S. W., and Kuczera, G.: Confronting Input Uncertainty in Environmental Modelling. In: Calibration of Watershed Models, American Geophysical Union, 2003.
Krzysztofowicz, R. and Maranzano, C. J.: Hydrologic uncertainty processor for probabilistic stage transition forecasting, Journal of Hydrology, 293, 57-73, 2004.
Lerat, J., Pickett-Heaps, C., Shin, D., Zhou, S., Feikema, P., Khan, U., Laugesen, R., Tuteja, N., Kuczera, G. T., M, and Kavetski, D.: Dynamic streamflow forecasts within an uncertainty framework for 100 catchments in Australia, 36th Hydrology and Water Resources Symposium: The art and science of water, Barton, ACT, 1396-1403, 2015.
Li, H. B., Luo, L. F., Wood, E. F., and Schaake, J.: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting, Journal of Geophysical Research-Atmospheres, 114, 2009.
Liu, Y. and Gupta, H. V.: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, Water Resources Research, 43, 2007.
Luo, J., Wang, E., Shen, S., Zheng, H., and Zhang, Y.: Effects of conditional parameterization on performance of rainfall-runoff model regarding hydrologic non-stationarity, Hydrological Processes, 26, 3953-3961, 2012.
Matte, S., Boucher, M. A., Boucher, V., and Fortier Filion, T. C.: Moving beyond the cost–loss ratio: economic assessment of streamflow forecasts for a risk-averse decision maker, Hydrol. Earth Syst. Sci., 21, 2967-2986, 2017.
Matthews, G. V. T.: The Ramsar Convention on Wetlands: its History and Development, Ramsar Convention Bureau, Gland, Switzerland, 1993.
McInerney, D., Thyer, M., Kavetski, D., Lerat, J., and Kuczera, G.: Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors, Water Resources Research, 53, 2199-2239, 2017.
Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., Stouffer, R. J., Dettinger, M. D., and Krysanova, V.: On Critiques of "Stationarity is Dead: Whither Water Management?", Water Resources Research, 51, 7785-7789, 2015.
Mount, N. J., Maier, H. R., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F. J., and Abrahart, R. J.: Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei Science Plan, Hydrological Sciences Journal, 61, 1192-1208, 2016.
Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., and Salamon, P.: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction, Journal of Hydrology, 522, 697-713, 2015.
Perrin, C., Michel, C., and Andreassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275-289, doi:210.1016/S0022-1694(1003)00225-00227, 2003.
Randrianasolo, A., Thirel, G., Ramos, M. H., and Martin, E.: Impact of streamflow data assimilation and length of the verification period on the quality of short-term ensemble hydrologic forecasts, Journal of Hydrology, 519, 2676-2691, 2014.
Refsgaard, J. C.: Validation and Intercomparison of Different Updating Procedures for Real-Time Forecasting, Hydrology Research, 28, 65-84, 1997.
Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resources Research, 46, 2010.
Schepen, A., Zhao, T., Wang, Q. J., and Robertson, D. E.: A new method for post-processing daily sub-seasonal to seasonal rainfall forecasts from GCMs and evaluation for 12 Australian catchments, Hydrol. Earth Syst. Sci. Discuss., 2017, 1-27, 2017.
Searcy, J. K., Hardison, C. H., and Langein, W. B.: Double-mass curves; with a section fitting curves to cyclic data, Report 1541B, 1960.
Shao, Q. and Li, M.: An improved statistical analogue downscaling procedure for seasonal precipitation forecast, Stochastic Environmental Research and Risk Assessment, 27, 819-830, 2013.
Sun, L., Seidou, O., and Nistor, I.: Data Assimilation for Streamflow Forecasting: State-Parameter Assimilation versus Output Assimilation, Journal of Hydrologic Engineering, 22, 04016060, 2017.
Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J. M., Viney, N. R., and Teng, J.: Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies, Journal of Hydrology, 394, 447-457, 2010.
Vrugt, J. A., Diks, C. G. H., Gupta, H. V., Bouten, W., and Verstraten, J. M.: Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, Water Resources Research, 41, 2005.
Wang, E., Zheng, H., Chiew, F., Shao, Q., Luo, J., and Wang, Q. J.: Monthly and seasonal streamflow forecasts using rainfall-runoff modeling and POAMA predictions, 19th International Congress on Modelling and Simulation (Modsim2011), 2011. 3441-3447, 2011.
Wani, O., Beckers, J. V. L., Weerts, A. H., and Solomatine, D. P.: Residual uncertainty estimation using instance-based learning with applications to hydrologic forecasting, Hydrol. Earth Syst. Sci., 21, 4021-4036, 2017.
Westra, S., Thyer, M., Leonard, M., Kavetski, D., and Lambert, M.: A strategy for diagnosing and interpreting hydrological model nonstationarity, Water Resources Research, doi: 10.1002/2013WR014719, 2014. 2014.
Wu, W., May, R. J., Maier, H. R., and Dandy, G. C.: A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks, Water Resources Research, 49, 7598-7614, 2013.
Yang, T., Asanjan, A. A., Welles, E., Gao, X., Sorooshian, S., and Liu, X.: Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information, Water Resources Research, 53, 2786-2812, 2017.
Ye, W., Bates, B. C., Viney, N. R., Sivapalan, M., and Jakeman, A. J.: Performance of conceptual rainfall-runoff models in low-yielding ephemeral catchments, Water Resources Research, 33, 153-166, 1997.
Yihdego, Y. and Webb, J.: An Empirical Water Budget Model As a Tool to Identify the Impact of Land-use Change in Stream Flow in Southeastern Australia, Water Resour. Manag., 27, 4941-4958, 2013.
Zhang, H., Huang, G. H., Wang, D., and Zhang, X.: Multi-period calibration of a semi-distributed hydrological model based on hydroclimatic clustering, Advances in Water Resources, 34, 1292-1303, 2011.

State Updating and Calibration Period Selection to Improve Dynamic Monthly Streamflow Forecasts for ~~a Wetland~~an Environmental Flow Management Application

Matthew S. Gibbs[1,2], David McInerney[1], Greer Humphrey[1], Mark A. Thyer[1], Holger R. Maier[1], Graeme C. Dandy[1], Dmitri Kavetski[1]

[1]School of Civil, Environmental and Mining Engineering, The University of Adelaide, North Terrace, Adelaide, South Australia, 5005, Australia
[2]Department of Environment, Water and Natural Resources, Government of South Australia, PO Box 1047, Adelaide, 5000.

Correspondence to: Matthew Gibbs (matthew.gibbs@adelaide.edu.au)

Abstract. ~~Sub-seasonal~~Seasonal streamflow forecasts provide useful information for a range of water resource management and planning applications. This work ~~has focused~~focuses on improving ~~forecasts for one~~ such ~~application:~~forecasts by considering the ~~management of water available in an open channel drainage network to maximise environmental and social outcomes in a region in southern Australia. Conceptual rainfall-runoff models with a postprocessor error model for uncertainty analysis were applied to provide forecasts of monthly streamflow. Two~~following two aspects ~~were considered to improve the accuracy of the forecasts~~: 1) state updating to force the models to match observations from the start of the forecast period, and 2) selection of a shorter calibration period that is more representative of the forecast period~~. Five metrics were~~, compared a longer calibration period traditionally used. The analysis is undertaken in the context of using streamflow forecasts for environmental flow water management of an open channel drainage network in southern Australia. Forecasts of monthly streamflow were obtained using a conceptual rainfall-runoff model combined with a post-processor error model for uncertainty analysis. This model setup is applied to two catchments, one with stronger evidence of non-stationarity than the other. A range of metrics are used to assess ~~forecast~~ different aspects of predictive performance~~, representing the~~ including reliability, ~~precision~~sharpness, bias and ~~skill of the forecasts produced, using both observed and forecast climate data.~~accuracy. The results indicate that~~ assimilating observed streamflow data into the model, by~~, for most scenarios and metrics, state updating ~~the storage level at the start of a forecast period, improved the~~improves predictive performance ~~of the forecasts across the metrics when compared to an approach that "warmed up" the storage levels using historical climate data. The~~ for both observed rainfall and forecast rainfall sources. Using the shorter calibration period ~~improved the~~also improves predictive performance~~ of the forecasts~~, particularly for ~~a~~the catchment ~~that was expected to have experienced a change in the rainfall-runoff relationship in the past.~~with stronger evidence of non-stationarity. The results highlight ~~the importance~~ that a traditional approach of ~~identifying a~~using long calibration ~~record representative~~period can degrade predictive performance when there is evidence of ~~the expected forecast conditions,~~non-stationarity. The techniques presented can form the basis for operational seasonal

streamflow forecasting systems and ~~if this step is ignored degradation of predictive performance can result.~~ provide support for environmental decision-making.

1 Introduction

Predictions of streamflow a month or a season ahead are essential information required by water resource managers for subsequent planning (Wang et al., 2011). This is particularly true in unregulated catchments with no capacity for storage and a highly variable flow regime that can be difficult to predict from historical data. ~~This is the case for the Drain M catchment in southern Australia, where streamflow can be diverted to support a range of environmental and social outcomes, and improved information on future water availability can assist in maximising the benefits realised from the water available.~~ A number of approaches have been developed to provide streamflow predictions with lead times from a month to a season ahead. These include "dynamic" hydrological modelling approaches (Demargne et al., 2014; Wood and Schaake, 2008), statistical approaches (Bennett et al., 2014; Robertson and Wang, 2013), or a combination of the two (Robertson et al., 2013).

In this work, a dynamic hydrological modelling based approach ~~has been~~is adopted to provide streamflow forecasts for an environmental management application~~, as this~~. The dynamic approach can often provide a better capture of catchment dynamics ~~that the simple indices used in~~than statistical models ~~cannot~~based on simple climatic indices (Robertson et al., 2013). In forecast mode, a ~~hydrologic~~hydrological model ~~that has been~~ calibrated ~~with~~using historical data is run forward in time, with input data ~~representing forecast~~provided by forecasted climate forcings~~, to predict the streamflow in an upcoming period~~. The following three major factors control forecasting performance (Luo et al., 2012): (1) the ability of the ~~hydrologic~~hydrological model to predict streamflow with actual forcings; (2) the accuracy of the assumed initial conditions ~~adopted~~ (e.g., soil moisture stores); and (3) the accuracy of the forecasts of the climate inputs. The focus of this paper is on the first two factors ~~related to the practical application of climate forecasts~~, in the context of a user need for seasonal streamflow forecasts to ~~assist with answering~~support environmental management ~~questions. The accuracy of the climate forecast is an important component of streamflow forecasts,~~ and ~~the effect of climate forecasts on streamflow forecasts is also considered in this work, but not approaches to improve climate forecasts.~~

~~Ability of the hydrologic model to predict streamflow~~decision-making.

Conceptual rainfall-runoff (CRR) models are widely used to simulate streamflow, due to their simplicity and accuracy ~~(Tuteja et al., 2011). As these models are conceptual, they require calibration to observed streamflow (if available). It is generally considered that longer calibration periods produce more robust parameter estimates (Brigode et al., 2013). However, the use of long calibration periods assumes time invariant catchment characteristics and processes, and that the parameter values derived from the calibration period are representative for the prediction period (Vaze et al., 2010).~~ (Li et al., 2015a; Tuteja et al., 2011). The parameters of these models have a limited relationship to measureable catchment attributes (e.g. soil horizon depth) (e.g. Fenicia et al., 2014), and typically require calibration to observed streamflow data (noting that physical models also require some calibration (Mount et al., 2016; Pappenberger and Beven, 2006)). The use of long calibration periods assumes

time-invariant catchment characteristics and processes, and that the parameter values derived from the calibration period are representative of the prediction period (Vaze et al., 2010). It is generally considered that longer calibration periods produce more robust parameter estimates, as a longer period exposes the model to a more diverse range of catchment conditions and flow events (Wu et al., 2013), however this is not always the case (for example Brigode et al., 2013).

~~The fitting of constant parameter values based on a given set of observed data can result in decreased model performance over time if the conditions that the model is exposed to in the forecast period are different from those used during model calibration (Coron et al., 2012). For the purposes of this work, catchments where physical changes are expected to have occurred, and there is evidence to reject the hypothesis of stationarity of the physical system, have been termed "non-stationary". Examples of non-stationary changes in catchment conditions that could~~The assumption that parameters are constant in time can result in decreased model performance if the conditions encountered in the forecast period are different from those in the calibration period (Bowden et al., 2012; Coron et al., 2012). In this work, the term "non-stationary" is used to refer to situations where physical changes are expected to have occurred in a catchment, and where there is evidence to reject the hypothesis of stationarity. In practice, catchments may have different "degrees" of non-stationarity, depending on the evidence available to reject the hypothesis of stationarity, the degree of change in a catchment, and the time scales over which the changes take place. Examples of catchment non-stationarity that can be expected to change the rainfall-runoff relationship include changes in land use or land-cover (e.g., deforestation, urbanization), land drainage, interception (e.g. dams or diversions), groundwater ~~depletion~~abstractions or responses to changes in climate (Milly et al., 2015). This definition of ~~non-stationarity, or~~ catchment non-stationarity~~, is used in this work, in contrast~~ can be contrasted to a broader definition of "hydrological model non-stationarity~~""~~", which ~~is used when~~ refers to temporal changes in hydrological model parameters ~~vary in time~~ for any reason (e.g. systematic data errors, poor calibration procedures, model structural deficiencies~~), and are therefore dependent on the period of record used for their estimation (for example, Westra et al., 2014).~~ , etc.) (for example Westra et al., 2014).

~~Degradation~~The degradation in model predictive performance due to catchment non-stationarity can ~~have a significant influence on the streamflow simulated by the model, and hence~~ impact on the decisions informed by ~~subsequent~~these forecasts. ~~In order to~~To address this concern, a number of studies have ~~proposed calibrating~~calibrated model parameters to subsets of the available data, by ~~finding analogous~~ attempting to find periods in the historical record that are analogous to conditions expected in the prediction time period, and by tailoring the time period selection to compensate for deficiencies in the model structure or input data (Brigode et al., 2013; de Vos et al., 2010; Luo et al., 2012; Vaze et al., 2010; Wu et al., 2013; Zhang et al., 2011). ~~The impact~~ Often there is a trade-off between the benefits of a longer calibration period ~~on~~ which exposes the model ~~results, including predictive uncertainty, is one aspect considered~~to a more diverse range of conditions and tends to improve parameter identifiability, versus the benefits of a shorter calibration period that exposes the model to the most recent – and hence often the most relevant – dynamics in the catchment. Demonstrating and understanding the impact of this trade-off on model predictive performance is a key research gap pursued in this study.

For the purposes of predicting streamflow, it is desirable for the hydrological model to not only be able to suitably predict streamflow with actual forcings, but also quantify the uncertainty associated with the prediction.Predictive uncertainty quantification is another major aspect of practical streamflow prediction. Many approaches are available to quantify thepredictive uncertainty in model predictions, from identifyingapproaches that identify a range of model parameters that represent the behaviour of the catchment using approaches such as generalised likelihood uncertainty estimation (GLUE) (Beven et al., 2008)(Beven and Binley, 1992), to post-processor approaches (e.g. Krzysztofowicz and Maranzano, 2004), to disaggregation approaches that attempt to characterise each individual source of error explicitly (e.g. Kavetski et al., 2003; Vrugt et al., 2005). In this work, predictive uncertainty has beenis estimated using a postprocessoran aggregated post-processor residual error model. ThisThe residual error model provides a statistical description ofrepresents the differences between the hydrological model predictions and observed data, without trying to identify the contributing sources (Evin et al., 2014). This approach is generally less data intensive and less complex than disaggregation approaches, which attempt to characterise each individual source of error. AThe post-processor approach to error modellingis chosen because it can also lead to more robust estimates of the modelpredictive uncertainty compared to jointly assessing joint calibration of all parameters (i.e. estimating CRR model and error model parameters concurrently) (Evin et al., 2014).

Adequate representations of errors is necessary to obtain reliable and precise hydrological predictions. The term "reliable" is used to define predictions that are statistically consistent with the distribution of the observations, and "precise" defined as predictions with a tight range of uncertainty. If poor assumptions are made about the properties of errors, unreliable or highly uncertain predictions can be obtained (Thyer et al., 2009). For example, errors in hydrological applications are generally heteroscedastic (i.e. the variance in the model errors is not constant, and increases with the magnitude of the flow). (McInerney et al., 2017) provided guidance on approaches most suitable for accounting for heteroscedasticity for different catchment types.

1.2 Accuracy of the initial conditions

Much of the skill in seasonal streamflow forecasts over periods following rainy seasons is commonly attributed to accurately representing initial catchment conditions (Koster et al., 2010; Pagano et al., 2004; Wang et al., 2009), whereas. In contrast, forecast skill over periods following dry seasons is generally attributed to both initial catchment conditions and meteorological inputs (Maurer and Lettenmaier, 2003; Wood and Lettenmaier, 2008). The impact of the 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)., although this time frame is generally catchment dependent.

In CRR models, catchment conditions are represented by (usually multiple) model storages, which are referred to as state variables. The values of these storages at the start of forecast period are typically determined using a "warmup" period, which allows the internal model states to reach reasonable values (e.g. Wang et al., 2011). Given the expected influence of the initial conditions on the simulated streamflow, model storage updating using observed data can be adopted to determine the initial state of the model storages (Wöhling et al., 2006). However, updating model storages of conceptual rainfall-runoff models is

~~not straightforward, as the model states are at best approximate representations of the real world as seen by the model (Berthet et al., 2009). A number of potential observed data sources could be used to update model storages. Li et al. (2015) suggested that although advanced remote sensing techniques provide an opportunity to improve hydrologic simulation, gauged discharge data assimilation is a more effective way to improve short term forecasts and is still preferred for operational streamflow forecasting purposes.~~

~~The impact of assimilating observed data into CRR model state variables at the start of a forecast on probabilistic seasonal forecasts 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.~~

In CRR models, catchment conditions are represented by (usually multiple) model storages, referred to as "state variables". The values of these storages at the start of a forecast period are typically determined using a warm-up period, which allows the internal model states to reach reasonable values. Given the expected influence of the initial conditions on the simulated streamflow, observed data can be assimilated into the model to update the state of the model storages. The most commonly used approaches in hydrological data assimilation include direct updating of storages (for example Demirel et al., 2013), Kalman filtering, particle filtering, and variational data assimilation (see Liu and Gupta, 2007). Berthet (2010) considered a number of tests for different updating approaches for the GRP model, a CRR model commonly used in short term streamflow forecasting applications.

