Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2019-189-RC2, 2019

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Hybrid climate datasets from a climate data evaluation system and their impacts on hydrologic simulations for the Athabasca River basin in Canada" *by* Hyung-Il Eum and Anil Gupta

Anonymous Referee #2

Received and published: 29 June 2019

This paper developed a methodological framework to generate a hybrid dataset by ranking multiple climate datasets according to six performance measures. The hybrid dataset was assessed by driving a hydrologic model as proxy validation and comparing the simulation results to some existing available climate datasets. Using the Athabasca River basin in Canada as an example, the framework was applied to produce three different hybrid datasets based on five climate datasets and the VIC model was used to simulate streamflow of five sub-basins.

General Comments:

The objective of this paper is straightforward and this paper has a clear structure with a solid methodology. Given the low density of station network in Canada and the high impacts of climate forcing on hydrologic modelling, the proposed framework is innovative and essential in providing an alternative to produce potentially better quality of input data for improving hydrologic simulations. Having said this, there are three major concerns needed to be addressed in this paper:

(1) Performance of multiple climate datasets against the ground stations

It seems to me that the performance of the climate datasets could be affected by the interpolation method used to estimate the values at the AHCCD stations. The authors used the inverse distance squared weighting method to obtain the estimated values from all the gridded products (P8L4-5), and the Township data was shown to outperform other climate datasets for all performance measures except P_bias. I am struggling to square away in my mind that the interpolation method might favour towards the Township data because the Township data also employed inverse distance weighting and used the same (or similar) set of ECCC stations to generate the data. Thus, the Township data would most likely rank first among the climate datasets because the major deficiency of the data lies from the difference between the raw station data it used and the adjusted data in AHCCD, while the deficiencies of other climate datasets come from interpolation method, numbers of stations used, and the errors arising from the use of additional information/numerical models.

(2) Superior performance of hybrid dataset over multiple existing climate datasets

I am a bit skeptical about the claim that the performance of hybrid datasets was 'superior' when compared to other five climate datasets (P1L30-31). By saying 'superior' the results should be far better than the others (e.g. a NSE value of 0.8 as compared to 0.5). In this study, I would argue that the overall performance of hybrid datasets was only marginally better than some of the existing climate datasets in most of the sub-basins. The performance of hybrid dataset, Hybrid(Rind), was even worse than

ANUSPLIN at Hinton station (Figure 11). Overall, the hybrid datasets only provided comparably good NSE values as the other climate datasets.

(3) Creditability of hybrid dataset in improving hydrologic simulations

Even though the hybrid datasets provided comparably good NSE values as the other climate datasets or even higher NSE values, when examining the hydrograph in Figure 12, one can find that there are four obvious large underestimation of the peaks in 2009, 2010, 2014, and 2015 simulated by using the hybrid datasets (purple lines and potentially green lines as well). Could the authors explain what happened at Hinton station? Could the authors also show the hydrographs at other stations to see whether similar situations happened in other sub-basins?

The claim that the two hybrid datasets performed better in terms of accuracy and precision in the proxy validation (P18L28-29) could be a bit misleading. In this study, it was coincidentally that the hybrid datasets (either based on single or multiple variables) were dominantly generated from one particular climate dataset in all sub-basins (except Clearwater when using precipitation as the variable). If the authors show the breakdown of the first ranked number of grid cells for each climate dataset in each sub-basin (just like in Table 5), I would guess that over 90% of the grid cells at Hinton came from ANUSPLIN when considering the performance measures of multiple variables (Figure 9c) and almost 99% of grid cells at Pembina came from the Township data. In this regard, I would argue that the performance of the hybrid datasets shown in Figure 11 was highly resemble to the performance of the climate dataset that was dominantly generated from. I would also argue that the optimal parameter sets of the hybrid datasets would be the same (or very similar) as that of dominant climate dataset. Have the authors checked the optimal parameter sets of the hybrid datasets and the five climate datasets? Will the calibrated parameter sets of the hybrid dataset (Hybrid(Rmul)) the same as the parameter sets of Township data at Pembina, for instance?

