

---

Dear Dr Ramos,

Please find point-by-point responses to the referees' comments below. We have also appended a version of the manuscript with all changes tracked. Note that the page/line references we list in the responses refer to the revised manuscript without tracked changes. We hope that our responses and changes satisfy your request for minor revisions.

Kind regards,

James Bennett

(On behalf of the other authors)

### **Response to interactive comment by anonymous referee #1**

#### **General comment**

*I found the paper very well written and it is well-arranged. In my opinion, the manuscript fits pretty well into this special issue and its content is relevant for publication in HESS. The extensive experiment of the continent-wide forecasting system is described in depth regarding the underlying methodology (including the statistical and rainfall-runoff models), the verification procedure as well as the numerous results. This work is a consistent step to continue previous work reported e.g. by Bennett et al. (2016) and Turner et al. (2017).*

**Response:** Thanks very much for the careful review and the positive feedback. We've included a list point-by-point responses below.

#### **Major comment**

*I suggest trying to condense section 2.2 (especially its subsections) in order to strengthen the role of the Bayesian prior, which is assessed in experiment 3 and which seems to be most promising / sensitive to improve FoGSS output. The error model approach is very interesting, but as a reader, I slightly lost the focus on the relevant aspect (the prior) for your study.*

**Response:** We have shortened this section, as suggested (Page 3, line 26 – Page 4, line 18). We have reorganised it to emphasise the prior, by separating the hydrological model (Section 2.2) from the error model (Section 2.3), and moving the description of parameter estimation after the description of the error model (Section 2.3.4). This means the reader arrives much more directly at the description of the bias-correction (Section 2.3.2), and we also more directly flag its use in the experiments with the prior: "This property is exploited in our experiments with the use of a prior, described in Section 4.4" (Page 4, Line 7).

#### **Minor / Technical comments**

*Page 2, line 18: I suggest adding a short explanation, why BJP is not (or even may be cannot be) suitable for those long-range forecasts in order to assist readers, who aren't familiar with the BJP approach.*

---

**Response:** We now note that the BJP has no mechanism for generating hydrographs, which are required for streamflow forecasts for long time series: “But the BJP is not well suited to generating time series forecasts to long (12-month) time horizons, because it has no mechanism for simulating the shape of hydrographs over long lead-times.” (Page 2, lines 16-18).

5

*Page 2, line 25: It might be beyond the scope of this paper, but did you experience that the preference of some water agencies to use stochastic scenarios (instead of seasonal forecasts) might be based on the fact that they still have to gain confidence to this “new” source of information (they are used to the scenarios, they comprehend it, ..)? So the “practical” advantages you mentioned might also contain such more psychological aspects instead of purely technical ones?*

10

**Response:** This is a thought-provoking question, and one we can only speculate on as the FoGSS forecasting system is not yet operational. (At present, stochastic scenarios or climatology are really the only option for many water agencies looking for 12-month panning scenarios.) We based our comments on the use of stochastic scenarios partly on responses of water agencies to a formal (but unpublished) survey conducted by the Bureau of Meteorology, and partly on our own (more anecdotal) interactions with water agencies. Neither source of information gives clear indication of possible barriers to adopting a new forecasting service. In some cases water agencies show a great deal of faith in the existing 3-month ensemble forecasts (generated with the BJP), and they simply wanted longer range forecasts: in these cases we feel that they would probably be quite happy to use long-range monthly forecasts if they were available (notwithstanding computational barriers to using large ensembles, which is another discussion). In other cases, however, we agree that the resistance to using forecasts is not purely technical, and it may be due to ‘institutional inertia’ (which is possibly the result of psychological preferences of key staff). As the reviewer surmises, however, we feel that this discussion is outside the scope of the paper. It is probably worthy of a more detailed discussion in its own right in a paper directed at barriers to adoption of new forecasting services.

15

20

25

*Page 3, line 8: I recommend mentioning how the three variants of the POAMA model are generated (variation of model parameter)?*

30

**Response:** We note that these variants are generated by changing model physics:” POAMA reforecasts are available as a 33-member ensemble; comprised of 11 members each from three variants of the model; each variant has slightly different model physics.” (Page 3, lines 5-6).

*Page 3, line 29: I suggest to split section 2.2 and to add a separate chapter “Hydrological model” (or something similar). Even this chapter might be relatively short, I suggest to have a separate section for each of the main FoGSS components, which correspond to the three experiments described in section 3 of this paper.*

35

**Response:** We have followed this suggestion, as noted above (see new Section 2.2).

*Page 3, line 35: Have you thought about / tested using data assimilation techniques to reduce hydrological model errors?*

40

**Response:** To date, we have focussed our research on improving the hydrological error model. This has a similar effect to state updating, but we feel it is a more approach direct (streamflow errors are directly measurable, while

---

states are not). Importantly, this allows us to correct reliability more directly. Additional updating of states may possibly improve our forecasts. However, existing methods of state updating would be difficult to implement alongside our error model (at least, as it is currently formulated). However, we may consider state updating in future (if we get the time!).

5

*Page 5, line 8: As you state that the upper limit of  $d$  is arbitrary, it would be interesting to know, if you have tested other thresholds before (and you ended up with 2)?*

**Response:** We also considered upper limits of 1 and 1.5. We found these too restrictive in some catchments (some of the biases are very strongly non-linear), but we felt that values greater than 2 could cause unrealistically large bias corrections under cross-validation (indeed, our use of the prior showed this to be true even for the upper limit of 2). The use of the prior makes the choice of the upper limit less important.

10

*Page 11, line 1: I guess that “perennial” and “ephemeral” have to be switched?*

**Response:** These are correct as they are, but we can see that this is confusingly phrased. We have rephrased to avoid the confusion: “Instances of strongly negative skill (<-15%) are rare in perennial catchments, and also absent in a substantial number of ephemeral catchments.” (Page 8, line 37-38). Note also that this discussion has been moved from section 5.2 to 5.1, to accord with the creation of the new Figure 3 (as suggested by referee #2).

15

*Page 12, line 4-6: Could you please explain, why you are planning to improve Wapaba instead of using the GR2M model, as the latter one seems to perform better, especially in drier catchments? I think you give some kind of explanation on page 13 (line 13-18), but I suggest to add a link or to explain your decision to the reader already at the end of section 4.3.*

20

**Response:** We agree that this is not a strong justification for future research, and we have removed this statement.

25

*Appendix A: As the state is a relevant information in your list, I suggest to explain the acronyms used, as several reader might not be familiar with the different Australian states.*

**Response:** We have added explanations of these acronyms.

*Figure 1: Why does the arrow linking “rainfall-runoff model” and “Climatology PE” point in the direction of the climatology? Isn’t climatology potential evaporation an input to the rainfall-runoff model?*

30

**Response:** Thanks for picking this up – we have changed the direction of the arrow.

*Figure 3 + 13: I suggest to explicitly mentioning the CRPSS as “skill measure”. Figure 5, 6, 7, 8, 9, 10, 11, and 12: I suggest adding the total number of catchments in each panel (e.g. in brackets behind the title).*

35

**Response:** We have added the information, as suggested.

### Typographical corrections

*Page 4, line 5: Shouldn’t it be “homoscedastic” in this context?*

*Page 4, line 24: I suggest to insert a comma before “: : : a and b are parameters.”*

*Page 5, line 1: I think “takes” should be deleted.*

---

*Page 5, line 3: The word-wrap slipped (comma at the beginning of line 4 should be in line 3).*

**Response:** Thanks for reading our manuscript to closely – these are errors that we have corrected.

---

## Response to interactive comment by anonymous referee #2

### General comments:

Overall I really enjoyed reading this paper as it is very well written and guides the reader nicely from the introduction to the conclusions. The paper fits very well within this HESS special issue, as it presents a newly developed ensemble seasonal streamflow forecasting system for Australia as an alternative to stochastic scenarios for decision-makers. Furthermore, the paper contains a rich number of relevant experiments aimed at improving ensemble seasonal streamflow forecasting, especially in very dry catchments, with some clear conclusions as to the benefits and limitations of each methods. Additionally, this paper refers to and builds nicely on relevant and previous work in this field.

5  
10 **Response:** Thanks very much for the careful review and the positive feedback. We've included a list point-by-point responses below. Note that the page/line references we list in the responses refer to the revised manuscript.

### Specific comments and technical corrections:

-P2, L19-22: Where available, other references for these statements would be good.

15 **Response:** We have added references to support the statements from Yuan et al. (2013), Fundel et al. (2013) and Wood and Schaake (2008), which give examples of negative skill, bias and poor reliability, respectively:

“Other seasonal forecasting systems generally have some combination of short-comings with respect to stochastic scenarios: they may not produce reliable ensembles (e.g., Crochemore et al., 2016; Wood and Schaake, 2008); the ensembles may be biased with respect to climatology (e.g., Fundel et al., 2013; Wood and Schaake, 2008); and/or the forecasts may be less skillful than climatology for certain months or lead times (Yuan et al., 2013).” (Page 2, lines 18-21).

20

-P3, L30: It would be good to also mention the other forcing variables of the rainfall-runoff model here, i.e., climatology PE. Is temperature not a forcing of the model?

25 **Response:** We have mentioned PE as well as rainfall: “Rainfall forecasts and climatology potential evaporation are used to force an initialised monthly rainfall-runoff model.” (Page 3 line 23). Wapaba does not require temperature for forcing: the catchments assessed have negligible influence from snow (in most cases, none) – as is true for the vast majority of Australia.

30 -2.2 Hydrological error model: I agree with reviewer 1 in that the paper could benefit from condensing this section. This would keep the readers more focused on the three experiments nicely described in section 3 and of central importance to the paper's results and conclusions.

35 **Response:** We have shortened this section, as suggested (Page 3, line 26 – Page 4, line 18). We have reorganised it to emphasise the prior, by separating the hydrological model (Section 2.2) from the error model (Section 2.3), and moving the description of parameter estimation after the description of the error model (Section 2.3.4). This means the reader arrives much more directly at the description of the bias-correction (Section 2.3.2), and we also more directly flag its use in the experiments with the prior: “This property is exploited in our experiments with the use of a prior, described in Section 4.4” (Page 4, Line 7).

-P4, L5: Maybe briefly explain what “heteroscedastic” means as not every reader might be familiar with it.

---

**Response:** For clarity, we have rephrased this to use its antonym (homoscedastic) and added a brief note explaining that this means the variance is constant: “The data transformation (Stage 1) allows us to assume that residuals,  $\varepsilon$ , are normally distributed and homoscedastic (i.e., the variance does not change).” (Page 4, lines 22-23)

5 -P4, Equation 1: Please mention here what TF stands for.

**Response:** We have explicitly noted that this term denotes the transformation: “The log-sinh transformation ( $TF$ ) is given by:

$$z = TF(q) = \frac{1}{b} \log(\sinh(a + bq)) \quad (1)”$$

(Page 3, line 32-34)

10

-P4, Equation 2: This equation does not seem vital to mention here so I suggest to remove it.

**Response:** Thanks for this suggestion – we will remove it.

-P5, L19: How is the error in the original domain at  $t-1$  calculated?

15 **Response:** The error at  $t-1$  is given by  $q_o(t-1) - q_2(t-1)$ , where  $q_o$  is observed streamflow and  $q_2$  is the back-transformed value of  $z_2$ . We have added this explanation: “This restriction corrects the forecast by whichever is smaller: the correction proposed by Eq. 3, or the error in the original domain at  $t-1$  given by  $q_o(t-1) - q_2(t-1)$  (where  $q_o$  is observed streamflow and  $q_2$  is the back-transformed value of  $z_2$ ).” (Page 4, Lines 16-18)

20

-3.1.2 Verification scores: Please mention the range of all the scores later displayed (e.g., a CRPSS of 100% corresponds to a perfect forecast).

**Response:** We have added this, as follows: “CRPSS ranges from  $-\infty$  (least skillful) to 100% (perfectly skillful).” (Page 5, Line 35)  
25 “Absolute bias ranges from 0 (unbiased) to  $+\infty$  (worst bias).” (Page 6, Line 7)  
“ $\pi_t$  takes values from zero to 1.” (Page 6, Line 12)  
“The alpha index essentially reflects the divergence of PIT values from the 1-1 line in PIT plots, ranging from 1 (perfectly reliable) to 0 (worst reliability).” (Page 6, Lines 19-20)

30 -P7, L21: Could you please state briefly which interpolation method was used here.

**Response:** The method they used is called ‘Barnes successive correction analysis’. We have noted this in text: “Rainfall and potential evaporation data are taken from the gridded AWAP dataset, which interpolates gauged observations with a Barnes successive correction analysis” (Page 6, Lines 23-25)

35 -4.1 Continent-wide performance of the base FoGSS model: I like the focus on the six case study catchments as it allows looking at the results and their differences into more details. However, and since this section is called “continent-wide performance” I think that it could be very beneficial to this section to quickly describe the overall performance of the forecasts for all 63 catchments prior to looking at the six individual case studies. This could

---

be done simply by adding a boxplot of the FoGSS CRPSS for all lead times and target months combined on Figure 3.

**Response:** Thanks for this suggestion. We have added a figure as the reviewer suggests (Figure 3). While this duplicates information in the following figure (now Figure 4), we agree with the reviewer that this makes the paper easier to follow.

-P10, L10-12: This criteria for FoGSS to be characterised as performing well should be stated before describing any results.

**Response:** We have moved this criterion earlier in the paper when introducing CRPSS: “FoGSS forecasts need not necessarily outperform climatology to function as a viable alternative to stochastic scenarios, but they do need to be at least similarly skillful to climatology. We term forecasts with skills near zero *neutrally skillful*, defined as  $-5\% > CRPSS < 5\%$ .” (Page 5, Line 35 – Page 6, Line 2)

-P10, L12-14: Is the negative skill in the Herbert catchment due to the large catchment memory then?

**Response:** In essence, yes, though this statement really only applies to the receding limb of the annual hydrograph. Hence our explanation is more specific: negative skills occur “because slight mispredictions of flow issued in wetter months (e.g. February) can result in proportionally larger errors in drier months at longer lead times.” (Page 9, Lines 10-11)

-P10, L14-15: You mention the positive or neutral skill for the Fitzroy catchment, although there are light orange colours (i.e. slightly negative skill) in the plot for this catchment. Could you rephrase this or define “neutral skill”.

**Response:** Thanks for identifying this ambiguity. We have defined neutral skill as  $-5\% > CRPSS < 5\%$  when we introduce

CRPSS: “We term forecasts with skills near zero *neutrally skillful*, defined as  $-5\% > CRPSS < 5\%$ .” (Page 6, Line 2)”

As the reviewer points out, this means our statement is no longer strictly true (the light orange colours). We have amended our statement to note these instances of slightly negative skill: “FoGSS also performs well in the ephemeral Fitzroy catchment, returning largely positive or neutral skill, with only a few isolated instances of slightly negative skill.” (Page 9 Lines 11-13)

-P10, L22-23: It is interesting that forecasts are also not reliable for September in the Eppalock catchment. Why is that?

**Response:** Thanks very much for reading the manuscript so closely. That figure was incorrect (the error was introduced just prior to submission, which is why the text does not agree with the content of the figure). In fact, the Eppalock forecasts for September are reliable, as implied in the text. We have corrected the figure in the revised manuscript (now Figure 5). The other panels differ negligibly from the figure incorrectly included in the original manuscript, and so are consistent with the other text and conclusions.

-P10, L39: Please state for which catchments forecasts are generally neutrally skillful by lead-6, i.e. is it for all catchments?

---

**Response:** Yes, this is true for all catchments – we have noted this: ” . By lead-6, forecasts are generally neutrally skilful for all catchments.” (Page 8, Lines 37-38 – NB, this has been moved to Section 5.1, to accommodate the creation of the new Figure 3)

5 -P11, L1: *I think that perennial and ephemeral should be swapped here.*

**Response:** These are correct as they are, but we can see that this is confusingly phrased. We have rephrased to avoid the confusion: “Instances of strongly negative skill (<-15%) are rare in perennial catchments, and also absent in a substantial number of ephemeral catchments.” (Page 8, line 36-37). Note also that this discussion has been moved from section 5.2 to 5.1, to accord with the creation of the new Figure 3.

10

-P11, L17: *I am not sure what is meant by “irrespective of forcing” here. Please explain further or rephrase.*

**Response:** We mean that it does not matter which forcing – ESP-type inputs or POAMA-CBaM – we use, FoGSS forecasts of accumulated volumes can be skillful to long accumulation periods in perennial catchments. We have rephrased this to be clearer: “We note that FoGSS forecasts for perennial catchments generally exhibit positive skill for accumulation periods up to 6 months, whichever rainfall forcing is used. This is clear evidence that FoGSS forecasts hold more useful information than stochastic scenarios.” (Page 9 Line 41 – Page 10 Line 2).

15

-P12, L5: *It is not obvious why a Budyko-based structure would remain attractive. Could you please argue this slightly for the reader to understand your plan to improve Wapaba instead of using GR2M despite its obvious benefits over the latter.*

20

**Response:** We agree that this is not a strong justification for future research, and we have removed this statement.

-P12, L15: *Could you please mention that the smaller the  $d$  values, the stronger the prior (if this is indeed the case), as it was not obvious to me at first.*

25

**Response:** We have note this here: “The stronger the prior (i.e., the smaller the value of  $\sigma_d$ ), the greater the removal of negative skills, with the effect of the prior becoming negligible for  $\sigma_d \geq 2.0$ .” (Page 10 Lines 34-35), and also added this explanation to the figure captions (Figure 11: Page 28 Line 3; Figure 13: Page 30 Line 2).

-P12, L24: *Please explain what is meant by “sensibly” here or choose another adjective, i.e. skilfully, reliably, etc.*

30

**Response:** We feel that ‘sensibly’ conveys our meaning effectively, and we would prefer to keep it. We clarify its meaning as follows: “...strongly negative skills generally only occur in very dry months, where there may be only a few non-zero observations on which to optimise the hydrological and error models. In these cases, it is sensible to encourage FoGSS to return a climatology-like forecast” (Page 11 Lines 3-5).

35

-P12, L31: *This is questionable for perennial catchments for some experiments.*

**Response:** We agree. We have added the qualification “although these changes were sometimes very slight” (Page 11, Lines 9-10).

40

-P13, L5-8: *Wouldn't we expect drier months to be better? This needs explaining if so.*

---

**Response:** Yes, we do expect this, exactly as the reviewer points out. In most cases drier months are improved. July, August and December in the Ranken catchment are examples where the prior on  $d$  did not work well. To illustrate how this happens, we'll focus on July. Flow in July is zero (24 observations) or close to zero (two observations of  $<0.02$  mm) in our evaluation period, except in one year where flows are dramatically larger ( $>8$  mm). Shrinking  $d$  means the model is much less able to cover this large event - to compensate, the variance of the error becomes large. This leads to persistent overpredictions in years with very small observed flows. In short, allowing a larger value of  $d$  allowed the bias-correction to handle this strongly non-linear case better (indeed, a value of  $d > 2$  would have worked even better). These cases are very challenging, and, in the context of this study, unusual: in most instances, the prior improved (or did not greatly impact) forecast skill in drier months. We have added a brief explanation of this problem as the reviewer suggests, to better acknowledge the difficulties we face in these catchments:

“In the very dry Ranken catchment, negative skill in wetter months (Jan-Apr) is largely removed, in favour of climatology-like forecasts. Conversely, skill in Jul, Aug and Dec has changed from neutral/positive in the base case to be substantially negative. All three of these months are very dry in the Ranken catchment (most flows are zero), but feature a single very large event in the record ( $>10x$  larger than the next largest measured flow). In these cases, the prior on  $d$  has a deleterious effect: allowing  $d$  to take a larger range of values better corrects the extremely non-linear biases in these months. When the prior is applied, the bias-correction is not as effective, forcing the error model to take larger values of  $\sigma^2$ . This leads to persistent overestimations of streamflow in the other (very dry) years, leading to negative skill.” (Page 11 Lines 21-28)

-P13, L10: *In the Fitzroy catchment the skill is however diminished for longer lead times for forecasts for JAS.*

**Response:** Yes, this is true. For longer lead-time forecasts in JAS, it would have been better to ‘trust’ the model more, as it offers some information. The prior on  $d$  does not result in universal improvements, but we believe the amelioration of strongly negative skills (e.g. in the Eppalock catchment) outweighs slight reductions in positive skill in some cases, such as this one. We have added a note on the reduction in skill in the Fitzroy in JAS: “There is little change to skill in the Fitzroy catchment (slight reductions in skill in Jul-Sep at long lead-times)” (Page 11 Lines 30-31).

-P14, L21-22: *I strongly agree with your belief in the inclusion of seasonal rainfall forecasts in FoGSS. You can however here make this argument stronger as you showed in the paper that the skill from climate forecasts can accumulate to produce skilful long-range total inflow forecasts (mentioned on P12, L36-37). These forecasts being valuable information for reservoir operations in Australia.*

**Response:** We have strengthen this argument as suggested: “However, this comes at the cost of including useful information in rainfall forecasts, information that we show can accumulate over multiple lead times.” (Page 12 Lines 31-33).

-Figure 1: *State which rainfall-runoff model is used in the FoGSS system.*

**Response:** We have added ‘Wapaba’ to the figure (Figure 1, Page 18)

-Figure 2: *Very nice plots! Adding rainfall on these plots could be a nice and useful addition.*

---

**Response:** Thanks - we have added a small panel above each plot showing rainfall statistics (Figure 2, Page 19)

5 -Figures 3 and 13: In the results you mention that FoGSS performs adequately when  $CRPSS \geq 0$ . Considering this, wouldn't it make sense to modify the colour bar and split the current +5 to -5 range in two sections: +5 to 0 and 0 to -5?

**Response:** Skill scores are somewhat noisy, and will sometime dip slightly below zero by random chance. We don't think it's reasonable to penalise forecasts for being within 5% of zero – essentially, we believe this to be 'neutrally skillful'. As noted in a previous response, we have added a formal definition of what we mean by 'neutral skill' (basically, within 5% of zero) when we introduce CRPSS, and noted that we consider performance neutral or positively skillful to be a requirement of FoGSS: "FoGSS forecasts need not necessarily outperform climatology to function as a viable alternative to stochastic scenarios, but they do need to be at least similarly skillful to climatology. We term forecasts with skills near zero *neutrally skillful*, defined as  $-5\% > CRPSS < 5\%$ ." (Page 5, Line 35 – Page 6, Line 2)

15 -Figures 3, 5, 6, 8, 10, 13: Change CRPS skill scores to CRPSS in the captions and on the plots.

**Response:** Thanks for picking up this inconsistency – we have corrected this as suggested.

-Figures 3, 4, 13: It would be good to be reminded on the plots or in the captions which of these case study catchments are perennial vs ephemeral.

20 **Response:** We have indicated which catchments are ephemeral in all these figures.

-Figures 5 to 12: It is sometimes hard to see the difference between two boxplots that you are comparing in the results. Adding notches on the boxplots could make it easier to see for the reader.

**Response:** We have added notches.

25

-Figures 7 and 11: Both plots for perennial catchments are hard to see, I would recommend rescaling the y-axes.

**Response:** We have rescaled the middle panels on these plots to show more detail.

30 -Figures 10 and 11: Are the numbers 4 to 0.25 the  $d$  values? This is slightly confusing and might be worth changing in the legend.

**Response:** These are the prior values ( $\sigma_d$ ) – we have now indicated this explicitly in the legend, and added an explanatory note that smaller values of  $\sigma_d$  indicate a stronger prior (Figure 11, page 28; Figure 13 Page 30).

