<<General Comments>>

(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.

((Reply)) The authors appreciate the reviewer's valuable comments. This study investigated the performance of the five gridded climate datasets at the AHCCD stations. Among the gridded climate datasets, station-based datasets (i.e., ANUSPLN and Alberta Township) employed different numbers of observed (raw) station data depending on data availability in a given year except for PNWNAmet that set a common period from 1945 to 2012 for all stations included in the interpolation. While ANUSPLIN used the Canada-wide archive (raw) station data collected only by ECCC, the Alberta Township data has been produced on the basis of the archive (raw) station data collected by ECCC and other agencies including Alberta Environment and Parks (AEP), and Alberta Agriculture and Forestry (AF) over Alberta. Therefore, one of the possible reason for outperformance of Township dataset might be the difference in the numbers of stations (i.e. station density) employed to produce the gridded climate datasets. This point has been added to the discussion section of the manuscript, as follows:

"Among the station-based gridded climate datasets, the Township dataset outperformed other station-based gridded climate datasets. As PNWNAmet set a common period from 1945 to 2012 for all stations included in the interpolation, many stations might be left out in the data generation processes. While ANUSPLIN used the Canada-wide archive (raw) station data collected by only ECCC, the Alberta Township data has been produced on the basis of the

archive (raw) station data collected by ECCC, AEP, and AF over Alberta. Therefore, one of the possible reason for outperformance of Township dataset might be the difference in the numbers of stations (i.e. station density) employed to produce the gridded climate datasets." (P23L2-P23L9)

(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.

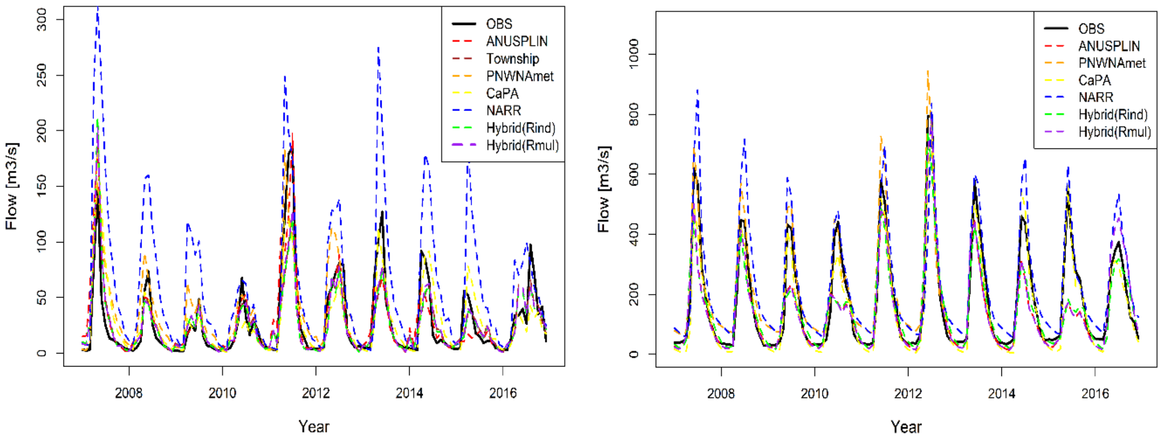
((Reply)) The authors agree with the reviewer's comment and agreed that 'superior' word may not be suitable in this context. In Table 6, the two hybrid climate datasets performed well with comparably better NSE values than other climate datasets, especially at Pembina, Clearwater, and Firebag located in the middle and lower reaches. From multiset-parameter hydrologic simulations shown in Figure 11, however, the hybrid climate datasets provided higher precision and accuracy in most of the stations except for Hinton as the reviewer pointed out. Therefore, the authors replaced the word "superior" to "utility" in the modified manuscript.

(3) Creditability of hybrid dataset in improving hydrologic simulations

(3-1) 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?

