<<Referee 1>>

(1) The study evaluated different climate dataset source against climate stations using multiple indices and generated a synthetic dataset based on the ranks. Afterwards, the VIC model is applied as proxy validation tool to evaluate multiple datasets and generated datasets. The research is innovative and the structure of the paper is clear. Methods are valid. My only concern is about results. The performance of the VIC model in the study. It is not like what author stated "most of the climate datasets performed well". On the contrary, in Christina and Firebag the NSE is below 0.45 for any datasets, and the worst is even below 0 which is in Pembina with NARR. The results of the model seemed unreliable. Please check the model and improve the performance of hydrological modeling.

((Reply)) We calibrated the parameters of the VIC model for the seven historical gridded climate datasets (i.e., ANUSPLIN, Alberta Township, PNWNAmet, CaPA, NARR, and two hybrid climate datasets) individually using an auto calibration method (dynamic dimensional search algorithm). Table 6 shows the Nash-Sutcliffe Efficiency (NSE) for the calibration and validation periods. Except for NARR, most of the NSE values during calibration period for Christina and Firebag are above 0.50 which is a threshold of satisfactory performance in hydrologic models as suggested by Moriasi et al. (2007). However, as indicated by the reviewer, model performance is not satisfactory for Christina and Firebag during the validation period. Accordingly, sentence has been revised in the manuscript (section 4.4). Figure 11 also shows box-whisker plots resulting from multiset-parameter hydrologic simulations that employed seven different model parameter sets (obtained through model calibration with individual climate datasets) and the same climate dataset as a forcing input data. In Figure 11, the averaged NSE values for Christina and Firebag were below 0.45 as pointed by the reviewer. However, these NSE values are different than the NSE values for calibration and validation shown in Table 6. The authors addressed more clearly how the biases in each climate dataset were estimated indirectly by the proxy validation as below.

"Under the assumption of REFRES that all of the existing climate datasets are of equal quality for hydrologic simulations, all of the calibrated parameter sets can be considered as mostly plausible parameter sets for the selected sub-basins. However, as mentioned above, intrinsic

biases exist temporally and spatially in all of the gridded climate datasets, e.g., discrepancies in the amount and spatial distribution of precipitation between the gridded climate datasets and observations. Therefore, the similarity of the gridded climate datasets in terms of magnitude, sequence, and spatial distribution of climate events relative to observations is crucial to reproduce historically observed streamflows. In addition to climate forcings, streamflows are mainly affected by geographic characteristics and physical land surface processes (e.g., infiltration and evapotranspiration), which are represented by model parametrization related to infiltration and soil properties (Demaria et al., 2007). In a hydrologic simulation, the biases in climate datasets can be compromised by model parameters that adjust hydrologic processes to observations (Harpold et al., 2017; Kirchner, 2006). That is, a calibrated parameter set may imply biases in a climate dataset. Under the assumption that the calibrated parameter sets are suitable for hydrologic simulations in each sub-basin, this study applied a multiset-parameter hydrologic simulation approach that employs all parameter sets calibrated by the seven climate datasets and the same climate dataset as a forcing input data to assess the sensitivity of the climate dataset to all feasible parameter sets. From the multiset-parameter hydrologic simulations, the bias in a climate dataset can be estimated indirectly by quantifying the variability in hydrologic simulations derived from the feasible calibrated parameter sets under a climate forcing dataset. In other words, lower variability in the hydrologic simulations indicates higher reliability in the climate forcing dataset. The suitability of the hybrid climate dataset for improving historical hydrologic simulations was also tested by directly comparing the performances of calibration and validation for each climate dataset. Proxy validations were carried out by conducting 49 hydrologic simulations (7 climate forcing × 7 parameter sets) for the Pembina and Christina catchment areas, whereas only 36 simulation runs were possible for Hinton, Firebag, and Clearwater sub-basins, as one of the gridded data sets (i.e., Township) did not cover the entire catchment areas of these three hydrometric stations." (P16L11-P17L10)

(2) Section 2.1 What's is the time duration of the climate observation data at AHCCD stations?