### Typographical corrections

35 -P1, L23: "catchments that experience" instead of "catchments that in experience".

-P3, L18: Remove "and" after "rainfall forecasts".

-P4, L27: "transformed domain" instead of "transform domain". Same on P5, L16.

-P5, L1: "takes" should be deleted.

-P5, L15: It should probably be "zo is the transformed observed streamflow"; "the" is missing.

---

-P7, L17: I think that “Catchments and data” should be section 3.1.3 and not 3.1.1.

-P8, L12: “alpine” should have a capital “A”.

**Response:** Thanks for reading our manuscript to closely – these are errors that we have corrected.

## 5 References

Crochemore, L., Ramos, M. H., and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, *Hydrology and Earth System Sciences*, 20, 3601-3618, doi: 10.5194/hess-2016-78, 2016.

10 Fundel, F., Jörg-Hess, S., and Zappa, M.: Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices, *Hydrol. Earth Syst. Sci.*, 17, 395-407, doi: 10.5194/hess-17-395-2013, 2013.

15 Wood, A. W., and Schaake, J. C.: Correcting Errors in Streamflow Forecast Ensemble Mean and Spread, *Journal of Hydrometeorology*, 9, 132-148, doi: 10.1175/2007jhm862.1, 2008.

Yuan, X., Wood, E. F., Chaney, N. W., Sheffield, J., Kam, J., Liang, M., and Guan, K.: Probabilistic Seasonal Forecasting of African Drought by Dynamical Models, *Journal of Hydrometeorology*, 14, 1706-1720, doi: 10.1175/jhm-d-13-054.1, 2013.

20

---

# Assessment of an ensemble seasonal streamflow forecasting system for Australia

James C. Bennett<sup>1,2</sup>, Quan J. Wang<sup>3</sup>, David E. Robertson<sup>1</sup>, Andrew Schepen<sup>4</sup>, Ming Li<sup>5</sup>, Kelvin Michael<sup>2</sup>

5 <sup>1</sup>CSIRO Land & Water, Clayton, Victoria, Australia

<sup>2</sup>Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania, Australia

<sup>3</sup>Department of Infrastructure Engineering, The University of Melbourne, Parkville, Victoria, Australia

<sup>4</sup>CSIRO Land & Water, Dutton Park, Queensland, Australia

<sup>5</sup>CSIRO Data61, Floreat, Western Australia, Australia

10 *Correspondence to:* James C. Bennett (james.bennett@csiro.au)

**Abstract.** Despite an increasing availability of skillful long-range streamflow forecasts, many water agencies still rely on simple resampled historical inflow sequences (stochastic scenarios) to plan operations over the coming year. We assess a recently developed forecasting system called *forecast guided stochastic scenarios* (FoGSS) as a skillful alternative to standard stochastic scenarios for the Australian continent. FoGSS uses climate forecasts from a coupled ocean-land-atmosphere prediction system, post-processed with the method of calibration, bridging and merging. Ensemble rainfall forecasts force a monthly rainfall-runoff model, while a staged hydrological error model quantifies and propagates hydrological forecast uncertainty through forecast lead times. FoGSS is able to generate ensemble streamflow forecasts in the form of monthly time series to a 12-month forecast horizon. FoGSS is tested on 63 Australian catchments that cover a wide range of climates, including 21 ephemeral rivers. In all perennial and many ephemeral catchments, FoGSS provides an effective alternative to resampled historical inflow sequences. FoGSS generally produces skillful forecasts at shorter lead times (<4 months), and transits to climatology-like forecasts at longer lead times. Forecasts are generally reliable and unbiased. However, FoGSS does not perform well in very dry catchments (catchments that experience zero flows more than half the time in some months), sometimes producing strongly negative forecast skill and poor reliability. We attempt to improve forecasts through the use of i) ESP rainfall forcings, ii) different rainfall-runoff -models, and iii) a Bayesian prior to encourage the error model to return climatology forecasts in months when the rainfall-runoff model performs poorly. Of these, **only** the use of the prior offered **the clearest** benefits in very dry catchments, where it moderated strongly negative forecast skill and reduced bias in some instances. However, the prior did not remedy poor reliability in very dry catchments. Overall, FoGSS is an attractive alternative to historical inflow sequences in all but the driest catchments. We discuss ways in which forecast reliability in very dry catchments could be improved in future work.

**Keywords.** Seasonal streamflow forecasting; ensemble prediction; CGCM; hydrological uncertainty; error modelling

---

## 1 Introduction

Recent years have seen a proliferation of experimental long-range ensemble streamflow forecasting systems (examples from this issue: Meißner et al., 2017; Beckers et al., 2016; Candogan Yossef et al., 2016; Bell et al., 2017; Greuell et al., 2016), and, to a lesser extent, the operationalization of these systems as forecasting services that are available to water agencies and the public. In Australia, the Bureau of Meteorology runs a freely available seasonal streamflow forecasting service that predicts total streamflow for the coming 3 months at more than 200 sites across Australia ([www.bom.gov.au/water/ssf/](http://www.bom.gov.au/water/ssf/)). While the Bureau's service has been well received by Australian water agencies, a number of agencies still rely primarily on resampled historical inflow sequences, not forecasts, to plan operations for the coming year. Resampled historical inflow sequences (termed stochastic scenarios in this paper) have some appeal for water agencies: they are unbiased, they are available as time series, they are easy to generate to long time horizons, and, presuming a long observation record is available from which to sample, the ensemble of inflows is inherently statistically reliable (either taken at individual months or when individual ensemble members are summed, e.g., to produce an ensemble of 6-month total inflow). The Bureau's service is based on a statistical method, the Bayesian joint probability (BJP) modelling approach (Wang and Robertson, 2011), which uses information from current streamflow conditions and climate indices to produce skillful streamflow forecasts. The BJP is able to produce skillful, unbiased forecasts with highly reliable ensembles, and can be used to generate monthly volume forecasts to short (e.g. 3-month) forecast horizons (Zhao et al., 2016). But the BJP is not well suited to generating time series forecasts to long (e.g. 12-month) time horizons, because it has no mechanism for simulating the shape of hydrographs over long lead times. Other seasonal forecasting systems generally have some combination of short-comings with respect to stochastic scenarios: they may not produce reliable ensembles (e.g., Crochemore et al., 2016; Wood and Schaake, 2008); the ensembles may be biased with respect to climatology (e.g., Fundel et al., 2013; Wood and Schaake, 2008); and/or the forecasts may be less skillful than climatology for certain months or lead times (Yuan et al., 2013). Any of these can be a serious barrier to their use by water agencies to plan ~~future~~ operations.

Of course, stochastic scenarios have a major short-coming of their own: they take no account of information from current catchment and climate conditions, and thus offer no skill to water agencies. To ~~attempt to~~ combine the practical advantages of stochastic scenarios with useful information contained in forecasts, we recently proposed a new streamflow forecasting system called *forecast guided stochastic scenarios*, or FoGSS. FoGSS uses ~~statistically~~ post-processed climate forecasts from a coupled climate forecasting system to force a monthly rainfall-runoff model, ~~together with a~~ hydrological error model ~~that is then used to~~ update forecasts, corrects biases, and propagates forecast uncertainty through the lead times. FoGSS ~~is designed to offer~~ produces time series forecasts to long time horizons (12 months). As forecast skill declines with lead time, FoGSS is designed to return forecasts that converge to climatology. Each ensemble member in the forecast is a realistic 12-month hydrograph at a monthly time step. In a previous paper (Bennett et al., 2016), we described the theoretical underpinnings of FoGSS and showed that it performed well for two high-rainfall Australian catchments, producing skillful and reliable ensemble forecasts. We noted that the viability of FOGSS as a continent-wide forecasting system remained to be tested. In particular, FOGSS needs to be tested for ephemeral rivers, which are an important source of water (e.g. for agriculture) in many Australian regions. (Note that catchments that cease to flow are variously termed 'intermittent' and 'ephemeral' in other studies; we will refer to all rivers that cease to flow >4% of the time as 'ephemeral' in this paper.)

---

The aim of this paper is to test FoGSS on a wide range of Australian catchments that encompass different climatic and hydrologic conditions. We then vary components of the system – rainfall forcings, rainfall-runoff modelling, and the hydrological error model – to assess to what extent, if any, forecasts can be improved. The paper is structured as follows. We give an overview of the FoGSS model in Section 2, and describe our [experiments-setup and data in Section 3](#), and [describe our experiments](#) to vary elements of FoGSS in Section [3.4](#). We present and discuss our results in Section [4.5](#), and we summarise and conclude our findings in Section [5.6](#).

## 2 The FoGSS Model

A schematic of the FoGSS model is shown in Figure 1.

### 2.1 Ensemble rainfall forecasts

10 Rainfall and sea-surface temperature (SST) predictions are taken from the POAMA M2.4 seasonal climate forecasting system (Hudson et al., 2013; Marshall et al., 2014). POAMA reforecasts are available as a 33-member ensemble; comprised of 11 members each from three variants of the model; [each variant has slightly different model physics](#). We use forecasts issued at the start of each calendar month (12 forecasts a year) from 1982–2010. These forecasts are then post-processed with the method of calibration, bridging and merging (CBaM; Schepen and Wang, 2014; Schepen et al., 2014). While POAMA produces skillful rainfall forecasts in some months and seasons in parts of Australia, it suffers from deficiencies common to many dynamical climate forecasting models: forecasts are often biased at the scale of catchments; forecast ensembles tend to be overconfident; and forecasts may be substantially less skillful than climatology in certain months and seasons (Schepen et al., 2016).

15 We have shown elsewhere that it is only possible to correct all these deficiencies by calibration, rather than applying a simple bias-correction (Zhao et al., 2017). Accordingly, POAMA rainfall reforecasts are calibrated to each catchment with the BJP, ~~which uses Bayesian inference, data transformation (Wang et al., 2012b) and a bivariate normal distribution to establish a statistical relationship between observations and rainfall forecasts and for each month and lead time. The statistical relationship is then used to generate forecasts.~~ This approach is [highly](#) effective at removing bias, correcting ensemble spread, and ensuring forecasts are ‘coherent’ - that is, never less skillful than climatology forecasts (Hawthorne et al., 2013; Schepen et al., 2016; Peng et al., 2014). To maximise the skill of rainfall forecasts, we use ‘bridging’ to build statistical relationships between POAMA [forecasts of SST indices \(e.g. Niño3.4\)-SST forecasts](#) and catchment rainfall, again with the BJP. Bridging ~~also~~ allows us to generate forecasts to 12-month forecast horizons: ~~POAMA produces forecasts only to 9 months in advance; we use bridging to establish~~ [lagged](#) relationships between 9-month SST forecasts and 10-, 11- and 12-month forecast horizons. To merge the calibration and bridging forecasts we use Bayesian model averaging (Wang et al., 2012a) to produce a forecast ensemble of 1000 members. Finally, realistic temporal patterns are instilled in each forecast ensemble member with the Schaake shuffle (Clark et al., 2004).

### 2.2 Hydrological model

35 [Precipitation forecasts and climatology potential evaporation are used to force an initialised monthly rainfall-runoff model. In the original conception of FoGSS, we used the Wapaba model \(Wang et al., 2011\). In this study, we also test two other rainfall-runoff models, GR2M and ABCD, and we describe all three each rainfall-runoff models in Section \[3.4.4.2\]\(#\).](#)

---

### 2.2.3 Hydrological eError model

~~Precipitation forecasts are used to force an initialised monthly rainfall runoff model. In the original conception of FoGSS, we used the Wapaba model (Wang et al., 2011). In this study, we test two other rainfall runoff models, and we describe all three rainfall runoff models in Section 3.4. Forcing a hydrological model with ensemble precipitation forecasts results in overconfident ensemble streamflow forecasts ~~that are overconfident~~, as uncertainty in the hydrological model is not incorporated into the forecast. In addition, hydrological models, even when optimised, are usually subject to errors and bias. To address these issues, FoGSS employs a 3-stage error model. The model is broken into stages to avoid undesirable interaction between parameters when they are optimizedestimated.~~

~~, with each stage described separately (sections 2.2.1, 2.2.2 and 2.2.3).~~

~~Parameters for each stage are estimated sequentially using maximum likelihood estimation (MLE). We give only a brief overview of the estimation procedure here, and refer the reader to Bennett et al. (2016) for a detailed description, including full likelihood equations. Stage 1 parameters are estimated and fixed, followed by Stage 2 and then Stage 3. We employ data transformation (Stage 1) so that residuals are normally distributed and heteroscedastic. For ephemeral rivers, this is not enough to satisfy the requirement of MLE for continuously distributed data. To handle zero values, we treat observations of zero as censored values in the likelihood, a technique established previously (Li et al., 2013).~~

~~A notable aspect of the estimation of hydrological and error model parameters is that we take no account of lead time in the parameter estimation. That is, parameters are estimated only from rainfall runoff simulations (forced by observed rainfall and potential evaporation) and observed streamflow, as with a conventional rainfall runoff model calibration. This is a key difference with approaches that post process streamflow forecasts separately at each lead time (e.g., Yuan, 2016), as it means that each FoGSS time series forecast is a continuous hydrograph that can be summed to produce reliable ensembles of, e.g., seasonal inflow totals. However, the FoGSS error model will not correct problems associated with ensemble rainfall forecasts (e.g. overconfident ensembles). FoGSS requires ensemble rainfall forecasts that are unbiased and reliable in order to produce unbiased and reliable streamflow forecasts.~~

#### 2.2.3.1 Stage 1: data transformation and hydrological modelling

~~It is widely recognised that errors from hydrological models are neither normally distributed nor homoscedastic (e.g., Schaeffli et al., 2007; Smith et al., 2015), and therefore difficult to model using conventional statistical methods. One method for addressing these problems is to use data transformation. We use the log-sinh transformation (Wang et al., 2012b); to normalise data and homogenise variance, which has proven highly effective for normalising hydrological data and homogenising variance (e.g., Del Giudice et al., 2013). The log-sinh transformation is given by:~~

$$z = TF(q) = \frac{1}{b} \log(\sinh(a + bq)) \quad (1)$$

~~where  $q$  is streamflow and  $a$  and  $b$  are parameters. ~~The inverse transformation (back transformation) is given by~~~~

$$q = TF^{-1}(z) = \max\left(\frac{1}{b}(\operatorname{argsinh}(e^{bz}) - a), 0\right), \quad (2)$$

where  $z$  is any streamflow in the transformed domain. For clarity, we will refer to the domain in which  $q$  exists as the original domain to differentiate it from the transform domain of  $z$ .

The parameters in Eq. (1) are estimated from observed streamflow. Once transformation parameters are obtained, hydrological model parameters (Section 3.4) are estimated with MLE.

### 2.23.2 Stage 2: bias-correction

Transformed streamflow is bias-corrected at each time- $t$  month  $i = 1, 2, \dots, 12 = \text{month}(t)$  by

$$z_2(t) = d(i)z_1(t) + \mu(i), \quad (32)$$

where  $z_1$  is the raw streamflow forecast after transformation with Eq. 1, and  $d(i)$  and  $\mu(i)$  are parameters that vary by month,

$$i = 1, 2, \dots, 12 = \text{month}(t), \quad (4)$$

Although Eq. (3) takes is a simple linear regression, because it is applied to transformed flows it is able to correct highly non-linear biases in the original domain. An important feature of Eq. (32) is that the  $d$  parameter can go to zero. That is, in months where the hydrological simulation performs poorly, the error model can return  $z_2 \approx \mu$

, a constant akin to a climatology. As we shall see, this is a particularly important property in ephemeral catchments. This property is exploited in our experiments with the use of a prior, described in Section 4.4.

We limit  $d$  to the range  $0 \leq d \leq 2$ . Values less than zero imply a negative correlation between simulations and observations, and in these cases it is more sensible to ignore the simulation (i.e., to allow  $d=0$ ). The upper limit of 2 is arbitrary, and is imposed to avoid unrealistically large corrections that could result in overfitting of the bias-correction being applied under cross-validation.

### 2.23.3 Stage 3: Autoregressive model and stochastic updating

FoGSS applies a ~~restricted~~ first-order autoregressive (AR1) model (Li et al., 2015) to improve accuracy of forecasts and to propagate hydrological uncertainty through the forecast lead times. The AR1 model is applied to transformed, bias-corrected flows by:

$$z_3(t) = z_2(t) + \rho(i)(z_o(t-1) - z_2(t-1)), \quad (53)$$

where  $z_o$  is the transformed observed streamflow and  $\rho(i)$  is the autoregression parameter, varied by calendar month. To avoid corrections that are too large, Because the AR1 model is applied in the transform domain, the magnitude of the correction can be greatly amplified in the original domain if it is applied to a rising hydrograph, leading to unrealistically large streamflows. To avoid this type of overcorrection, we apply the restriction

proposed by Li et al. (2015). This restriction corrects the forecast by whichever is smaller: the correction proposed by Eq. 3, or the error in the original domain at  $t-1$  (given by  $q_o(t-1) - q_2(t-1)$ , where  $q_o$  is observed streamflow and  $q_2$  is the back-transformed value of  $z_2$ ), or the correction proposed by Eq. (5). As with the previous stages, when estimating Stage 3 parameters with MLE we assume errors,  $\epsilon$ , are normally distributed:

$$\begin{aligned} z_o(t) &= z_3(t) + \varepsilon(i) \\ \varepsilon(i) &\sim N(0, \sigma^2(i)) \end{aligned} \quad (6)$$

where  $\sigma^2(i)$  is the variance of  $\varepsilon$  at each calendar month.

### 2.3.4 Estimating parameters

Parameters for each stage are estimated sequentially using maximum likelihood estimation (MLE), as detailed by Bennett et al. (2016). The data transformation (Stage 1) allows us to assume that residuals,  $\varepsilon$ , are normally distributed and homoscedastic (i.e., the variance can be represented with a single value). At Stage 3, the residuals are given by:

$$\begin{aligned} z_o(t) &= z_3(t) + \varepsilon(i) \\ \varepsilon(i) &\sim N(0, \sigma^2(i)) \end{aligned} \quad (4)$$

where  $\sigma^2(i)$  is the variance of  $\varepsilon$  at each calendar month.

To handle zero values in ephemeral catchments, we treat observations of zero as censored values in the likelihood, a technique established previously (Li et al., 2013).

A notable aspect of the estimation of hydrological and error model parameters is that we take no account of lead time in the parameter estimation. Parameters are estimated only from rainfall-runoff simulations (forced by observed rainfall and potential evaporation) and observed streamflow, as with a conventional rainfall-runoff model calibration. This is a key difference with approaches that post-process streamflow forecasts separately at each lead time (e.g., Yuan, 2016), as it means that each FoGSS time series forecast is a continuous hydrograph that can be summed to produce reliable ensembles of, e.g., seasonal inflow totals. However, the FoGSS error model will not correct problems associated with ensemble rainfall forecasts (e.g. overconfident ensembles). FoGSS requires ensemble rainfall forecasts that are unbiased and reliable in order to produce unbiased and reliable streamflow forecasts.

### 2.3.5 Generating a forecast: stochastic updating

Hydrological uncertainty is propagated with stochastic updating. At the first lead time,  $l=0$ , this is straightforward: we have an observation available when the forecast is issued, and hence we can apply Eq. (53) directly, and then add noise according to Eq. (64) to produce a forecast value  $z_F$ . At longer lead times  $l=1, \dots, 11$  we do not have observations available with which to update the forecast. Instead, we substitute the forecast value,  $z_F$ , for the observation,  $z_o$ , in Eq. (53), and forecasts are generated by:

$$\begin{aligned} z_F(t+l) &= z_2(t+l) + \rho(i) \left( z_F(t+l-1) - z_2(t+l-1) \right) + \varepsilon(i) \quad | \quad l=1, \dots, 11 \\ \varepsilon(i) &\sim N(0, \sigma^2(i)) \end{aligned} \quad (75)$$

In this way hydrological uncertainty grows through the forecast, as expected (i.e., forecasts become less certain at longer lead times). As with Eq. (53), the restriction described (see Section 2.3.3) above is applied to Eq. (75).

---

### 3. Experiments General setup and data

#### 3.1 General setup

##### 3.1.1 Forecast cross-validation

5 Thorough validation of forecast systems requires a large population of reforecasts to allow testing over a variety of conditions and to be able to calculate robust probabilistic verification scores. ~~Unfortunately, r~~ Reforecasts are often limited in number, in our case because ~~of the availability of~~ POAMA reforecasts are only available for 1982-2010 (see Section 2.1). Rigorous cross-validation is a vital element of robust forecast validation. We use the following scheme:

1. The post-processing of rainfall forecasts is cross-validated using leave-3-years out cross-validation
- 10 2. Hydrological and error models are cross-validated using leave-5-years out cross-validation.

A more stringent cross-validation is required for hydrological models because catchment memory is more persistent than memory in seasonal weather patterns or SST (i.e., current catchment conditions can influence streamflow ~~over the next~~ for 2 or more years in some catchments).

To estimate parameters and to generate forecasts, the hydrological model is initialised by running it from January, 15 1970.

##### 3.1.2 Verification scores

In accordance with most studies of ensemble forecasting systems, we are chiefly concerned with two aspects of forecast performance: forecast skill and forecast reliability. To measure forecast skill, we use the well-known continuous ranked probability score (CRPS; see, e.g., Gneiting and Katzfuss, 2014). Skill is measured against ~~“climatology”, or historical frequency, of~~ streamflow climatology. Forecast skill is given by the continuous ranked probability skill score (CRPSS):

$$CRPSS = \frac{CRPS_{Ref} - CRPS_F}{CRPS_{Ref}} \times 100\%, \quad (86)$$

where  $CRPS_F$  and  $CRPS_{Ref}$  are CRPS values for FoGSS and ~~cross-validated~~ climatology forecasts, respectively. To generate the climatology ~~reference~~ forecasts, a log-sinh transformed (Eq. 1) normal distribution is fitted to the observed streamflow data for each month. When fitting the distribution, zero values are handled with data censoring as described by Wang and Robertson (2011). This ensures the climatology reference forecasts correctly replicate the observed incidence of zero values. 1000 samples are drawn from the transformed normal distribution to generate a climatology (values below zero are set to zero). Climatology is generated using observed data for the period 1982-2009, applying the same leave-5-years-out cross-validation procedure as described for the hydrological modelling (Section ~~33.21~~). ~~Zero values are handled through data censoring, as described by Wang and Robertson (2011); that is, the climatology reference forecasts correctly replicate the incidence of zero flows. In some very dry catchments, some months recorded only zero flow, and~~ ~~In some very dry catchments, some months recorded only one or no non-zero flows for the period 1982-2009. In in~~ these cases it is not possible to fit a distribution. Here, we take a pragmatic approach: we simply assign a reference forecast of zero. CRPSS ranges from  $-\infty$  (least skillful) to 100% (perfectly skillful). FoGSS forecasts need not necessarily outperform climatology to function as a viable alternative to stochastic scenarios, but they do need to be at least similarly skillful to

climatology. We term forecasts with skills near zero ~~(we term this neutrally skillful/skillful,~~ defined as  $-5\% > CRPSS < 5\%$ ).