((Reply)) The authors appreciate the reviewer's valuable comment. The two hybrid climate datasets were produced by combining with the existing gridded climate datasets based on the performance measures. Therefore, it has an intrinsic limitation that the performance of the hybrid dataset for a basin may resemble that of a climate dataset that is dominantly ranked first for the basin. As commented in (3-2) below, ANUSPLIN was dominantly ranked first for Hinton, consequently the hydrographs of ANUSPLIN and the hybrid datasets were similar to each other as shown in the figure below. In addition, the authors present a monthly hydrograph for Pembina where the Township data was dominantly ranked first for this basin. The hydrograph of the two hybrid climate datasets (green and purple dashed lines) are highly similar to that of Township (brown dashed line). The authors addressed the limitation in the discussion section.

"*In Figure 12, the hybrid climate datasets underestimated the peak flows (in 2009, 2010, 2014, and 2015) at Hinton, and hydrograph is similar to the hydrograph produced by ANUSPLIN data set that dominantly ranked first in this watershed. On contrary, the hydrograph of the hybrid climate datasets at Pembina resembles that of Township that is dominantly ranked first in Pembina (refer to Table 5). These results indicate that the hybrid climate dataset has the intrinsic limitation that the performance of the hybrid dataset for a basin may closely resemble that of the climate dataset that is dominantly ranked first for the basin. However, the utility of the hybrid climate dataset can be clearly found at a whole-basin scale for a large watershed, as the added values of the hybrid climate dataset in sub-basins can be cumulated to the main stem at the downstream in the watershed." (P23L18-P24L2)*



(a) Hinton (b) Pembina

Figure 12. Monthly observed and simulated hydrographs from the gridded climate datasets at (a) Hinton and (b) Pembina

(3-2) 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.

((Reply)) The authors appreciate the reviewer's excellent comment. As mentioned in (3-1) above, the performance of the hybrid climate dataset is similar to that of an existing climate dataset which is dominantly ranked first for a sub-basin, and the utility of the hybrid climate dataset can be clearly demonstrated when it is applied for simulations at the whole basin scale. However, this study confirmed that the hybrid climate dataset provides a better representation of historical climatic conditions as different watersheds have different dominant gridded climate data and the proposed methodology helps to identify the appropriate dominant climate data in derived hybrid dataset. Further, as suggested by reviewer, we calibrated the VIC model for larger watersheds (i.e. Fort McMurray and Eymundson) to provide additional simulation results. The table below shows the NSE values calculated for ANUSPLIN and Hybrid (R_{ind}) at few hydrometric stations in the main stream of the Athabasca River. The result shows that as the size of watershed increases, hybrid climate dataset start performing better than the existing gridded climate dataset (in this case ANUSPLIN). This is mainly due to the fact that as watershed area increases, the derived hybrid climate data set is no longer dominate by a single grided dataset. Due to the limitation of computational capacity, initially only five sub-basins were selected for proxy validation.

Performance in Nash-Sutcliffe Efficiency (NSE)

No	Station name/ID	Drainage area (km^2)	ANUSPLIN		Hybrid	
			Calibration	Validation	Calibration	Validation
1	Hinton / 07AD002	9,760	0.85	0.82	0.83	0.76
2	Windfall / 07AE001	19,600	0.80	0.72	0.80	0.76
3	Athabasca / 07BE001	74,600	0.78	0.69	0.77	0.78
4	Fort McMurray / M07DA001	133,000	0.77	0.66	0.78	0.75
5	Eymundson / S24	147,086	0.77	0.67	0.79	0.75

<<Specific Comments>>

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

((Reply)) Four points were used for the inverse distance squared weighting method.

(2) 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?

((Reply)) Yes, as the data lengths are different at each AHCCD station, we selected a common period between each AHCCD station and climate datasets, and neglected missing values to estimate performance measures (P6L22-24).

(3) P8L21-24: please also define i

((Reply)) Yes, we have defined *i* in the modified manuscript, as follows:
"G_i and O_i represent gridded and observed climate datasets at i^{th} time step, respectively" (P11L16-L17)

(4) 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?
((Reply)) There are two steps to select the nearest neighbors in RM. Firstly, 20% (of all AHCCD) stations are selected based on the nearest distance criteria. Then, the five nearest stations from them is finally selected by the minimum elevation difference criteria. Accordingly Figure 2 has been modified in the revised manuscript.