Updating the states of conceptual rainfall-runoff models is not straightforward, as any environmental model is at best an approximate representations of the real catchment (Berthet et al., 2009). A number of observed data sources can be used to update model storages, including observed streamflow and in-situ or remotely sensed soil moisture. From these options, Li et al. (2015b) suggests that gauged discharge data assimilation is a more effective way to improve short-term forecasts and is still preferred for operational streamflow forecasting purposes.

Studies on observed data assimilation and CRR model state updating have focused primarily on flood forecasting with short lead-times. The benefits at longer lead-times (e.g. seasonal) to forecast water availability have received less attention in the published literature.

~~1.3~~1.1 Study Aims ~~and Objective~~

~~The objective of this~~This work ~~is to determine~~focuses on determining the degree to which ~~post-processing~~state updating and the selection of calibration ~~periods~~period length can enhance monthly streamflow ~~forecasting skill for~~ predictions in the ~~management of water availability for~~context of an environmental flow management application. ~~The specific~~More specifically, the aims ~~to achieve~~of this ~~objective~~study are:

~~0.~~1. Evaluate ~~a data assimilation approach for CRR model~~ the ability of state updating in a daily CRR model to ~~enhance~~improve predictive performance when forecasting ~~skill,~~streamflow volume for the upcoming month.

~~1.~~2. ~~Demonstrate that~~ Assess the degree to which using a shorter calibration period ~~can affect~~, that is more representative of the forecast ~~skill, and can be considered as a mechanism to account for~~period, can improve predictive performance, in particular when there is evidence of catchment non-stationarity ~~in a catchment~~.

- ~~Demonstrate a postprocessor error model for uncertainty analysis and quantify forecasting skill (reliability, precision, bias).~~

~~In the following section, a~~The paper is organized as follows. Section 2 outlines the user need for seasonal forecasts to manage a drainage network for environmental and social outcomes in ~~Southern~~southern Australia ~~is outlined, including,~~ describes the case study catchments ~~of interest~~ and data available. ~~This is followed by~~Section 3 describes the model setup and forecasting framework, as well as the methodology ~~adopted~~designed to ~~demonstrate~~achieve the aims ~~outlined~~ above. Sections 4 and 5 present and discuss the case study results, and Section 6 summarizes the key conclusions.

2 Environmental Flow Management Case Study ~~for Streamflow Forecasts~~

2.1 Catchment location and characteristics

The location considered in this study is a component of an extensive drainage network (exceeding 2500 km of open channels) in the ~~South East of South~~southern Australia. ~~A large proportion of the region's runoff is generated in the higher rainfall areas of the Lower South East.~~ (Figure 1). Historically, ~~this~~ runoff flowed in a northerly direction, along the watercourses adjacent to ranges, parallel to the coastline. Over the past 150 years, these flow paths ~~were~~have been diverted through a series of cross-country drains, constructed to provide flood relief and improve the agricultural productivity of the region by draining water in a south-westerly direction, creating outlets to the ocean. The largest of these cross-country drains is Drain M, (Figure 1) which conveys water from Bool Lagoon to ~~the ocean near Beachport (from station A2390541 to A2390512,~~ Figure 1~~), is the largest of the cross-country drains. This drain collects flow from several drainage systems located to the east and south of the drain, as outlined in Figure 1.~~

~~Recently, the ability to divert flow from this drain to the Upper South East and toward Ramsar listed wetlands (the Coorong) has been established. This diversion point is the Callendale regulator, located just upstream of gauge A21390514 which can divert flow to the north and restore a more natural flow path (as indicated by the drainage line in Figure 1). However, the termination of the Drain M system also contains~~ Lake George~~, a water body of high importance. Lake George has a high social value, including a fishery and public access, and can suffer from reduced water quality (hyper salinity) without sufficient volumes from the drainage network.~~

~~Decisions must be made throughout the year (mainly in the high flow season from late winter and throughout spring) regarding diversions from Drain M; if water should continue to flow to Lake George and the ocean, or if it can be diverted to the north, and be used to support inundation at numerous wetlands. Such decisions aim to meet the following competing objectives:~~

- ~~Ensure the water requirements of Lake George are met. Freshwater is necessary to maintain the estuarine ecology of the lake, to support its significance as a biological resource, and as a resource for recreational fishing.~~
- ~~Maintain some flow to the ocean to restrict sediment from entering the lake and to maintain connectivity to the sea to allow for fish movement and aide fish recruitment.~~
- ~~Minimise possible impacts on sea grasses, caused by fresh water flows, often with high nutrients, out to sea.~~
- ~~Maximise water diverted to the north, to benefit the wetlands in the Upper South East, subject to the downstream objectives at Lake George.~~

~~Runoff~~. Monthly runoff volumes from Drain M are highly variable, ranging from close to zero to more than is required to support Lake George, with the historical volumes varying over 3-4 orders of magnitude for a given month~~.~~ (Figure 2). This variability makes it difficult to maximise the use of water, as the seasonal pattern described by the historical record alone provides little guidance.~~ Bool and Hacks Lagoons (between stations A2390519 and A2390541 in Figure 1), act as water storages in this region as well as being significant Ramsar listed wetlands. With reduced inflows and the need to meet the water requirements of Bool and Hacks Lagoons, only minor releases from the lagoons at station A2390541 have occurred in the past 15 years. With the many competing demands on the water resources available in this system, it is desirable to forecast future flows at key locations along Drain M to maximise the environmental and social outcomes achieved from the water available.~~

The streamflow in the case study region is seasonal to ephemeral, with very low flow over the summer and autumn months (Figure 2). Runoff coefficients are low, with annual runoff in the range of 0.01-0.1 of annual rainfall (Gibbs et al., 2012). The predominant land use in the region is dry land pasture with some flood irrigation as well as plantation forestry; there is no major urbanization in the catchments. The topography of the region is very flat, with mainstream slopes in the order of 0.005. The hydrogeology of the catchment includes shallow aquifers with major karstification of limestone, which may be suggestive of non-conservative catchments with appreciable groundwater exchanges across their boundaries.

Mosquito Creek flows into Bool Lagoon (Catchment C1 in Figure 1, area 1002 km^2). Drain M commences at the outlet of Bool Lagoon, and a large catchment flows into Drain M between Bool Lagoon and a diversion point at Callendale (Catchment C3 with an area of 2200 km^2). Finally, the Drain M local catchment contributes flow downstream of the Callendale diversion point, flowing into Lake George (Catchment C2, area 383 km^2).

In the region where the case study catchments are located, plantation forestry expanded substantially in the late 1990s. ~~To assess potential benefits and impacts of different diversion scenarios on the wetlands in the Drain M system, a water balance model has been developed in eWater Source (Welsh et al., 2013). The model also simulates salinity in Lake George based on a constant inflow of salinity from Drain M, and modelling the subsequent flushing and evapoconcentration of the lake. This model provides the functionality to take forecast streamflow volumes in the drainage network and allow different management scenarios to be considered for the main operational locations in the system: Bool Lagoon and Callendale regulator. Given a~~

~~probabilistic streamflow forecast, the impact of different diversions and releases on the forecast range of water level in Bool Lagoon, water level and salinity in Lake George and volume diverted along the drain can be estimated. This capability is intended to provide greater information to assist the decision-making process, extending the probability of a certain volume of water occurring in the network to enable the influence of operational decisions on the key variables (volume diverted, water levels and Lake George salinity) in the system to be considered. The model has been calibrated to provide a suitable representation of historical water levels, flows and salinities.~~

~~To use this model for to inform operations, forecast climate and streamflow are required model inputs. The focus of this work is to derive suitable streamflow~~Changes in the relationship between rainfall and runoff also occurred during this period, evidenced by the reduced slope in the plot of cumulative runoff against cumulative rainfall (double-mass analysis) in Figure 3 (Searcy et al., 1960; Yihdego and Webb, 2013). The runoff ratio in catchment C1 is approximately 0.045 before year 2000, but reduces by 70% to 0.013 after 2000. The runoff ratio in catchment C2 is around 0.088 before year 2000, but reduces by 30% to 0.061 after 2000. This comparison provides stronger evidence of non-stationarity in catchment C1 than in catchment C2. Other studies have also investigated the link between changes in the hydrology and changes in land use in the region (Avey and Harvey, 2014; Brookes et al., 2017). These changes have implications on the choice of calibration data period, as data from the 1970s may not be representative of hydrological conditions in the 2000s.

It is evident from Figure 3 that catchment C3, despite having the largest catchment area (2200 km^2) of the three catchments, generates very little runoff. This behaviour is due to a number of factors, including the very flat terrain and depression storage, substantial vegetation cover (both plantation and natural) and irrigation extractions from the shallow underlying aquifer. Given its limited streamflow volume, catchment C3 is excluded from further analysis in this study. From a practical perspective, it is assumed that in the years where there is substantial yield from this catchment there will already be surplus flow from the upstream catchments.

2.2 Management Issues

Drain M serves multiple competing demands on the water resources available in this catchment system. These demands influence the decision to use the regulators along the system:

a) Bool Lagoon has water requirements that influence releases from the lagoon into Drain M.
b) Lake George has water requirements to maintain the estuarine ecology of the lake, to support its significance as a biological resource, and as a resource for recreational fishing.
c) The ocean outlet requires some flow to prevent sediment from entering Lake George and to maintain connectivity to the sea (which allows fish movement and aids fish recruitment). However, high flows may impact on sea grasses, due to their low salinity and high nutrient load.

d) The wetlands of the Upper South East to the north typically benefit from as much water as possible from the Drain M system.

Decisions to undertake diversions from Drain M must be made throughout the year (mainly in the high flow season from late winter and throughout spring). It is expected that forecasts ~~for this application. A forecast period~~of future flows at key locations will assist in maximising the environmental and social outcomes achieved from the available water. Forecasts of monthly volume with a ~~leadtime~~lead time of one month ~~has been~~ahead are considered ~~as~~ most appropriate for this application, ~~as~~because: 1) the main quantities of interest in this application are volume and the overall water balance, rather than the size or timing of daily peak flows, and 2) one month lead time provides sufficient time to undertake any changes in ~~management, and 3) reasonable forecast skill is expected to be possible compared to longer forecast horizons.~~diversions to satisfy the competing demands on the system.

~~2.1~~2.3 ~~Climate and~~ Climate Data

The mean annual rainfall for the region ~~is in the range~~varies from 600 ~~-~~mm in the north to 675 mm ~~and~~in the south. The mean annual FAO56 potential evapotranspiration (PET) (Allen et al., 1998) is approximately 1000 mm. The highest rainfalls are experienced in the winter months, with rainfall exceeding evapotranspiration in ~~the months~~ May ~~-~~September. The SILO Patched Point Dataset (Jeffrey et al., 2001) was used for the observed rainfall and the FAO56 evapotranspiration data was adopted, with the climate stations used shown in Figure 1. ~~A Thiessen polygon approach was used to combine stations and produce one time~~Time series ~~each~~ of rainfall and evapotranspiration ~~for~~in each catchment~~.~~ were obtained using a Thiessen polygon approach. This weighting approach ~~has been~~is considered appropriate for the region, due to the flat terrain being unlikely to lead to significant topographic effects on the spatial distribution of rainfall.

Rainfall forecasts from the Australian Bureau of Meteorology's ~~(BoM) seasonal forecast system, POAMA-2, were used to represent the effect of climate during the forecast period. POAMA~~seasonal forecast system, POAMA-2 (Hudson et al., 2011), were used. POAMA-2 is a dynamical climate forecasting system designed to produce multi-week to seasonal forecasts of climate for Australia based on a coupled ocean/atmosphere model and ocean/atmosphere/land observation assimilation systems. ~~POAMA-2 consists of both a seasonal and a~~ In this paper, we use a 30-member ensemble of monthly/multi-week ~~forecast system using~~ forecasts from version 2.4 of the ~~model. Both systems are multi-model ensembles consisting of three different model configurations. Furthermore, each configuration is used to generate an ensemble of 10 forecasts by perturbing initial conditions, resulting in a 30-member ensemble.~~ POAMA-2. POAMA-2 ~~produces large (~250 km) spatial scale~~ predictions ~~over Australia,~~have a coarse spatial resolution (~250 km), which ~~do~~does not capture the ~~large degree of~~ spatial variability in catchment-scale rainfall. ~~For~~ For the purposes of this application ~~in this study~~, the ~~large scale~~ POAMA-2 rainfall hindcasts (i.e. forecasts developed ~~using~~by applying the modelling system ~~for~~to the historical period) at the relevant pixel were downscaled to ~~the single rainfall gauge scale. The~~each climate station in the study region (Figure 1) using the statistical downscaling method detailed in Shao and Li (2013) ~~was used, resulting in rainfall hindcasts at each of the rain gauges shown in Figure 1. As an example Figure 2 shows that the mean rainfall hindcasts at station 26000 provide a reasonable estimate of~~

~~the observed rainfalls at this station, while the upper and lower hindcast bounds capture the majority of observed rainfalls.~~ Further details of the downscaling approach are provided in Humphrey et al. (2016).

~~2.2~~2.4 Streamflow ~~and Streamflow~~ Data

~~The streamflow in the region is seasonal to ephemeral, with flow ceasing over the summer and autumn months and runoff coefficients are low, with annual runoff in the range of 2–10% of annual rainfall (Gibbs et al., 2012). The predominant land use in the region is dry land pasture with some flood irrigation as well as plantation forestry; there is no major urbanization in the catchments. The topology of the region is very flat, with mainstream slopes in the order of 0.005. The hydrogeology of the catchment includes shallow aquifers with major karstification of limestone, which may be suggestive of groundwater exchange across the boundaries of the catchments.~~

~~The details of the three catchments contributing to Drain M are as follows. Mosquito creek flows into Bool Lagoon (catchment area of 1002 km^2), with the inflows gauged at station A2390519 (referred to as C1, as shown in Figure 1). The Southern Bakers Range catchment flows into Drain M between Bool Lagoon and the diversion point at Callendale (2200 km^2 – C3). The flow from this catchment has been determined from that gauged at Callendale (A2390514) minus that released from Bool Lagoon (A2390541). Travel times along this section of drain are less than one day, and for the purposes of this study the daily flows have been subtracted to determine the flow from the Southern Bakers Range catchment. It should also be noted that releases from Bool Lagoon have occurred infrequently in the recent past, with only four short release events since 1997 (Gibbs et al., 2014). Finally, the Drain M local catchment contributes flow downstream of the Callendale diversion point, flowing into Lake George (383 km^2). This is gauged at station A2390512 (C2). As with the Southern Bakers Range catchment, the flow contributing to the drain from this local catchment has been derived as the difference between the flow at stations A2390514 and A2390512, as travel times are again less than one day.~~

Daily streamflow data are available from the South Australian Department of Environment, Water and Natural Resources Surface Water Archive (https://www.waterconnect.sa.gov.au/Systems/swd), with the flow stations used shown in Figure 1. Three of the flow stations have data available from the early 1970s, with the exception being the station at the outlet of Bool Lagoon (site A2390541), where data were available from 1985. Travel times along Drain M between flow stations are typically less than one day. To determine the flow generated within catchment C2, the daily flows recorded at upstream flow station A2390514 where subtracted from the downstream flow station A2390512.

~~Streamflow measurements in the three catchments were obtained from fixed concrete weirs. The stations at A2390512 and A2390519 have v-notch cross-sections to improve sensitivity at low flows, whereas stations A2390514 and A2390541 have flat cross sections and therefore could be expected to be less sensitive at low flows. All stations are free flowing downstream and are constructed drains upstream with a constant slope and cross section, which would be expected to provide a reliable relationship between stage and discharge. Gauging records for the four stations are extensive, with between 78 and 166 flow~~

~~gaugings at each station, with a 90th percentile deviation of 10.5% for stations A2390512 and A2390519, and 16.3% at station A2390514. Given the regular cross sections and high number of gaugings available to develop stage-discharge relationships, output data uncertainty is expected to be low for the stations considered. This identification of high quality data was considered an important process, as biases and systematic changes in the measurement of hydrological data can significantly affect model calibration and can lead to non-stationarity in the estimated model parameters (Westra et al., 2014).~~

The identification of high quality data is important because biases and systematic changes in the measurement of hydrological data can significantly affect model calibration and lead to non-stationarity in the estimated model parameters (Westra et al., 2014). Analysis of the data and monitoring stations suggested that streamflow data uncertainty is expected to be low, given the regular cross sections of the weirs used for monitoring stage and upstream drains, and the high number of gaugings (between 78 and 166 flow gaugings at each flow station) available to develop stage-discharge relationships.

~~In the region where the case study catchments are located, plantation forestry expanded substantially in the late 1990s. Changes in the relationship between rainfall and runoff also occurred during this period, as seen as the reduced slope in the double-mass analysis (Searcy et al., 1960; Yihdego and Webb, 2013), i.e. the plot of cumulative runoff against cumulative rainfall in Figure 3. These changes have implications for the period most relevant for forecasting future flows in the catchments, as data from the 1970s may not be representative of what could be expected in future years. It is also evident from Figure 3 that, despite the contributing area, very little runoff is generated from C3 due to a number of factors, including the very flat terrain and substantial depression storage, substantial vegetation cover (both plantation and natural) and irrigation use from the shallow underlying aquifer. Given the limited volumes generated within this catchment, it has been excluded from further analysis. From a practical perspective, in years where there is substantial yield from this catchment, it is assumed that there will already be surplus flow from the upstream catchments. Based on the case study outlined, the two main challenges considered in this paper are: (i) estimation of predictive uncertainty to inform management, and (ii) determination of approaches to adequately represent the non stationarity in the catchments.~~

3 Methodology

3.1 CRR Model ~~and Parameter Calibration~~

The GR4J model (Perrin et al., 2003) ~~was~~is a parsimonious daily CRR model, selected ~~as the CRR model~~ for this study~~, as it is a parsimonious model that~~ because it explicitly accounts for non-conservative (or 'leaky') catchments~~,~~ (relevant for the study area, see Section 2.1) and has demonstrated good performance for Australian conditions (Coron et al., 2012; Guo et al., 2017; Westra et al., 2014)~~. The ability to represent non-conservative catchments is important in the study area (see section 2.2).~~ The standard form of the GR4J model has four calibration parameters~~,~~: the maximum capacity of a production (soil) store, X1, a catchment water exchange coefficient, X2, the maximum capacity of a routing store, X3, and a time base for a unit hydrograph, X4. Further details of the model structure and parameters can be found in Perrin et al. (2003). ~~In addition to the four standard~~

parameters, the split between the routing store and the direct runoff has also been considered a calibration parameter in this study. The resulting additional parameter is termed *split*; the standard GR4J formulation adopts a default value of *split*=0.9.