The creditability of generating a hybrid dataset might not be fully assessed at sub-basin

scale, especially when the hybrid datasets were generated mainly from one particular climate dataset. I think a better assessment to reveal the usefulness of the hybrid datasets was to calibrate the model at whole-basin scale for this particular basin (e.g. calibrating at Fort McMurray using 07DA001 station). In this case, the hybrid dataset is better mixed by different climate datasets for different parts of the whole basin, thus reducing the chance of one particular climate dataset being dominant in the data generation process.

The authors should, therefore, discuss the limitations of such data generation system (based on ranking) and comment on the worthiness of generating a hybrid dataset when one particular climate dataset being dominant in the composition of the hybrid dataset.

Specific Comments:

P8L4: How many grid points were used in the inverse distance squared weighting?

P8L5-6: The AHCCD stations have different starting and ending points and percentage of missing values. How did the authors take care of these? Did the authors calculate the performance measures using a common period?

P8L21-24: please also define i

P9L5: The authors mentioned 20% of all AHCCD stations were selected here but five nearest AHCCD neighbours were shown in Figure 2. Which one is correct?

P11L27-29: What did the authors mean by “the number of gridded climate datasets was optimized”? Please elaborate.

P12L3: Why were only two hybrid datasets from the Rind and Rmul? Didn’t the authors rank for precipitation and temperature separately (Rind)? (P10L12-13) I think there would be two sets of hybrid datasets based on Rind, one for precipitation only and one for temperature only, as shown in Figures 9 and 10.

P12L5: I assume that in this study the authors used the same version and the same VIC setup as described in Eum et al. (2017). Could the authors clarify the sources of the other meteorological variables (e.g. wind speed) required in the VIC model? Did the authors use the meteorological variables from NARR for all the climate datasets and the hybrid datasets? Did the authors use the wind speed data of the Township data itself, for instance?

P12L21: What were the calibration and validation periods in this study?

P13L3-7: Table 3 shows the 'average' performance of each climate datasets. How did the results indicate under- or over-estimation of 'extreme' precipitation? Please explain.

P13L25: Should it be >800 mm/year?

P14L16-19: It would be better to show the breakdown of the first-ranked number of grid cells and their percentages for each sub-basin as well because the authors calibrated and validated the VIC model at sub-basin scale.

P15L12: Again, I think there should be three different hybrid datasets.

P15L19: Same as the above comment. If only two hybrid datasets were implemented, could the authors clarify which Rind was used?

P15L20-22: It was shown that NARR did not perform well in temperature (Section 3.2). Why did the authors still combine CaPA precipitation with NARR temperature for the proxy validation? Would such combination be unfair to CaPA performance? The performance of CaPA should be assessed by combining with the temperature data of all other climate datasets.

P16L4-9: What was the validation period for other climate datasets? For better comparison with CaPA, I think the authors could show the NSE results calculated from 2010 to 2016 for all the climate datasets.

P16L12: The VIC performance using NARR did not get positive NSE even after calibration. This means that no optimal parameter sets could be identified using NARR and the parameter sets could be anywhere in the parameter space. I wonder how such unidentified parameter sets could still produce fair NSE values when it was used with other climate datasets (Figure 11). I would expect a long lower whisker (just like the case in CaPA). Otherwise, I would think that the errors from the climate dataset were greatly compensated by the parameter uncertainties during the calibration. Could the authors explain what happened at Pembina?

Remarks:

P2L20: should be "may not produce" not "may not produces"

P4L4: should be "the aims of this study are" not "the aims of this study is"

P4L32: should be "Peace River" not "Peasce River"

P9L5: should be "criteria" not "citeria"

P19L19-21: please update the reference. Christensen and Lettenmaier (2007) has been published in HESS already, not HESSD.

P20L16-18: missing the name of journal

P20L19: should be "Dibike, Y." not "Yonas, D."

Table 6: should there be two hybrid datasets of Rind?

Figure 1: should be "precipitation" not" preciptation"

Figure 3: this figure could be combined with Figure 8 to reduce the numbers of figures (or the other way round). Otherwise, the authors should provide the geographical information about the basin on the map to facilitate the understanding of the international readers (e.g. elevation, latitude and longitude, a mini map showing the geographical location of the basin in Canada). Also, it would be better to show the river network of

the basin.

Figure 9: there are too much unnecessary white space between the labels, the figures, and the legend. Consider squeezing the white space to make the figure more compact.

Figure 11: should there be two hybrid datasets of Hybrid(Rind)?

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-189, 2019.