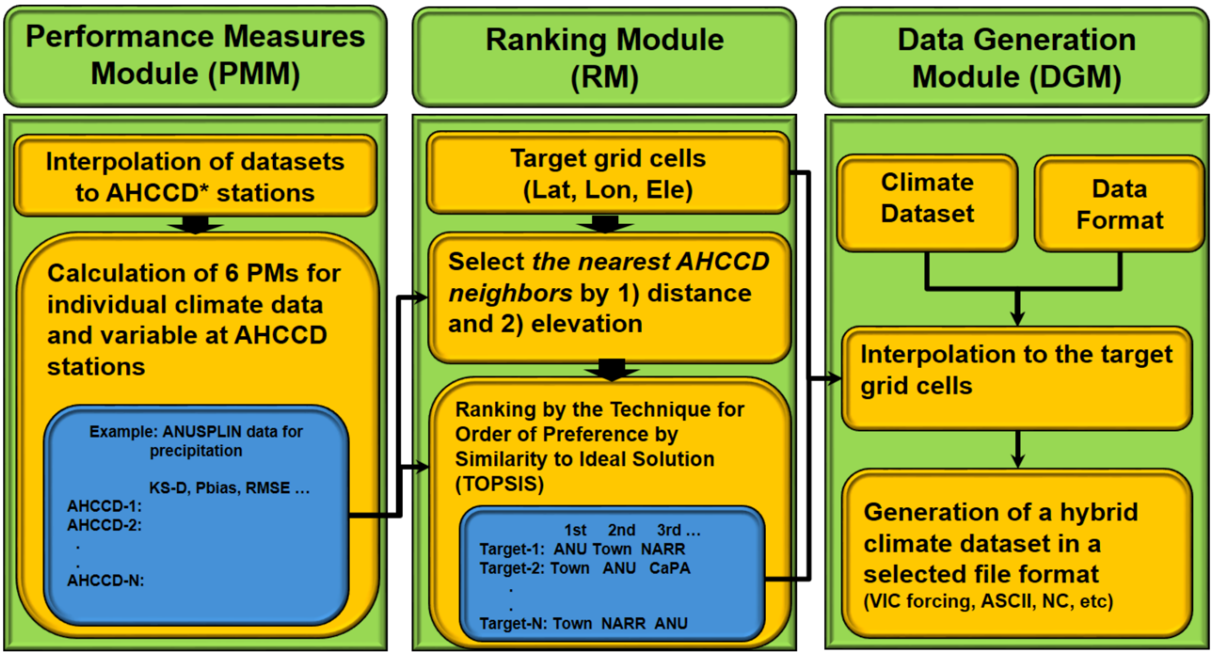

Performance Measures Module (PMM)
Interpolation of datasets to AHCCD* stations
Calculation of 6 PMs for individual climate data and variable at AHCCD stations
Example: ANUSPLIN data for precipitation
KS-D, Pbias, RMSE …
AHCCD-1:
AHCCD-2:
AHCCD-N:
Ranking Module (RM)
Target grid cells (Lat, Lon, Ele)
Select the nearest AHCCD neighbors by 1) distance and 2) elevation
Ranking by the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)
1st 2nd 3rd …
Target-1: ANU Town NARR
Target-2: Town ANU CaPA
Target-N: Town NARR ANU
Data Generation Module (DGM)
Climate Dataset
Data Format
Interpolation to the target grid cells
Generation of a hybrid climate dataset in a selected file format
(VIC forcing, ASCII, NC, etc)

Figure 2. Structure of REFRES comprised of three modules; 1) Performance Measure Module (PMM), 2) Ranking Module (RM), and 3) Data Generation Module (DGM)

(5) P11L27-29: What did the authors mean by "the number of gridded climate datasets was optimized"? Please elaborate.
((Reply)) It has been modified as below,

"In other words, a higher number of gridded climate datasets contributing to the hybrid climate dataset within a catchment was selected to evaluate the utility of the hybrid climate data relative to the existing gridded climate datasets." (P15L22-L24)

(6) 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.

((Reply)) In this study, a climate dataset consists of three variables, i.e., daily precipitation, minimum temperature, and maximum temperature. Considering the ranks from R_{ind} and R_{mul}, that is, two hybrid climate datasets was produced to be used in the proxy validation as a forcing data of the VIC model.

(7) 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?