Note that the catchments considered have a relatively slow streamflow response. Consequently, the pre-specified split to the routing store of 0.9 in the original specification of the GR4J model may be too low for these catchments. To mitigate this potential deficiency, we have modified the GR4J model so that the split between the routing store and the direct runoff is included as an explicit calibration parameter termed *split*.

3.2 Parameter estimation

The GR4J parameters wereare inferred using Bayes equation. The posterior probability density of the parameters given daily observed streamflow data $\tilde{\boldsymbol{q}}$ and climate data $\mathbf{X}$, $p(\boldsymbol{\theta}|\tilde{\boldsymbol{q}}, \boldsymbol{X})$, is given by:

$$p(\boldsymbol{\theta}|\tilde{\boldsymbol{q}}, \boldsymbol{X}) \propto p(\tilde{\boldsymbol{q}}|\boldsymbol{\theta}, \boldsymbol{X})\, p(\boldsymbol{\theta}) \quad (1)$$

where $p(\boldsymbol{\theta})$ is the prior distribution and $p(\tilde{\boldsymbol{q}}|\boldsymbol{\theta}, \boldsymbol{X})$ is the likelihood function.

A standard least squares likelihood function wasis adopted (see, for example, Thyer et al., 2009). The function was calculated using the daily time series, and, which is derived from ana residual error model assumingthat assumes independent, homoscedastic residuals. While the assumptions required by this function (i.e. constant variance, independent) are somewhat unrealistic for This likelihood function is adopted for the calibration of the daily hydrological applications, statistics based on sum of squared errors (Nash Sutcliffe Efficiency, RMSE) are often used for model calibration and assessment, and hence this function has been selected. This function has also been adopted given that the focus of this study is on the estimation of monthly runoff volumes, as this functionbecause it provides a focus onbetter fit to the highesthigh daily flows in the time series, where the majority of the runoff volume occurs (Wright et al., 2015).

, which make a big contribution to monthly volumes of interest in our study. Uniform prior distributions were adoptedare used for all parameters, with the assumed bounds given in Table 1. A number of trial runs were used to ensure the bounds of the prior distribution were suitable and well-removed from the high-density areas of the posterior distribution.

.

The posterior distribution in equationEq. (1) wasis sampled using the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm (Vrugt et al., 2009). DREAM provides a sample of The sampled parameter sets that representsare then used to approximate the posterior parameter distribution, and the trends in these distributions for different time periods are considered in the results. For the purposes of residual error modelling, only the parameter set resulting in the maximum posterior was used. The a given calibration period. Computations were carried out using the Hydromad R package implementation of the DREAM algorithm and the GR4J model (Andrews et al., 2011) implementation. A total of 25,000 iterations of the DREAM algorithm and GR4J model have been adopted. A maximum of 25,000 evaluations were usedcarried out, including a "burn-

in" period ~~where the initial samples were discarded before~~of 6250 iterations to allow the Markov Chain ~~was assumed~~ to ~~have stabilised~~stabilise. The number of parallel chains was set equal to the number of parameters~~, i.e. five parallel chains were adopted~~ (Vrugt et al., 2009)~~.~~, which, for the modified GR4J model used in this work (Section 3.1), led to five parallel chains being used.

The posterior distributions obtained for different calibration time periods are investigated for evidence of trends and changes over time. For the purposes of developing streamflow predictions using the post-processing approach (Section 3.5), only the single parameter set resulting in the maximum posterior probability is used.

3.3 Calibration ~~Periods to account for Non-stationarity~~Approach

A rolling calibration approach ~~has been~~is used to ~~minimize~~account for the impact of ~~external influences~~non-stationarity on the inferred CRR model parameters~~.~~. This rolling calibration approach is similar to ~~that of~~the approach used by Luo et al. (2012) and Wagener et al. (2003). ~~External influences include model structural limitations, physical changes in the catchments influencing the rainfall-runoff relationships (land use change, change in groundwater level), or over-representation~~It consists of ~~particular hydroclimatic conditions. In the rolling~~choosing a calibration ~~approach,~~length and then moving it forward year by year, while recalibrating the model parameters ~~were recalibrated~~ to each ~~year based on the data from the preceding period. The~~ such calibration "window". The calibrated parameter values ~~identified were~~are used to simulate the following one year of data ~~(following that used in the calibration process),~~ before recalibrating the model ~~to include these "new" data, while discarding the earliest year of data, to maintain the same period length for calibration. The recalibration~~and repeating the process. This methodology ~~adopted~~ allows ~~for~~ the ~~change~~identification of changes in parameter distributions over time ~~to be included.~~, without the need to identify specific periods when changes in the rainfall-runoff response may have occurred.

Calibration period lengths of CPL = 10 ~~year~~ years and CPL = 20 ~~year calibration periods have been~~ years length are considered, to assess the trade-off between using a longer calibration period to ~~reduce the~~expose the model to more diverse catchment conditions and improve parameter ~~uncertainty, and~~ identifiability, versus using a shorter calibration period ~~representing~~length to expose the ~~most~~model to more recent hydrological dynamics ~~observed in the catchment.~~.

As an example, consider a 10 year calibration period ~~was~~ from 1/5/1995-30/4/2005, after a one year ~~warmup period. The~~ warm-up period. Predictions are computed for the following one year "prediction period ~~was then considered to be the monthly forecasts over the following year (~~", i.e. 1/5/2005-30/4/2006~~).~~. The process ~~was~~is then repeated each year, ~~where in this example~~i.e., the next calibration period ~~becomes~~is 1/5/1996-30/4/2006, and the ~~prediction~~ calibrated model is used to predict the period 1/5/2006-30/4/2007. The ~~start~~starting month of May corresponds to the start of the flow season ~~in autumn. Through this rolling approach, the model predictions assessed were independent of the data used for calibration in all the results presented. For the case of using forecast rainfall, the model was run with observed rainfall up to the forecast month, before using forecast rainfall only for the forecast month. This was repeated for each month in the prediction period.~~(Figure 2).

3.4 State Updating in GR4J

~~An~~The approach ~~to directly correct~~used for the ~~states~~state updating of ~~a rainfall runoff model with observed data has been adopted in this work,~~GR4J is similar to ~~that~~the approach of Demirel et al. (2013). ~~The model was run for a one year warmup period, and the simulated level in the production store at the end of this period was maintained. As model states generally do not reflect reality directly (Berthet et al., 2009), the production store level has been maintained "as seen" by the model.~~

~~In order to force the model to simulate the flow observed at the start of the month, the necessary routing store level to produce the observed flow after accounting for the modelled direct flow was calculated~~State updating is set to take place at the start of each month within the one year prediction period, using the observed streamflow at the start of each month. GR4J has two stores, namely the production store and the routing store. Following the procedure of Demirel et al. (2013), the routing store level is updated such that the GR4J simulation of streamflow matches the observed flow. This procedure is undertaken after accounting for the modelled direct flow from the production store (Demirel et al., 2013)~~. In GR4J, total simulated streamflow on a given day q_t^{θ} is calculated as the sum of the flow direct from the production store (after applying a unit hydrograph), $q_{t,d}^{\theta}$, and the flow from the routing store, $q_{t,r}^{\theta}$, i.e.:~~

More specifically, the following procedure is used. In GR4J, the total simulated streamflow on a given day q_t^{θ} is defined by the sum of the direct flow from the production store (after applying a unit hydrograph), $q_{t,d}^{\theta}$, and the flow from the routing store, $q_{t,r}^{\theta}$.

$$q_t^{\theta} = q_{t,d}^{\theta} + q_{t,r}^{\theta} \tag{2}$$

~~The necessary~~ Let $q_{t,r}^{SU}$ denote the flow from the routing store ~~for~~that yields q_t^{θ} ~~to~~ equal to the observed flow~~,~~ $\tilde{q}_t$~~, defined as $q_{t,r}^{SU}$~~. This quantity is ~~then~~ calculated as~~:~~

$$q_{t,r}^{SU} ~~= \min(\tilde{q}_t - q_{t,d}^{\theta}~~ = \max(\tilde{q}_t - q_{t,d}^{\theta}, 0) \tag{3}$$

The routing store level, R, can then be obtained by setting $q_{t,r}^{\theta} = q_{t,r}^{SU}$ ~~and inverting~~, and solving (using the bisection method) the equation used by the GR4J model to calculate the outflow from this storage:

$$q_{t,r}^{\theta} = ~~R\left(1 - \left(1 + \left(\frac{R}{X3}\right)^4\right)^{-1/4}\right)~~ R\left(1 - \left(1 + \left(\frac{R}{X3}\right)^4\right)^{-1/4}\right) \tag{4}$$

Equs. (2) – (4) can be ~~where~~ X3 ~~is an estimated runoff model parameter.~~

~~More complex approaches for data assimilation and state updating are available, for example, the ensemble Kalman filter and particle filter approaches are commonly used (He et al., 2012; Lei et al., 2014; Li et al., 2013; Plaza Guingla et al., 2013; Spaaks and Bouten, 2013; Xie and Zhang, 2013). However, particularly when used to update both model state variables and model parameters, these approaches impose a substantial computational burden for obtaining accurate results, and accuracy~~

~~can decrease if incorrect correlations between states and parameters are identified, due to biased model error quantification and a large degree of freedom for high-dimensional vectors of the augmented state (Xie and Zhang, 2013).~~

used to update R given the observed streamflow flow $\tilde{q}_t$.

3.5 Estimation of Predictive Uncertainty

~~A post-processor error model was used to estimate predictive uncertainty in forecasts. After the hydrological model has been calibrated, a statistical model was fitted to the residuals given by the difference between monthly observations ($\tilde{\boldsymbol{Q}}$) and predictions from the hydrological model ($\boldsymbol{Q}^{\boldsymbol{\theta}}$) using the maximum posterior parameter estimates. In this case, the streamflow at the monthly time scale was considered, i.e., the sum of the streamflow at the daily time scale each month, $\widetilde{Q_t} = \sum_{t=1}^{n} \tilde{q}_t$, for the n days in month t. Depending on the case, the hydrological model predictions were either those forced by observed rainfall, or for the case when the model was forced by the ensemble of forecast rainfall, 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. Henceforth, the resulting hydrological model prediction monthly time series is referred to as $\boldsymbol{Q}^{\boldsymbol{\theta}}$.~~

~~Heteroscedasticity~~The monthly streamflow forecasts are obtained by aggregating the daily GR4J simulations. In order to quantify predictive uncertainty using a residual error model, the monthly-aggregated GR4J simulations, $\boldsymbol{Q}^{\theta}$, are compared to observed monthly streamflow volumes, $\tilde{\boldsymbol{Q}}$. The quantification of error is based on residuals errors, defined by the differences between observed and simulated monthly streamflow. Separate error models are estimated for the GR4J predictions for each catchment and for each type of forcing data (observed or forecast rainfall), as follows:

- When observed rainfall is used as input to GR4J, the daily streamflow time series simulated using G4J are aggregated to produce monthly time series of hydrological model predictions, $\boldsymbol{Q}^{\theta}$.
- When forecast rainfall is used as input to GR4J, an ensemble of daily streamflow forecasts is produced (with a single GR4J streamflow time series per rainfall forecast time series). Each such "individual" daily GR4J time series is then aggregated to a monthly time step. The time series $\boldsymbol{Q}^{\theta}$ is constructed from the time series of medians of the individual monthly streamflow time series. This use of the median streamflow forecasts from the multiple ensembles of the meteorological ensemble forecasts may result in some information loss, but aggregation approach of streamflow forecast ensembles is commonly used in operational applications (e.g. Lerat et al., 2015; Matte et al., 2017; Schepen et al., 2017; Wani et al., 2017). Further work is needed to more fully utilise the information from ensemble forecasts when developing post processing models.

The heteroscedasticity (i.e. larger residuals for larger flows) and skewness ~~in these residuals are captured by~~ of forecast errors is accounted for using the Box Cox transformation~~:~~, by defining normalized residuals as

$$\eta_t = \mathrm{Z}(\widetilde{Q_t}) - \mathrm{Z}(Q_t^{\theta}) \quad (6)$$

$$Z_{\mathrm{BC}}(Q;\lambda,A) = \begin{cases} \dfrac{(Q+A)^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(Q+A) & \text{otherwise} \end{cases} \quad (5)$$

where λ is where a transformation parameter, and A is an offset parameter (often important when transforming low flows). Based on the findings of McInerney et al. (2017), $\lambda = 0.5$ was selected, which was shown to produce good predictive performance in ephemeral catchments, especially in terms of improving precision and reducing bias. A value of $A = 1\times10^{-5}$ mm/month was selected.

The Box Cox transformation is applied to observed and predicted flows, and

$$Z(Q;\lambda,A) = \begin{cases} \dfrac{(Q+A)^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(Q+A) & \text{otherwise} \end{cases} \quad (6)$$

with a transformation parameter λ and an offset parameter A (often important when transforming low flows).

$\lambda = 0.5$ was used, as this setting was shown to produce good predictive performance (especially in terms of sharpness and bias) in ephemeral catchments by McInerney et al. (2017). The offset is set as $A = 1\times10^{-5}$ mm/month.

The normalized residuals are calculated as:

η_t in Eq. (7) ~~$\eta_t = \mathrm{Z}(\widetilde{Q_t}) - \mathrm{Z}(Q_t^{\theta})$ (6)~~

These normalized residuals η_t are assumed to be Gaussian with mean μ_η μ_η and variance σ_η^2 σ_η^2, i.e. $\eta_t \sim N(\mu_\eta, \sigma_\eta^2)$.

..

$$\eta_t \sim N(\mu_\eta, \sigma_\eta^2) \quad (7)$$

The parameters μ_η μ_η and σ_η σ_η are then estimated using the method of moments, i.e. as the sample mean and sample standard deviation of η computed with the maximum posterior parameter set the time series of η. The same rolling calibration approach outlined in Section 3.3 for the GR4J model is also applied for the calibration of the post-processor error models.

Once the residual error model ~~has been~~is calibrated, replicates from the predictive distribution, ~~$\mathbf{Q}^{(r)}$~~ $\mathbf{Q}^{(r)}$ for ~~$r = 1..N_r$,~~ ~~are~~ $r = 1..N_r$, can be generated for ~~the independent evaluation~~any time period of interest, as follows:

1. Sample the normalized residual at time step *t*, ~~$\eta_t^{(r)} \leftarrow N(\mu_\eta, \sigma_\eta^2)$~~ $\eta_t^{(r)} \leftarrow N(\mu_\eta, \sigma_\eta^2)$ (~~7~~8)

2. Rearrange ~~Equation~~ Eq. (6) to yield:

$$Q_t^{(r)} = Z^{-1}\left(Z(Q_t^\theta) + \eta_t^{(r)}\right) \quad (\text{~~8~~9})$$

3. Truncate negative values to zero.

~~The~~ Equs. (5) – (8) are used to generate replicates ~~represent~~from the predictive distribution (PD) of the forecasts.

The assumptions of the post-processor residual error model used to estimate predictive uncertainty for monthly volumes are different to the assumptions of the residual error model used in the likelihood function for calibrating the daily GR4J model. As outlined in Section 3.2, the GR4J model is calibrated at the daily scale to observed streamflow using the standard least squares likelihood function, because it better captures the high daily flows, important for estimating the monthly volumes. The post-processing error model for the monthly volumes is designed to capture the predictive uncertainty in these monthly volumes, in particular the heteroscedasticity and skew of the residuals (McInerney et al., 2017; Refsgaard, 1997). These choices of residual error models at the daily and monthly time scales contribute to the study objectives of reliable forecasts at the monthly time scale, and are common in forecasting applications (for example, Lerat et al., 2015).

3.6 Model Configurations and Implementation

Two options for state updating (with versus without) and two options for calibration period length (CPL = 10 years versus CPL = 20 years) are considered. The combination of these options leads to four model configurations. Four different cases are considered for each model configuration, given by the combinations of two catchments (C1 and C2) and two sources of climate data (observed and forecast). This results in a total of 16 scenarios considered.

Twelve sets of one month ahead predictions are generated during the one year prediction period. For all scenarios observed rainfall is used as input to the hydrological model prior to the start of each set of one month ahead predictions. When state updating is used, the GR4J state is updated at the start of this month using the procedure outlined in Section 3.4. During the one month ahead predictions, either observed or forecast rainfall are used, depending on the scenario considered.

~~3.6~~3.7 Performance Metrics

Five metrics ~~were~~are used to evaluate ~~different~~distinct aspects of predictive performance. ~~-~~These include metrics for reliability, ~~precision~~sharpness, volumetric bias, the cumulative ranked probability score (CRPS) and the Nash Sutcliffe Efficiency (NSE). ~~Depending on the data available, the metrics are calculated over different periods. When observed climate data was used to drive the CRR models, the periods commences 21 years after the start of the streamflow record (after a 1 year warmup and 20 year calibration period), that is 1/5/1992 for C1 and 1/5/1994 for C2. In all cases, the period ends on 30/4/2010. The same period is used within each case considered, e.g. different calibration lengths, with/without state updating, and as such are comparable within these cases for a given catchment. For display purposes enabling a straight forward comparison across all the cases considered, the value for each metric has been normalised by linearly scaling the worst value to a value of 0.05 and the best value to 0.95.~~

Reliability refers to ~~whether~~ the degree to which the observations (of streamflow) over a series of time steps can be considered to be statistically consistent with ~~being samples from~~ the predictive distribution. ~~Reliability has been~~In this work, reliability is assessed using predictive quantile-quantile (PQQ) plots, and ~~the plot has been summarized~~quantified using the reliability metric of Renard et al. (2010) ~~which is~~ based on the area between the PQQ plot and the 1:1 line. A value of 0 represents perfect reliability, while a value of 1 represents the worst ~~case of~~reliability, i.e., all observations lying ~~either~~ outside (above or below) the PD.

