Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC3
https://doi.org/10.5194/hess-2021-69-RC3, 2021

Comment on hess-2021-69

Anonymous Referee #3

Referee comment on "Bias-correcting input variables enhances forecasting of reference crop evapotranspiration" by Qichun Yang et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-69-RC3, 2021

Title: Bias-correcting individual inputs prior to combined calibration leads to more skillful forecasts of reference crop evapotranspiration

Author(s): Qichun Yang et al.

MS No.: hess-2021-69

This paper focuses on the comparison of two calibration strategies to provide short-term reference crop evapotranspiration (ETo). ETo forecasting is still a relatively new area of research, in Australia and elsewhere, and has received more attention in the past few years. Skilful ETo forecasts in Australia would help support efficient water use and water management. Two strategies to calibrate ETo forecasts have emerged: i) the calibration of raw ETo forecasts and ii) bias-correcting input variables first before calibrating ETo forecasts. Little work to date compares the two approaches, it is unclear which method might be more advantageous or skilful. This paper therefore addresses a topical subject with a large audience interest.

I have some reservations regarding some methodological choices and justifications (purpose and inclusion of experiment 3 and 4), as well as a lack of interpretations of the results overall. I recommend revision to strengthen this paper.

The authors re-grid the weather forecast variables of ACCESS-G2 to match the timeframe and resolution of the gridded data AWAP. They perform four experiments: experiments 1) and 2) are based on the ETo anomaly and climatological mean, whereas experiment 3 and 4) use the ETo values directly. Furthermore, experiment 1) and 3) use raw inputs to calculate and calibrate ETo forecasts whereas experiments 2) and 4) first bias-correct inputs before ETo calibration. The SCC calibration method is used for ETo forecast while a quantile mapping method is used to bias-correct input forecasts. The authors evaluate the forecasts using three metrics for the theoretical assessment of bias, reliability and accuracy. Overall results suggest that the second strategy (bias-correction of inputs before ETo calibration) provides more skilful forecasts.

Major comments:

Methodology:

- P4 section 2.3: Why not compare the calibration method used SCC to other methods tested in the literature which would enable to place this work in context to other studies on ETo forecasting?
- Presentation of summary statistics. Why not use boxplots to present overall statistics and across lead times (for example next to figure 4 and so on)? Reliability diagrams for particular ETo thresholds would be helpful to communicate when the forecasts are reliable.
- Authors present experiments 1-4 in the method but then only present some results one experiment 3) and 4) in the last section of results (CRPSS in 3.5). No explanation are provided of why calibration 3) and 4) are only briefly introduced. Why is there a big gap with no results on calibration 3) and 4) on the bias and reliability results? Could the authors please expand on the purpose of including these at all in? At p17 l350-354, 'a further evaluation based on a different way of implementing the calibration demonstrate similar improvements in calibrated ETo forecasts with the adoption of bias-correction to input variables'. Is the purpose of including experiment 3) and 4) to test the generalisation of the method? If so, it needs to be clearly stated and justified earlier.

Methodological choices for evaluation:

- P7 l 180-185 : why choosing the absolute bias and over a relative measure e.g. percentage bias? This choice makes it difficult to compare the magnitude of the errors in the results across different variables and studies. For example, figure 1 shows a bias between -2 to 2mm/day which does not seem like much compared to other input variables such as precipitation. Figure 3 with a range of -0.1 to 0.1 seems very small. Conversely, percentages are used for the correlation coefficient in Figure 6 so why not use it for the bias?
- P8 l205-2015: why is climatology used as reference forecast for the skill score? In hydrological forecasting persistence is typically used for short lead times, whereas climatology would be used for longer lead times, see fore example (Pappenberger, Ramos et al. 2015). Could you please expand and justify the choice of reference forecast used and implication of interpretation of results?
- P8 l214. Why is the definition of CRPSS using percentage? As far as I am aware, most studies do not present the CRPSS in terms of percentage, could you please comment on the reason of this choice with references that also use percentages and if there is any advantages?

Analysis and interpretation of results:

- P11 l259-261: why the higher difference in bias in approaches for the Nothern Territory? How does this relate to the biases, errors and assumptions of the NWP? Is it correlated to the biases of specific input variables? How is it correlated to the nonlinear relationship in calculatint ETo? Why are the biases most pronounced for shorter lead times? Please comment.
- P13 l282-285: Why lowest score of correlation coefficient in northern Territory? Is it linked to the NWP (and if so how?) or is it linked to observations? E.g. differneces in observations compared to rest of country?
- P14 l294-297: The geographical patterns of the correlation performance is very similar to the patterns of the bias performance. Could you please comment why and if the reasons are the same? Are these related to either the NWP or observations?
- P16 l320-328. Please comment on why the accuracy has larger differences in terms of geographical patterns than for the bias and PIT performance which had very strong localised performance.
- P16 l329: Results on calibration 2 and 4: what is the comparison between 2 and 4? Why are these only addressed in the evaluation of forecast accuracy section? Why is there no mention of these for the bias and reliability evaluation? I suggest changing the section order and moving this section first. Then, add a sentence in the bias and reliability section to explicitly communicate what results of experiment 3) and 4) are not presented and why.

Discussion:

- There are little to no direct comparison of results and calibration work presented here to any previous methods or studies (which were mentioned in the introduction). To address a research closure, please put the work presented in this paper in context with other studies applying strategy 1 and strategy 2.
- It is unclear whether authors recommend the use of experiment 2) or 4), when and why. In that sense, I question again the inclusion of these experiments without further elaborating and discussing these results.

Structure:

- The introduction is well structured and appropriately present previous work studies and existing strategies.
- The title is a bit lengthy, authors could consider shortening it.
- As noted above, I suggest authors consider the order of results presented in the context of results from experiment 3) and 4).

Minor comments:

P4 l106: I suggest adding a diagram clearly explaining steps and differences of procedure between the calibration experiments.

P3 l68: '…pressing need to investigate.' Please expand why it is pressing?

P3 l74: Calibrate should be calibrate with small cap letter.

P3 l80-84: There are many efforts to develop downscaling methods, please comment on

what was been done here to downscale ACCESS-G2 to the AWAP grid. Why not scaling AWAP to the match the forecast grid?

P4 l100: please add a comment that SCC model will be described in section 2.3.2

P5 l134 climatological means or mean? Please rephrase and clarify this sentence.

P6 l165: Why are only 100 members drawn, is there any difference with a varying number of ensemble members for forecast reliability? Is there a need or a reason to verify accumulated Eto forecast values across lead times (as is often the case for streamflow forecasting)? Please comment.

P8 l225: 'wind speed is higher than 1m/s than the reference in Australia'. Could you please translate that in terms of percentage so that this statement can be more easily compared to other locations.

P18 l380 'NWP outputs have been increasingly used for ETo forecasting.' For which applications? Please finish the sentence.

P18 l385 Addition 'of' in … skill 'of' the calibrated ETo forecasts.

References:

Pappenberger, F., M. H. Ramos, H. L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller and P. Salamon (2015). "How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction." Journal of Hydrology **522**: 697-713.