((Reply)) The AHCCD stations have different record lengths. For example, the longest record period is from 1840 to 2016 while the shortest period is from 1967 to 2004. As the data length are different at each AHCCD station, we selected a common period between AHCCD stations

and gridded climate datasets to estimate performance measure. The authors added this information in section 2.1.

(3) Method: Is the evaluation carried out on the whole time period and could be regarded as the average performance over the time? Is there any temporal variation of the performance for different observation dataset at different stations, and how do you consider the temporal variation of the performance?

((Reply)) The aim of REFRES is to choose a suitable climate dataset among the existing multiple historical gridded climate datasets based on the performance measures selected in this study. Each performance measure was evaluated over a whole common period at each AHCCD station. As the data lengths are different at each AHCCD station, it is not possible to consistently evaluate the temporal variation of the performance over the domain. In addition, consideration of temporal variation in performance may require a common period that covers a whole period of the hybrid climate dataset to be produced by choosing the most suitable climate dataset for a selected period. Therefore, this study only evaluated the performance averaged over a whole period to simplify the method and also to make sure that the methodology is computationally efficient.

(4) Section 3.1.3 It is not clear how the dataset is generated. Do you just choose the best one based on the evaluation over time or make a combination of several good ones?

((Reply)) Two things were considered in generating the hybrid climate data set: (i) the ranking of all datasets at each grid cell and (ii) a period of record or the availability of the gridded climate data sets. For each grid cell, the data were extracted by following the ranking (higher to lower) and data availability. For example, see the table below:

Dataset	RANK	Period of record	Time period contributing to the hybrid climate dataset
ANUSPLIN	2	1950-2015	1950-1959
Township	1	1961-2016	1960-2016
PNWNAmet	3	1945-2012	x
CaPA	4	2002-2017	x

NARR	5	1979-2017	x

In the above table, the hybrid climate dataset should be a period from 1950 to 2016 which is covered by the existing climate datasets. Although Township is ranked first, Township cannot cover the period from 1950 to 1959. In this case, the data generation module in REFRES chooses the second ranked climate dataset, i.e., ANUISPLIN, to produce the hybrid climate dataset and the first ranked data for the remaining period from 1960 to 2016.

The authors addressed more clearly how the hybrid climate datasets are generated using the ranking information in DGM.
"As each climate dataset has different data periods shown in Table 1, the first ranked dataset cannot fully cover a whole target period to be extracted from a set of climate data candidates. DGM provides a systematic procedure to identify the most reliable dataset for a target region and extracts the data from the inventory of climate datasets considering the ranking and availability of each dataset for a desired period. For instance, if CaPA and ANUSPLIN ranked first and second for precipitation and the desired period is 1950 to 2016, DGM starts searching for the availability of precipitation in 1950. As CaPA is only available between 2002 to 2016, DGM reorders the rank to select ANUSPLIN as the best climate dataset available in 1950. In this way, a hybrid dataset over the period 1950 to 2016 is generated by extracting from ANUSPLIN from 1950 to 2001 and CaPA from 2002 to 2016 in this particular case." (P14 L18-P15L2)

(5) 3.2 proxy validation "it is questionable if the hybrid climate dataset can represent a historical climate better than the individual gridded climate dataset. Utilizing a proxy validation approach (Klyszejko, 2007), this study applied streamflow records to confirm the superiority of the derived hybrid climate dataset over other existing climate datasets." The underlying assumption is that the better input data could derive a more realistic streamflow simulation. The VIC model is calibrated against different dataset, so the calibration of parameters could offset the error from the input data. Judging the superiority through the output of a hydrological model is not straightforward and could even be misleading. How to consider the propagation of the error from the input through calibration?
((Reply)) The authors appreciate the valuable comment on the propagation of the error from the input climate data in hydrologic simulation. As the reviewer pointed, biases in climate data can be compromised or compensated by model calibration. This study indirectly estimated the impacts of the biases in climate datasets by a multiset-parameter hydrologic simulation

approach that employs all seven feasible parameter sets (obtained through calibration with the seven climate datasets separately) and seven climate dataset as a forcing data in the VIC model (i.e. 49 simulations; 7 climate forcing × 7 parameter set). From the multiset-parameter hydrologic simulations, the bias in a climate dataset can be estimated indirectly by quantifying the variability in hydrologic simulations derived from the feasible calibrated parameter sets under a climate forcing dataset. In other words, the lower variability in the hydrologic simulations indicates higher reliability in the climate forcing dataset as shown in Figure 11. This point has been clarified in the draft manuscript as follows:

"In a hydrologic simulation, the biases in climate datasets can be compromised by model parameters that adjust hydrologic processes to observations (Harpold et al., 2017; Kirchner, 2006). That is, a calibrated parameter set may imply biases in a climate dataset. Under the assumption that the calibrated parameter sets are suitable for hydrologic simulations in each sub-basin, this study applied a multiset-parameter hydrologic simulation approach that employs all parameter sets calibrated by the seven climate datasets and the same climate dataset as a forcing input data to assess the sensitivity of the climate dataset to all feasible parameter sets. From the multiset-parameter hydrologic simulations, the bias in a climate dataset can be estimated indirectly by quantifying the variability in hydrologic simulations derived from the feasible calibrated parameter sets under a climate forcing dataset. In other words, lower variability in the hydrologic simulations indicates higher reliability in the climate forcing dataset." (P16L20-P17L5)

(6) Could you specify what input you used here for the VIC model?

((Reply)) The VIC model requires several input data, i.e., climate forcing, soil, vegetation, and routing. This study used the same soil, vegetation, and routing input data as described in previous publications (Eum et al., 2014; 2017). The additional data sets used are the new climate forcing data (i.e. hybrid climate data generated in this study) comprised of daily precipitation, minimum temperature and maximum temperature climate variable.

(7) The number of Results should be 4.

((Reply)) Corrected.

(8) 3.1 Precipitation performance measures in Alberta, could you explain why ANUSPLIN and Township underestimate extreme precipitation?

((Reply)) The main reason that ANUSPLIN and Township underestimate extreme precipitation is that they employed raw station data instead of the adjusted precipitation data which are higher than the raw station data by 5 % to 20%. The authors addressed this as below,
"Interestingly, two station-based gridded climate datasets, ANUSPLIN and Township, show negative P_{bias} while PNWNAmet, CaPA, and NARR datasets have positive P_{bias}. This indicates that ANUSPLIN and Township may underestimate extreme precipitation, as they employed the raw station data instead of the adjusted precipitation data which is higher than the raw station data by 5%-20%. In contrast, other climate datasets (especially multiple sources and reanalysis data) overestimate extreme precipitation." (P17L20-L25)

(9) Figure 10 is it a maximum, minimum or mean temperature in this figure?
((Reply)) The ranking was determined based on the performance of precipitation and temperature (minimum and maximum) individually by TOPSIS. The performance measures for both minimum and maximum temperature were employed into TOPSIS and the ranks were presented in Figure 10 (b). Figure 10 (c) showed the ranking when the performance measures for all variables were considered in TOPSIS. Please also see the following clarification text in the manuscript:

"To alleviate the erroneous output that minimum temperature is higher than maximum temperature on a certain day when producing the hybrid climate dataset by the ranking of temperature values individually, the performance measures of both minimum and maximum temperature are employed together to rank the climate datasets for temperature. " (P14L5-L8)

(10) Page 15 line 24-26 "Over the five hydrometric stations, most of the climate datasets performed well with the exception of NARR in the Pembina catchment." Please explain why NARR in Pembina performs bad which only got -0.85 for NSE. The criterial of well or not well is quite subjective. In Hinton the model performance could be acceptable. However, in Christina and Firebag the NSE is even below 0.45 for any cases and In Pembina and Clearwater NSE below 0.7. This is not a behavioral model honestly. Is the model suitable for the river basin? If it is suitable why the NSE is low? I suggest to check the calibration of the model. Otherwise the proxy validation is not reliable.
((Reply)) 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 discontinued and consequently, NARR overestimates precipitation (refer to section 4.1) and has