As noted in the introduction, a key attribute of stochastic scenarios is that they are inherently unbiased and thus can be used directly in planning models by water agencies. To be a viable alternative to stochastic scenarios, FoGSS forecasts should be unbiased. Absolute relative bias of forecasts (hereafter referred to as 'bias') is calculated at each lead time,  $l$ , by

$$Bias(l) = \left| \frac{\overline{q_f(l)} - \overline{q_o}}{\overline{q_o}} \times 100\% \right|, \quad (97)$$

where  $\overline{q_f(l)}$  is the mean of all ensemble forecasts at each lead time. Bias ranges from 0 (unbiased) to  $+\infty$  (worst bias). ~~(For brevity, absolute relative bias is simply referred to as bias throughout the paper.)~~

The statistical reliability of ensemble forecasts is assessed with probability integral transform (PIT) - uniform probability plots (shortened to *PIT plots*). Given the cumulative distribution function (CDF) of a forecast at time  $t$ ,  $C_t$ , the PIT of the accompanying observed value,  $q_o(t)$ , is given by:

$$\pi_t = C_t(q_o(t)). \quad (108)$$

$\pi_t$  takes values from zero to 1. When a set of forecasts is reliable, the set of  $\pi_t$  values is uniformly distributed between 0 and 1, and the resulting PIT plot will follow the diagonal one-to-one line. In catchments with zero values, the CDF in Equation (108) will not be continuous (and therefore cannot be expected to follow a uniform distribution). In these catchments, if  $q_o(t) = 0$  -we generate a *pseudo-PIT* value,  $\pi_t$ , randomly sampled from a uniform distribution in the range  $[0, C_t(0)]$ .

To compare reliability for many catchments we summarise information from PIT plots with the alpha index (Renard et al., 2010)

$$\alpha = 1 - \frac{2}{n} \sum_{t=1}^n \left| \pi_t^* - \frac{t}{n+1} \right|, \quad (119)$$

where  $\pi_t^*$  is the sorted  $\pi_t$  - in increasing order, and  $n$  - the number of forecasts. The alpha index essentially reflects the divergence of PIT values from the 1-1 line in PIT plots, ranging from 1 (perfectly reliable) to 0 (worst reliability).

### 3.1.13 Catchments and data

We assess FoGSS forecasts on 63 Australian catchments ranging in size from  $<100 \text{ km}^2$  to  $>200,000 \text{ km}^2$  (Appendix A). Catchments are distributed across the continent, encompassing temperate, desert, subtropical and tropical climates. Rainfall and potential evaporation data are taken from the gridded AWAP dataset, ~~which interpolates gauged observations~~ (Jones et al., 2009; Raupach et al., 2008), which interpolates gauged observations with a Barnes successive correction analysis (<http://www.bom.gov.au/jsp/awap>). Streamflow data are mainly from

gauges, but we have also included several ‘inflow sites’, which are not directly gauged. The inflow site records give total inflow to storages, and are calculated from a combination of streamflow gauge records, storage levels, and discharge from storages. We include these sites because they are of good quality, and often of central importance to water agencies. All streamflow data records have been supplied and checked for quality by the Bureau of Meteorology.

Of the catchments-rivers we assess, one third – 21 catchments – are ephemeral rivers (defined as having zero flows in >4% of their records), occurring in both temperate and tropical climates. As ephemeral rivers tend to be very difficult to predict – they can exhibit strongly non-linear responses of rainfall-runoff to runoffrainfall; and they often experience highly sporadic rainfall – we pay particular attention to these catchments. To illustrate different aspects of the performance of FoGSS, we choose a subset of six catchments (Table 1). The streamflow characteristics of these rivers is-are shown in Figure 2, and each is bBriefly descriptions of each catchment are described:

- **Fitzroy River** (Western Australia): Has Ephemeral river with a large catchment area, and that ceases to flow only occasionally (Figure 2). Like all northern, tropical regions in Australia, the Fitzroy receives most rainfall in the monsoon period (Nov-Mar), and very little rainfall at other times of the year.
- **Ranken River** (Northern Territory): An extremely dry catchment that ceases to flow for long periods, flowing regularly only in Mar. Can record zero flows at any time of year, and is usually dry from Apr-Dec. Over the period 1982-2009, the river never flowed in September.
- **Herbert River** (Queensland): Perennial River-river that receives the bulk of its rainfall in the northern monsoon period (Nov-Mar).
- **Lake Eppalock inflows** (Victoria): Lake Eppalock receives inflow from the temperate and seasonally ephemeral Campaspe River, largely in the period-during Jul-Nov. This is a high-quality inflow series synthesised from stream gauge and storage level records. Often receives zero inflow in late summer to early autumn (Jan-Apr).
- **Goobarragandra River** (New South Wales): Perennial River-river that receives most rainfall in the winter and spring (Jun-Nov). This catchment generally exhibits strong catchment memory.
- Ringarooma River (Tasmania): alpineAlpine, temperate river that receives regular, winter dominant rainfall (Jun-Aug), but has little catchment memory.

## 4 Experiments

### 34.2-1 **Base case: continent-wide performance assessment of FoGSS**

To establish whether FoGSS is a system capable of being deployed across the Australian continent, we test FoGSS as it was described by Bennett et al. (2016): that is, as described in Section 2, using the Wapaba rainfall-runoff model. This constitutes the base case, against which the following variations will be tested. The performance of the base case is assessed by skill, reliability and bias (Section 3.4.2).

### 34.3-2 **Experiment 1: Contribution-contribution of rainfall forecasts to skill**

To assess the contribution of rainfall forecasts to overall streamflow forecast skill, we compare our base case to ESP-like forecasts (extended streamflow predictions). Traditional ESP methods use resampled historical rainfall

---

to force an initialised hydrological model (Day, 1985). An ensemble of historical rainfall forcings is reliable and unbiased but completely uninformative, so any forecast skill remaining will be due to catchment memory (Wood and Lettenmaier, 2008). We use a similar approach, except that we also apply the FoGSS hydrological error model. By comparing streamflow forecasts generated with ESP-like historical rainfall forcings to those generated with the full FoGSS system, we can determine the relative contribution of post-processed POAMA forecasts to overall forecast skill. ~~Rainfall observations~~ Historical rainfall forcings are resampled from ~~the period~~ observations from 1982-2009, using a leave-4-years-out cross-validation scheme. (The leave-4-years-out scheme was chosen in part for computational convenience: it results in a forcing ensemble of 25 members, which divides evenly into 1000, the size of the FoGSS ensemble.) To produce a 1000-member ensemble, we run each historical rainfall sequence through the FoGSS hydrological and error models forty times, using a different random seed at the start of each run. To keep the distinction clear, we refer to the post-processed POAMA forcings as *forecast rainfall* to distinguish them from the ESP-like *historical rainfall* forcings.

### **3.4.4.3 Experiment 2: ~~hydrological~~ hydrological modelling**

As already noted, the original conception of FoGSS made use of the Wapaba rainfall-runoff model (Wang et al., 2011). Wapaba is a five-parameter conceptual hydrological model based on the Budyko curve, which casts the water balance as a competition between available water and available energy. Its parameters and a schematic of its structure are given in Appendix B. Wapaba performed well in a study that compared it to other rainfall-runoff models for simulating 331 (largely) perennial Australian rivers (Wang et al., 2011). However, as we will see, Wapaba's performance is more equivocal for forecasts ~~for~~ of ephemeral rivers.

To test whether performance can be improved using alternative rainfall-runoff models, we substitute two alternative monthly rainfall-runoff models, ABCD and GR2M, into the FoGSS system. ABCD (Thomas, 1981; Alley, 1984) is a four-parameter monthly water balance model and GR2M (Mouelhi et al., 2006) is a simpler model with ~~only~~ two parameters. Parameters and structures of the two models are shown in Appendix B. In general, ABCD and Wapaba are more similar to each other than to GR2M. ABCD and Wapaba each have two parameters to control the apportionment of water between the surface water store and groundwater/direct runoff, while GR2M simply relies on an empirical equation for this apportionment. All three models have two conceptual soil moisture stores, but they function slightly differently in each case. The surface stores in ABCD and Wapaba can lose water only to evaporation or when the storage spills. GR2M's production store loses water to evaporation and spill, but also drains to the routing store at a non-linear rate in relation to the level of the production store. ABCD and Wapaba both have groundwater stores of unlimited capacity and both have parameters to control the (linear) rate of discharge from the groundwater store. GR2M has a finite (and fixed) groundwater storage capacity, and uses a fixed (non-linear) relationship to govern discharge from its routing store. In both Wapaba and ABCD, catchment losses are entirely controlled by evaporation. In GR2M water can be lost to, or gained from, an unlimited conceptual groundwater store outside the catchment. Wapaba and ABCD differ in the way that they apportion water between soil moisture stores and groundwater and direct runoff, and have different methods to calculate actual evaporation from the surface store.

Rainfall-runoff model parameters are estimated using maximum likelihood. Parameters of the subsequent stages of the error model (stages 2 and 3) are then estimated, as described in Section [2.22.3.1](#). That is, only the rainfall-

---

runoff models and Stage 2 and Stage 3 error model parameters change in this experiment: all other elements of FoGSS remain the same.

### 3.5.4.4 Experiment 3: encouraging the error model to return climatology forecasts

As we shall see, the FoGSS system is outperformed by climatology in some very dry catchments. ~~FoGSS forecasts need not necessarily outperform climatology to function as a viable alternative to stochastic scenarios, but they do need to be at least similarly skillful to climatology (we term this *neutrally skillful*).~~ One way to achieve ‘coherent’ forecasts (i.e., where forecasts are at least neutrally skillful) ~~this~~ is to encourage the error model to return climatology forecasts in instances where there are few non-zero streamflows. We do this by ~~encouraging pushing~~ the  $d$ - parameter in the bias-correction (Eq. 32) ~~to wards go to~~ zero. That is, we encourage the error model to discount information from the forecast and to return a climatology ( $z_2 \approx \mu$ ). This is achieved by placing an informative Bayesian prior on the  $d$  parameter:

$$d \sim N(0, \sigma_d^2).$$

(4410)

where the standard deviation,  $\sigma_d$ , controls the strength of the prior: smaller values encourage  $d$  to take values closer to zero. We test the values  $\sigma_d = 0.25, 0.5, 1.0, 2.0, 4.0$ . Because of the use of the prior, this estimation approach is no longer formally MLE, but a *maximum a posteriori* (MAP) estimation. The posterior density used to estimate the parameters is given in Appendix C (Equation C.3).

## 4.5 Results and discussion

### 4.5.1 Continent-wide performance of the base FoGSS model

Forecast skill for all catchments is summarised in Figure 3. At very short lead times (e.g. lead-0) FoGSS forecasts generated are very often skillful. Skill at lead-0 is overwhelmingly positive in perennial catchments, but is generally also positive in ephemeral catchments. Skill subsides with lead time, with forecast skill in ephemeral catchments declining more rapidly. By lead-6, forecasts are generally neutrally skillful for all catchments. Instances of strongly negative skill (<-15%) are rare in perennial catchments, and also generally not present in absent longer lead times in a substantial number of ephemeral catchments. Strongly negative skills do occur in a few very dry ephemeral catchments, as described for the Ranken catchment below. Analysis of bias and statistical analysis of reliability and bias for all 63 catchments will be described in the results of the three experiments- (sections 5.2-5.4) ~~we have conducted, below.~~

~~We To~~ illustrate the overall performance of the FoGSS base case, ~~we by reviewing~~ skill and reliability for the six case study catchments.

~~In general, The strong performance of~~ FoGSS ~~performs well in in~~ perennial catchments, ~~and this~~ is reflected in the Herbert, Goobarragandra and Ringarooma rivers, shown in Figure 34. Forecasts are generally skillful at shorter lead times (typically <3 months), and thereafter become neutrally skillful. ~~In some~~ There is, however, considerable variation in performance perennial catchments; forecasts can be strongly skillful to long lead times (e.g. 6 months or more in the Goobarragandra River), while catchments with little catchment memory (e.g. Ringarooma River) may only be skillful to lead-0 (i.e., in the first month). ~~Regardless of the level of skill, we consider FoGSS to~~

~~perform its role adequately when it does not return negatively skillful forecasts for any month or lead time.~~ Some moderately negative skills do occur in the Herbert catchment, in low-flow months at longer lead times (e.g. August); ~~These are caused by~~ because ~~slight mispredictions of flow issued in wetter months (e.g. February),~~ which can result in proportionally larger errors in drier months at longer lead times. FoGSS also performs well in the ephemeral Fitzroy catchment, returning ~~either largely~~ positive or neutral skill, with only a few isolated instances of slightly negative skill ~~for all months and lead times.~~

In more strongly ephemeral catchments, performance can be poor. In the seasonally ephemeral Eppalock catchment, forecast skill is strongly negative in the dry months from January to April, although the forecasts perform well at other times of the year. In the Ranken catchment, which experiences high incidences of zero flows year-round, performance is poor for a majority of months and lead times.

The cause of the poor forecast skill in the Ranken and Eppalock catchments is evident when we consider PIT plots (Figure 45). Forecasts are highly reliable for the perennial Herbert, Ringarooma and Goobarragandra catchments, as well as the Fitzroy catchment, for all months and lead times. Forecasts are not reliable for the dry months of the Eppalock catchment (see February in Figure 34), and particularly unreliable for drier months in the Ranken catchment (e.g. September). The bowed shape of the PIT plots in Eppalock is evidence of a persistent bias – a tendency to overestimate flows – in the drier months, driven by an underestimation of the incidence of zero flows. The same problem exists in the Ranken catchment, but to a stronger degree. We have established in earlier work that post-processing rainfall forecasts with CBaM is able to produce highly reliable forecast rainfall ensembles (e.g., Peng et al., 2014; Schepen et al., 2012), meaning the problem lies with the hydrological error model. In catchments where more than half of streamflow observations are zero, FoGSS will always underestimate the incidence of zero flows. This is because the error model is assumed to follow a symmetrical distribution (Gaussian after transformation) about the value of the forecast. Even if the forecast is zero before the error model is applied, randomly drawing from a symmetrical distribution will yield ~50% of values greater than zero. We will see in the following experiments that this can have a particularly strong influence on bias.

## 45.2 Experiment 1: contribution of rainfall forecasts to skill ESP forecasts

Figure 56 summarises how forecast skill varies with lead time for all 63 catchments with both forecast and historical rainfall forcings. ~~As described above (Section 4.1), at very short lead times (e.g. lead 0) FoGSS forecasts generated with forecast rainfall are very often skilful. Skill at lead 0 is overwhelmingly positive in perennial catchments, but is generally also positive in ephemeral catchments. Skill subsides with lead time, with forecast skill in ephemeral catchments declining more rapidly. By lead 6, forecasts are generally neutrally skilful. Instances of strongly negative skill (< 15%) are rare in perennial catchments, and generally not present in longer lead times in ephemeral catchments. Strongly negative skills do occur in very dry catchments, as with the Ranken catchment described above.~~

~~Figure 5 shows that~~ Skill at individual lead times is generally not strongly influenced by changing the rainfall forcing to ESP. This ~~generally~~ highlights the predominant role catchment memory plays in generating skilful forecasts. Forecast rainfall tends to produce slightly more skilful forecasts at lead-3 and lead-6 in perennial catchments, but ~~tends to produce~~ more instances of negative skill at longer lead times (e.g. lead-9). Conversely, in ephemeral catchments historical rainfall forcings tend to ~~be~~ produce slightly more skilful streamflow forecasts than forecast rainfall forcings at all lead times.

---

Forecast rainfall shows slightly more evident benefits, however, when we consider forecasts of accumulated volumes. Figure 67 shows forecast skill calculated for forecasts of total streamflow volume accumulated over 1-, 3-, 6-, 9- and 12-month periods. In ephemeral catchments, ESP forecasts are slightly better, with fewer instances of strongly negative skill, particularly for shorter accumulation volumes. In perennial catchments, however, forecast rainfall produces slightly, but noticeably, more skilful streamflow forecasts for accumulation periods of 6 months or more. We note that FoGSS forecasts for perennial catchments generally exhibit positive skill for accumulation periods up to 6 months, whichever rainfall forcing is used. This is clear evidence that FoGSS forecasts hold more useful information than stochastic scenarios, giving clear evidence of the information content of these forecasts over that of stochastic scenarios, irrespective of forcing.

Historical rainfall forcings do, however, have a clear advantage in reducing bias, particularly in ephemeral catchments (Figure 78). Bias is calculated using the mean of the forecast ensemble. Because the BJP models used to post-process POAMA make use of data transformation, the forecasts are unbiased in the transform domain. However, the back-transformation means that forecast ensemble means become separated from (and larger than) ensemble medians, resulting in positive biases. These positive biases are often slight (~5%), but can be amplified by the rainfall-runoff model. This amplification is particularly prevalent in ephemeral catchments, where the responses of runoff to rainfall can be highly non-linear. We note, however, that even with historical rainfall forcings, streamflow forecasts can be heavily biased. In very dry catchments this is partly due to the underestimation of the incidence of zero flows, as described in Section 45.1, above.

Streamflow forecasts generated from historical rainfall forcings show similar reliability to those generated with forecast rainfall forcings (not shown for brevity).

### 45.3 Experiment 2: hydrological modelling

Figures 89 and 910 show how forecast skill and bias vary with the choice of rainfall-runoff model. In general, the skill is similar for all three models, but both GR2M and Wapaba are noticeably less biased than ABCD. Wapaba and GR2M are similarly skillful and exhibit similar biases in perennial catchments. GR2M moderates some of the very negative skill scores and high catchment biases produced by Wapaba in very dry ephemeral catchments, which suggests that Wapaba's infinite groundwater store is not well suited to ephemeral rivers. Like many models, Wapaba can underestimate flows in wet seasons by pushing too much water into groundwater stores and diverting too little through direct runoff. These underestimations have little impact on forecast skill in high flow months. However, the excess water that is pushed into the infinite groundwater store cannot be lost, so it eventually drains out in dry seasons. This can result in substantial overestimation of streamflow in very dry seasons, which causes high proportional errors and biases. While we apply a bias-correction in the error model, Wapaba's overestimation in dry months is caused by isolated (i.e., rare) events, which are difficult to capture under cross-validation. GR2M's ability to destroy water held in its groundwater store appears to be important for accounting for the high losses that can occur in drylands. GR2M requires the error model to do less work, making the system less prone to errors/bias under cross-validation in ephemeral rivers. ~~Despite the benefits of GR2M over Wapaba, we note that Wapaba's Budyko-based structure remains theoretically attractive. We plan to explore ways to improve Wapaba's simulation of ephemeral rivers in future research.~~

A noteworthy finding of this experiment is that the choice of rainfall-runoff model did not have a major impact on forecast skill in perennial catchments. While a considerable amount of effort is often expended on selecting

---

rainfall-runoff models for particular purposes, our results suggest that, at least at the monthly time step, a well-designed error model can mitigate various deficiencies in rainfall-runoff models for wide-scale application to perennial rivers.

#### 45.4 Experiment 3: encouraging the error model to return climatology forecasts

5 As we expect, the application of a prior on the  $d$ - parameter has negligible effect on the skill of forecasts in perennial rivers at all lead times (Figure 4011). However, applying the prior did reduce some of the strongly negative skills experienced in ephemeral catchments at all lead times. The stronger the prior (i.e., the smaller the value of  $\sigma_d$ ), the greater the removal of negative skills~~the more that negative skills were removed~~, with the effect of the prior becoming negligible for  $\sigma_d \geq 2.0$ . Similarly, bias is greatly reduced by applying a strong prior to 10 ephemeral rivers (Figure 4112), as the forecasts have a reduced tendency to overestimate flows in very dry months. Interestingly, applying a strong prior also reduced biases in perennial catchments. This indicates that the prior is guarding against over-fitting of the bias-correction in these instances, with virtually no reduction in positive forecast skill. The reduction in bias has a slight positive impact on reliability in ephemeral rivers at longer lead times, as shown by the alpha index in Figure 4213. However, the prior is unable to address the fundamental 15 inability of FoGSS to generate a sufficient number zero flows in months where more than half of observed flows are zero, as discussed in Section 45.51.

In summary, the prior encourages FoGSS to behave sensibly. As already noted, strongly negative skills generally only occur in very dry months, where there may be only a few non-zero observations on which to optimise the hydrological and error models. In these cases, it is sensible to encourage FoGSS to return a climatology-like 20 forecast. Conversely, when there are sufficient data to inform the ~~optimisation estimation of the models parameters~~ and the models perform well, the system should use the models. Using a prior in a MAP optimization enforces this sensible behaviour in the ~~model~~FoGSS system.

#### 45.5 Synthesis

In each experiment, variations on the base case resulted in changes in forecast performance, although these changes were sometimes very slight. The use of historical rainfall forcings (Experiment 1) is the least beneficial of the changes. Historical rainfall forcings can reduce bias, and this leads to fewer strongly negative skills, largely in very dry months and catchments. We note, however, the use of a strong prior has a stronger ability to remove bias in dry months than historical forcings (not shown), thus nullifying the benefits of the historical forcing. The use of historical forcings comes at the cost of removing information available from climate forecasts. We have shown 30 that skill from climate forecasts can accumulate to produce skillful long-range total inflow forecasts. In addition, the POAMA model is being upgraded to a much higher resolution climate forecasting system by the Bureau of Meteorology (ACCESS-S), and this should result in stronger skill. On balance, the inclusion of climate forecasts is beneficial, both for the additional skill available in some months/catchments with post-processed POAMA forecasts, but also for the prospect of including better climate forecasts in future.

35 To show the effects of the other variations (experiments 2 and 3), we combine forecast rainfall forcings with the GR2M model and a strong prior on  $d$  ( $\sigma_d = 0.25$ ), and show forecast skill for our six example catchments in Figure 4314. There are some key differences between Figure 4314 and forecast skill of the base case (Figure 34).

---

In the very dry Ranken catchment, negative skill in wetter months (Jan-Apr) is largely removed, in favour of climatology-like forecasts. Conversely, skill in ~~Jun~~Jul, Aug and Dec has changed from neutral/positive in the base case to be substantially negative. All three of these months are very dry in Ranken catchment (most flows are zero), but feature a single very large event in the record (>10x larger than the next largest measured flow). In these cases, the prior on  $d$  has a deleterious effect: allowing  $d$  to take a larger range of values better corrects the extremely non-linear biases in these months. When the prior is applied, the bias-correction is not as effective, forcing the error model to take larger values of  $\sigma^2$ . This leads to persistent overestimations of streamflow in the other (very dry) years, leading to negative skill. In the Eppalock catchment, the variations on the base case have an unequivocal benefit: negative skill in the dry months of Jan-Apr is completely removed. In the other five catchments, the changes generally either improve or have little impact on base case forecasts. There is little change to skill in the Fitzroy catchment (very slight reductions in skill in Jul-Sep at long lead-times), negative skill in the Herbert catchment in August is largely eliminated, and there are no discernible differences in skill in the Ringarooma and Goobarragandra catchments.

As already noted, the GR2M model's main benefit is in ephemeral catchments. In our example catchments in Figure ~~13~~14, GR2M acts mainly to reduce negative skills in the Eppalock catchment in Feb and Mar by reducing bias, with little differences in other catchments. As with the benefits of historical rainfall forcings, however, the ability of GR2M to reduce bias is largely subsumed by the use of a strong prior on the  $d$  parameter: similar reductions in negative skill in Eppalock are achieved when a prior on  $d$  is applied with the Wapaba model (not shown).