~~**Precision** refers to the width of the predictive distribution, and is otherwise known as resolution or sharpness. We quantify precision using the following metric from McInerney, et al (2017):~~

$$\text{Precision} = \frac{1}{N_t}\sum_{t=1}^{N_t} \text{sdev}\,\mathbf{Q}_t \Big/ \frac{1}{N_t}\sum_{t=1}^{N_t} \tilde{Q}_t \quad (1)$$

(**Sharpness** refers to the width of the predictive distribution, and is otherwise known as "resolution" or "precision". We quantify sharpness using the following metric from McInerney et al. (2017):

$$\text{Sharpness} = \sum_{t=1}^{N} \text{sdev}(\boldsymbol{Q}_t) \Big/ \sum_{t=1}^{N} \widetilde{Q_t} \quad (9\text{10})$$

where N is the number of months and sdev() is the sample standard deviation.

Volumetric bias measures the overall water balance error of the predictions relative to the observations. It is calculated as:

$$\text{VolBias} = \left| \frac{\sum_{t=1}^{N} \text{mean}(\boldsymbol{Q}_t) - \sum_{t=1}^{N} \widetilde{Q_t}}{\sum_{t=1}^{N} \widetilde{Q_t}} \right| \left| \frac{\sum_{t=1}^{N} \text{mean}(\boldsymbol{Q}_t) - \sum_{t=1}^{N} \widetilde{Q_t}}{\sum_{t=1}^{N} \widetilde{Q_t}} \right| \quad (10\text{11})$$

where mean() is the sample mean.

CRPS is a widely used probabilistic performance metric that ~~summarises and evaluates~~combines in a single measure multiple aspects of predictive performance ~~(~~, including reliability, ~~uncertainty and~~ sharpness~~) in a single measure~~ and bias (Hersbach, 2000). The CRPS is calculated by comparing the cumulative distribution of the ~~forecast~~predictions with the cumulative distribution of the observation at each time step. ~~For each~~At a single time step, the CRPS is ~~calculated~~defined as:

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} \left[F_f(Q_t) - F_o(Q_t)\right]^2 dQ \int_{-\infty}^{\infty} \left[F_{p,t}(Q) - F_{o,t}(Q)\right]^2 dQ \quad (11\text{-}12)$$

where ~~F_f~~$F_{p,t}$ and $F_{o,t}$ are the cumulative distributions of the streamflow ~~forecast~~predictions ($\boldsymbol{Q}_t$) and observation~~, respectively~~ ($\widetilde{Q_t}$), at time step t. The average value of the CRPS is then calculated over all time steps t. Note that the cumulative distribution of the observations is a step function. A CRPS of 0 corresponds to ~~a~~the perfect ~~forecast~~prediction, while larger CRPS values correspond to worse performance.

To normalize CRPS metric values across catchments, the CRPS metric for the ~~forecast (CRPS$_f$~~predictions (CRPS$_P$) is expressed as a skill score with respect to the CRPS metric of a "reference" distribution for that catchment (CRPS$_R$)

$$\mathrm{CRPS_{SS}} = \frac{\mathrm{CRPS_R - CRPS_P}}{\mathrm{CRPS_R}} \quad (13)$$

~~A~~ CRPS$_{SS}$ ~~of~~ values below zero indicate forecasts with worse performance than the reference distribution, a CRPS$_{SS}$ of 0 corresponds to the ~~forecast~~predictions having ~~similar~~the same performance as the reference distribution, ~~while~~and a CRPS$_{SS}$ of 1 corresponds to a perfect ~~forecast.~~ prediction.

The reference distribution for each month is calculated as the empirical distribution of all observed data in that month~~. Note that since all of the observed data is used to estimate the reference distribution (including data in the forecast verification period), this distribution has an unfair advantage over the models which use only previously observed data. This means that there may be cases where the CRPS$_{SS}$ is less than zero, corresponding to the reference distribution having better performance than the forecast.~~, using the entire set of observed data (including data from the prediction period). This approach provides a stringent baseline for the CRPS normalization in Eq. (13).

NSE is a commonly used metric for the assessment of the accuracy hydrological ~~models~~model predictions, and is calculated as:

~~$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T}\left(Q_t^{\theta} - \widetilde{Q_t}\right)^2}{\left(\widetilde{Q_t} - \overline{\widetilde{Q_t}}\right)^2} \quad ($$~~

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T}\left(Q_t^{\theta} - \widetilde{Q_t}\right)^2}{\sum_{T=1}^{T}\left(\widetilde{Q_t} - \overline{\widetilde{Q}}\right)^2} \quad (12\text{-}14)$$

~~NSE can range from $-\infty$ to 1, with NSE = 1 corresponding to perfect predictions of the observed data, and NSE < 0 indicates the observed mean is a better predictor than the model. While limitations in the NSE have been identified (see, for example,~~

~~Gupta et al., 2009), the metric has been included as it provides a measure of performance for the hydrological model, and has interpretation of given the broad application of the metric in the literature.~~

~~3.7 Model Configurations under Consideration~~

~~Based on the methodology outlined, a number of model configurations have been compared: CRR models with and without state updating, and each calibrated to a rolling 10 year or 20 year window. Each configuration has been applied to the C1 and C2 catchments, and driven by both observed and forecast rainfall. Hence, there were 64 different cases considered in this work.~~

~~Two different influences of the calibration period on the models were analysed. The first was changes in the predictive distribution with changes in the length of data record used to calibrate the model parameters. The second was the rolling approach used to calibrate the model parameters, where the model was recalibrated at the start of the water year to the most recent data, which allowed model parameters to change over time. Each influence is considered further in the following section.~~

NSE can range from $-\infty$ to 1, with NSE = 1 corresponding to perfectly accurate predictions of the observed data, and NSE < 0 indicating the observed mean is a better predictor than the model.

To ensure a consistent comparison of multiple model scenarios, the metrics are computed as follows:

- the same period is used to calculate the metrics in all cases. This period was determined by the availability of the forecast rainfall, from May 2001 to April 2011.
- the performance metrics are normalized by linearly scaling the worst value to a value of 0 and the best value to 1,

$$M_r = \frac{M - M_w}{M_b - M_w} \quad (15)$$

where the worst and best values for each metric, M_w and M_b, respectively, are listed in Table 2. The remainder of the presentation, in particular Figure 4, reports the normalized metrics computed using Eq. (15).

4 Results

The ~~results for each of the cases considered for each of the~~ performance metrics ~~outlined above~~for all model configurations are summarised in Figure 4. ~~The original values for each metric that correspond to the minimum and maximum bounds in Figure 4 can be seen in Table 2. From Figure 4, in general, the model that included state updating and was calibrated to the shorter period produced the best results across the metrics considered. The~~These changes ~~due adopting the state updating approach, for the different calibration periods,~~ are considered in ~~more~~ detail below. We begin by comparing the predictive performance of model configurations with and without state updating (Objective 1), and then investigate the influence of

calibration period length in the context of catchment non-stationarity (Objective 2), considering changes in both the predictive performance and changes in CRR parameter values over time.

4.1 Impact of State Updating

The ~~two options for initialising the model's routing store, a continuous 'warm up' period, or updating the routing store level based~~ impact of state updating on ~~the observed flow at the start of the forecast month, are shown as the~~ predictive performance can be seen in Figure 4, by comparing the red and ~~green~~blue bars ~~in Figure 4~~ (darker ~~colours~~shading indicating results for the 10 year calibration period length, and lighter ~~colours~~shading indicating results for the 20 year calibration period~~).~~ length). It is clear that state updating improves the sharpness, bias, $CRPS_{SS}$ and NSE metrics.

~~The models with state updating improved the precision and bias in all cases considered (Figure 4). This can be seen in Figure 5, where the 90th percentile predictive limits for the models with and without state updating, and a 10 year calibration period and observed rainfall data, are presented for a subset of the record considered. From Figure 5 it can be seen that the predictive limits were generally more precise when state updating was adopted compared to without state updating, particularly during the periods when there was some flow occurring in the winter-spring months. However, in dry months, state updating can be seen to slightly widen the predictive limits (for example at the start of 2001 for C2), potentially capturing the low flows that occurred at this time more accurately.~~

~~State updating also improved the reliability metric in 7 of the 8 comparative cases, with the only exception being for C2 and the 20 year calibration period with forecast rainfall. As an example of the improved reliability from state updating, PQQ plots with and without state updating, for the 10 year calibration period observed rainfall cases, can be seen in Figure 6. A line for a given predictive distribution closer to the 1:1 line represents a more reliable distribution, and it can be seen this is the case for the models that include state updating for both catchments. Any detrimental impacts on the predictive distribution from state updating tended to occur 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 Figure 5.~~

The improvement in predictive performance achieved by state updating to the observed flow data is tentatively attributed to being able to correct the model for any systematic overestimation of simulated streamflow. Consider Figures 5 and 6, which show the 90th percentile predictive limits for each model configuration, for catchments C1 and C2, respectively. The longer 20 year calibration period length without state updating is considered the "typical approach", and is shown in grey on each panel. A representative time period is shown, with the full time series for each case provided as Supplementary Material. Figures 5 and 6 show that state updating sharpens the predictive limits, especially during low flow months. For example, this behaviour can be seen for the 20 year CPL by comparing the predictions in panels (a) to (b) for the case of forecast rainfall and the predictions in panels (e) to (f) for the case of observed rainfall.

In terms of reliability, Figure 4 shows that state updating provides improved predictions for catchment C1. However, for catchment C2, Figure 4 shows that the reliability of all model configurations is relatively high compared to the reliability achieved in catchment C1, and state updating can lead to a slight loss of reliability.

4.2 Impact of Calibration Period Length

4.2.1 Differences in ~~predictive distribution~~Predictive Distribution

The changes in the predictive distribution due to changes in the ~~different~~ calibration period ~~lengths~~length can be seen in Figure 4, by comparing the darker to the lighter ~~shades~~shade of each colour (darker colour for 10 year calibration period length, lighter colour for 20 year calibration period~~) in Figure 4.~~ length). The following findings can be seen:

- ~~Most~~When state updating is not used (comparing dark blue versus light blue in Figure 4), all metrics improved ~~for~~when the shorter 10 year calibration periods length was used.
- When state updating is used (comparing the dark red versus light red in Figure 4), the impact of the shorter 10 year calibration period length depends on the catchment. In catchment C1, which provided stronger evidence of non-stationary than catchment C2 (Section 2.1), the use of the 10 year calibration period ~~where state updating was not used. The reduction in performance for the longer~~ length improves all metrics compared to the use of the 20 year calibration period ~~could be due to the model being calibrated to data that represent higher yielding conditions from the past, where the catchment displayed a higher runoff coefficient (i.e. the start of the 1990s in Figure 3) than in the forecast~~length. In contrast, in catchment C2, the length of the calibration period~~.~~ had little impact on the NSE and $CRPS_{SS}$ values; and only small improvements in the reliability, sharpness and bias metrics are obtained when the 10 year period is used.

~~However, where state updating was used, it was able to compensate for some of the changes, with smaller~~The differences between the ~~dark and light green bars compared to the dark and light red bars in Figure 4. This improvement is potentially due to the observed flow data being able to correct the model for any systematic overestimation in the simulated streamflow. The differences were more pronounced for the more practically relevant case with forecast rainfall, which introduced further errors to be compensated for by the state updating approach, and the postprocessor error model.~~

~~One aspect of the impact of different calibration periods that was not obvious from the changes in values for the metrics was trends in the model outputs when compared to the observed streamflow over time. These trends~~streamflow predictions obtained in the two catchments C1 and C2 (for the case of GR4J forced with observed rainfall) are illustrated ~~by the time series at the end of the period considered, presented in Figure 7. As noted in Section 2.2, catchment C1 had been identified to have a substantial reduction in the rainfall runoff relationship over time (Figure 3). From Figure 7, it can be seen that the model calibrated to the longer period overestimated~~in Figure 7 for the most recent period 2009-2011. In catchment C1, using a longer calibration period length tends to yield wider prediction limits and an overestimation of the observed flow in 2009 and 2010

~~in catchment C1~~, whereas ~~the model calibrated to~~using the shorter ~~period provided~~calibration length provides a better ~~reflection~~capture of the catchment response~~.~~ in these two years. In contrast, in catchment C2, which ~~had a much more consistent relationship between rainfall and runoff over time in~~has less evidence of non-stationarity (Section ~~Figure 3, the differences in the 90th percentile predictive limits due to~~2.1), the calibration period ~~in Figure 7 were minimal~~length makes very little difference on the resulting streamflow predictions.

4.2.2 Differences in ~~trends~~Trends in ~~parameter values~~Parameter Values

The rolling calibration approach ~~of recalibrating the CRR model parameters each year to the preceding data relevant for the calibration period considered (i.e. the previous 10 or 20 year period) allows for any~~(see Section 3.3) enables temporal trends in the parameter distributions to be investigated ~~(see Section 3.2).~~ Figure 8 presents the median~~,~~ and 90th percentile prediction limits of these distributions for each parameter for each catchment, ~~for~~with the 10 year and 20 year calibration ~~periods, as~~period lengths shown in different colours.

~~In catchment C1, for the first 10-12 years of the record, the median value for each parameter was similar for both calibration periods, with slightly wider bounds for the shorter calibration period, likely due to the reduced data available to infer representative parameter values. After this period, but particularly in the last 5 years, the shorter calibration period can be seen to identify values that result in reduced runoff, in particular more negative values for X2, the groundwater exchange coefficient. Reducing values for X4, the time base of unit hydrograph, may appear a strange result because it corresponds to a modification of the travel time within the catchment. However, the reduced value of X4 may represent a reduced contributing catchment area through interception in the upper reaches of the catchment. Cross-correlation analysis of the rainfall-runoff data supports this reduction in travel time based on the lag between rainfall and runoff producing the maximum correlation value, when comparing between the start and end of the dataset (Gibbs et al., 2017).~~

~~In catchment C2, the parameter values resulting from the different calibration periods were similar to each other. The distinct change in parameter values for the 10 year calibration period in 1999 appears to be a model fitting anomaly resulting from a shorter calibration period. This finding highlights that longer calibration periods would generally help identify more robust parameter values. There were no substantial differences in parameter values for the calibration periods considered, which may be expected given the more consistent rainfall runoff relationship.~~ In catchment C1, up until year 2005 (representing models calibrated from 1995 – 2004 for the 10 year calibration period length) the calibration period length has little impact on the median value for each parameter. Slightly wider parameter bounds are obtained when the shorter calibration period length is used, likely due to the reduced data available to infer representative parameter values. Post 2005, the parameter values obtained using the shorter calibration period length respond to the distinct non-stationarity of the catchment discussed in Section 2.1. The more pronounced negative values of the groundwater exchange coefficient X2 estimated in the 1994-2005 calibration period are consistent with the reduced runoff ratio in the period post 2000. In contrast, parameter values estimated from the

longer calibration period length, which includes data from the 1980s even when predicting the 2000s, do not exhibit this distinct change.

In catchment C2, the median values of parameter estimated from each calibration period length were similar over ~~time (Figure 3).~~the record. This result agrees with the lack of strong evidence of non-stationarity in this catchment. However, there ~~were~~is some evidence of a reduction in streamflow in this catchment, with the post 2000 period being characterized by a reduction in the runoff ratio from 0.088 to 0.061 (Section 2.1). This reduction is weaker in catchment C2 than in catchment C1, yet appears to be supported by the trends in the median parameter values ~~over time for both calibration periods. Considering the~~. Analysis of results from the 20 year calibration period~~, the~~ length suggests statistically significant trends ($p<0.05$) in the median values of the model parameters ~~were~~, namely $\Delta X1 = 3.96$ mm/year and $\Delta X3 = -5.17$ mm/year. ~~Both trends are in~~An exception to the ~~direction~~pattern of ~~parameters values that reduce runoff for a given amount~~the median parameter values being insensitive to calibration period lengths can be seen in 1999, where the use of ~~rainfall, through increased evaporation (increase~~the 10 year calibration period length produces higher values of X4 and lower values of X2 and the *split* introduced in ~~X1, maximum capacity of the production store) and increased loss to inter-basin transfers (reduction in X3, maximum capacity of the routing store).~~ this study (Section 3.1). This exception could represent a model fitting anomaly resulting from a shorter calibration period length.

5 Discussion

5.1 ~~The~~ Beneficial Impact of State-Updating on Forecast Performance

Most previous studies have used state updating in a short term flood forecasting context, and found limited effect of the initial conditions after a number of days (e.g. Berthet et al., 2009; Randrianasolo et al., 2014; Sun et al., 2017). However, forecasting of flood peak and timing is a different application to the forecasting of streamflow volumes. A number of data driven modelling studies have demonstrated that monthly streamflow lagged by one (or more) months provided some useful information for forecasting at a one month lead time (e.g. Bennett et al., 2014; Humphrey et al., 2016; Yang et al., 2017). This study demonstrates that these benefits also hold when CRR models ~~were~~, rather than data-driven approaches, are used as the forecasting model.

State updating is found to improve ~~the~~ predictive ~~distributions produced, across the~~ performance ~~metrics for the~~in both catchments considered~~.~~, for the majority of the multiple performance metrics considered. State updating ~~would be~~is expected to reduce ~~the simulated~~predictive bias, as errors in the simulated streamflow during the warm up period are corrected at the start of the forecast period. State updating ~~would be~~is also expected to increase the ~~precision~~sharpness of the predictive distribution, as the range ~~in~~of model predictions ~~should be reduced~~is generally tightened by forcing the model to simulate the observed streamflow at the start of the forecast period. ~~However, the trade-off for~~

The only metric where state updating did not show an ~~increase in precision would typically be a reduction in~~ improvement is for the reliability of ~~the predictive distribution (e.g. McInerney et al., 2017). This was not the case in this work, where in all but one of the cases considered, reliability also improved.~~predictions for catchment C2. However, the reliability of all model configurations in this catchment is already relatively high without state updating. All other metrics (sharpness, bias, CRPS and NSE) show improvements from state updating in catchment C2, suggesting potential trade-offs in performance, similar to that reported by McInerney et al. (2017). This slight reduction in reliability is not considered to have a significant detrimental impact of the PD produced for this practical application.