((Reply)) This study used VIC version 4.1.2 that has the MT-CLIM package to estimate required climate variables in VIC. Hydrologic simulations were forced by only the three daily climate variables (i.e., precipitation, minimum temperature, and maximum temperature) for the proxy validation and other climate variables including wind speed were estimated by the MT-CLIM package in VIC. Next stage of this study is to expand the number of climate variables, such as wind speed, solar radiation, etc, for further improving hydrologic simulations.

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

((Reply)) The calibration and validation periods were added to the modified manuscript: *"The calibration period is 1985-1997 as in Eum et al., (2017), except for CaPA that uses the period of 2003-2009 for calibration, as CaPA covers the period from 2002 to 2016. The remaining period of total record length for each climate dataset is used for validation" (P16L7-L10)*

(9) 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.

((Reply)) The authors addressed the impacts of biases in precipitation (resulting in under or over estimation of extreme precipitation) in the discussion section of the manuscript, as follows:

"Among the station-based gridded climate datasets, the Township dataset outperformed other station-based gridded climate datasets. As PNWNAmet set a common period from 1945 to 2012 for all stations included in the interpolation, many stations might be left out in the data generation processes. While ANUSPLIN used the Canada-wide archive (raw) station data collected by only ECCC, the Alberta Township data has been produced on the basis of the arc hive (raw) station data collected by ECCC, AEP, and AF over Alberta. Therefore, one of the possible reason for outperformance of Township dataset might be the difference in the numbers of stations (i.e. station density) employed to produce the gridded climate datasets. In addition, PNWNAmet showed a positive P_{bias} for precipitation, especially in the mountainous areas, while ANUSPLIN, which employs similar thin plate spline interpolation, generated negative P_{bias}. PNWNAmet overestimated precipitation over the mountainous area, which considerably affects simulated low flows at Hinton in the ARB. Figure 12 shows the observed and simulated hydrographs from gridded climate datasets at (a) Hinton and (b) Pembina. It clearly shows that PNWNAmet highly overestimated the low and high, which is caused by overestimated precipitation in the drainage area of the sub-basins. As with PNWNAmet, NARR also overestimated the low and high flows, which is induced by the combined effects of overestimating precipitation and warm biases in cold temperature. The temperature bias of NARR is thus further confirmed and is consistent with the earlier finding of Eum et al., (2014) and Islam and Dery (2016).

In Figure 12, the hybrid climate datasets underestimated the peak flows (in 2009, 2010, 2014, and 2015) at Hinton, and hydrograph is similar to the hydrograph produced by ANUSPLIN data set that dominantly ranked first in this watershed. On contrary, the hydrograph of the hybrid climate datasets at Pembina is similar to that of Township that is dominantly ranked first in Pembina (refer to Table 5). These results indicate that the hybrid climate dataset has the intrinsic limitation that the performance of the hybrid dataset for a basin may closely resemble that of the climate dataset that is dominantly ranked first for the basin. However, the utility of the hybrid climate dataset can be clearly found at a whole-basin scale for a large watershed, as the added values of the hybrid climate dataset in sub-basins can be cumulated to the main stem at the downstream in the watershed" (P23L2-P24L2)

(10) P13L25: Should it be >800 mm/year?

((Reply)) The authors addressed this clearly as below.

“(e.g., 300 mm/year higher than the observation at the station ID 3050519)”

(11) 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.

((Reply)) The authors modified Table 5 to add the information on the first ranked climate datasets for the five sub-basins and the whole Athabasca River basin.

Table 5. First ranked number of grid cells in the five sub-basins and the whole Athabasca River Basin (ARB) and their percentage for each climate dataset considering the performance measures of individual (Case A and Case B) and multi-variables (Case C, i.e., precipitation and temperature in this study). Total number of grid cells is 22,372 at 1/32° (2~3 km)