The use of a strong prior on  $d$  results in neutral to positive impacts on skill in most cases shown in Figure ~~13~~14. The exception is the very dry Ranken catchment, where the benefits of the prior are equivocal. The prior removes the base case's negative skills in the Ranken catchment in Jan-Mar, but also introduces negative skill in the drier months of Jun, Aug and Dec. We note that, on balance, this may have practical benefits: in another study (Turner et al., 2017, this issue) we show that FoGSS forecasts can benefit reservoir operations in cases where forecasts are not skillful in very dry months, but positively or neutrally skillful at other times of the year. This is because the dry months contribute little to the annual inflow volume, so small positive bias in dry months (the cause of negative skill) does not have a strong influence on the value of forecasts. Conversely, a strong prior is responsible for removing negative skill in August in the Herbert catchment, and also removes the strongly negative skills in the Eppalock catchment in Jan-Apr. At the same time, the prior has little effect on the good performance of the base case in the Fitzroy, Ringarooma and Goobarragandra catchments.

We reiterate that the prior does not correct reliability problems in dry catchments, with PIT plots giving almost identical results to the base case (not shown for brevity). To mitigate the inherent tendency of the FoGSS error model to underestimate the occurrence of zero flows, we need to change its fundamental function. One approach for doing this would be to censor both simulations and observations in the MLE, and carry this approach through to generating forecasts. This would effectively change the assumption of a symmetrical error distribution about forecasts of zero, and offset the error distribution to increase the incidence of zeros. We will explore this approach in future research.

---

## 5-6 Summary

We assess a new seasonal streamflow forecasting system called Forecast Guided Stochastic Scenarios (FoGSS) for continent-wide application in Australia. FoGSS uses post-processed climate model forecasts to force a monthly rainfall-runoff model, and applies a staged error model to quantify and propagate hydrological model uncertainty.

5 FoGSS is intended to provide a skillful alternative to resampled inflows for water agencies to use in operational planning: it is designed to extract skill from climate and catchment conditions, to produce unbiased and reliable ensemble predictions to 12-month forecast horizons, and to produce ‘coherent forecasts’ when forecast skill is not available – that is, forecasts that are similarly skillful to climatology. FoGSS is assessed on 63 Australian catchments, of which 21 are ephemeral rivers. FoGSS performs well in all but the driest catchments. Skill is generally positive at shorter lead times in both perennial and ephemeral catchments, and transitions to neutral (zero) skill with respect to climatology at longer lead times. Forecast ensembles are generally reliable. However, in very dry catchments forecasts can be strongly negatively skillful and biased, in many cases because the ensembles are not reliable.

We conduct 3 experiments to establish whether components of the FoGSS system can be improved:

1. We use historical rainfall forcings – similar to ESP forecasts - to assess the contribution of forecast rainfall forcings to forecast skill
2. We assess three monthly rainfall-runoff models (Wapaba, GR2M, ABCD)
3. We use a Bayesian prior in our parameter estimation procedure to encourage the FoGSS error model to return climatology forecasts in months where the hydrological model performs poorly

20 Historical rainfall forcings sometimes improve forecasts (largely in very dry catchments) by reducing bias. However, this comes at the cost of including useful information in rainfall forecasts, information that we show can accumulate over multiple lead times. On balance we believe the inclusion of seasonal rainfall forecasts in the FoGSS system is beneficial.

25 Wapaba and GR2M clearly outperform the ABCD rainfall-runoff model, and GR2M performs slightly better than Wapaba in ephemeral catchments. However, the advantages of the GR2M model are overshadowed by the use of the Bayesian prior. The prior reduces the instances of negative forecast skill and reduces bias in ephemeral catchments, and has little effect on performance in perennial catchments. The use of the prior does not, however, result in reliable forecast ensembles in catchments where zero flows occur more than half the time. We point to future research that could improve reliability in these very dry catchments.

## 30 6-7 Acknowledgements

This research has been supported by the Water Information Research and Development Alliance (WIRADA) between the Bureau of Meteorology and CSIRO Land & Water. Thanks to Senlin Zhou, Julien Lerat, Paul Feikema and Daehyok Shin (all Bureau of Meteorology) for supplying data and for fruitful discussions on the development of FoGSS.

---

## References

- Alley, W. M.: On the Treatment of Evapotranspiration, Soil Moisture Accounting, and Aquifer Recharge in Monthly Water Balance Models, *Water Resources Research*, 20, 1137-1149, doi: 10.1029/WR020i008p01137, 1984.
- 5 Beckers, J. V. L., Weerts, A. H., Tjeldeman, E., and Welles, E.: ENSO-conditioned weather resampling method for seasonal ensemble streamflow prediction, *Hydrol. Earth Syst. Sci.*, 20, 3277-3287, doi: 10.5194/hess-20-3277-2016, 2016.
- Bell, V. A., Davies, H. N., Kay, A. L., Brookshaw, A., and Scaife, A. A.: A national-scale seasonal hydrological forecast system: development and evaluation over Britain, *Hydrol. Earth Syst. Sci. Discuss.*, 2017, 1-19, doi: 10.5194/hess-2017-154, 2017.
- 10 Bennett, J. C., Wang, Q. J., Li, M., Robertson, D. E., and Schepen, A.: Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model, *Water Resources Research*, 52, 8238–8259, doi: 10.1002/2016wr019193, 2016.
- Candogan Yossef, N., van Beek, R., Weerts, A., Winsemius, H., and Bierkens, M. F. P.: Skill of a global forecasting system in seasonal ensemble streamflow prediction, *Hydrol. Earth Syst. Sci. Discuss.*, 2016, 1-39, doi: 10.5194/hess-2016-521, 2016.
- 15 Clark, M. P., Gangopadhyay, S., Hay, L., Rajagopalan, B., and Wilby, R.: The Schaake shuffle: a method for reconstructing space–time variability in forecasted precipitation and temperature fields, *Journal of Hydrometeorology*, 5, 243–262, doi: 10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2, 2004.
- 20 Crochemore, L., Ramos, M. H., and Pappenberger, F.: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, *Hydrology and Earth System Sciences*, 20, 3601-3618, doi: 10.5194/hess-2016-78, 2016.
- Day, G. N.: Extended streamflow forecasting using NWSRFS, *Journal of Water Resources Planning and Management*, 111, 157–170, doi: 10.1061/(ASCE)0733-9496(1985)111:2(157), 1985.
- 25 Fundel, F., Jörg-Hess, S., and Zappa, M.: Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices, *Hydrol. Earth Syst. Sci.*, 17, 395-407, doi: 10.5194/hess-17-395-2013, 2013.
- Gneiting, T., and Katzfuss, M.: Probabilistic forecasting, *Annual Review of Statistics and Its Application*, 1, 125-151, doi: 10.1146/annurev-statistics-062713-085831, 2014.
- 30 Greuell, W., Franssen, W. H. P., Biemans, H., and Hutjes, R. W. A.: Seasonal streamflow forecasts for Europe – I. Hindcast verification with pseudo- and real observations, *Hydrol. Earth Syst. Sci. Discuss.*, 2016, 1-18, doi: 10.5194/hess-2016-603, 2016.
- Hawthorne, S., Wang, Q. J., Schepen, A., and Robertson, D. E.: Effective use of GCM outputs for forecasting monthly rainfalls to long lead times, *Water Resources Research*, 49, 5427–5436, doi: 10.1002/wrcr.20453, 2013.
- 35 Hudson, D., Marshall, A. G., Yin, Y., Alves, O., and Hendon, H. H.: Improving intraseasonal prediction with a new ensemble generation strategy, *Monthly Weather Review*, 141, 4429-4449, doi: 10.1175/mwr-d-13-00059.1, 2013.
- Jones, D. A., Wang, W., and Fawcett, R.: High-quality spatial climate data-sets for Australia, *Australian Meteorological and Oceanographic Journal*, 58, 233-248, 2009.

- 
- Li, M., Wang, Q. J., and Bennett, J.: Accounting for seasonal dependence in hydrological model errors and prediction uncertainty, *Water Resources Research*, 49, 5913–5929, doi: 10.1002/wrcr.20445, 2013.
- Li, M., Wang, Q. J., Bennett, J. C., and Robertson, D. E.: A strategy to overcome adverse effects of autoregressive updating of streamflow forecasts, *Hydrology and Earth System Sciences*, 19, 1-15, doi: 10.5194/hess-19-1-2015, 5 2015.
- Marshall, A. G., Hudson, D., Wheeler, M. C., Alves, O., Hendon, H. H., Pook, M. J., and Risbey, J. S.: Intra-seasonal drivers of extreme heat over Australia in observations and POAMA-2, *Climate Dynamics*, 43, 1915-1937, doi: 10.1007/s00382-013-2016-1, 2014.
- Meißner, D., Klein, B., and Ionita, M.: Development of a monthly to seasonal forecast framework tailored to inland 10 waterway transport in Central Europe, *Hydrol. Earth Syst. Sci. Discuss.*, 2017, 1-31, doi: 10.5194/hess-2017-293, 2017.
- Mouelhi, S., Michel, C., Perrin, C., and Andréassian, V.: Stepwise development of a two-parameter monthly water balance model, *Journal of Hydrology*, 318, 200-214, doi: 10.1016/j.jhydrol.2005.06.014, 2006.
- Peng, Z., Wang, Q. J., Bennett, J. C., Schepen, A., Pappenberger, F., Pokhrel, P., and Wang, Z.: Statistical 15 calibration and bridging of ECMWF System4 outputs for forecasting seasonal precipitation over China, *Journal of Geophysical Research (Atmospheres)*, 119, 7116–7135, doi: 10.1002/2013JD021162., 2014.
- Raupach, M. R., Briggs, P. R., Haverd, V., King, E. A., Paget, M., and Trudinger, C. M.: Australian Water Availability Project (AWAP), final report for Phase 3, CSIRO Marine and Atmospheric Research, Canberra, Australia, 67 pp, 2008.
- 20 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, W05521, doi: 10.1029/2009wr008328, 2010.
- Schepen, A., Wang, Q. J., and Robertson, D. E.: Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall, *Journal of Geophysical Research*, 117, D20107, doi: 25 10.1029/2012JD018011, 2012.
- Schepen, A., and Wang, Q.: Ensemble forecasts of monthly catchment rainfall out to long lead times by post-processing coupled general circulation model output, *Journal of Hydrology*, 519, 2920–2931, doi: 10.1016/j.jhydrol.2014.03.017, 2014.
- Schepen, A., Wang, Q. J., and Robertson, D. E.: Seasonal forecasts of Australian rainfall through calibration and 30 bridging of coupled GCM outputs, *Monthly Weather Review*, 142, 1758-1770, doi: 10.1175/mwr-d-13-00248.1, 2014.
- Schepen, A., Wang, Q. J., and Robertson, D. E.: Application to post-processing of meteorological seasonal forecasting, in: *Handbook of hydrometeorological ensemble forecasting*, 1 ed., edited by: Duan, Q., Pappenberger, F., Thielen, J., Wood, A., Cloke, H. L., and Schaake, J. C., Springer-Verlag Berlin Heidelberg, 1-29, 2016.
- 35 Thomas, H. A.: Improved methods for national water assessment, Harvard Water Resources Group, 1981.
- Turner, S. W. D., Bennett, J., Robertson, D., and Galelli, S.: Value of seasonal streamflow forecasts in emergency response reservoir management, *Hydrol. Earth Syst. Sci. Discuss.*, 2017, 1-26, doi: 10.5194/hess-2016-691, 2017.
- Wang, Q. J., Pagano, T. C., Zhou, S. L., Hapuarachchi, H. A. P., Zhang, L., and Robertson, D. E.: Monthly versus 40 daily water balance models in simulating monthly runoff, *Journal of Hydrology*, 404, 166-175, doi: 10.1016/j.jhydrol.2011.04.027, 2011.

- 
- Wang, Q. J., and Robertson, D. E.: Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences, *Water Resources Research*, 47, W02546, doi: 10.1029/2010WR009333, 2011.
- Wang, Q. J., Schepen, A., and Robertson, D. E.: Merging seasonal rainfall forecasts from multiple statistical models through Bayesian model averaging, *Journal of Climate*, 25, 5524-5537, doi: 10.1175/JCLI-D-11-00386.1, 2012a.
- Wang, Q. J., Shrestha, D. L., Robertson, D. E., and Pokhrel, P.: A log-sinh transformation for data normalization and variance stabilization, *Water Resources Research*, 48, W05514, doi: 10.1029/2011WR010973., 2012b.
- Wood, A. W., and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, *Geophysical Research Letters*, 35, L14401, doi: 10.1029/2008gl034648, 2008.
- Wood, A. W., and Schaake, J. C.: Correcting Errors in Streamflow Forecast Ensemble Mean and Spread, *Journal of Hydrometeorology*, 9, 132-148, doi: 10.1175/2007jhm862.1, 2008.
- Yuan, X., Wood, E. F., Chaney, N. W., Sheffield, J., Kam, J., Liang, M., and Guan, K.: Probabilistic Seasonal Forecasting of African Drought by Dynamical Models, *Journal of Hydrometeorology*, 14, 1706-1720, doi: 10.1175/jhm-d-13-054.1, 2013.
- Yuan, X.: An experimental seasonal hydrological forecasting system over the Yellow River basin – Part 2: The added value from climate forecast models, *Hydrology and Earth System Sciences*, 20, 2453-2466, doi: 10.5194/hess-20-2453-2016, 2016.
- Zhao, T., Schepen, A., and Wang, Q. J.: Ensemble forecasting of sub-seasonal to seasonal streamflow by a Bayesian joint probability modelling approach, *Journal of Hydrology*, 541, Part B, 839-849, doi: 10.1016/j.jhydrol.2016.07.040, 2016.
- Zhao, T., Bennett, J. C., Wang, Q. J., Schepen, A., Wood, A. W., Robertson, D. E., and Ramos, M.-H.: How suitable is quantile mapping for post-processing GCM precipitation forecasts?, *Journal of Climate*, doi: 10.1175/jcli-d-16-0652.1, 2017.

25

## Tables

**Table 1 Case study catchments**

<b>Gauge Name</b>	<b>Gauge number</b>	<b>State</b>	<b>Perennial/Ephemeral</b>	<b>Catchment area (km<sup>2</sup>)</b>	<b>Longitude</b>	<b>Latitude</b>	<b>Missing data (%)</b>
Goobarragandra River above Lacmalac	410057	NSW	Perennial	668	148.35	-35.33	0.3
Ranken River at Soudan Homestead	G0010005	NT	Ephemeral	4,360	137.02	-20.05	8.3
Herbert River above Abergowrie	116006B	QLD	Perennial	7,486	145.92	-18.49	0.0
Ringarooma River at Moorina Bridge	30	TAS	Perennial	517	147.87	-41.13	8.0
Lake Eppalock inflows (Campaspe River)	Inflows site	VIC	Ephemeral	1,749	144.56	-36.88	0.0
Fitzroy River at Fitzroy Crossing Bridge	802055	WA	Ephemeral	46,133	125.58	-18.21	0.3

Figures

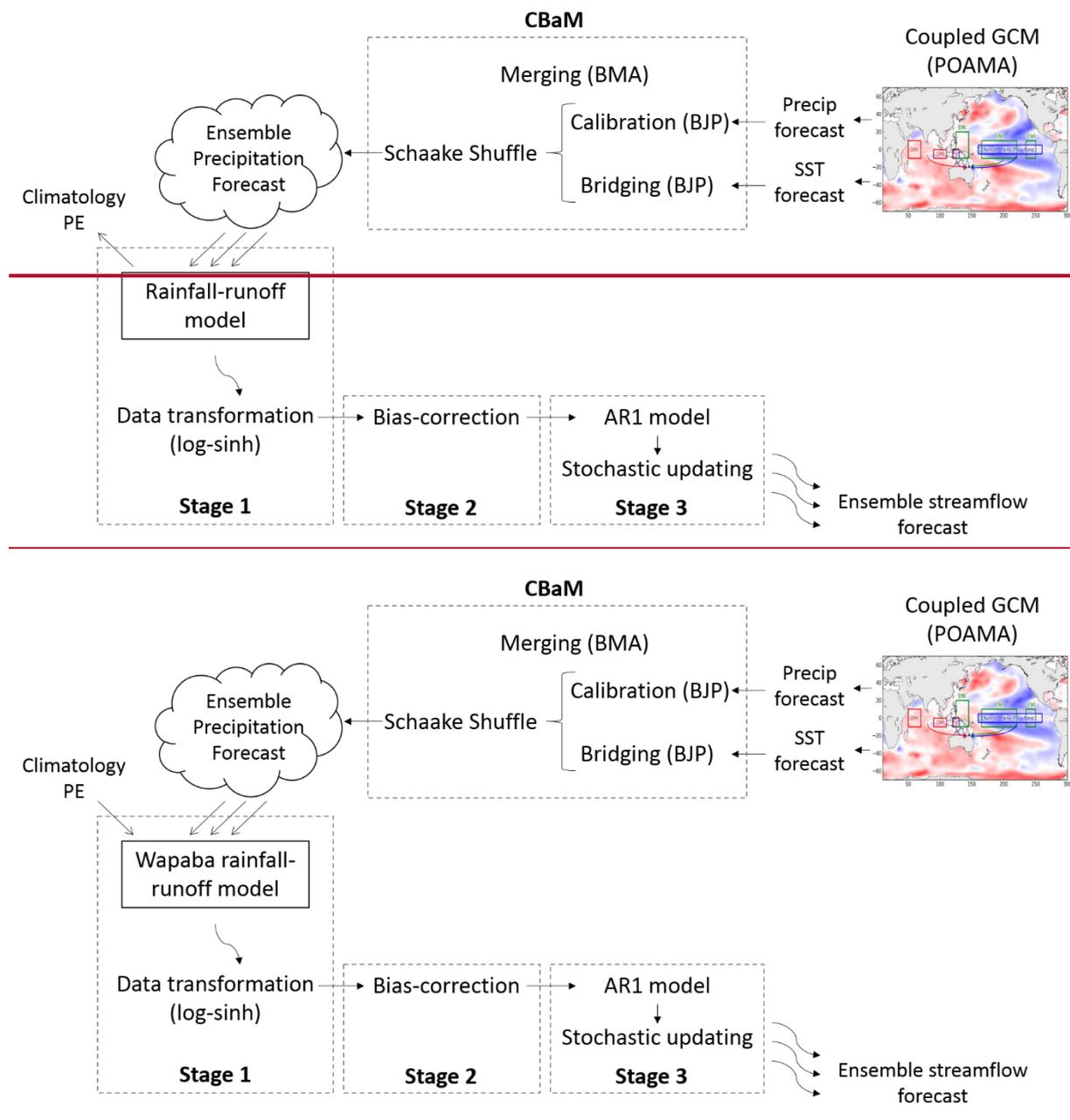
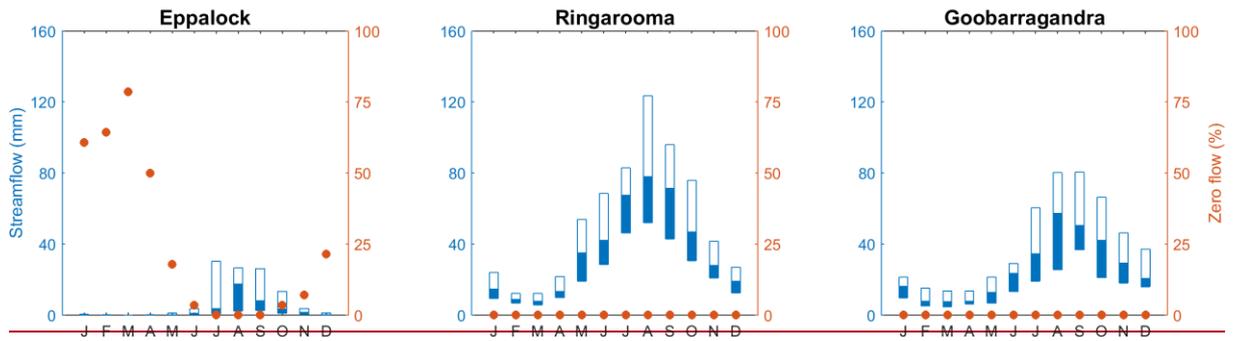
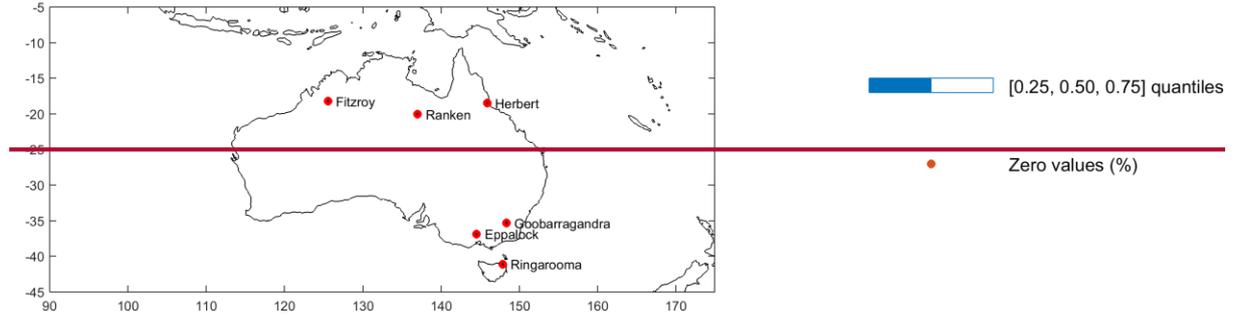
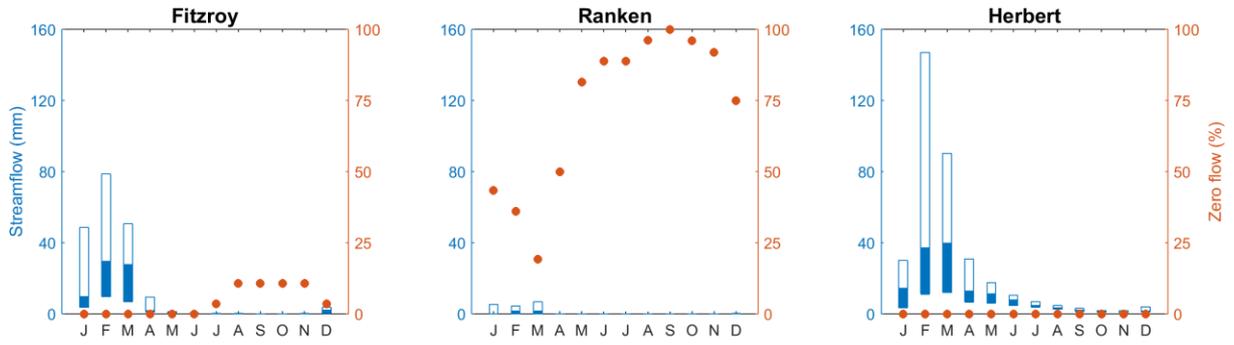