~~The state updating approach adopted could be modified to potentially further improve the reliability of the forecasts produced. For example, the approach could be extended to also update the GR4J production store along with the routing store. This could be expected to improve the results in the periods where detrimental changes were found, in months with very low flow and limited storage in the routing store to update. Given the main detrimental impacts of state updating occurred in the no/low flow period, the approach could also be adopted only in periods where there was some recorded flow. However, forcing the model to correctly represent an observed no flow may also be beneficial (for example, in late 2002 and into 2003 in Figure 5), as an approach such as this requires further investigation.~~

~~5.2~~ ~~Out of context, the calibration period results could be seen to suggest that the shorter calibration period provided better (or at least not detrimental) parameter estimates. However, this interpretation would not be expected to be a general result, nor would it indicate that even~~Importance of Choosing a Calibration Period that is Representative of Current Catchment Conditions.

Traditionally, long calibration periods are used to maximise the use of available data and increase parameter identifiability. The empirical results in this study suggest that the shorter calibration period can provide better (or at least not worse) predictive performance. The reduction in performance seen when the longer calibration period is used is likely due to the calibration data representing catchment conditions that are substantially different to those in the prediction period. For example, when the prediction period is 2009 (as shown for catchment C1 in Figure 7), a 20 year calibration period length corresponds to the period of 1989-2008, which includes a large portion of the pre-2000 period when catchment C1 displayed a much higher runoff coefficient (section 2.1). In contrast, a 10 year calibration period length corresponds to a calibration period of 1999-2008, which is likely to be more representative of the lower runoff hydrological regime seen in the post 2000 period.

The reported improvement in model performance with the 10 year calibration period length does not imply that shorter calibration periods would ~~further improve predictions. Ultimately, our~~result in further improvements. Shorter calibration period lengths will eventually reduce parameter identifiability (e.g., as manifested by greater parameter uncertainty in Figure 8), and may produce poor parameters estimates due to fitting only a small number of events and hence being unable to represent the full range of flow conditions.

The empirical findings highlight the benefits of identifying ~~the longest~~a calibration period of data that is representative of conditions of interest for a given model application, which is a task often overlooked in practical ~~application. The expected result that longer calibration periods improve parameter identifiability can be seen in Figure 8, however for some catchments, there may be a trade-off with calibrating a model to data that are no longer~~applications. Suitable representative ~~of the current or future rainfall-runoff regime. As such, data suitable for model calibration should be~~ periods can be identified through techniques such as trend analysis~~ and other~~, using knowledge of changes in a catchment (e.g. land use data, abstraction volumes~~). The~~), and testing predictive performance for different calibration period lengths (as done in this work). The empirical results ~~presented~~ indicate that, if ~~this step~~ the selection of calibration data is ~~ignored~~poorly implemented, and/or if the modeller naively assumes that longer calibration periods are inherently better for model development, ~~detrimental impacts to all aspects of the~~ predictive ~~distribution~~performance can ~~occur~~degrade appreciably.

5.3 Value of Forecasts for Improving Water Management

The forecasting approaches developed in this work can support improved water management in the drainage system considered. The approach currently used by the management authority is very conservative: streamflow forecasts are not attempted, and changes in water management are made only once downstream requirements have been met. With the forecasting models and methods developed in this work, it becomes possible to produce streamflow forecasts with a high reliability, improved sharpness and reduced bias. Thus it becomes possible to provide useful probabilistic estimates of how likely it is that the downstream flow requirements will be met in the next month. With this information, managers can more confidently consider increasing the frequency and duration of inundation for many of the wetlands in the region, and can make decisions on management changes much earlier in the season.

5.4 Future Research Work

The enhancements to predictive performance of streamflow forecasts from state updating and a shorter calibration period have been demonstrated on two catchments. These catchment were selected based on an established user need for seasonal forecasts to improve the water management of a channel drainage system with multiple competing demands. Importantly, the case study catchments in this work are ephemeral and dry, with low runoff ratios. These types of catchments are known to be challenging to model (McInerney et al., 2017; Ye et al., 1997). Future work will evaluate the proposed seasonal streamflow forecasting techniques over a wider range of catchments and environmental conditions.

6 Conclusions

~~This work investigated the ability to produce streamflow forecasts for an application in southern Australia, where streamflow can be diverted to support a range of environmental and social outcomes, and where improved information on future water~~

~~availability can assist in decision support and management. As part of meeting this objective, several methods related to improving the predictive ability of streamflow forecasts were investigated.~~

~~The first main contribution of this study was the implementation of the state updating approach to assimilate recently observed streamflow into the conceptual rainfall-runoff model GR4J, and to assess the implications of this approach on the forecast predictions. The results indicate that this simple state updating approach could improve all aspects of the predictive distribution, including the precision and reliability of the forecasts.~~

This work has focused on improving monthly streamflow forecasts by considering two aspects: 1) state updating to force the GR4J hydrological model to match observations from the start of the forecast period, and 2) investigating the trade-offs between using shorter versus longer calibration periods. The analysis was applied to two ephemeral catchments in southern Australia, which are part of a drainage network with competing environmental management demands.

The major findings from the empirical analysis are as follows:

1. State updating improves predictive performance in the case study catchments, for the majority of the multiple performance metrics considered. Previous studies focusing on the forecasting of flood peak and timing have typically found limited effect of initial conditions on predictive performance after a number of days. This study demonstrates that, when forecasting streamflow volumes, using state updating to more accurately represent initial conditions can have a benefit even at a one month lead time.
2. ~~1.~~ The ~~second main major contribution of this study was to establish the impact of the~~ length of the calibration period ~~on the forecast accuracy.~~has a major impact on predictive performance of a hydrological model. In the case study catchments, the shorter calibration period typically improves predictive performance, especially in the case study catchment with stronger evidence of non-stationarity. The ~~results indicated that the~~benefits of a shorter calibration ~~period improved model performance in most cases,~~length appears contrary to the standard approach of using as much data as possible for model calibration. The reduction in performance for the longer calibration period is likely due to the model being calibrated to data that represent higher yielding conditions from the past, ~~where the catchment displayed a higher runoff coefficient than~~which no longer hold true in the forecast period. This ~~result~~finding highlights that identifying a data set that is representative of the forecast period, through trend analysis and other knowledge of a catchment, is an important step in model development. ~~The results presented indicate that if~~If this step is ignored, and it is naively assumed that ~~the more~~longer calibration data ~~the~~is inherently better for model development, ~~detrimental impacts to~~ all aspects of ~~the~~ predictive ~~distribution can result~~performance may suffer.

The conclusions ~~drawn in~~of this empirical study are limited by the ~~one case study region considered. While the results are intuitive, further work should~~small number of catchments and single hydrological model used. Further work will consider a larger sample of catchments~~, encompassing~~ and a wider range of ~~flow regimes. Similarly, only one CRR model structure has been considered, and the state updating approach, and benefits thereof, may change for different~~hydrological model structures.

~~Further work should also consider the merits of updating other aspects of the models considered, for example similar or improved performance may be achieved through updating the error estimated by the postprocessor error model instead of, or as well as, the CRR model storage level. Nonetheless, this study has demonstrated a number of relatively simple~~ In general, we expect the techniques ~~for pre processing~~of state updating, post-processing uncertainty estimation, and usage of shorter calibration ~~approaches~~period length representative of future forecast conditions to ~~enhance monthly streamflow forecasting skill.~~ be of value to hydrologists and environmental modellers seeking to improve the predictive performance of their modelling systems.

Author Contribution

M. Gibbs performed the analysis and produced the manuscript, with contributions from all co-authors. H. Maier and G. Dandy assisted with the design of the project. D. McInerney undertook the postprocessor error modelling and analysis, with help from M. Thyer and D. Kavetski. G. Humphrey implemented the climate model forecast downscaling to generate the inputs for the hydrological models.

Acknowledgements

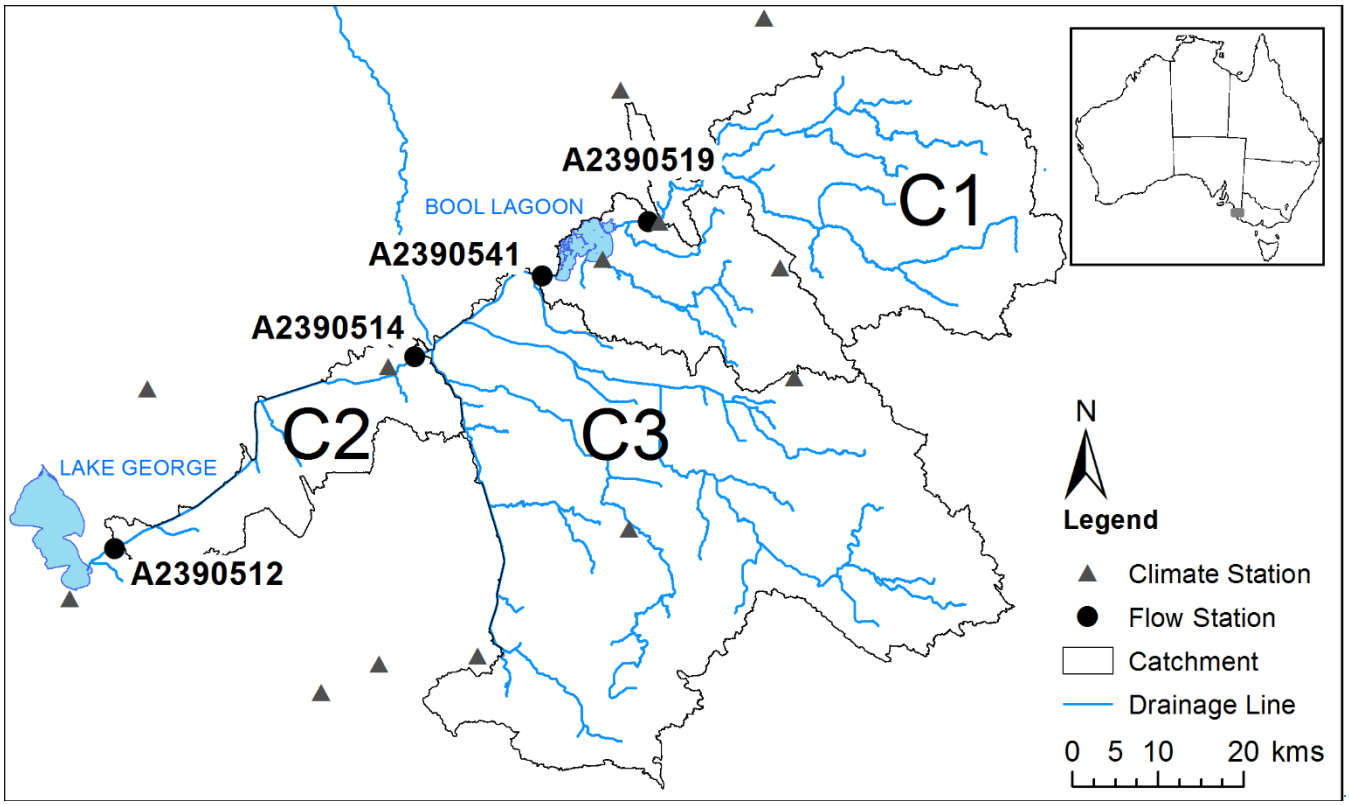
The flow data used in this paper is available from the South Australian Department for Environment, Water and Natural Resources Surface Water Archive (https://www.waterconnect.sa.gov.au/Systems/swd). The climate data used in this paper is available from the Queensland Department of Science, Information Technology, Innovation and the Arts SILO climate data archive (https://www.longpaddock.qld.gov.au/silo/). Access to forecast climate data from the ~~POAMA2~~POAMA-2 model was gratefully provided by the Bureau of Meteorology (http://poama.bom.gov.au/). M. Gibbs and G. Humphrey were supported by the Goyder Institute for Water Research, Project E.2.4. D. McInerney was supported by Australian Research Council grant LP140100978 with the Australian Bureau of Meteorology and South East Queensland Water. Input from South East Water Conservation and Drainage Board staff, in particular Senior Environmental Officer, Mark DeJong, is gratefully acknowledged. The authors would like to thank the three anonymous reviewers for their comments and suggestions, which improved the clarity and contribution of the manuscript.

Table 1 Bounds adopted for the uniform prior distribution on the GR4J parameters

Parameter	Name	Lower Bound	Upper Bound
X1	production store maximal capacity (mm)	100	600
X2	catchment water exchange coefficient (mm)	-15	5
X3	one-day maximal capacity of the routing reservoir (mm)	1	300
X4	unit hydrograph time base (days)	0.5	6
split	proportion of flow directed to the routing store	0.6	0.99

Table 2 ~~Minimum~~Best and ~~maximum~~worst values for each predictive performance metric across all ~~models and catchments considered. Higher values are better for~~model configurations. For CRPS$_{SS}$ and NSE, ~~lower~~higher values ~~are~~denote better performance; for the other metrics lower values denote better performance. The values in this table should be interpreted alongside Figure 4, where the worst and best values reported here ~~corresponding~~correspond to ~~the 0.05 and 0.95 relative~~metric values ~~in Figure 4.~~of 0 and 1, respectively.

	Reliability	~~Precision~~Sharpness	Bias	CRPS$_{SS}$	NSE
~~Minimum~~Worst	0.~~08~~41	~~0.45~~2.21	~~0.04~~1.49	-0.~~64~~65	-1.01
~~Maximum~~Best	0.~~39~~07	~~2.18~~0.45	~~1.48~~0.11	0.~~65~~57	0.~~91~~88


A2390519
BOOL LAGOON
A2390541
A2390514
C1
C2
C3
LAKE GEORGE
A2390512
N
Legend
Climate Station
Flow Station
Catchment
Drainage Line
0 5 10 20 kms

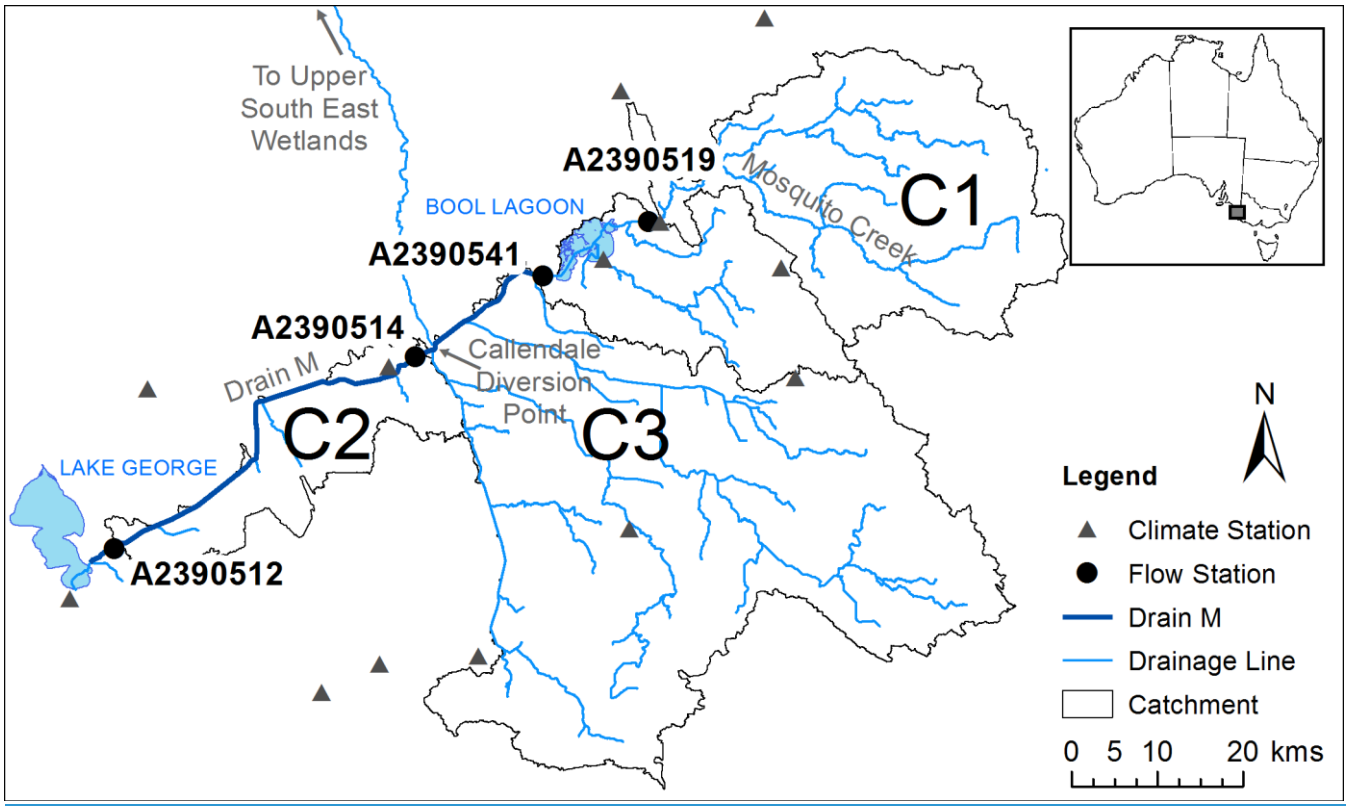
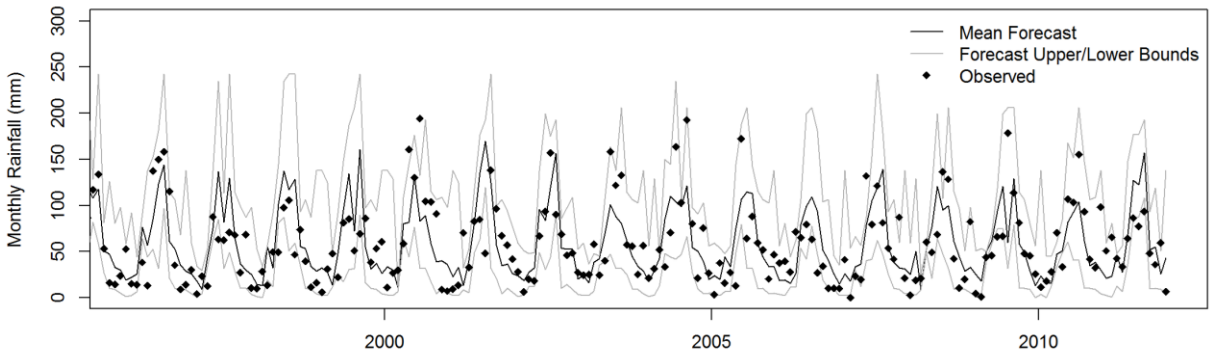


