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.

Response: Thanks very much for the careful review and the positive feedback.

Specific comments and technical corrections:

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

Response: We will add 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.

-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?

Response: We will mention PE as well as rainfall. 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.

-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.

Response: We will shorten this section, as suggested. We will also reorganise it to emphasise the prior, by separating the hydrological model from the error model, and moving the discussion of parameter estimation after the description of the error model. This means the reader arrives much more directly at the description of the bias-correction, and we will also more directly flag its use in the experiments with the prior.

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

Response: We will add a brief note explaining that this means the variance is not constant.

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

Response: We will explicitly note that this term denotes the transformation.

-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?

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 will add this explanation.

-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 will add this.

-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 will note this in text.

-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 will add a figure as the reviewer suggests. While this duplicates information in the following Figure 3, we agree with the reviewer that this makes the paper easier to follow. The new figure will appear as follows:

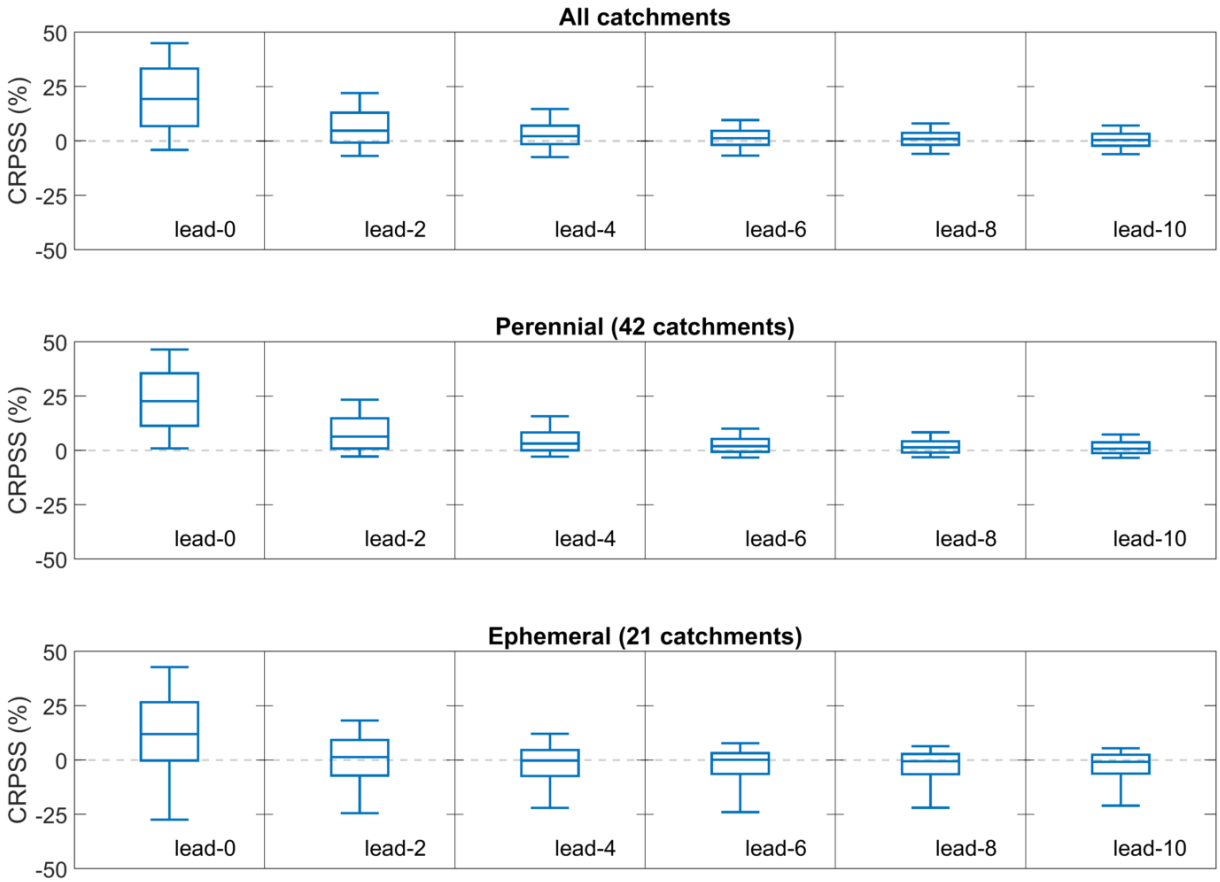

Figure 3: Forecast skill (CRPSS) for all 63 catchments by lead time for the FoGSS base case. For each lead time, forecast skill is summarised for all months and catchments with box and whisker plots. Boxes show interquartile range with the median, whiskers give 10th and 90th percentiles. Top panel shows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

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

Response: We will move this criterion earlier in the paper when introducing CRPSS.

-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."

-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 will define neutral skill as $-5\% > CRPSS < 5\%$ when we introduce CRPSS. As the reviewer points out, this means our statement is no longer strictly true (the light orange colours). We will amend our statement to note these instances of slightly negative skill.