Figure 1: Schematic of the FoGSS model





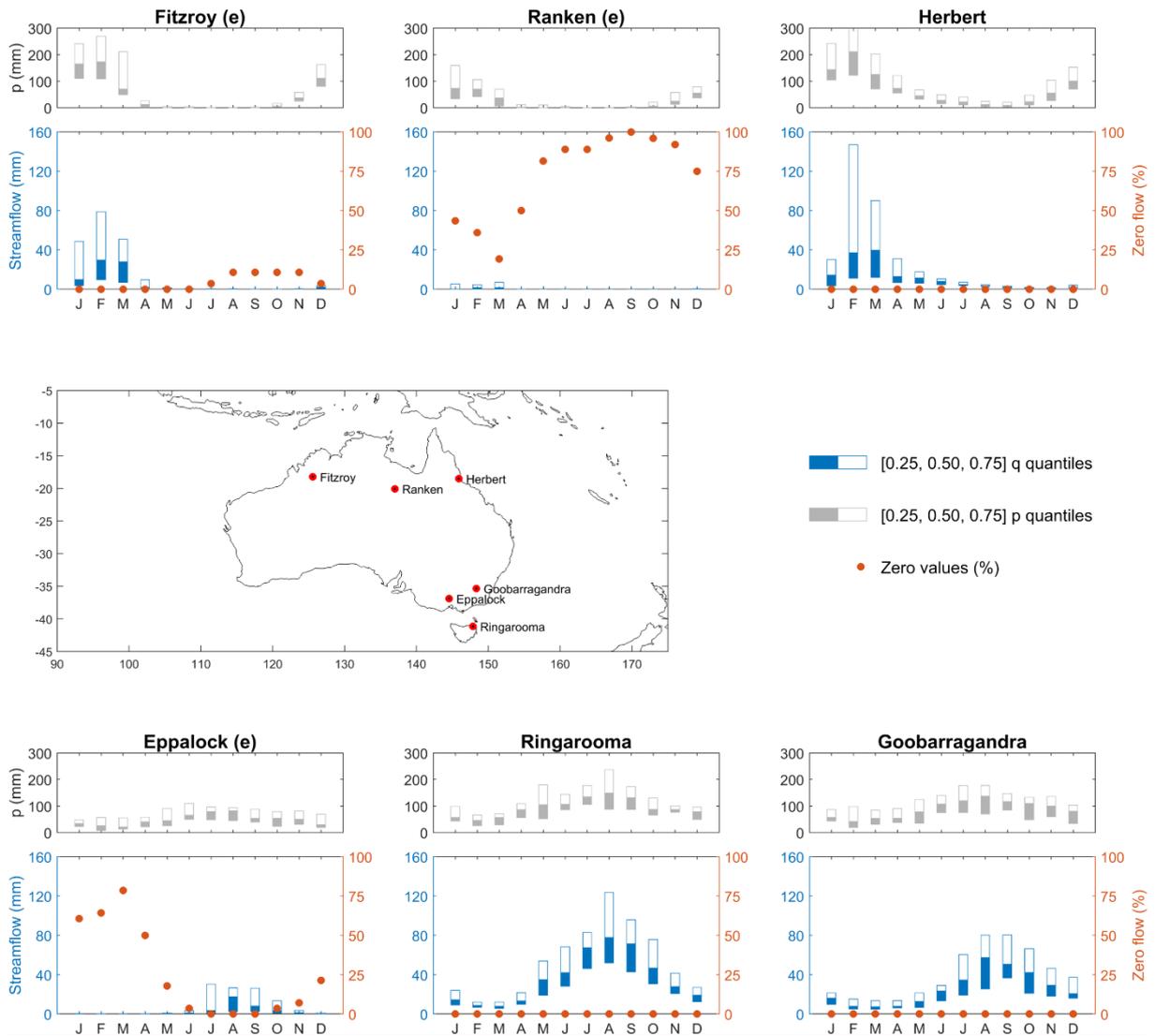
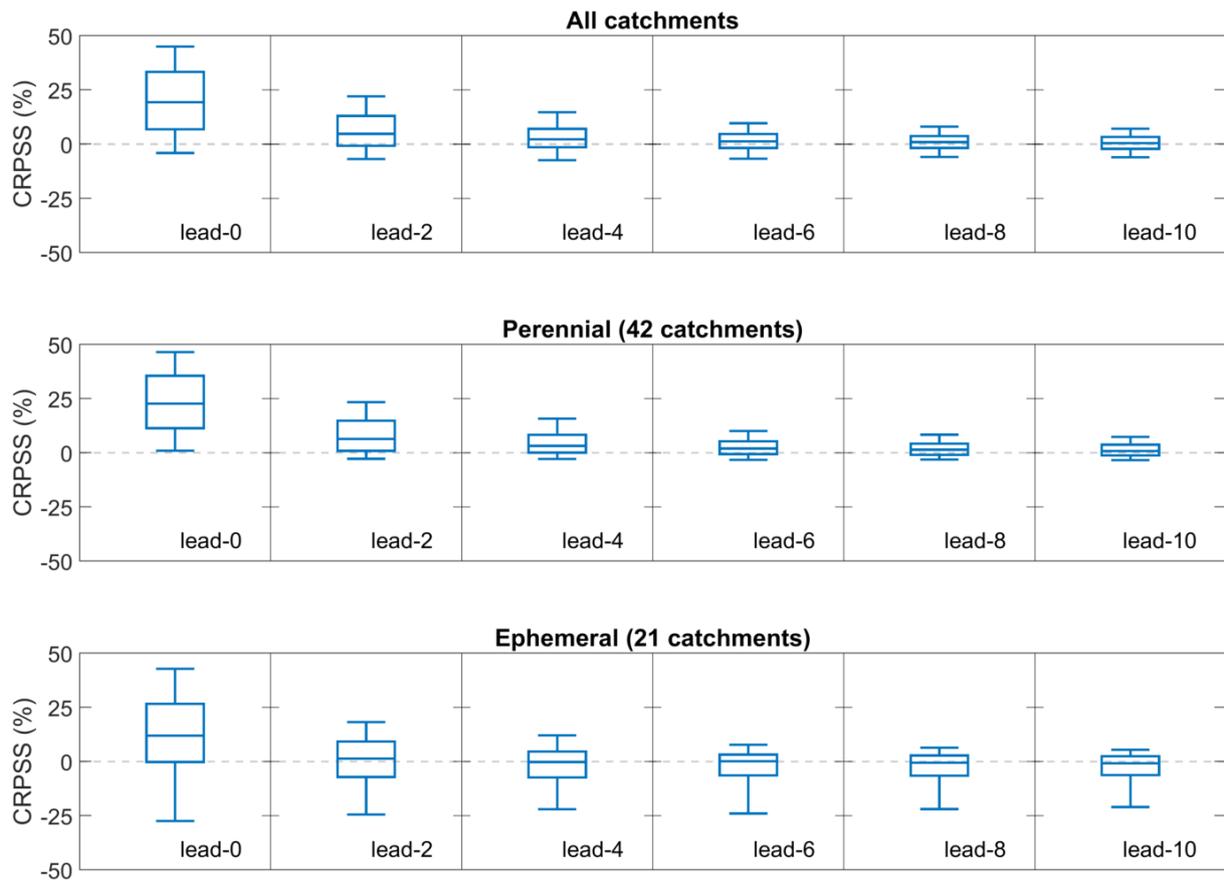


Figure 2: Catchment characteristics of six case study catchments. **Ephemeral catchments are denoted by (e)**. Left axis shows monthly streamflow (**q**) and rainfall (**p**) characteristics, with **blue** bars showing interquartile range and median flows for the period 1982-2009. Right axis shows proportion of zero flows (orange points) in each month for the period 1982-2009.

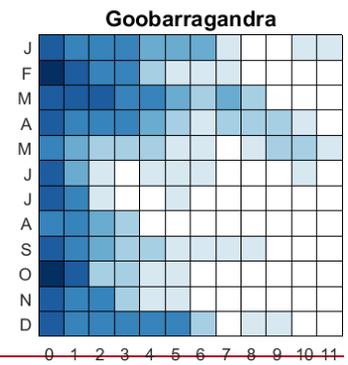
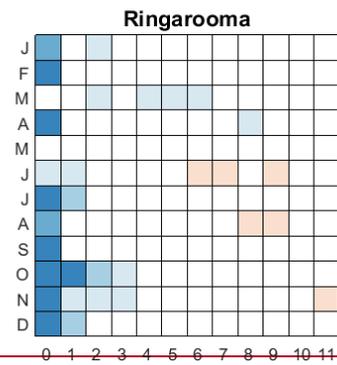
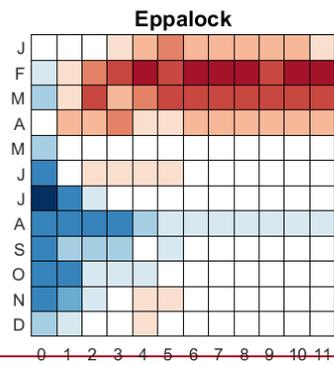
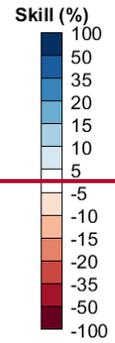
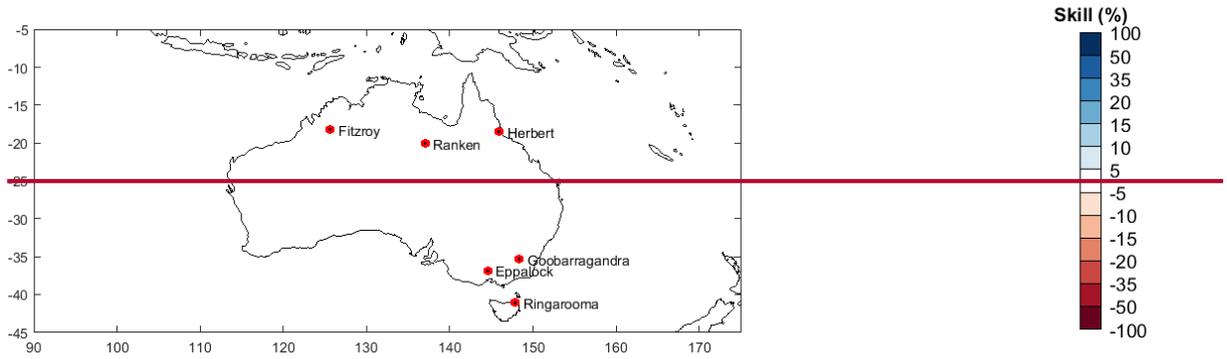
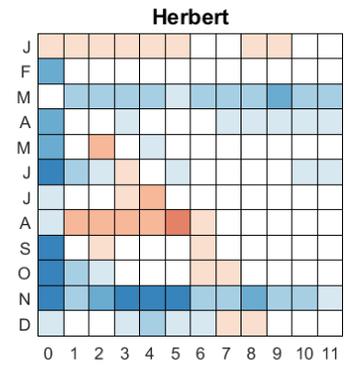
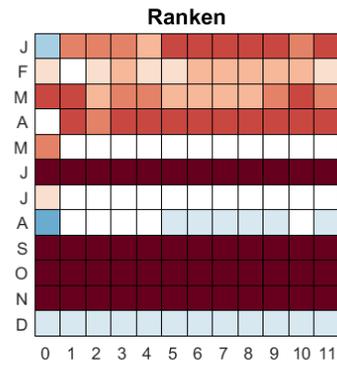
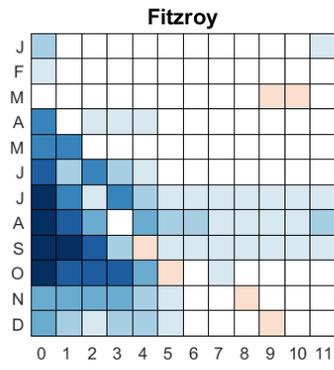
5

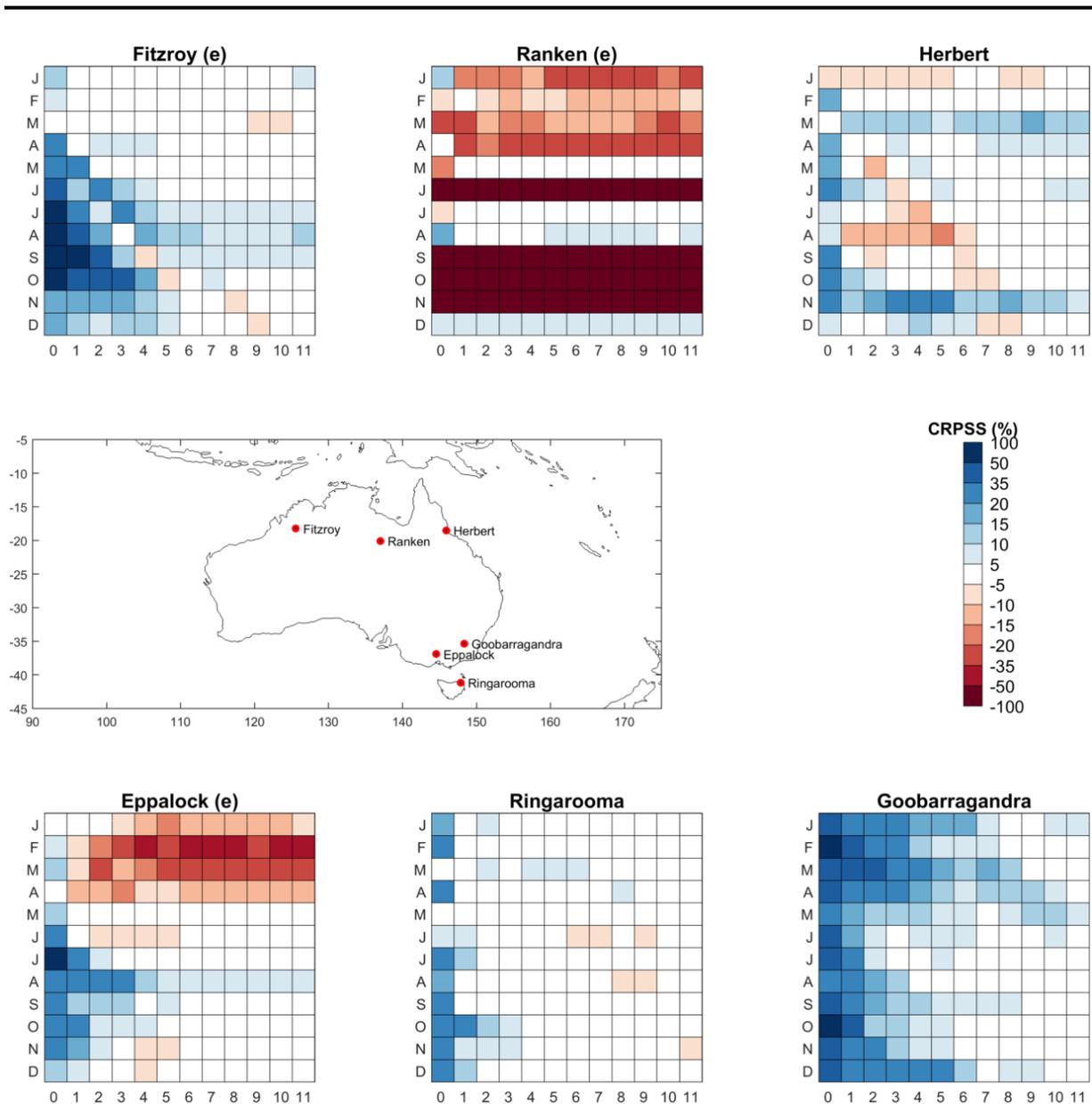


**Figure 3: Forecast skill (CRPSS) for all 63 catchments by lead time for the FoGSS base case. For each lead time, forecast skill is summarised for all months and catchments with box and whisker plots. Boxes show interquartile range with the median, whiskers give 10th and 90th percentiles. Top panel shows all catchments, middle panel shows perennial catchments, and bottom panel shows ephemeral catchments.**

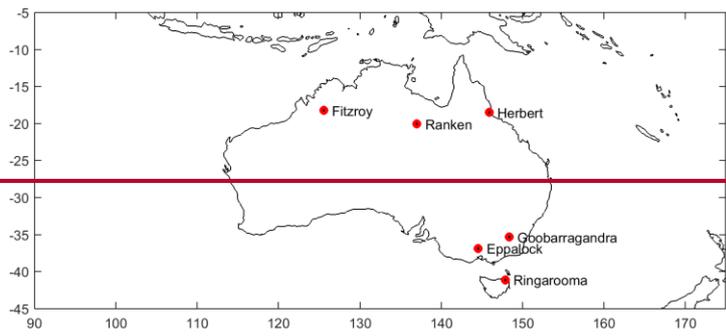
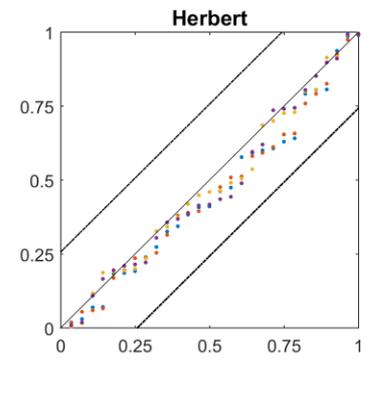
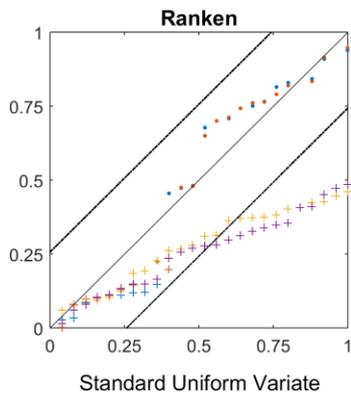
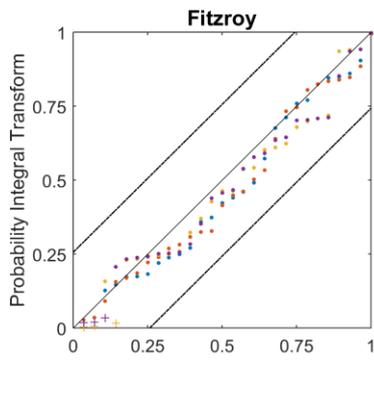
5



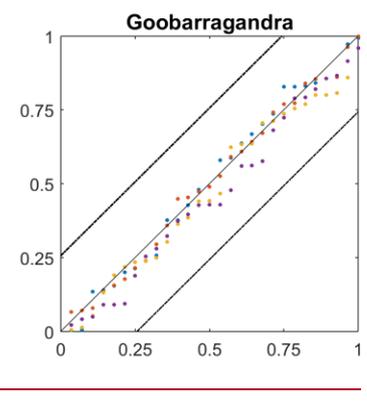
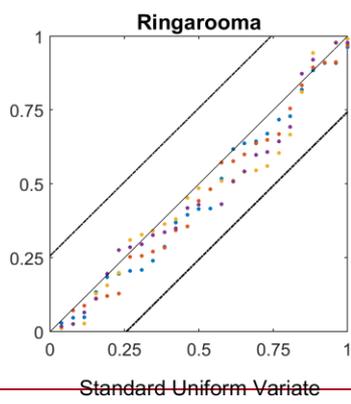
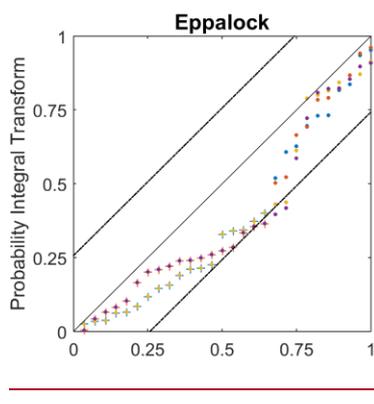


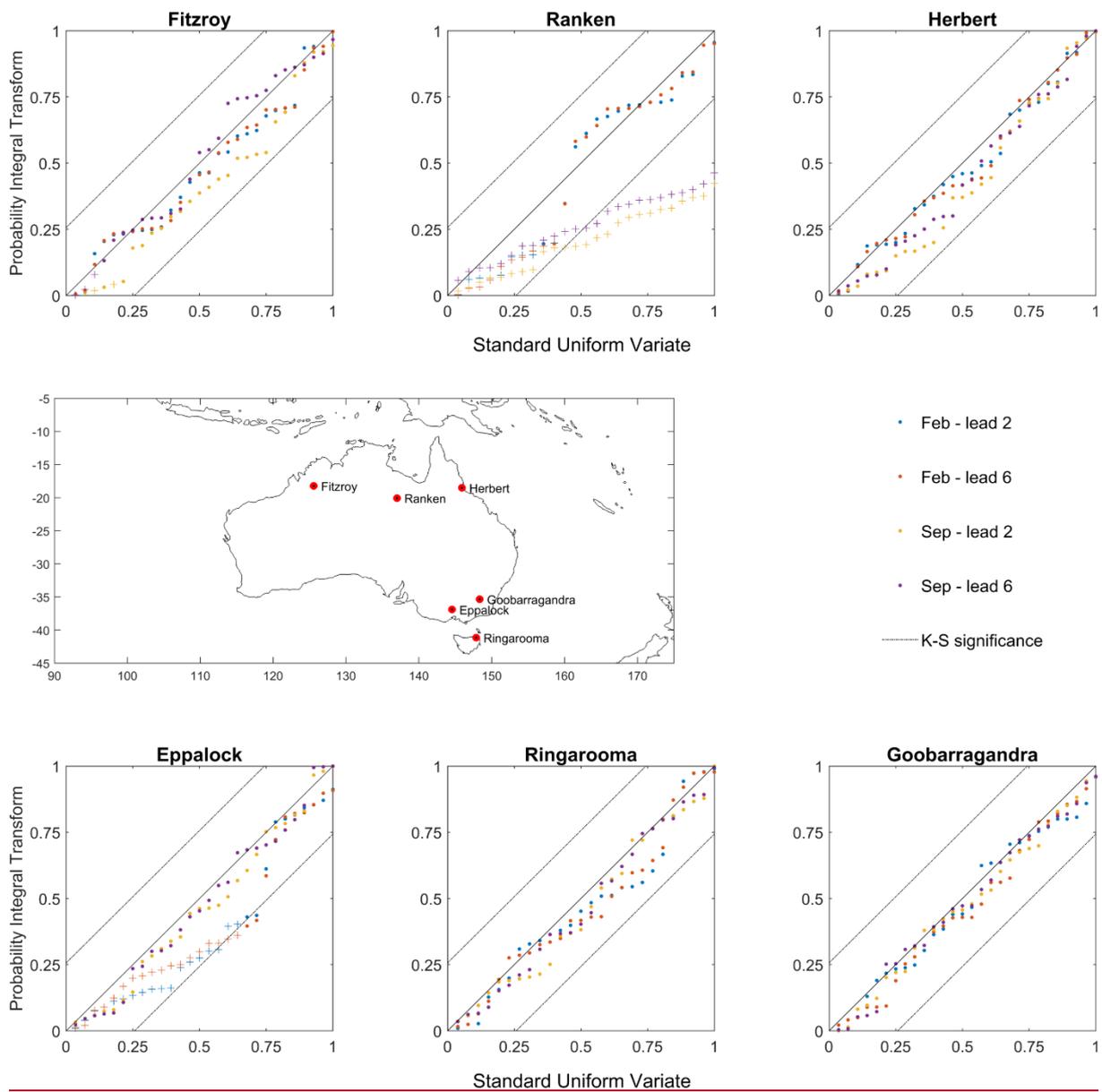


**Figure 34: Skill (CRPSS) skill-scores for FoGSS forecasts (base case). Ephemeral catchments are denoted with (e). Target months are shown on the vertical axes, and target lead times on the horizontal axes. Centre map gives catchment locations.**