Figure 1 Map of the ~~forecasting application~~case study region, in the ~~south east of South~~southern Australia



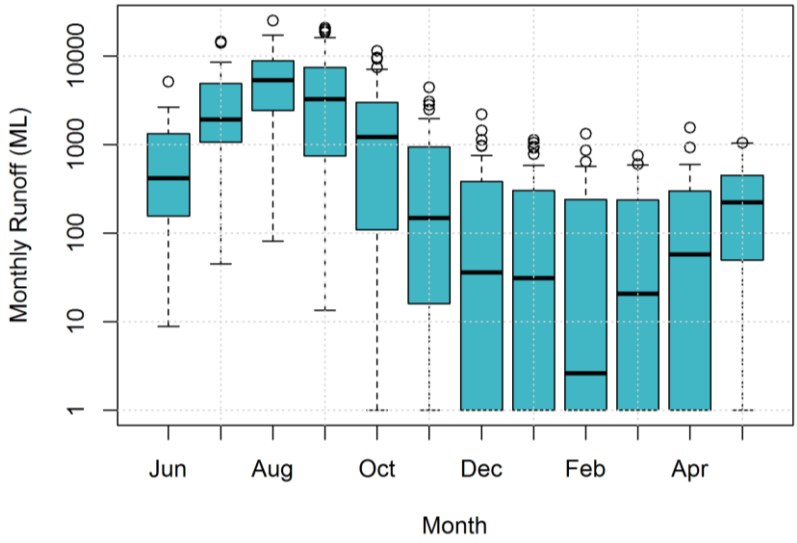
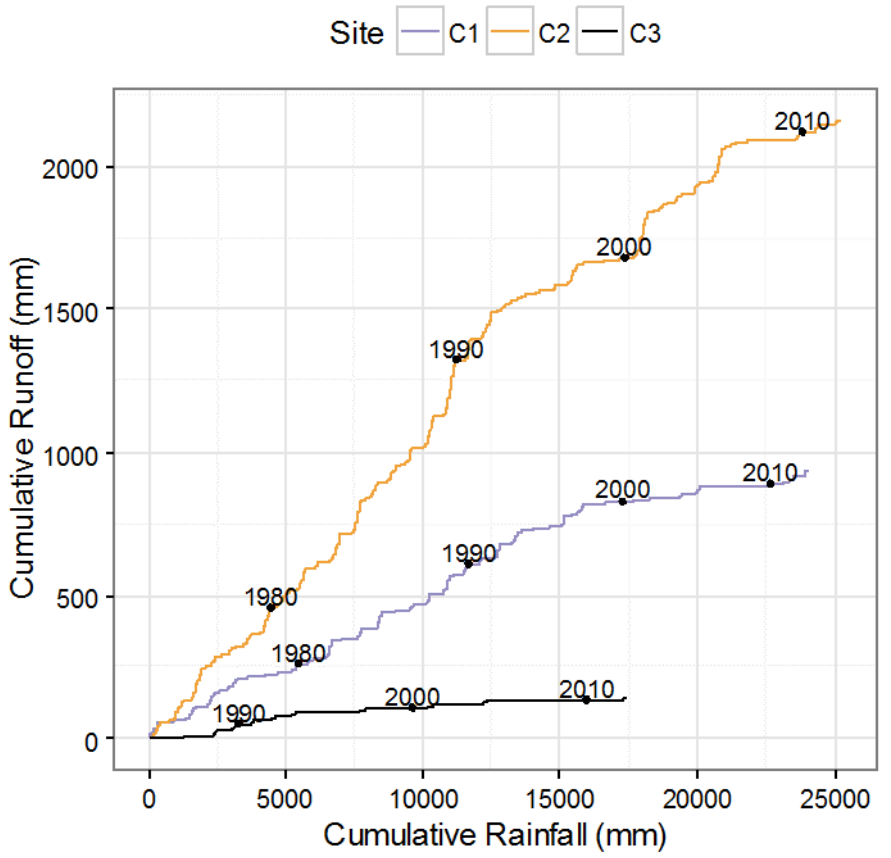


Figure 2 ~~Example of downscaled POAMA rainfall forecasts~~Variability in monthly runoff in Drain M **at** ~~station 26000 in comparison to observed rainfalls~~location at Flow Station A2390512.


Site
C1
C2
C3
Cumulative Runoff (mm)
Cumulative Rainfall (mm)
0
500
1000
1500
2000
0
5000
10000
15000
20000
25000
1980
1990
2000
2010
1980
1990
2000
2010
1990
2000
2010

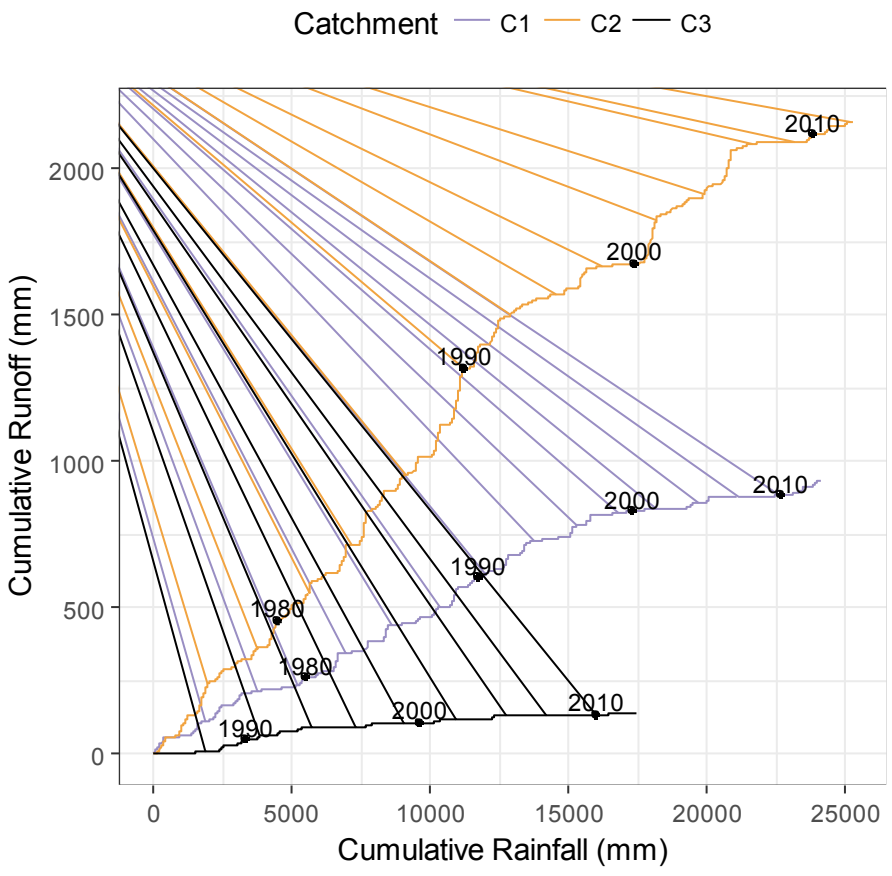
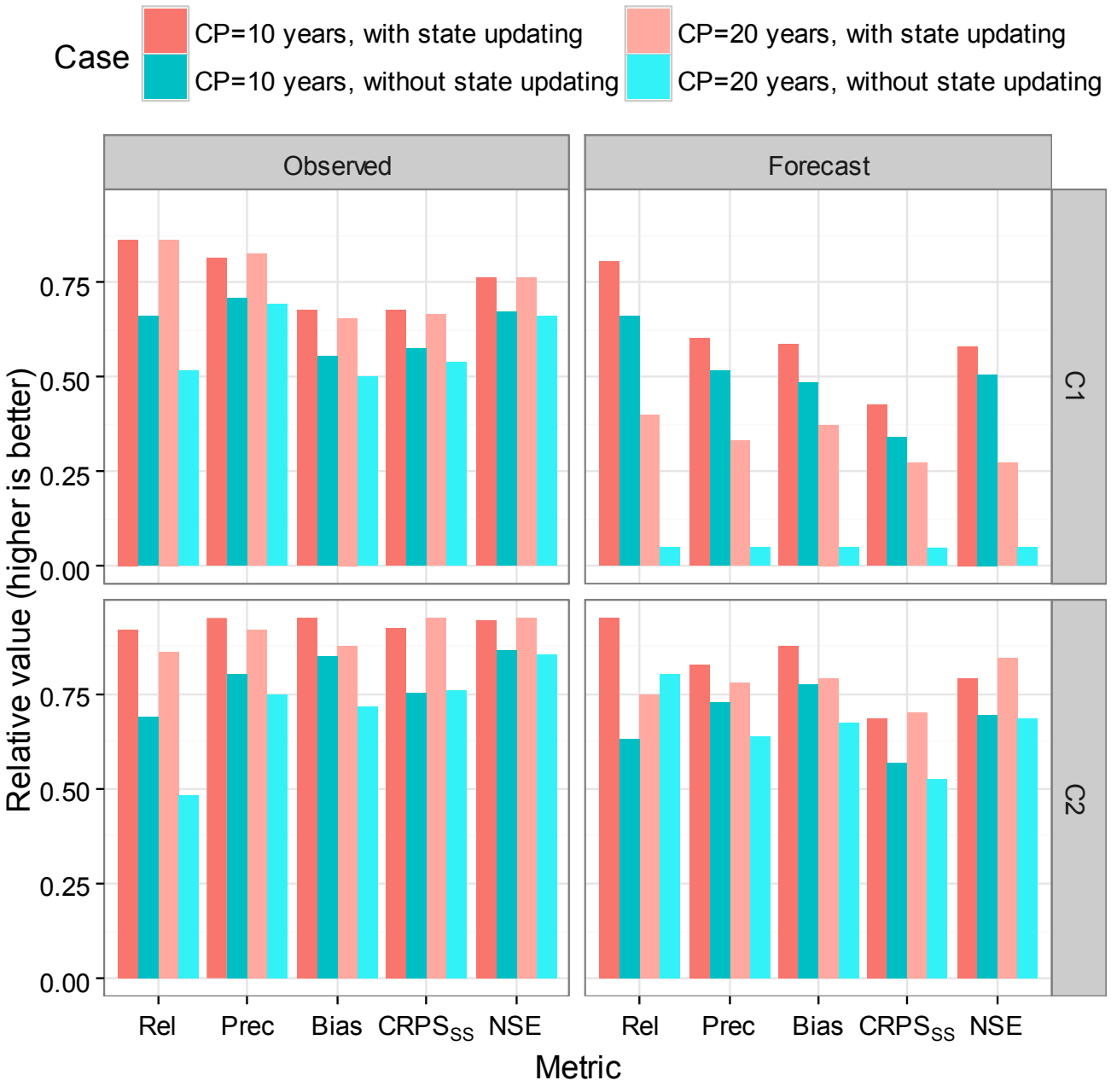


Figure 3 Double mass plot of the rainfall-runoff data in the three main catchments contributing to Drain M. It can be seen that 1) the volume of runoff for the same volume of rainfall has reduced in the latter decade, and 2) very little runoff is generated from ~~the C3~~ catchment C3.

5


Case
CP=10 years, with state updating
CP=10 years, without state updating
CP=20 years, with state updating
CP=20 years, without state updating
Observed
Forecast
C1
C2
Relative value (higher is better)
0.00
0.25
0.50
0.75
Rel
Prec
Bias
CRPSSS
NSE
Metric

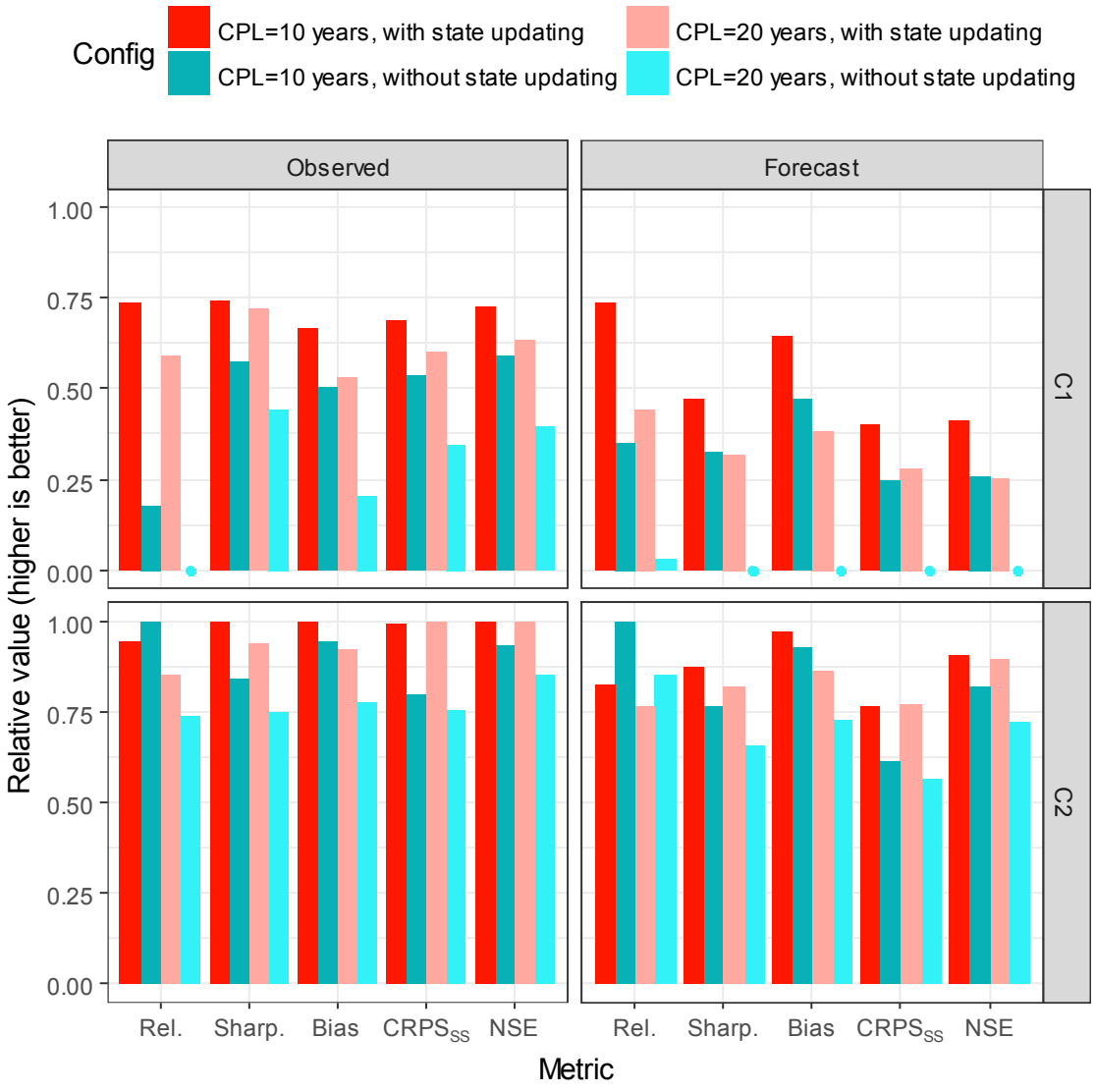
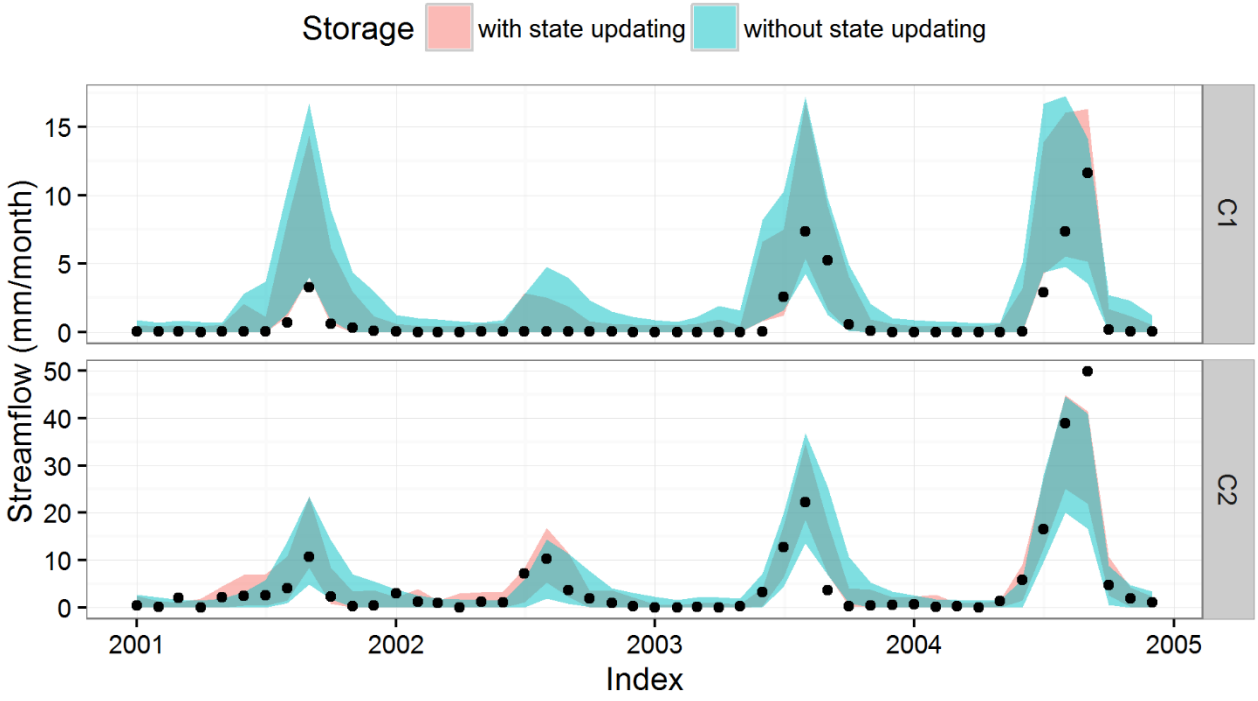


Figure 4 ~~Comparison of~~Predictive performance metrics ~~across~~for the two case study catchments (C1 and C2) and ~~source~~the two sources of rainfall forcing data~~. The~~ (observed and forecast). Relative metric values are presented (Section 3.7 and Table 2); higher values represent better performance. The impact of ~~models with and without~~ state updating can be seen ~~between the red and green bars (~~by comparing the ~~darker shades and lighter shades).~~red vs blue bars. The change in performance due to different calibration ~~periods (CP~~period lengths (CPL) can be ~~compared between~~seen by comparing the bars with darker ~~and~~vs lighter shading ~~(darker colour representing the 10 year period, and lighter colour 20 year period).~~.