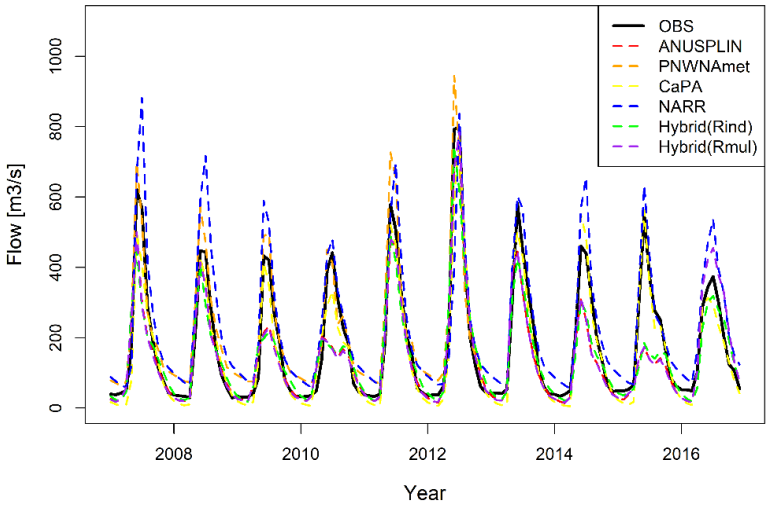
warm and cold biases in temperature (refer to section 4.2). In addition, Pembina has been recognized as a parameter-sensitive basin in Eum et al. (2014b)'s study, implying that selection of a calibration period is critical for the performance of hydrologic simulations in this watershed. These biases in NARR and the hydrologic characteristics of the basin may induce poor performance in the hydrologic simulation during the validation period in Pembina. A qualitative rating has been suggested by Moriasi et al. (2007) as shown in the table below.

Very Good	Good	Satisfactory	Unsatisfactory
$0.75 \leq NSE \leq 1.00$	$0.65 \leq NSE \leq 0.75$	$0.50 < NSE \leq 0.65$	$NSE \leq 0.50$

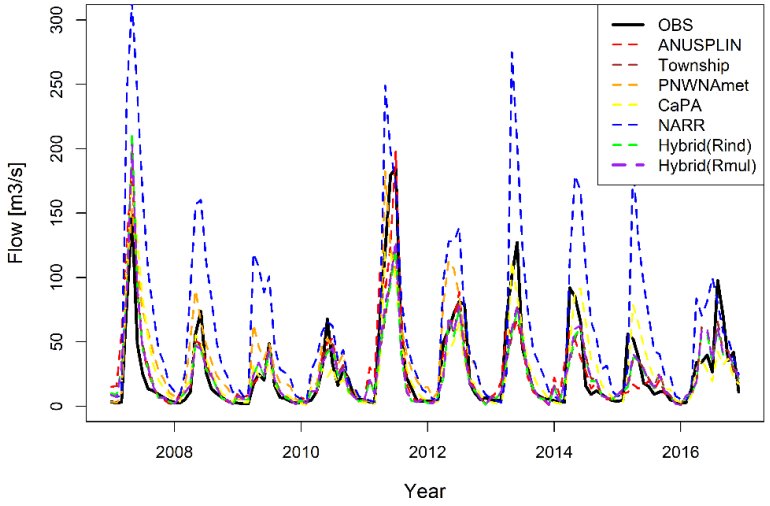
The goodness-of-fit statistics table shows modelling is satisfactory when NSE > 0.5. Table 6 presents Nash-Sutcliffe Efficiency (NSE) for the calibration and validation periods at the selected hydrometric stations (Hinton, Pembina, Christina, Clearwater, and Firebag) in the ARB to assess the suitability of each climate datasets as a climate forcing for hydrologic simulations. Over the five hydrometric stations, most of the climate datasets performed well with the exception of NARR in the Pembina catchment. That is, most of the NSE values in calibration for Christina and Firebag were above 0.50 which is a threshold of satisfactory performance in hydrologic models as suggested by Moriasi et al. (2007).

(11) Figure 12 is suggested to be refined it is hard to tell the difference between different experiments. Is it m^3/s in the label of Y axis? There is lack of label of X axis.