Criteria	Basin	Climate dataset				
		ANUSPLIN	Township	PNWNAmet	NARR	CaPA
(A) Precipitation	ARB	2985 (13%)	17515 (78%)	691 (3%)	499 (2%)	682 (3%)
	Hinton	1271 (91%)	126 (9%)	0 (0%)	0 (0%)	0 (0%)
	Pembina	0 (0%)	1791 (100%)	0 (0%)	0 (0%)	0 (0%)
	Christina	0 (0%)	658 (99.5%)	3 (0.5%)	0 (0%)	0 (0%)
	Clearwater	1474 (56%)	252 (9.6%)	10 (0.4%)	682 (26%)	215 (8%)
	Firebag	129 (14%)	750 (79%)	9 (1%)	0 (0%)	64 (6%)
(B) Temperature (Min & Max Temp.)	ARB	13809 (62%)	6924 (31%)	1639 (7%)	0 (0%)	-
	Hinton	63 (5%)	77 (6%)	1257 (89%)	0 (0%)	-
	Pembina	486 (27%)	1305 (73%)	0 (0%)	0 (0%)	
	Christina	492 (74%)	169 (26%)	0 (0%)	0 (0%)	-
	Clearwater	2593	40	0	0	-

		(98%)	(2%)	(0%)	(0%)	
	Firebag	924 (97%)	28 (3%)	0 (0%)	0 (0%)	-
(C) Multi- variables	ARB	8049 (36%)	14323 (64%)	0 (0%)	0 (0%)	-
	Hinton	1271 (91%)	126 (9%)	0 (0%)	0 (0%)	-
	Pembina	0 (0%)	1791 (100%)	0 (0%)	0 (0%)	-
	Christina	109 (16%)	552 (84%)	0 (0%)	0 (0%)	-
	Clearwater	2574 (98%)	59 (2%)	0 (0%)	0 (0%)	-
	Firebag	536 (56%)	416 (44%)	0 (0%)	0 (0%)	-

(12) P15L12: Again, I think there should be three different hybrid datasets.
((Reply)) Based on the response mentioned in (6), I believe the reviewer fully understands the definition of a climate dataset.

(13) P15L19: Same as the above comment. If only two hybrid datasets were implemented, could the authors clarify which Rind was used?
((Reply)) Please refer to the response provided for (6) and (12).

(14) 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.
((Reply)) As both CaPA and NARR data sets are produced from climate model-based outputs, authors thought that it will be more logical to supplement the CaPA precipitation data with temperature data from another similar type of data set (i.e., NAAR). The performance evalution of CaPA data when supplemented with different temperature data is beyond the scope of this stsudy.

(15) 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.

((Reply)) Please refer to the reply of (8) and P21L25-P22L5.

"The validation period of CaPA is only six years from 2010 to 2016, as CaPA data are only available between 2002 to 2016. This might be a reason why CaPA produced the highest NSE (accuracy) among the climate datasets used in this study. Therefore, the results of CaPA need to be considered carefully otherwise they might be misleading. In this context, the CaPA dataset was excluded from further assessment of the precision and accuracy even though all of the results of CaPA were included in Figure 11 for reference only." (P22L6-L11)

(16) 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?

((Reply))

The reviewer 1 has raised the same issue on the results in the performance of NARR in Pembina. In case of Pembina watershed with NAAR data set, the NSE value for calibration period (1985 to 1997) is 0.5 while it is -0.14 for the validation period (1998 -2016). There are some reasons of such a poor performance of NARR in most of the watersheds including Pembina. Since 2003, assimilation of observed precipitation data in to NARR has been discountinued And consequently, NARR overestimates precipitation (refer to section 4.1) and has warm and cold biases in temperature (refer to section 4.2), resulting in highly overestimating flows (refer to Figure 12). In addition, Pembina has been recognized as a parameter-sensitive basin in Eum et al. (2014b)'s study, implying that selection of a calibration period is critical for the performance of hydrologic simulations in this watershed. These biases in NARR and the hydrologic characteristics of the basin may induce poor performance in the hydrologic simulation during the validation period in Pembina. As the reviewer commented, the NARR parameter set produced fair NSEs in simulations forced by the other climate datasets except for CaPA and PNWNAmet. Such result indicates that 1) all of parameter sets used in this study were calibrated reasonably and 2) climate forcing input data plays a more crucial role in hydrolog simulations as any parameter sets did not produce a fair NSE value from NARR in Pembina. The authors addressed the impacts of NARR on hydrologic simulations in the discussion section of the manuscript, as follows:

"Literature has demonstrated that NARR, a reanalysis-based climate dataset, can be an alternative as a climate forcing dataset for hydrologic simulations in data sparse regions (Choi et al., 2009; Praskievicz and Bartlein, 2014; Islam and Dery, 2016). In this study, the NARR dataset performed quite well in high-elevation regions (Hinton in this study) while it did not perform so well in the middle and lower reaches, i.e., lower-elevation watersheds. NARR performed especially poorly in the Pembina sub-basin, a region where hydrologic simulations are highly sensitive to model parameters (Eum et al., 2014b). In Figure 11 (b), however, the NARR parameter set produced fair NSE values in hydrologic simulations forced by the other climate datasets except for CaPA and PNWNAmet. Such result indicates that 1) all of parameter sets used in this study were calibrated reasonably and 2) climate forcing input data plays a more crucial role in hydrolog simulations as any parameter sets did not produce a fair NSE value from NARR in Pembina." (P24L19-P25L3)

<<Remarks>>

(1) P2L20: should be "may not produce" not "may not produces"

((Reply)) Corrected

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

((Reply)) Corrected

(3) P4L32: should be "Peace River" not "Peasce River"

((Reply)) Corrected

(4) P9L5: should be "criteria" not "citeria"

((Reply)) Corrected

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

((Reply)) Corrected

(6) P20L16-18: missing the name of journal

((Reply)) Corrected

(7) P20L19: should be “Dibike, Y.” not “Yonas, D.”

((Reply)) Corrected

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

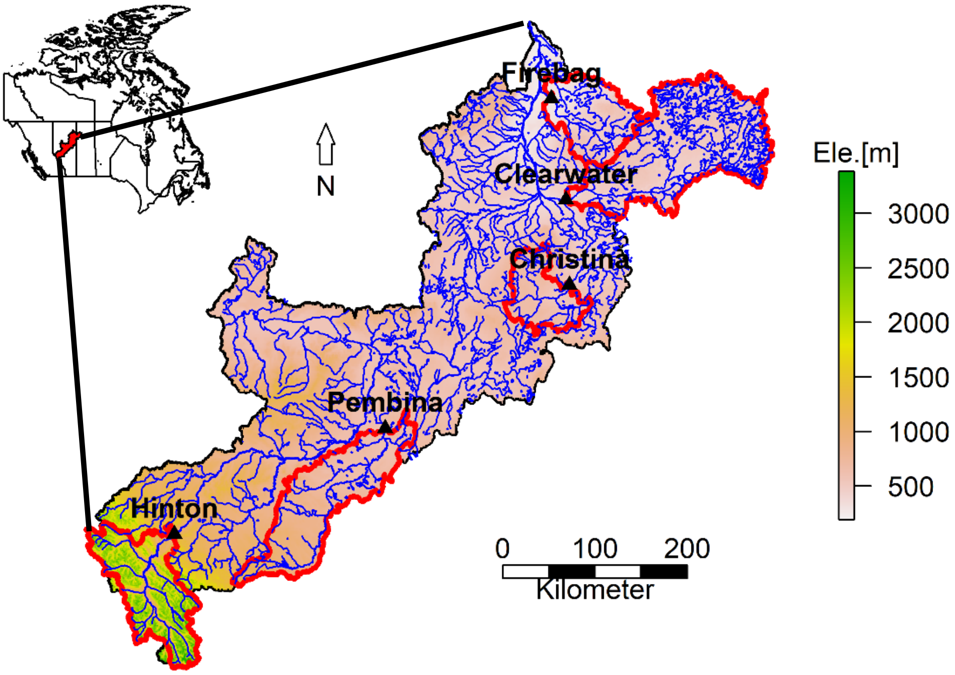
((Reply)) Based on the reply above, I believe the reviewer fully understands how the hydrologic simulations were conducted with two hybrid climate datasets (i.e., R_{ind} and R_{mul}).

(9) Figure 1: should be “precipitation” not” preciptation”

((Reply)) Corrected

(10) 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.

((Reply)) The authors modified Figure 3 to provide the geographical information of the ARB as the reviewer suggested.



(11) 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.

((Reply)) Corrected

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

((Reply)) Again, I believe the reviewer fully understands how the hydrologic simulations were conducted with two hybrid climate datasets.