Storage
with state updating
without state updating
Streamflow (mm/month)
15
10
5
0
50
40
30
20
10
0
C1
C2
2001
2002
2003
2004
2005
Index

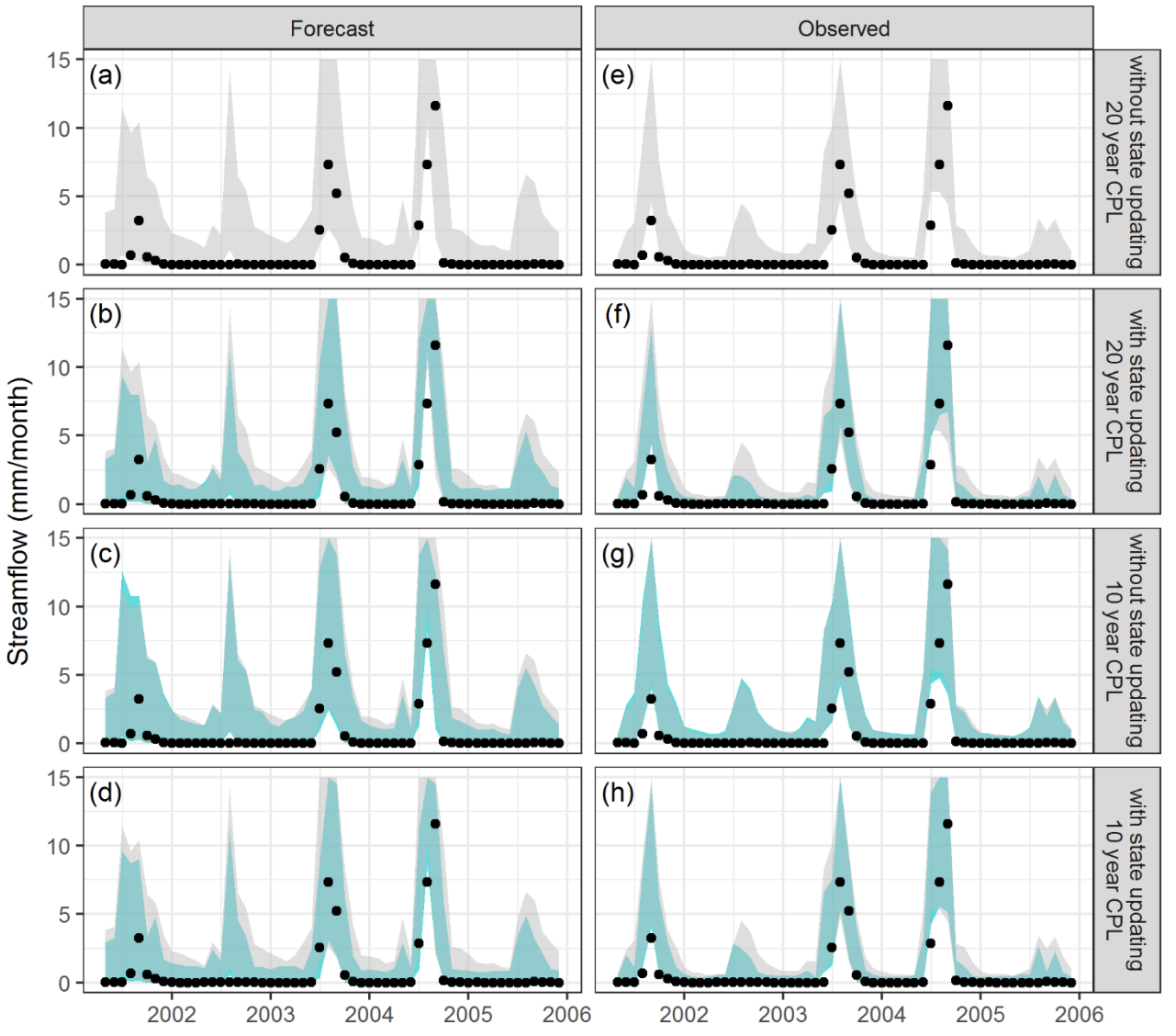
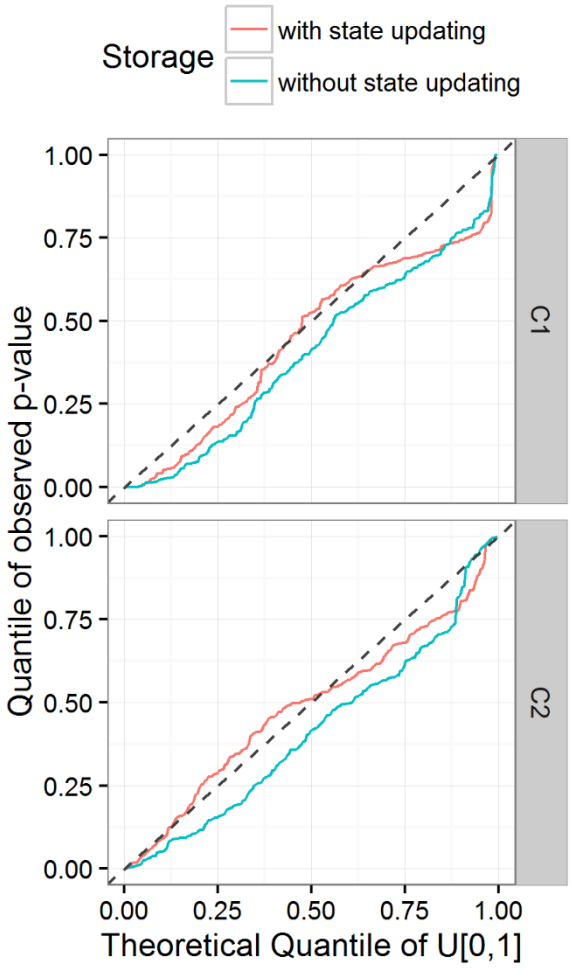


Figure 5 Representative streamflow **time series** ~~of the~~ in catchment C1 obtained using forecast rainfall (left) and observed rainfall (right). The shaded area represents the **90th percentile prediction limits**~~, with observed values as~~ and the **black dots**~~, for~~ the ~~10~~observed values. The "traditional approach" of the 20 **year calibration period** ~~observed rainfall models. State updating can be seen to increase precision compared to~~length (CPL) **without state updating** is showing in grey on each panel**.**


Storage
with state updating
without state updating
1.00
0.75
0.50
0.25
0.00
C1
C2
Quantile of observed p-value
0.00
0.25
0.50
0.75
1.00
Theoretical Quantile of U[0,1]

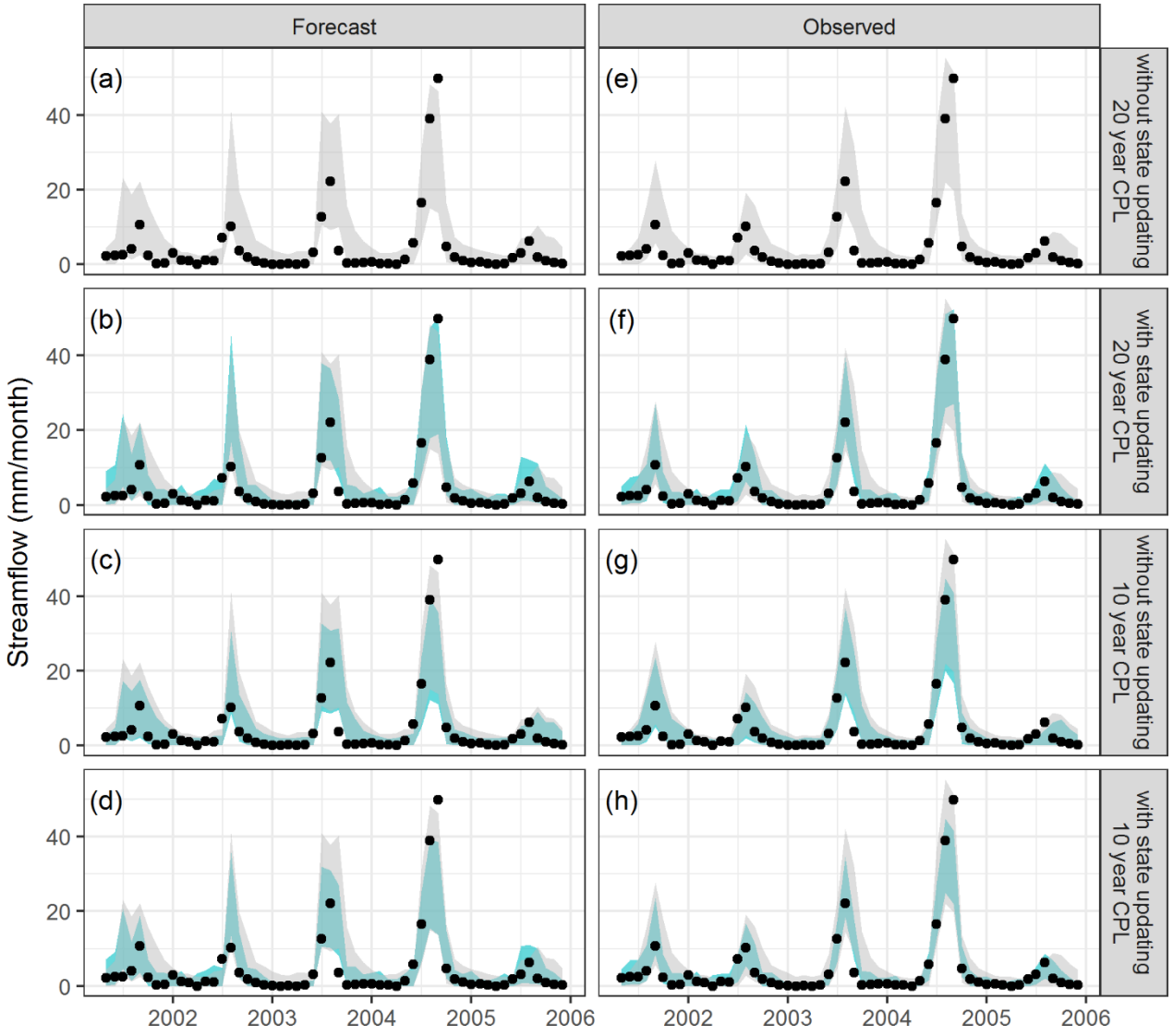
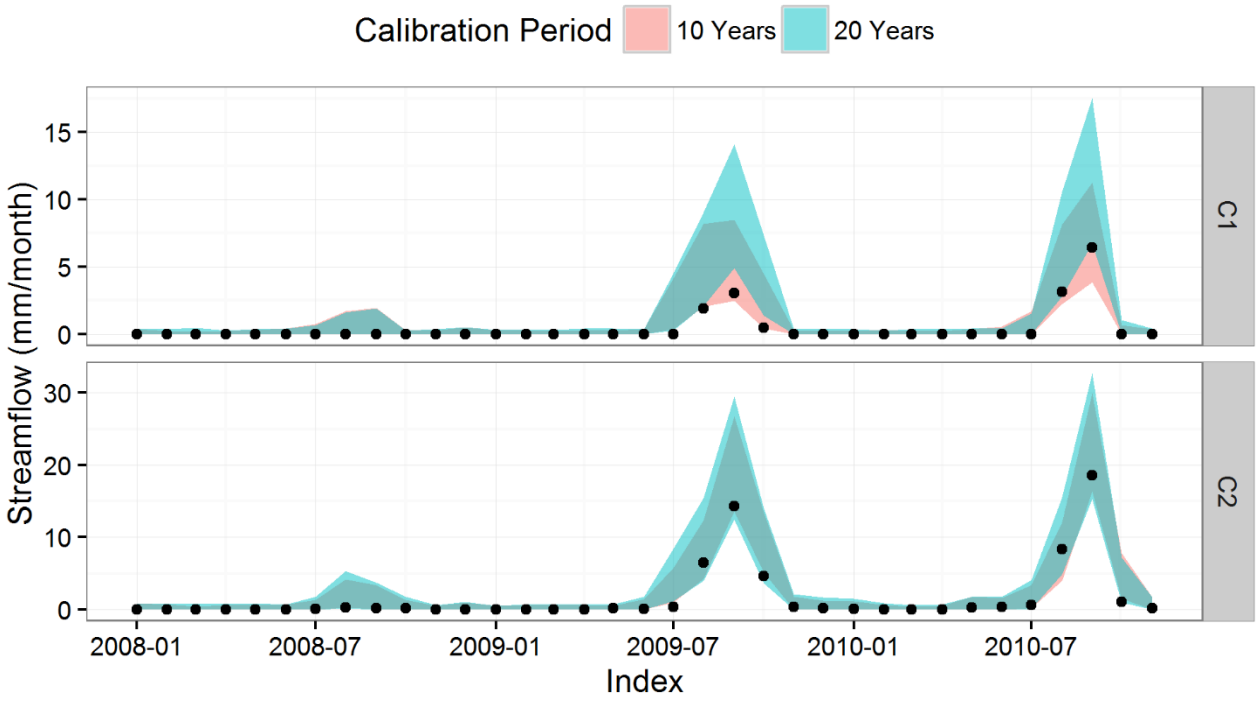


Figure 6 ~~Predictive Quantile – Quantile plots for models with the 10 year calibration period and forced by observed rainfall, comparing reliability with and without state updating. State updating can be seen to increase the reliability of the predictions.~~



~~Figure 7~~ Representative streamflow time series ~~of~~in catchment C2 obtained using forecast rainfall (left) and observed rainfall (right). The shaded area represents the 90th percentile prediction limits~~, with observed values as~~ and the black dots~~, for~~ the ~~models with state updating and~~ observed ~~rainfall. For the catchment with non-stationary changes (C1), the use of the longer calibration period results in overestimated flow at the end of the period, due to the influence of data earlier in the calibration period~~values. The "traditional approach" of the 20 year calibration period length (CPL) without state updating is showing in grey on each panel.

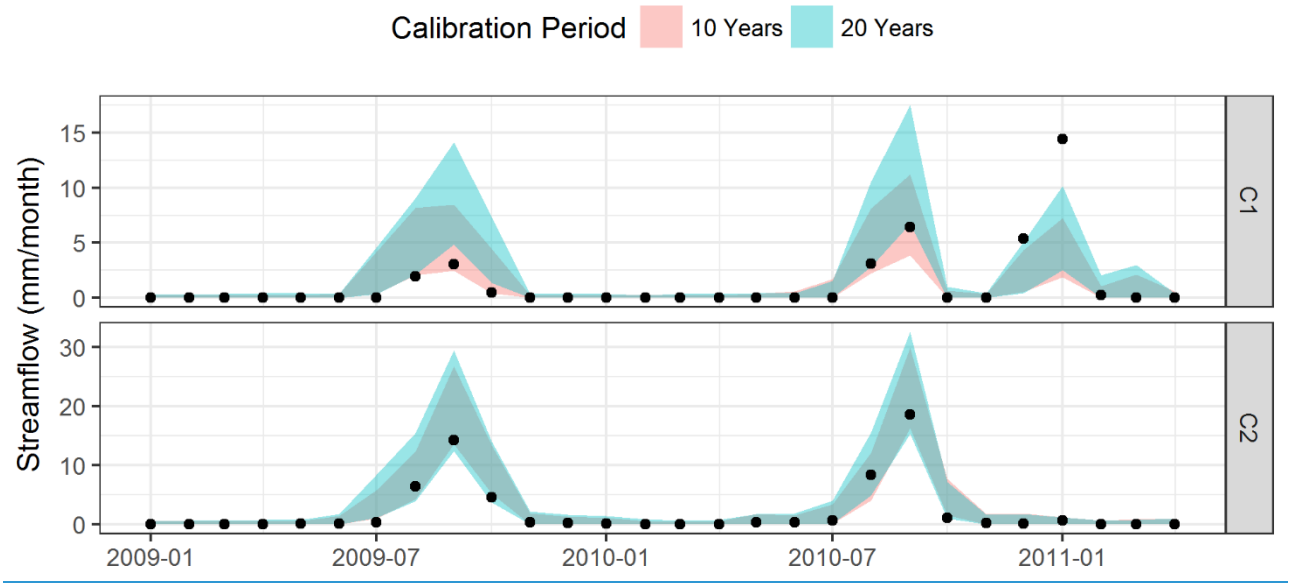


Figure 7 Streamflow predictions for catchment C1 (top) and C2 (bottom) for period 2009-2011 using observed rainfall. The shaded area represents the 90th percentile prediction limits and the black dots the observed values. For catchment C1, using shorter calibration periods (red) can be seen to produce lower streamflow predictions than using longer calibration periods (blue).