((Reply)) The authors have modified Figure 12 from daily to monthly hydrograph and added another hydrographs for Pembina and x-axis has been labeled to improve visualization.



(a) Hinton



(a) Pembina

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

<< Referee 2>>

<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 Pbias. 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.

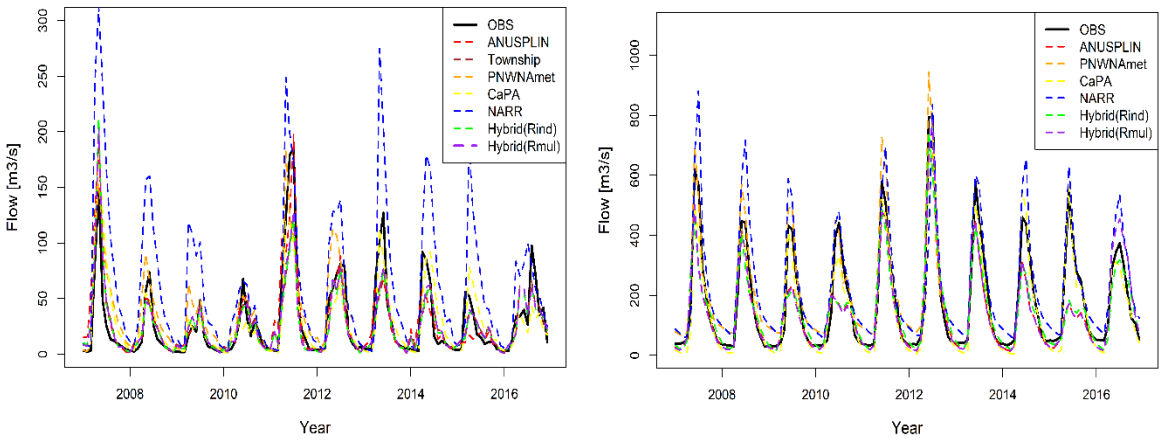
(3) Creditability of hybrid dataset in improving hydrologic simulations

(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 dataset that dominantly ranked first in this watershed. On the 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 the derived hybrid dataset. Further, as suggested by the 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 a 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 the watershed area increases, the derived hybrid climate dataset is no longer dominated by a single gridded dataset. Due to the limitation of computational capacity, initially only five sub-basins were selected for proxy validation.

Nash-Sutcliffe Efficiency (NSE) of ANUSPLIN and the hybrid climate datasets at the main stream of the Athabasca River

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.