-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 will include a corrected figure in the new manuscript. The other panels differ negligibly from the figure incorrectly included in the original manuscript, and so are consistent with the other text and conclusions. The corrected figure is as follows:

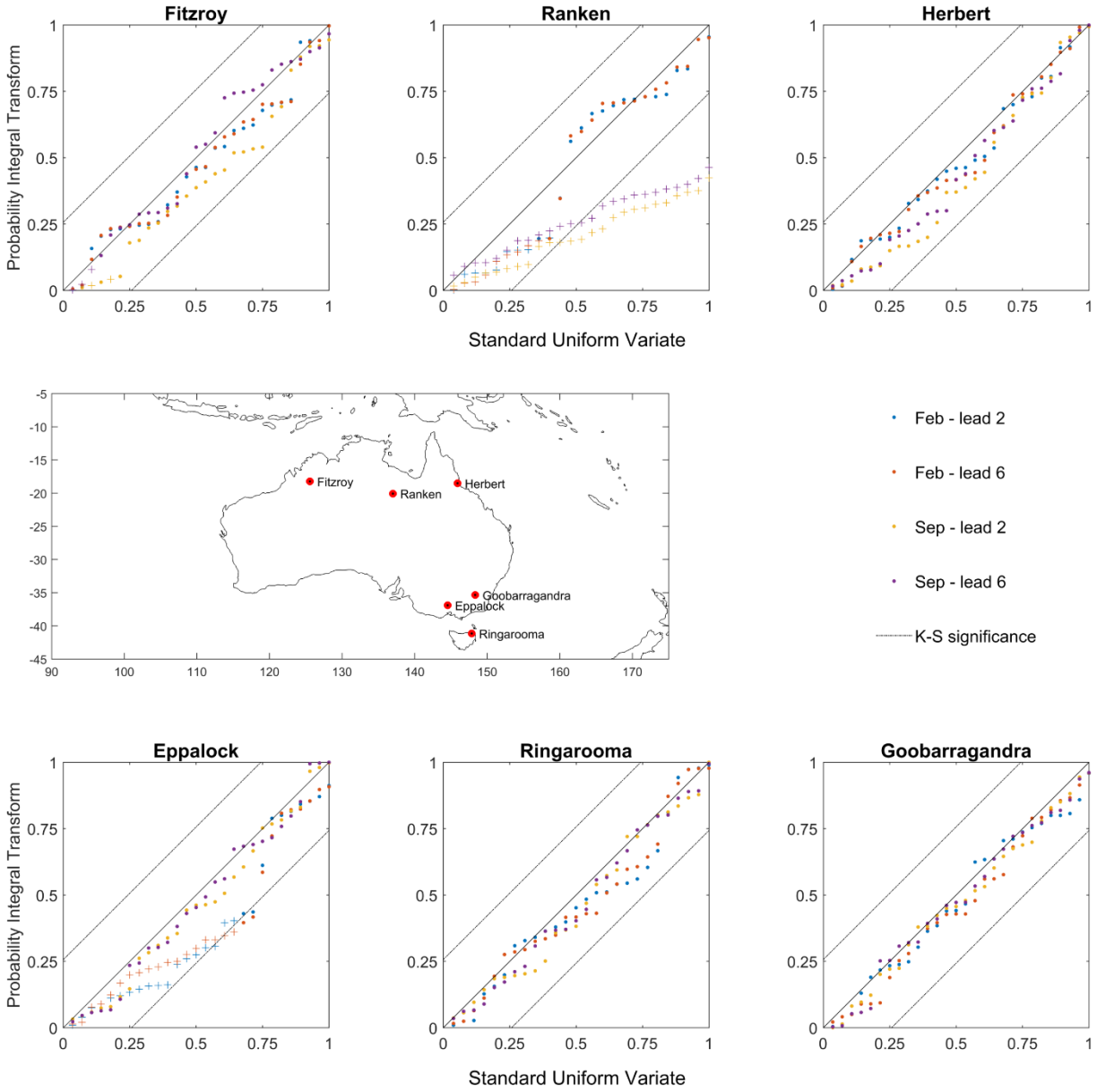

Figure 5: PIT plots for selected months and lead times (colours) for FoGSS forecasts (base case). Points are PIT values, crosses are pseudo-PIT values. Centre map gives catchment locations.

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

Response: Yes, this is true for all catchments – we will note this.

-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 was confusingly phrased. We will rephrase to avoid the confusion.

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

Response: We will rephrase this to be clearer. 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.

-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.

Response: We agree that this is not a strong justification for future research, and we will remove 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.

Response: We will note this here, and also add this explanation to the figure captions.

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

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".

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

Response: We agree. We will add the qualification "although these changes were sometimes very slight".

-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 will add a brief explanation of this problem as the reviewer suggests, to better acknowledge the difficulties we face in these catchments.

-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 will add a note on the reduction in skill in the Fitzroy in JAS.