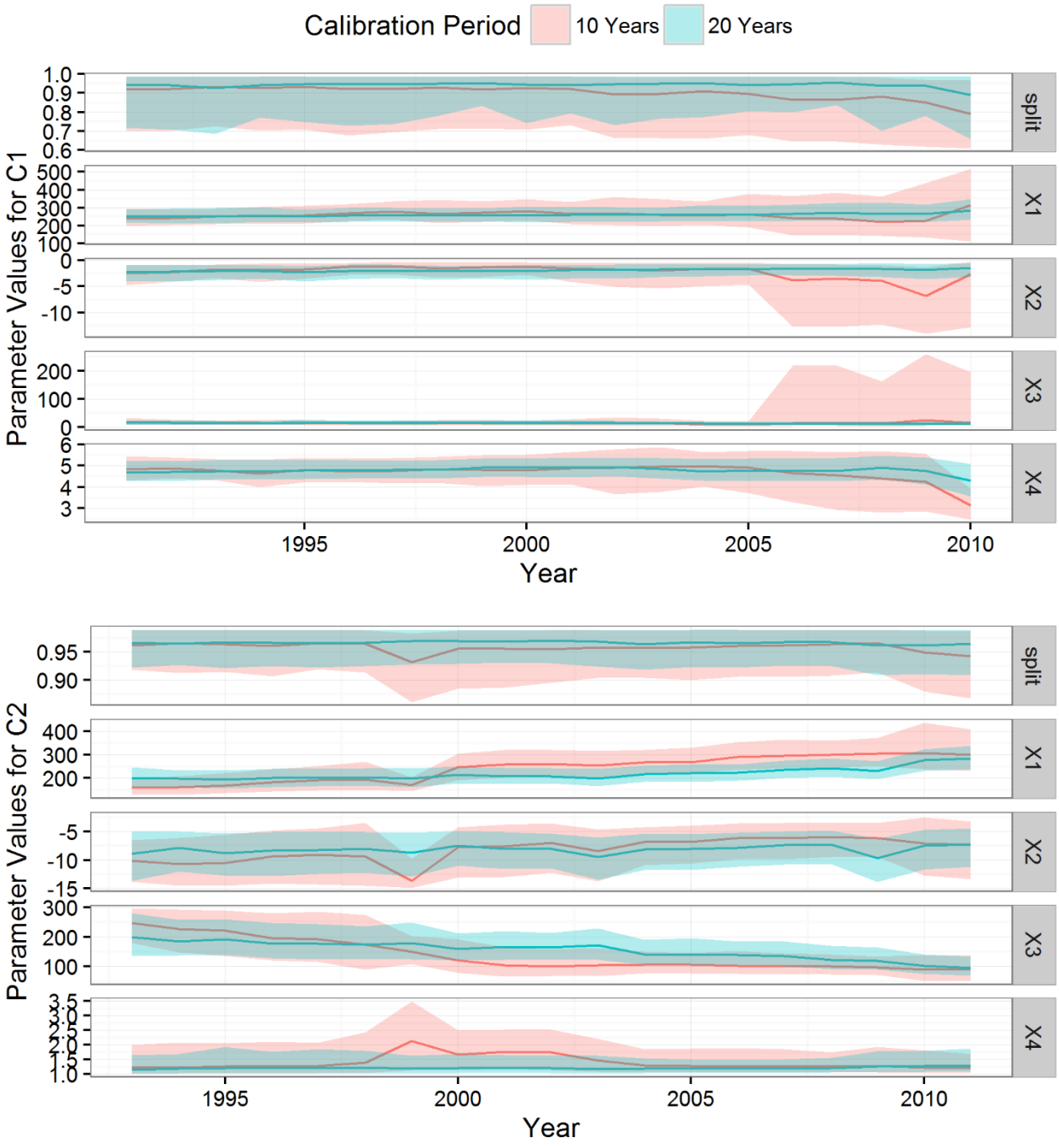


Figure 8 Temporal trends in posterior parameter distributions, for catchments C1 (top) and C2 (bottom). The median values are shown as the solid lines and the shaded area represent the 90th percentile prediction limits.

References

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M.: Crop evapotranspiration-Guidelines for computing crop water requirements, FAO, RomeFAO Irrigation and drainage paper 56, 300 pp., 1998.

Andrews, F. T., Croke, B. F. W., and Jakeman, A. J.: An open software environment for hydrological model assessment and development, Environmental Modelling & Software, 26, 1171-1185, 2011.

Avey, S. and Harvey, D.: How water scientists and lawyers can work together: A 'down under' solution to a water resource management problem, Journal of Water Law, 24, 25-61, 2014.

Bennett, J. C., Wang, Q. J., Pokhrel, P., and Robertson, D. E.: The challenge of forecasting high streamflows 1 & 3 months in advance with lagged climate indices in southeast Australia, Nat. Hazards Earth Syst. Sci., 14, 219-233, 2014.

Berthet, L.: Prévision des crues au pas de temps horaire : pour une meilleure assimilation de l'information de débit dans un modèle hydrologique, 2010. AgroParisTech, 2010.

Berthet, L., Andreassian, V., Perrin, C., and Javelle, P.: How crucial is it to account for the antecedent moisture conditions in flood forecasting? Comparison of event-based and continuous approaches on 178 catchments, Hydrology and Earth System Sciences, 13, 819-831, 2009.

Beven, K. ~~J., Smith, P. J., and Freer, J. E.: So just why would a modeller choose to be incoherent?, Journal of Hydrology, 354, 15-32, 2008~~and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrological Processes, 6, 279-298, 1992.

Bowden, G. J., Maier, H. R., and Dandy, G. C.: Real-time deployment of artificial neural network forecasting models: Understanding the range of applicability, Water Resources Research, 48, 2012.

Brigode, P., Oudin, L., and Perrin, C.: Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change?, Journal of Hydrology, 476, 410-425, 2013.

Brookes, J. D., Aldridge, K., Dalby, P., Oemcke, D., Cooling, M., Daniel, T., Deane, D., Johnson, A., Harding, C., Gibbs, M., Ganf, G., Simonic, M., and Wood, C.: Integrated science informs forest and water allocation policies in the South East of Australia, Inland Waters, 7, 358-371, 2017.

Coron, L., Andreassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F.: Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, Water Resources Research, 48, W05552, 2012.

de Vos, N. J., Rientjes, T. H. M., and Gupta, H. V.: Diagnostic evaluation of conceptual rainfall–runoff models using temporal clustering, Hydrological Processes, 24, 2840-2850, 2010.

Demargne, J., Wu, L. M., Regonda, S. K., Brown, J. D., Lee, H., He, M. X., Seo, D. J., Hartman, R., Herr, H. D., Fresch, M., Schaake, J., and Zhu, Y. J.: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service, Bull. Amer. Meteorol. Soc., 95, 79-98, 2014.

Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models, Water Resources Research, 49, 4035-4053, 2013.

Evin, G., Thyer, M., Kavetski, D., McInerney, D., and Kuczera, G.: Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, Water Resources Research, 50, 2350-2375, 2014.

~~Gibbs, M. S., Dandy, G. C., and Maier, H. R.: Assessment of the ability to meet environmental water requirements in the Upper South East of South Australia, Stochastic Environmental Research and Risk Assessment, 28, 39-56, 2014.~~

~~Gibbs, M. S., Dandy, G. C., Maier, H. R., and Thyer, M. A.: Representing Catchment Nonstationarity in Conceptual Rainfall Runoff Models Using Time-Varying Parameters and Groundwater Data, Water Resources Research, Submitted, 2017.~~

Fenicia, F., Kavetski, D., Savenije, H. H. G., Clark, M. P., Schoups, G., Pfister, L., and Freer, J.: Catchment properties, function, and conceptual model representation: is there a correspondence?, Hydrological Processes, 28, 2451-2467, 2014.

Gibbs, M. S., Maier, H. R., and Dandy, G. C.: A generic framework for regression regionalization in ungauged catchments, Environmental Modelling & Software, 27-28, 1-14, 2012.

Guo, D., Westra, S., and Maier, H. R.: Impact of evapotranspiration process representation on runoff projections from conceptual rainfall-runoff models, Water Resources Research, doi: 10.1002/2016WR019627, 2017. ~~n/a-n/a,~~ 2017.

~~Gupta~~Hudson, D., Alves, O., Hendon, H. ~~V., Kling,~~ H., ~~Yilmaz, K. K.,~~ and ~~Martinez,~~Marshall, A. G~~. F.: Decomposition of~~.: Bridging the ~~mean squared error~~gap between weather and ~~NSE performance criteria: Implications~~seasonal forecasting: intraseasonal forecasting for ~~improving hydrological modelling,~~Australia, Quarterly Journal of ~~Hydrology, 377, 80-91, 2009.~~

~~He, M., Hogue, T. S., Margulis, S. A., and Franz, K. J.: An integrated uncertainty and ensemble-based data assimilation approach for improved operational streamflow predictions, Hydrology and Earth System Sciences, 16, 815-831, 2012~~the Royal Meteorological Society, 137, 673-689, 2011.

Humphrey, G. B., Gibbs, M. S., Dandy, G. C., and Maier, H. R.: A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network, Journal of Hydrology, 540, 623-640, 2016.

Jeffrey, S. J., Carter, J. O., Moodie, K. B., and Beswick, A. R.: Using spatial interpolation to construct a comprehensive archive of Australian climate data, Environmental Modelling & Software, 16, 309-330, 2001.

Kavetski, D., Franks, S. W., and Kuczera, G.: Confronting Input Uncertainty in Environmental Modelling. In: Calibration of Watershed Models, American Geophysical Union, 2003.
Koster, R. D., Mahanama, S. P. P., Livneh, B., Lettenmaier, D. P., and Reichle, R. H.: Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow, Nature Geoscience, 3, 613-616, 2010.
Lei, F. N., Huang, C. L., Shen, H. F., and Li, X.: Improving the estimation of hydrological states in the SWAT model via the ensemble Kalman smoother: Synthetic experiments for the Heihe River Basin in northwest China, Advances in Water Resources, 67, 32-45, 2014.
Krzysztofowicz, R. and Maranzano, C. J.: Hydrologic uncertainty processor for probabilistic stage transition forecasting, Journal of Hydrology, 293, 57-73, 2004.
Lerat, J., Pickett-Heaps, C., Shin, D., Zhou, S., Feikema, P., Khan, U., Laugesen, R., Tuteja, N., Kuczera, G. T., M, and Kavetski, D.: Dynamic streamflow forecasts within an uncertainty framework for 100 catchments in Australia, 36th Hydrology and Water Resources Symposium: The art and science of water, Barton, ACT, 1396-1403, 2015.
Li, H. B., Luo, L. F., Wood, E. F., and Schaake, J.: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting, Journal of Geophysical Research-Atmospheres, 114, 10, 2009.
Li, Y., Ryu, D., Western, A. W., and Wang, Q. J.: Assimilation of stream discharge for flood forecasting: The benefits of accounting for routing time lags, Water Resources Research, 49, 1887-1900, 2013.
Li, L., Lambert, M. F., Maier, H. R., Partington, D., and Simmons, C. T.: Assessment of the internal dynamics of the Australian Water Balance Model under different calibration regimes, Environmental Modelling & Software, 66, 57-68, 2015a.
Li, Y., Ryu, D., Western, A. W., and Wang, Q. J.: Assimilation of stream discharge for flood forecasting: Updating a semidistributed model with an integrated data assimilation scheme, Water Resources Research, 51, 3238-3258, 20152015b.
Liu, Y. and Gupta, H. V.: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, Water Resources Research, 43, 2007.
Luo, J., Wang, E., Shen, S., Zheng, H., and Zhang, Y.: Effects of conditional parameterization on performance of rainfall-runoff model regarding hydrologic non-stationarity, Hydrological Processes, 26, 3953-3961, 2012.
Matte, S., Boucher, M. A., Boucher, V., and Fortier Filion, T. C.: Moving beyond the cost–loss ratio: economic assessment of streamflow forecasts for a risk-averse decision maker, Hydrol. Earth Syst. Sci., 21, 2967-2986, 2017.
Maurer, E. P. and Lettenmaier, D. P.: Predictability of seasonal runoff in the Mississippi River basin, Journal of Geophysical Research-Atmospheres, 108, 2003.
McInerney, D., Thyer, M., Kavetski, D., Lerat, J., and Kuczera, G.: Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors, Water Resources Research, 53, 2199-2239, 2017.
Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., Stouffer, R. J., Dettinger, M. D., and Krysanova, V.: On Critiques of "Stationarity is Dead: Whither Water Management?", Water Resources Research, 51, 7785-7789, 2015.
Mount, N. J., Maier, H. R., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F. J., and Abrahart, R. J.: Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei Science Plan, Hydrological Sciences Journal, 61, 1192-1208, 2016.
Pagano, T., Garen, D., and Sorooshian, S.: Evaluation of official western US seasonal water supply outlooks, 1922-2002, J. Hydrometeorol., 5, 896-909, 2004.
Pappenberger, F. and Beven, K. J.: Ignorance is bliss: Or seven reasons not to use uncertainty analysis, Water Resources Research, 42, 2006.
Perrin, C., Michel, C., and Andreassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275-289, doi:210.1016/S0022-1694(1003)00225-00227, 2003.
Plaza Guingla, D.Randrianasolo, A., De Keyser, R., De LannoyThirel, G. J., Ramos, M., Giustarini, L., Matgen, P., H., and Pauwels, V.: Improving particle filters in rainfall-runoff models: applicationMartin, E.: Impact of streamflow data assimilation and length of the resample-move step and verification period on the quality of short-term ensemble Gaussian particle filter, Water Resourceshydrologic forecasts, Journal of Hydrology, 519, 2676-2691, 2014.
Refsgaard, J. C.: Validation and Intercomparison of Different Updating Procedures for Real-Time Forecasting, Hydrology Research, 49, 4005-4021, 201328, 65-84, 1997.
Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resources Research, 46, 2010.
Robertson, D. E., Pokhrel, P., and Wang, Q. J.: Improving statistical forecasts of seasonal streamflows using hydrological model output, Hydrol. Earth Syst. Sci., 17, 579-593, 2013.
Robertson, D. E. and Wang, Q. J.: Seasonal Forecasts of Unregulated Inflows into the Murray River, Australia, Water Resour. Manag., 27, 2747-2769, 2013.
Schepen, A., Zhao, T., Wang, Q. J., and Robertson, D. E.: A new method for post-processing daily sub-seasonal to seasonal rainfall forecasts from GCMs and evaluation for 12 Australian catchments, Hydrol. Earth Syst. Sci. Discuss., 2017, 1-27, 2017.
Searcy, J. K., Hardison, C. H., and Langein, W. B.: Double-mass curves; with a section fitting curves to cyclic data, Report 1541B, 1960.
Shao, Q. and Li, M.: An improved statistical analogue downscaling procedure for seasonal precipitation forecast, Stochastic Environmental Research and Risk Assessment, 27, 819-830, 2013.

Spaaks, J. H. and Bouten, W.: Resolving structural errors in a spatially distributed hydrologic model using ensemble Kalman filter state updates, Hydrology and Earth System Sciences, 17, 3455-3472, 2013.
Sun, L., Seidou, O., and Nistor, I.: Data Assimilation for Streamflow Forecasting: State-Parameter Assimilation versus Output Assimilation, Journal of Hydrologic Engineering, 22, 04016060, 2017.
Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., and Srikanthan, S.: Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, Water Resources Research, 45, 2009.
Tuteja, N. K., Shin, D., Laugesen, R., Khan, U., Shao, Q., Li, M., Zheng, H., Kuczera, G., Kavetski, D., Evin, G., Thyer, M. A., MacDonald, A., Chia, T., and Le, B.: Experimental evaluation of the dynamic seasonal streamflow forecasting approach, Bureau of Meteorology, Australia, 2011.
Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J. M., Viney, N. R., and Teng, J.: Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies, Journal of Hydrology, 394, 447-457, 2010.
Vrugt, J. A., Diks, C. G. H., Gupta, H. V., Bouten, W., and Verstraten, J. M.: Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, Water Resources Research, 41, 2005.
Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., Hyman, J. M., and Higdon, D.: Accelerating Markov Chain Monte Carlo Simulation by Differential Evolution with Self-Adaptive Randomized Subspace Sampling, International Journal of Nonlinear Sciences and Numerical Simulation, 10, 273-290, 2009.
Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., and Gupta, H. V.: Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis, Hydrological Processes, 17, 455-476, 2003.
Wang, E., Zheng, H., Chiew, F., Shao, Q., Luo, J., and Wang, Q. J.: Monthly and seasonal streamflow forecasts using rainfall-runoff modeling and POAMA predictions, 19th International Congress on Modelling and Simulation (Modsim2011), 2011. 3441-3447, 2011.
Wang, Q. J., Robertson, D. E., and Chiew, F. H. S.: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites, Water Resources Research, 45, 2009.
Welsh, W. D., Vaze, J., Dutta, D., Rassam, D., Rahman, J. M., Jolly, I. D., Wallbrink, P., Podger, G. M., Bethune, M., Hardy, M. J., Teng, J., and Lerat, J.: An integrated modelling framework for regulated river systems, Environmental Modelling & Software, 39, 81-102, 2013.
Wani, O., Beckers, J. V. L., Weerts, A. H., and Solomatine, D. P.: Residual uncertainty estimation using instance-based learning with applications to hydrologic forecasting, Hydrol. Earth Syst. Sci., 21, 4021-4036, 2017.
Westra, S., Thyer, M., Leonard, M., Kavetski, D., and Lambert, M.: A strategy for diagnosing and interpreting hydrological model nonstationarity, Water Resour. Res., 50, 5090-5113,Resources Research, doi: 10.1002/2013WR014719, 2014. 2014.
Wöhling, T., Lennartz, F., and Zappa, M.: Technical Note: Real-time updating procedure for flood forecasting with conceptual HBV-type models, Hydrol. Earth Syst. Sci. Discuss., 3, 925-940, 2006.
Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, Geophys. Res. Lett., 35, 5, 2008.
Wood, A. W. and Schaake, J. C.: Correcting errors in streamflow forecast ensemble mean and spread, J. Hydrometeorol., 9, 132-148, 2008.
Wright, D. P., Thyer, M., and Westra, S.: Influential point detection diagnostics in the context of hydrological model calibration, Journal of Hydrology, 527, 1161-1172, 2015.
Wu, W., May, R. J., Maier, H. R., and Dandy, G. C.: A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks, Water Resources Research, 49, 7598-7614, 2013.
XieYang, T., Asanjan, A. A., Welles, E., Gao, X. H., Sorooshian, S., and Zhang, D.Liu, X.: A partitioned update scheme for state-parameter estimation of distributed hydrologic models based on the ensemble Kalman filterDeveloping reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information, Water Resources Research, 49, 7350-7365, 201353, 2786-2812, 2017.
Ye, W., Bates, B. C., Viney, N. R., Sivapalan, M., and Jakeman, A. J.: Performance of conceptual rainfall-runoff models in low-yielding ephemeral catchments, Water Resources Research, 33, 153-166, 1997.
Yihdego, Y. and Webb, J.: An Empirical Water Budget Model As a Tool to Identify the Impact of Land-use Change in Stream Flow in Southeastern Australia, Water Resour. Manag., 27, 4941-4958, 2013.
Zhang, H., Huang, G. H., Wang, D., and Zhang, X.: Multi-period calibration of a semi-distributed hydrological model based on hydroclimatic clustering, Advances in Water Resources, 34, 1292-1303, 2011.