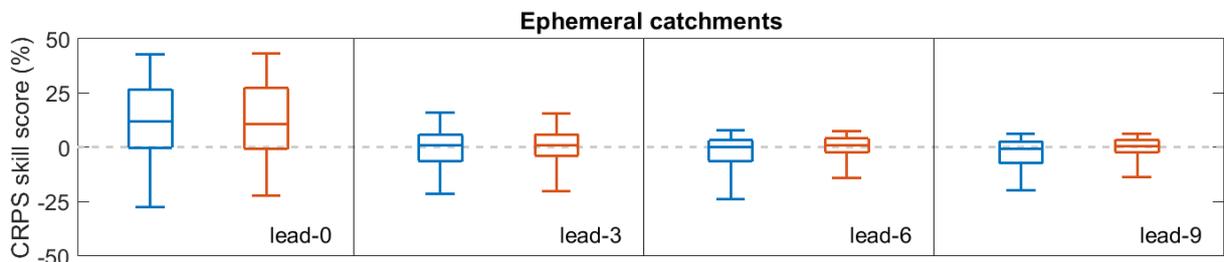
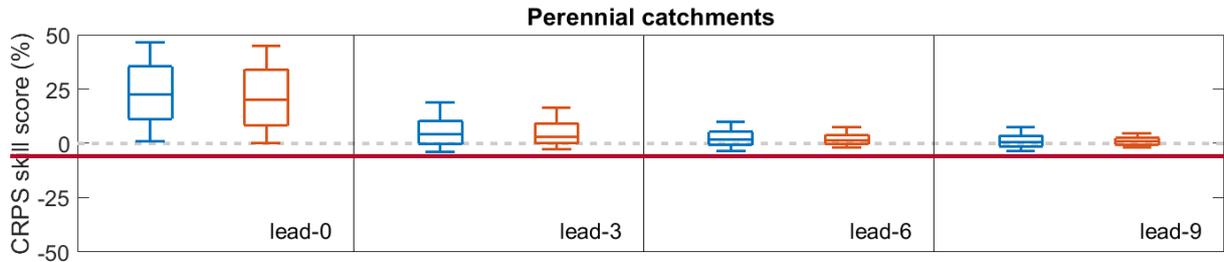
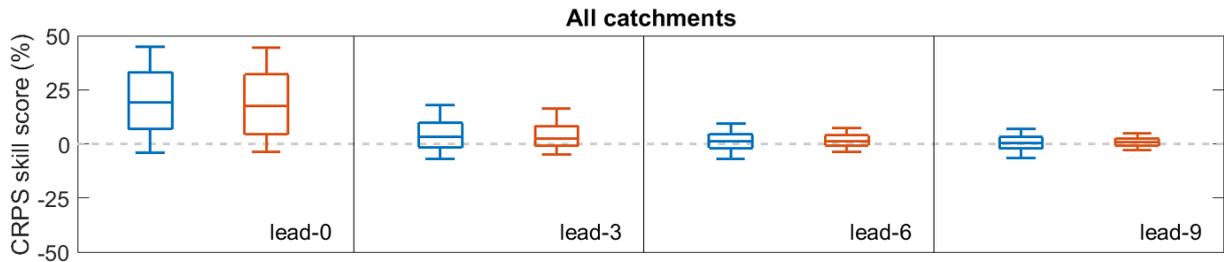


- Feb - lead 2
- Feb - lead 6
- Sep - lead 2
- Sep - lead 6
- K-S significance

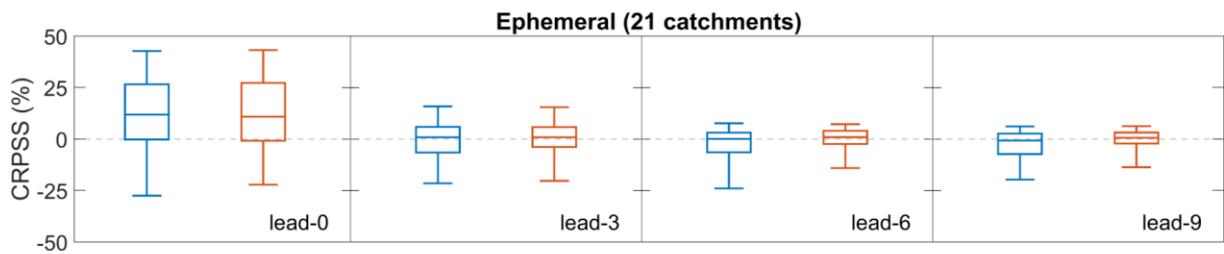
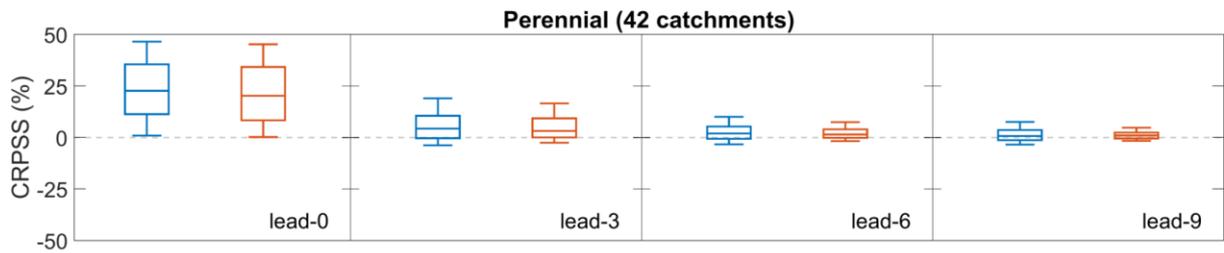
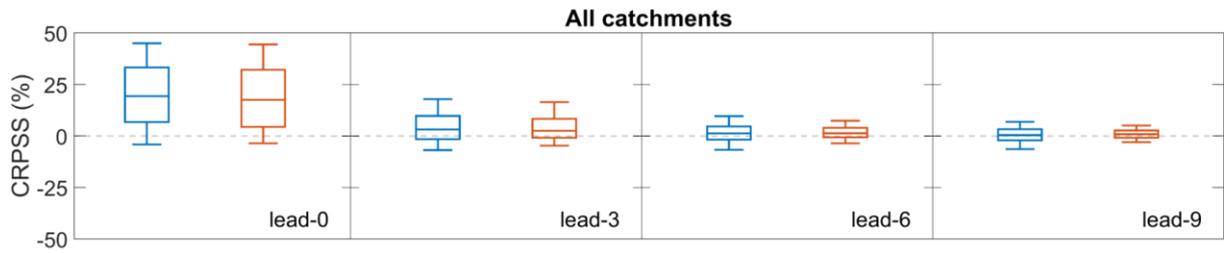




**Figure 45:** PIT plots for selected months and lead times (colours) for FoGSS forecasts (base case). Points are PIT values, crosses are pseudo-PIT values. Centre map gives catchment locations. Dashed lines give Kolmogorov-Smirnoff significance tests at 5%.



— POAMA-CBaM — ESP

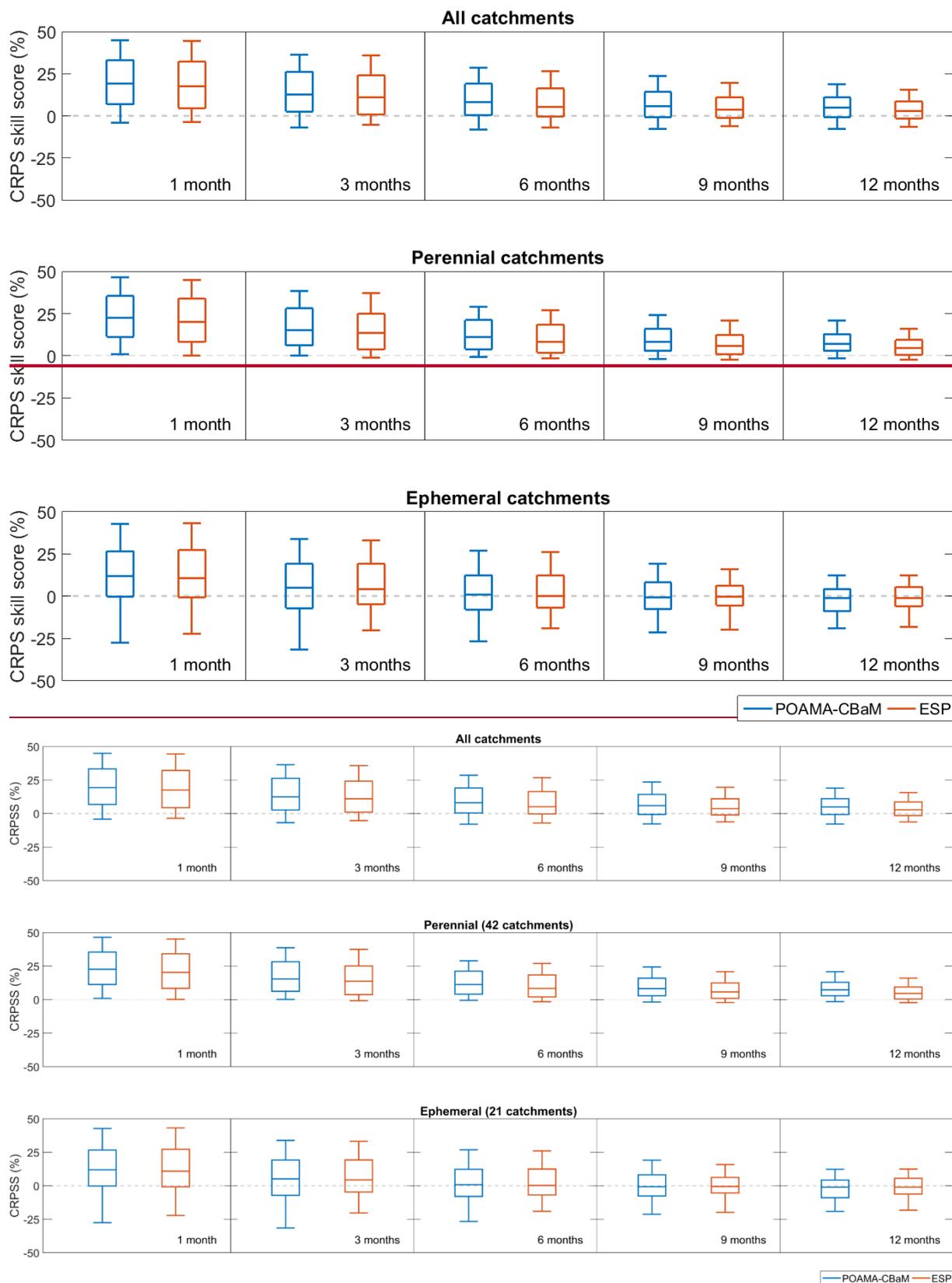


— POAMA-CBaM — ESP

---

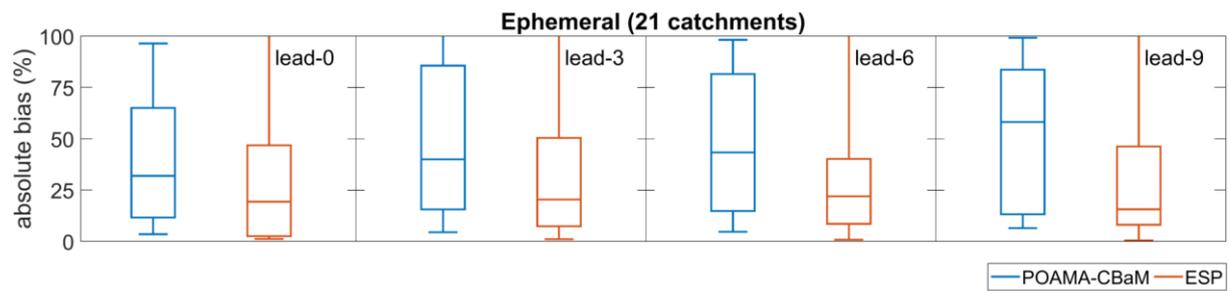
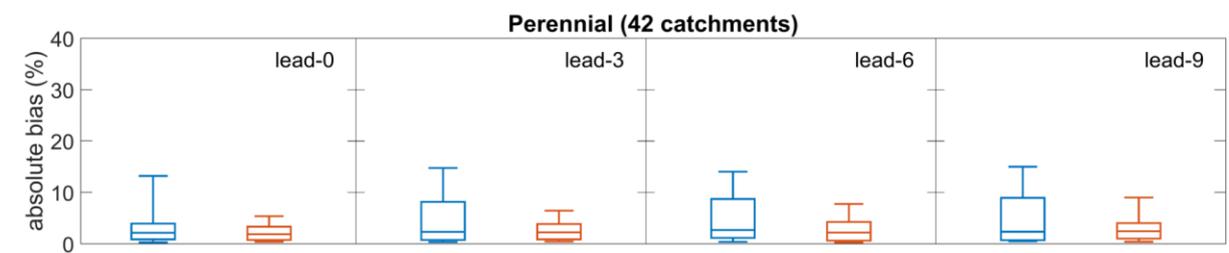
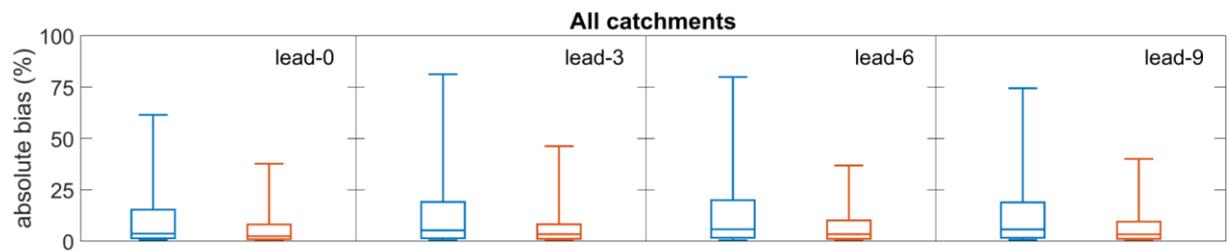
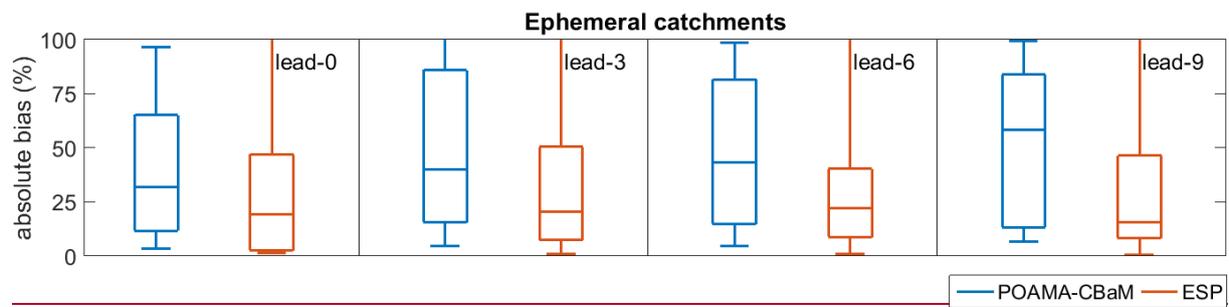
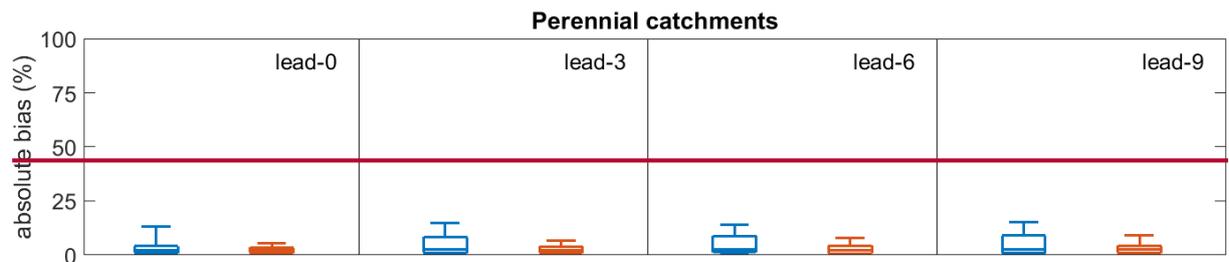
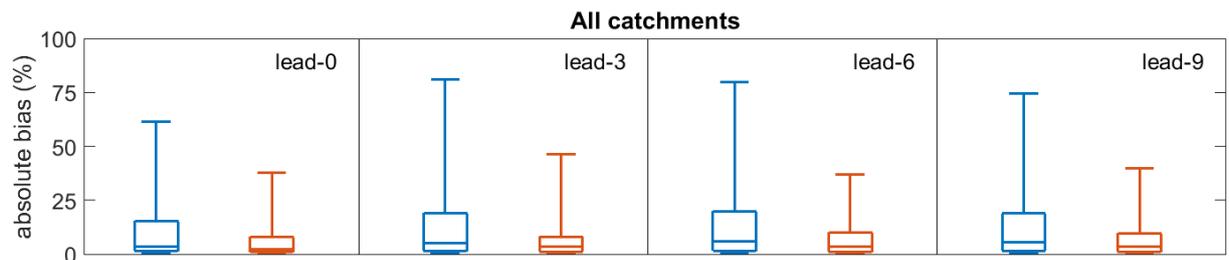
Figure 56: Forecast skill (CRPSS) for all 63 catchments by lead time for FoGSS forecasts forced by forecast (POAMA-CBaM) and historical (ESP) rainfall. For each lead time, forecast skill is summarised for all months and catchments with box and whisker plots. Boxes show interquartile range with the median, whiskers give 10<sup>th</sup> and 90<sup>th</sup> percentiles. Top panel shows all catchments, middle panel shows perennial catchments, and bottom panel shows ephemeral catchments.

5



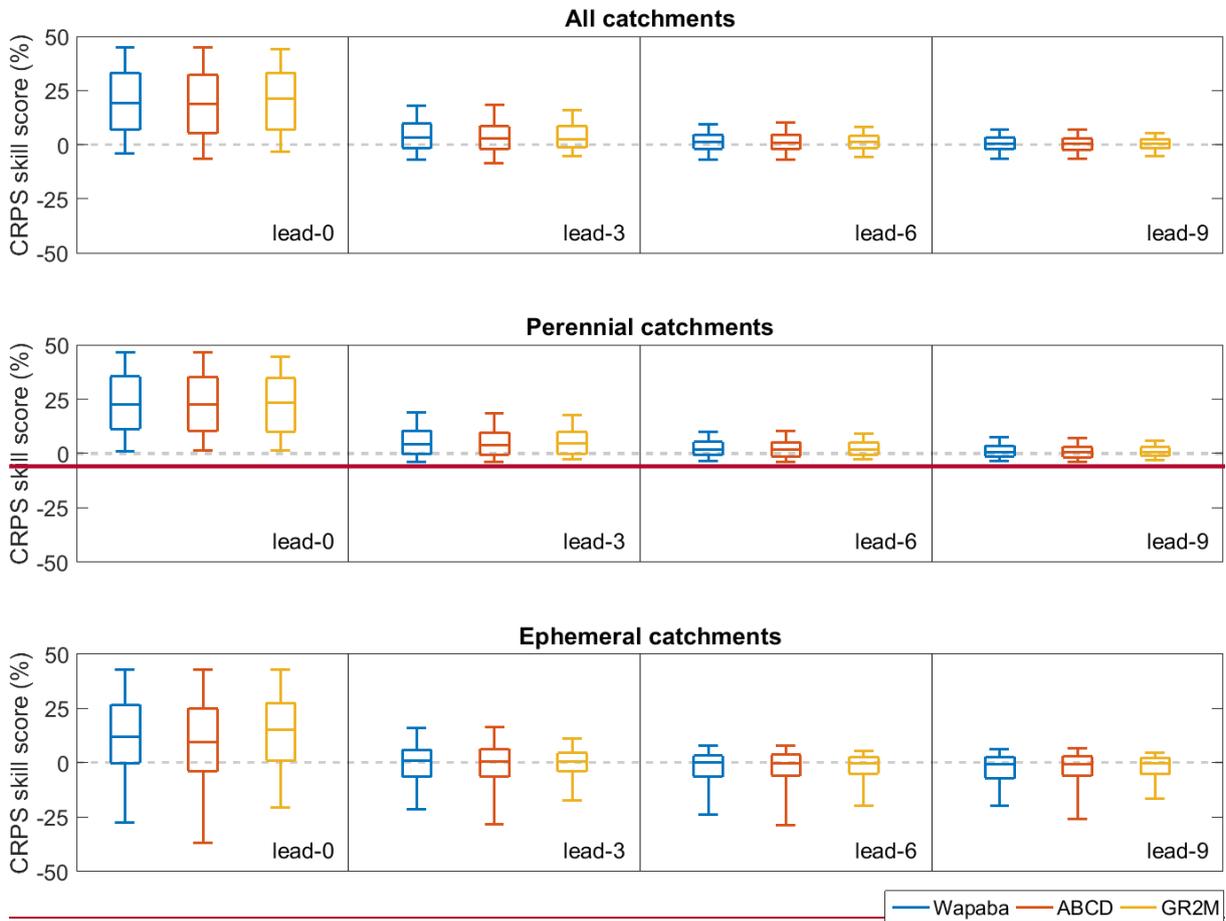
**Figure 67:** Forecast skill for 63 catchments by forecast accumulation period for FoGSS forecasts forced by forecast (POAMA-CBaM) and historical (ESP) rainfall. For each lead time, forecast skill is summarised for all months and catchments with box and whisker plots. Boxes show interquartile range with the median, whiskers give 10<sup>th</sup> and 90<sup>th</sup> percentiles. Top panel shows all catchments, middle panel shows perennial catchments, and bottom panel shows ephemeral catchments.

5



5

Figure 78: Absolute bias in 63 catchments by lead time for FoGSS forecasts forced by forecast (POAMA-CBaM) and historical (ESP) rainfall. For each lead time, absolute bias is calculated for all months, and then summarised for all catchments with box and whisker plots. Boxes show interquartile range with the median, whiskers give 10<sup>th</sup> and 90<sup>th</sup> percentiles. Top panel shows all catchments, middle panel shows perennial catchments, and bottom panel shows ephemeral catchments. Note the differently scaled vertical axis of middle panel.