-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 will strengthen this argument as suggested.

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

Response: We will add 'Wapaba' to the figure

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

Response: Thanks - we will add a small panel above each plot showing rainfall statistics, as follows:

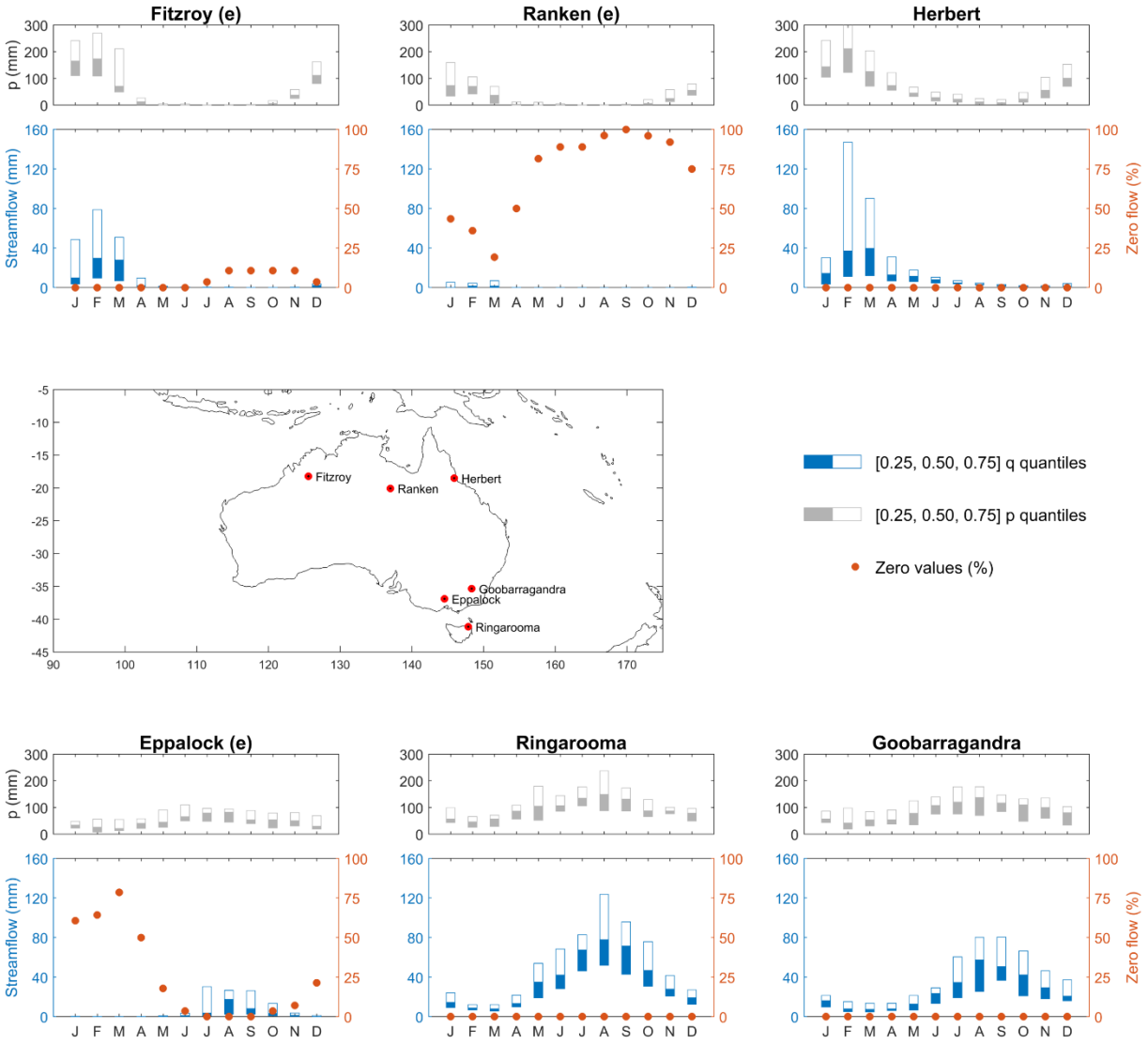

Figure 2: Catchment characteristics of six case study catchments. Ephemeral catchments are denoted by (e). Left axis shows monthly streamflow (q) and rainfall (p) characteristics, with bars showing interquartile range and median flows for the period 1982-2009. Right axis shows proportion of zero flows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

-Figures 3 and 13: In the results you mention that FoGSS performs adequately when CRPSS >= 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'. We will add a formal definition of what we mean by 'neutral skill' (basically, within 5% of zero) when we introduce CRPSS, and note that we consider performance neutral or positively skillful to be a requirement of FoGSS.

-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 will correct 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.

Response: We will indicate 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 will add notches.

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

Response: We will rescale the middle panels on these plots to show more detail.

-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 will indicate this explicitly in the legend, and add an explanatory note that smaller values of sigma_d indicate a stronger prior.

Typographical corrections

-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 will correct.

References

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.

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.