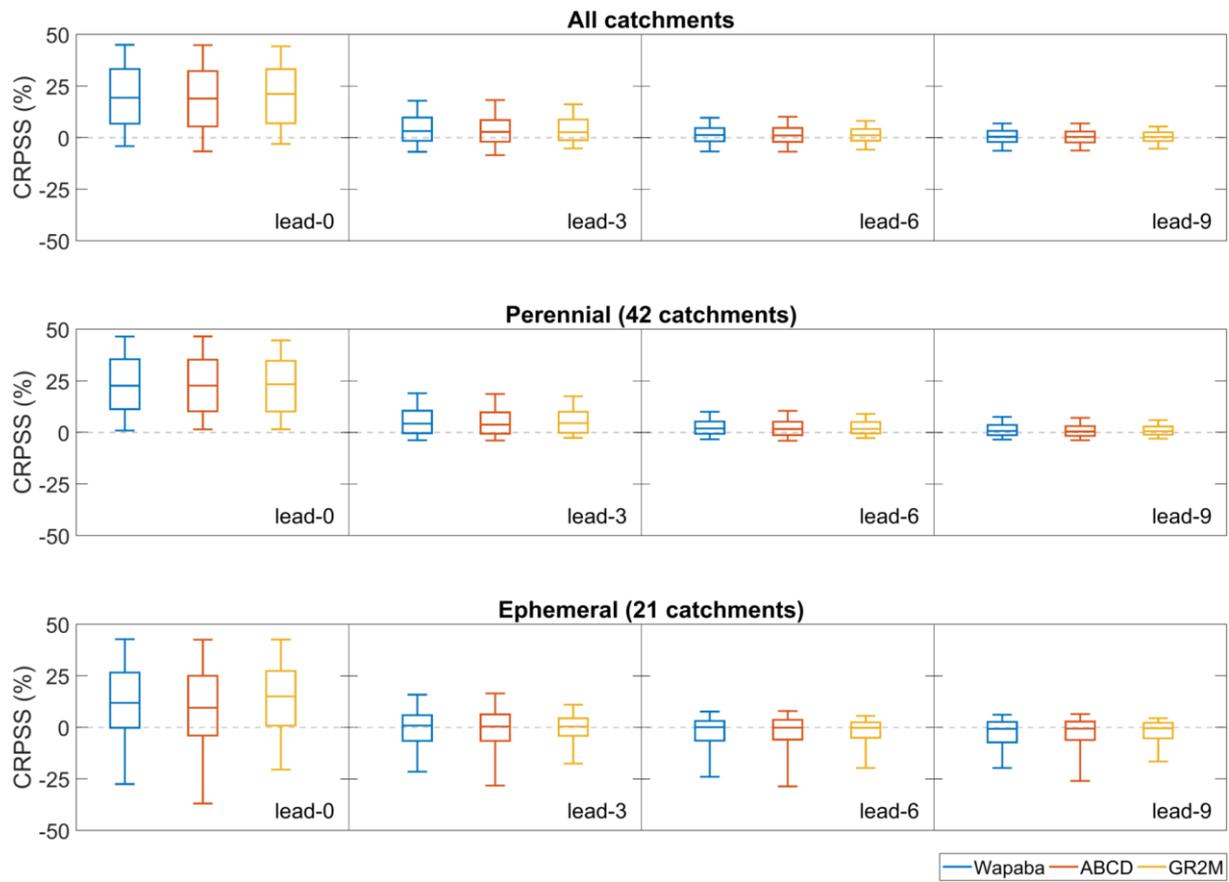
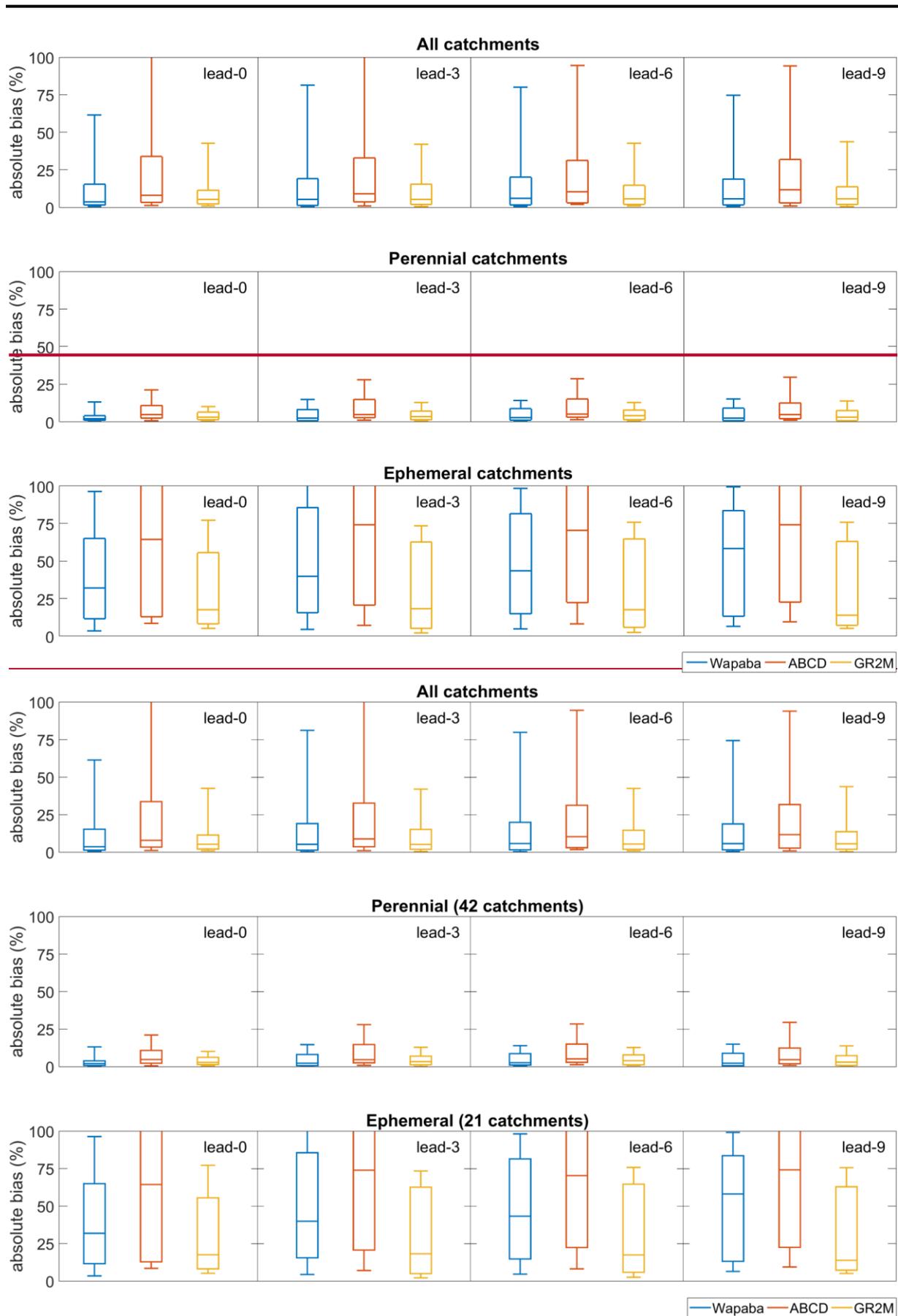


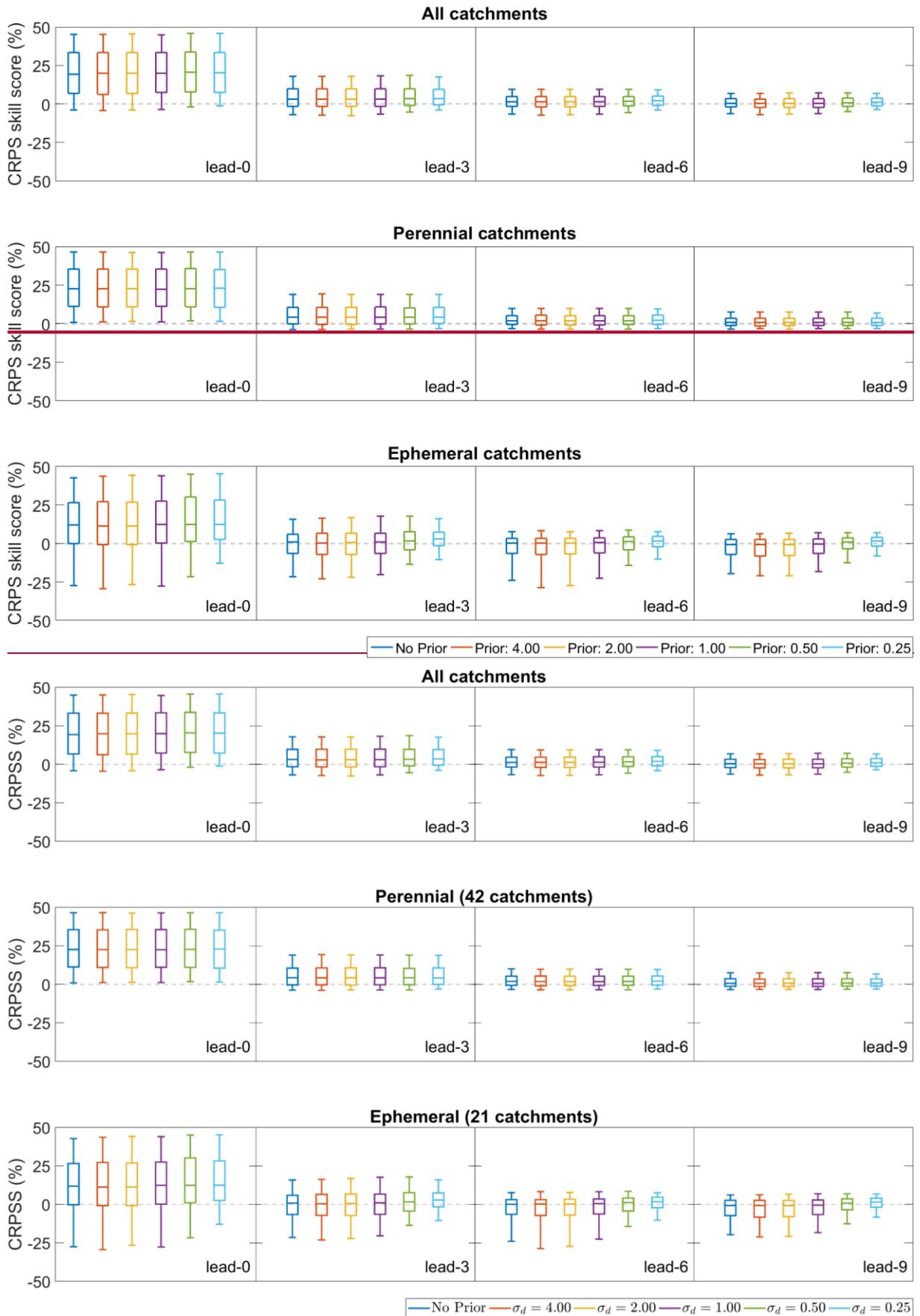
Figure 89: Forecast skill in 63 catchments by lead time for FoGSS forecasts with different rainfall-runoff models. For each lead time, forecast skill is summarised for all months with box and whisker plots. Boxes show interquartile range with the median, whiskers give 10<sup>th</sup> and 90<sup>th</sup> percentiles. Top panel shows all catchments, middle panel shows perennial catchments, and bottom panel shows ephemeral catchments.

5



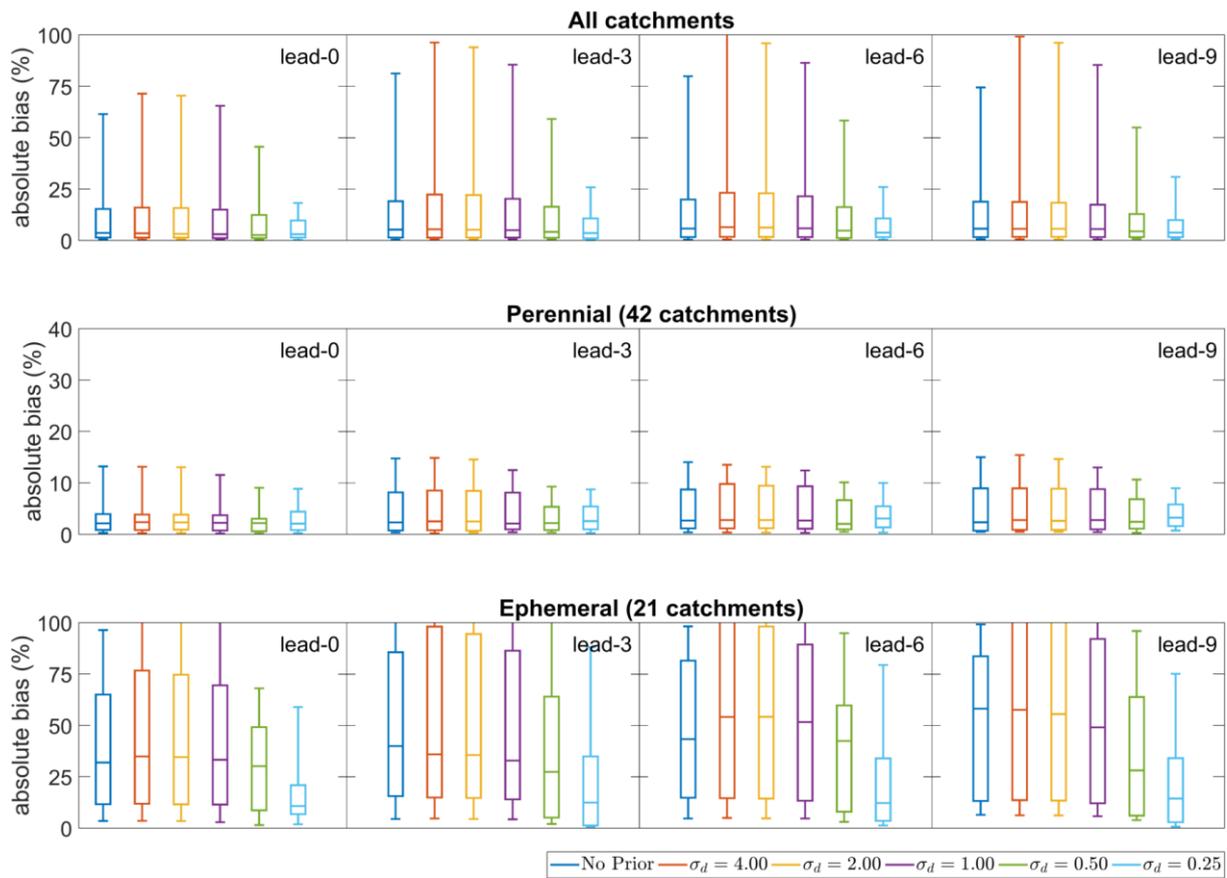
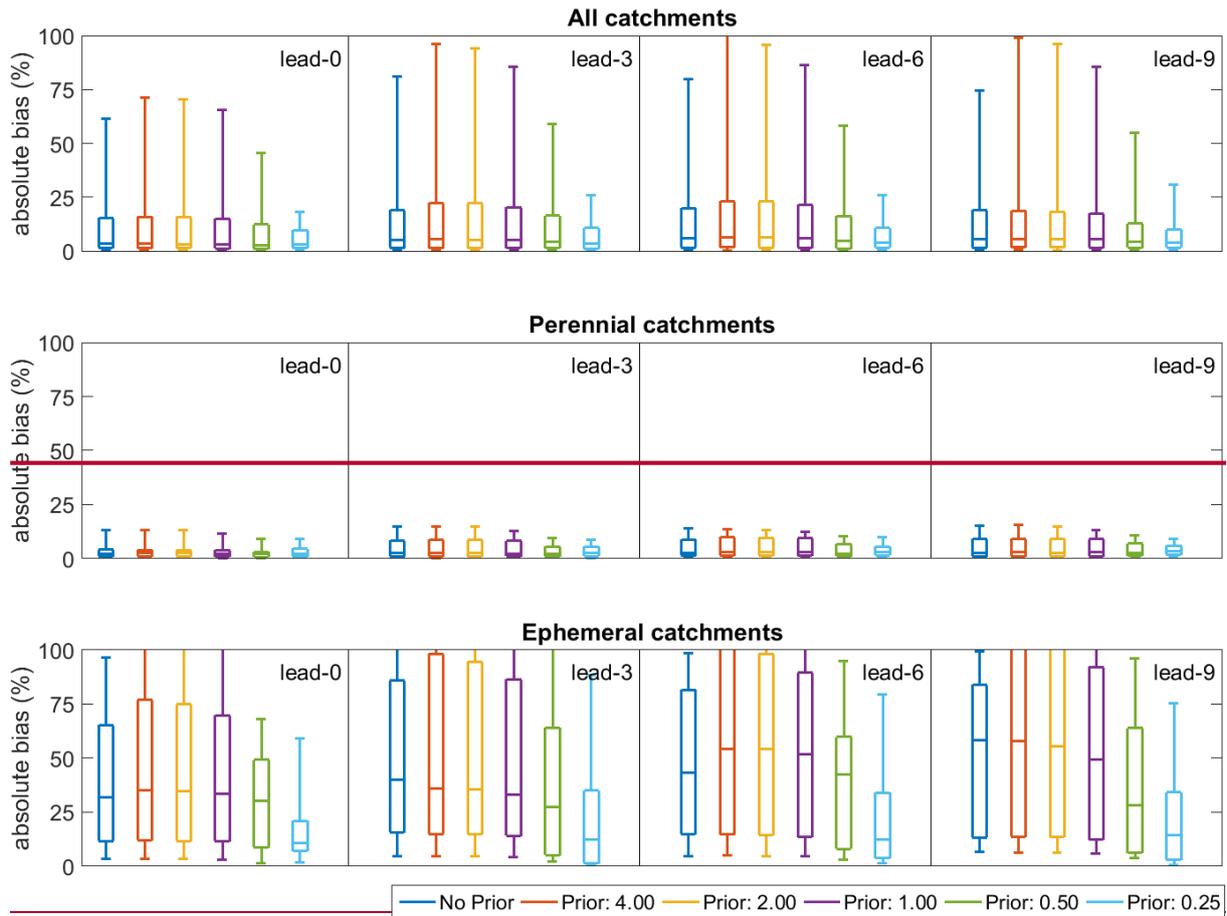
**Figure 910** Absolute bias in 63 catchments by lead time for FoGSS forecasts with different rainfall-runoff models. For each lead time absolute bias is calculated for all months, and then summarised for all catchments with box and whisker

plots. Boxes show interquartile range with the median, whiskers give 10<sup>th</sup> and 90<sup>th</sup> percentiles. Top panel shows all catchments, middle panel shows perennial catchments, and bottom panel shows ephemeral catchments.



---

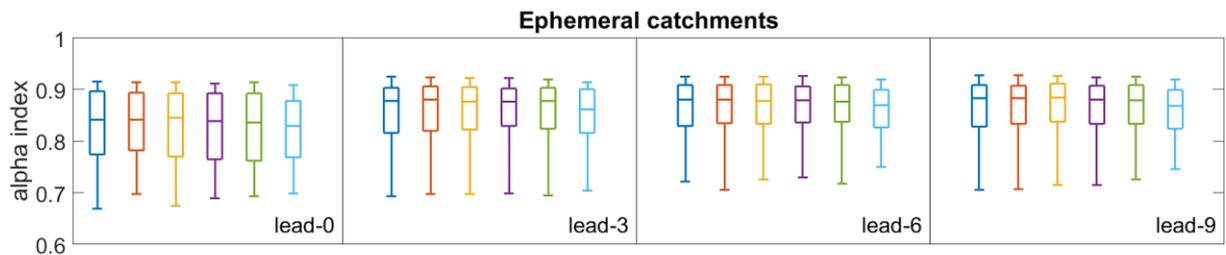
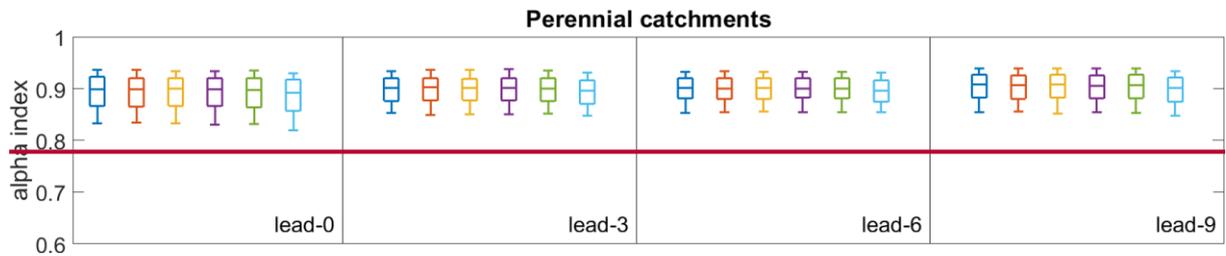
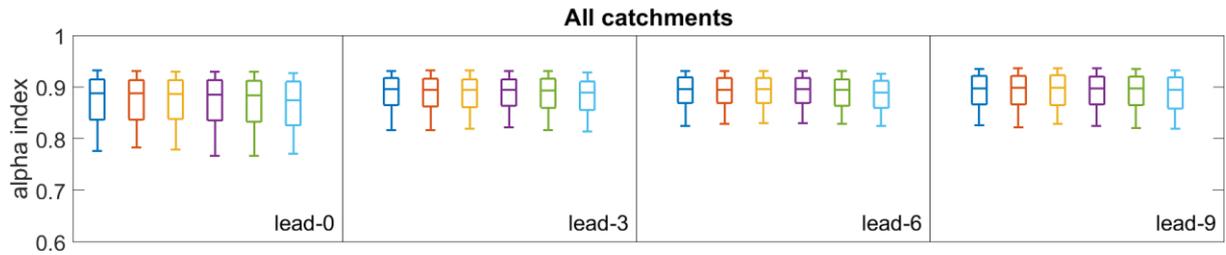
**Figure 1011:** Forecast skill in 63 catchments by lead time for FoGSS forecasts with different strength priors on the  $d$  parameter. For each lead time, forecast skill is summarised for all months with box and whisker plots. Boxes show interquartile range with the median, whiskers give 10<sup>th</sup> and 90<sup>th</sup> percentiles. Top panel shows results for all catchments, middle panel for perennial catchments only, and bottom panel for ephemeral catchments.



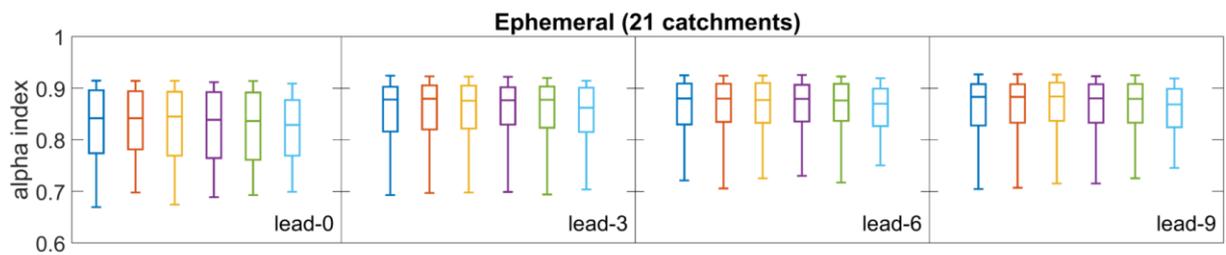
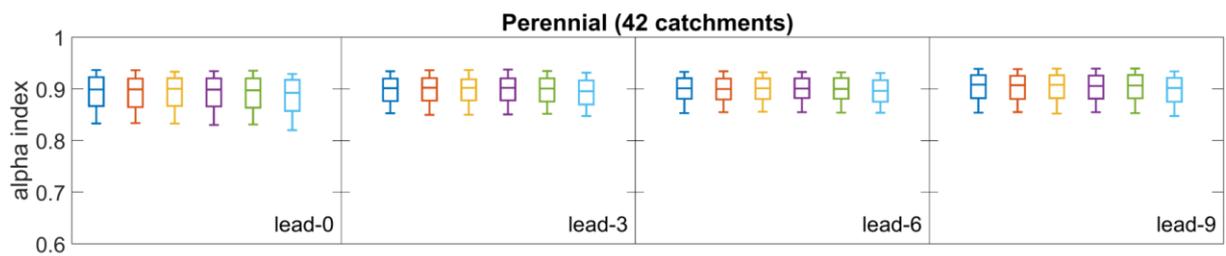
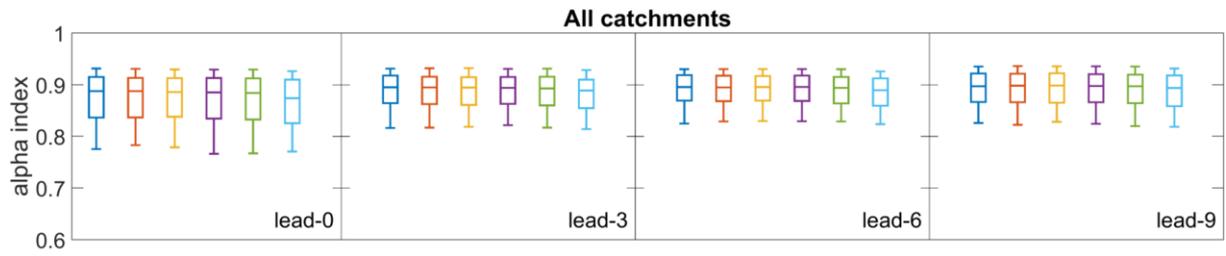
---

Figure 4412: Absolute bias in 63 catchments by lead time for FoGSS forecasts with different strength priors on the  $d$  parameter (smaller values of  $\sigma_d$  result in a stronger prior). For each lead time, forecast skill is summarised for all months with box and whisker plots. Boxes show interquartile range with the median, whiskers give 10<sup>th</sup> and 90<sup>th</sup> percentiles. Top panel shows results for all catchments, middle panel for perennial catchments only, and bottom panel for ephemeral catchments. Note the differently scaled vertical axis of middle panel.

5



— No Prior — Prior: 4.00 — Prior: 2.00 — Prior: 1.00 — Prior: 0.50 — Prior: 0.25



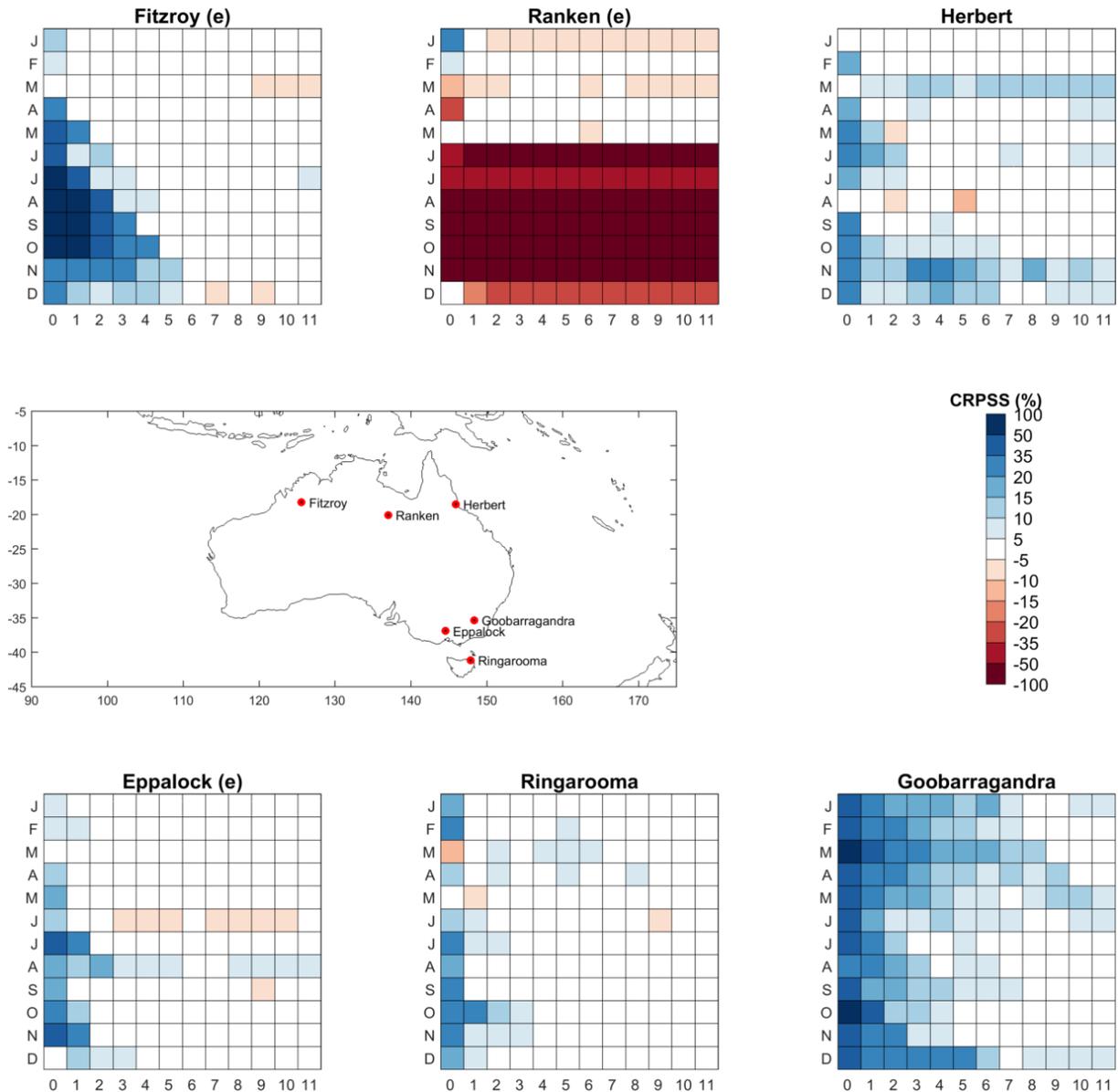
— No Prior —  $\sigma_d = 4.00$  —  $\sigma_d = 2.00$  —  $\sigma_d = 1.00$  —  $\sigma_d = 0.50$  —  $\sigma_d = 0.25$

---

Figure 12.13: Reliability (alpha index) in 63 catchments by lead time for FoGSS forecasts with different strength priors on the  $d$  parameter (smaller values of  $\sigma_d$  result in a stronger prior). For each lead time, reliability is summarised for all months with box and whisker plots. Boxes show interquartile range with the median, whiskers give 10<sup>th</sup> and 90<sup>th</sup> percentiles. Top panel shows all catchments, middle panel shows perennial catchments, and bottom panel shows ephemeral catchments.

5





**Figure 1314: Skill (CRPSS) skill scores for of FoGSS forecasts forced generated with the GR2M hydrological model and with a strong prior on the  $d$  parameter of  $\sigma_d = 0.25$ . Ephemeral catchments are denoted by (e). Target months are shown on the vertical axes, and target lead times on the horizontal axes. Centre map gives catchment locations.**

## Appendix A

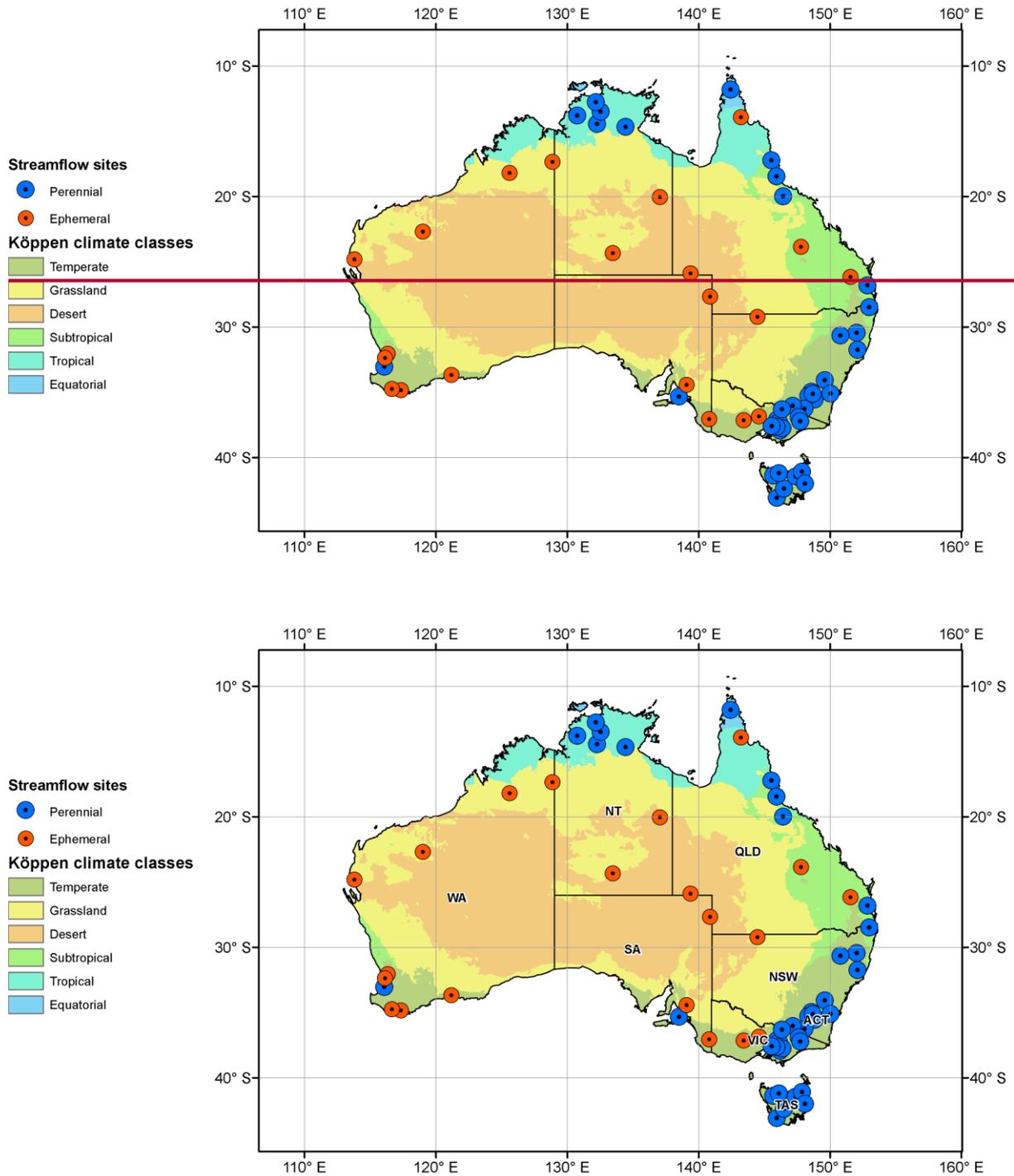
Table A1: List of catchments

Gauge Name	Gauge number	State*	Perennial/ Ephemeral	Zero flow s (%)	Area (km <sup>2</sup> )	Lon	Lat	Missin g data (%)
<del>Cotter River above Gingera Abererom bie River above Hadley No.2</del>	<del>4107304120 66</del>	<del>ACTNS W</del>	<del>PerennialPerenn ial</del>	<del>0.01- 8</del>	<del>1304634</del>	<del>148.82149- 6</del>	<del>- 35.59 - 34.11</del>	<del>1.22-7</del>
Abercrombie River above Hadley No.2	412066	NSW	Perennial	1.8	1631	149.6	- 34.11	2.7
Burrinjuck Dam inflows	Inflows site	NSW	Perennial	0.0	10,310	148.58	- 35.00	0.0
<del>Corang River at HockeysBurrinjue k Dam inflows</del>	<del>215004Inflo ws site</del>	<del>NSWNS W</del>	<del>PerennialPerenn ial</del>	<del>0.90- 0</del>	<del>16610,31 0</del>	<del>150.03148- 58</del>	<del>- 35.15 - 35.00</del>	<del>4.20-0</del>
Corang River at Hockeys	215004	NSW	Perennial	0.9	166	150.03	- 35.15	4.2
Cotter River above-Gingera	410730	NSW	Perennial	0.0	130	148.82	- 35.59	1.2
Goobarragandra River above Lacmalac	410057	NSW	Perennial	0.0	668	148.35	- 35.33	0.3
Goodradigbee River above Wee Jasper (Kashmir)	410024	NSW	Perennial	0.0	990	148.69	- 35.17	10.1
Murray River above Biggara	401012	NSW	Perennial	0.0	1,257	148.05	- 36.32	3.9
Nowendoc River above Rocks Crossing	208005	NSW	Perennial	0.0	1,893	152.08	- 31.78	1.8
Paroo River at Willarra Crossing	424002	NSW	Ephemeral	19.9	35,239	144.46	- 29.24	0.0
Wollomombi River above Coninside	206014	NSW	Perennial	0.0	377	152.03	- 30.48	3.0
Daly River at Mount Nancar	G8140040	NT	Perennial	0.0	47,100	130.74	- 13.83	4.8
Hugh River at South Road Crossing	G0050115	NT	Ephemeral	32.3	3,140	133.43	- 24.35	4.2
Katherine River at Railway Bridge	G8140001	NT	Perennial	0.0	8,640	132.26	- 14.46	3.3
Ranken River at Soudan Homestead	G0010005	NT	Ephemeral	72.4	4,360	137.02	- 20.05	8.3
Roper River at Red Rock	G9030250	NT	Perennial	0.0	47,400	134.42	- 14.70	14.6
South Alligator River at El Sherana	G8200045	NT	Perennial	0.0	1,300	132.52	- 13.53	7.7
West Alligator River at Upstream Arnhem Highway	G8190001	NT	Perennial	0.0	316	132.17	- 12.79	3.9

Barron River above Picnic Crossing	110003A	QLD	Perennial	0.0	239	145.54	-17.26	0.0
Burdekin River above Sellheim	120002	QLD	Perennial	0.6	36,230	146.43	-20.01	7.1
Coen River above Coen Racecourse	922101B	QLD	Ephemeral	5.1	170	143.2	-13.94	6.0
Diamantina River at Birdsville	A0020101	QLD	Ephemeral	26.8	119,034	139.37	-25.91	3.3
Dulhunty River at Dougs Pad	926002A	QLD	Perennial	2.3	332	142.42	-11.83	8.0
Herbert River above Abergowrie	116006B	QLD	Perennial	0.0	7,486	145.92	-18.49	0.0
Namoi River above North Cuerindi	419005	QLD	Perennial	0.0	2,532	150.78	-30.68	1.5
Nogoa River at Craigmore	130209A	QLD	Ephemeral	21.3	13,876	147.76	-23.88	13.4
Richmond River above Wiangaree	203005	QLD	Perennial	0.0	712	152.97	-28.51	0.6
Stuart River at Proston Rifle Range	136304A	QLD	Ephemeral	16.7	1,546	151.55	-26.18	41.1
Stanley River above Peachester	143303A	QLD	Perennial	0.0	102	152.84	-26.84	0.6
Cooper Creek at Cullyamurra Water Hole	A0030501	SA	Ephemeral	20.2	232,846	140.84	-27.7	0.0
Myponga US Dam and Road Bridge	A5020502	SA	Perennial	0.3	71	138.48	-35.38	4.5
North Para River at Penrice	A5050517	SA	Ephemeral	11.1	121	139.06	-34.46	3.3
Davey River above D/S Crossing Rv	473	TAS	Perennial	0.0	698	145.95	-43.14	0.9
Florentine above Derwent	304040	TAS	Perennial	0.0	445	146.5	-42.44	0.0
Hellyer River above Guilford Junction	61	TAS	Perennial	0.0	101	145.67	-41.42	0.3
Leven River at Bannons Bridge	314207	TAS	Perennial	0.0	499	146.09	-41.25	1.8
North Esk River at Ballroom	318076	TAS	Perennial	0.0	363	147.38	-41.49	0.9
Ringarooma River at Moorina Bridge	30	TAS	Perennial	0.0	517	147.87	-41.13	8.0
Swan River at the Grange	302200	TAS	Perennial	0.0	448	148.08	-42.05	7.4
Avoca River at Amphitheatre	408202	VIC	Ephemeral	9.5	83	143.4	-37.18	0.0
Lake Eildon	Inflows site	VIC	Perennial	0.0	3,877	145.97	-37.16	0.0
Lake Eppalock	Inflows site	VIC	Ephemeral	25.6	1,749	144.56	-36.88	0.0
Goulburn River above Dohertys	405219	VIC	Perennial	0.0	700	146.13	-37.33	4.5
Grace Burn Creek	Inflows site	VIC	Perennial	0.0	25	145.55	-37.64	0.0

Lake Hume	Inflows site	VIC	Perennial	1.5	11,754	147.15	- 36.08	0.0
Mosquito Creek above Struan	A2390519	VIC	Ephemeral	10.7	1,249	140.77	- 37.09	0.0
Mitta Mitta River above Hinnomunjie	401203	VIC	Perennial	0.0	1,518	147.61	- 36.95	4.5
O'Shannassy Reservoir	Inflows site	VIC	Perennial	0.0	127	145.81	- 37.68	0.0
Ovens inflows	Inflows site	VIC	Perennial	0.0	7,515	146.33	- 36.36	0.0
Tanjil Junction inflows	85266	VIC	Perennial	0.0	289	146.19	- 37.98	0.0
Thomson Reservoir	Inflows site	VIC	Perennial	0.0	487	146.37	- 37.79	0.0
Tambo River above Swifts Creek	223202	VIC	Perennial	0.0	899	147.72	- 37.26	3.9
Upper Yarra Reservoir	Inflows site	VIC	Perennial	0.0	337	145.92	- 37.68	0.0
Watts River inflows	Inflows Site	VIC	Perennial	0.0	104	145.55	- 37.64	0.0
Darkin River at Pine Plantation	616002	WA	Ephemeral	50.8	665	116.29	- 32.07	0.9
Denmark River at Mt Lindesay	603136	WA	Perennial	5.1	502	117.31	- 34.87	0.0
Deep River above Teds Pool	606001	WA	Ephemeral	17.0	468	116.62	- 34.77	0.0
Fitzroy River at Fitzroy Crossing Br	802055	WA	Ephemeral	4.2	46,133	125.58	- 18.21	0.3
Gascoyne River at Nine Mile Bridge	704139	WA	Ephemeral	60.1	74,432	113.77	- 24.83	0.0
Harvey River above Dingo Road	613002	WA	Perennial	0.6	148	116.04	- 33.09	2.4
Marillana Creek at Flat Rocks	708001	WA	Ephemeral	29.8	1370	118.97	- 22.72	0.0
Ord River at Old Ord Homestead	809316	WA	Ephemeral	26.2	19,513	128.85	- 17.37	3.3
Serpentine Reservoir	Inflows site	WA	Ephemeral	7.1	664	116.10	-32.4	0.0
Young River at Neds Corner	601001	WA	Ephemeral	42.6	1,893	121.14	- 33.71	0.0

\*Abbreviations of Australian state names: ACT = Australian Capital Territory; NSW = New South Wales; NT = Northern Territory; QLD = Queensland; SA = South Australia; TAS = Tasmania; VIC = Victoria; WA = Western Australia. Locations of states are shown in Figure A1.



**Figure A1: Distribution of gauge/inflows sites showing ephemeral/perennial streams**

## Appendix B

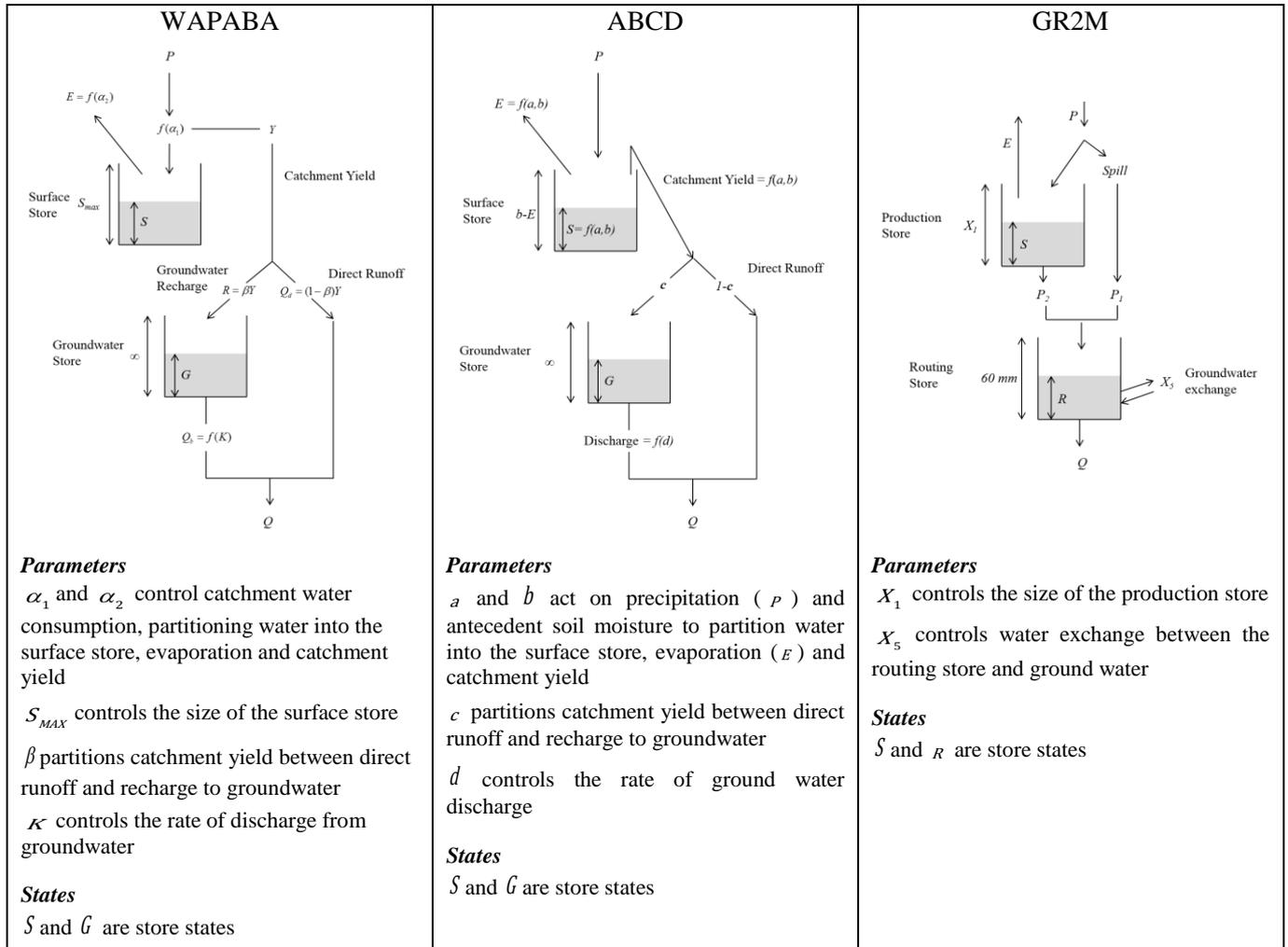


Figure B1: Hydrological model structures and parameters

## Appendix C

### C.1 Posterior density used for estimation of Stage 2 parameters

We assume that residuals are normally distributed. Because parameters in Eq. (32) vary by month,  $i = 1, 2, \dots, 12 = \text{month}(t)$ , it follows that the residual distribution also varies by month:

$$\begin{aligned} z_o(t) &= z_2(t) + \varepsilon_2(i) \\ \varepsilon_2(i) &\sim N(0, \sigma_2^2(i)) \end{aligned} \quad (C1)$$

The Stage 2 parameters to be estimated (from Eq. 32 and Eq. C1) are denoted as

$$\theta_2(i) = \{d(i), \mu(i), \sigma_2(i)\}. \quad (C2)$$

We maximize the posterior density

$$p(q_o(t) | \theta_2(i), q_1(t)) \propto p(d) \prod_{t \in T_i} J_{q \rightarrow z} N(z_o(t) | z_2(t), \sigma_2(i)), \quad (C3)$$

10 where  $q_1$  is the simulation produced with Stage 1, the Jacobian (from the log-sinh transformation),  $J_{z \rightarrow q}$ , is given by

$$J_{z \rightarrow q} = \frac{1}{\tanh(a + bq_o(t))} \quad (C4)$$

and  $p(d)$  is the a prior on the  $d$  parameter (Section 34.5.4),

$$p(d) = d \sim N(0, \sigma_d^2). \quad (C5)$$

If  $q_o(t) = 0$ , then the likelihood term  $J_{q \rightarrow z} N(z_o(t) | z_2(t), \sigma_2(i))$  in Eq. (C3) is substituted with the normal cumulative

15 probability  $\Phi\left(\frac{z_c - z_2(t)}{\sigma_2(i)}\right)$ , where  $z_c = TF(0)$  is the log-sinh transformed value of zero (see Eq. 1).