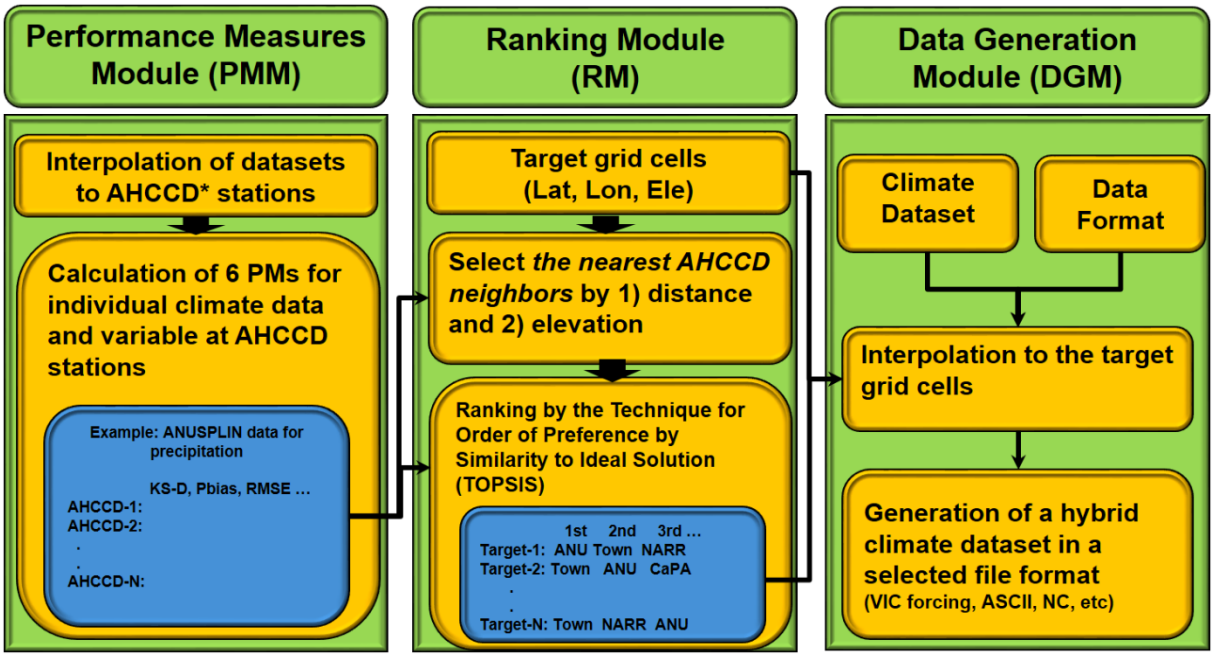


Figure 2. Structure of REFRES comprised of three modules; 1) Performance Measure Module (PMM), 2) Ranking Module (RM), and 3) Data Generation Module (DGM)

(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 4.2.d 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 dataset that dominantly ranked first in this watershed. On the 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 (98%)	40 (2%)	0 (0%)	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 datasets 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 dataset (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 the case of Pembina watershed with NAAR dataset, the NSE value for the calibration period (1985 to 1997) is 0.5 while it is -0.14 for the validation period (1998 -2016). There are some reasons for 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 hydrologic 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?

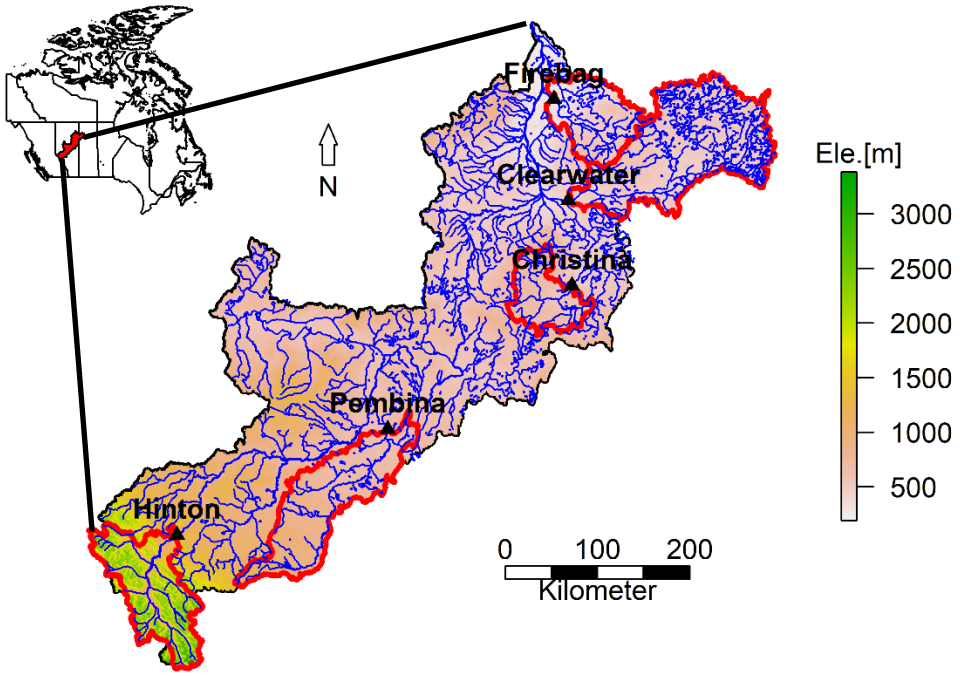
((Reply)) Based on the reply above, I believe the reviewer fully understands how the hydrologic simulations were conducted with two hybrid climate datasets (i.e., R_{ind} and R_{mul}).

(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.

<<Short comment from David Thompson >>

<General Comments>

The study evaluates five climate datasets; ANUSPLIN, Alberta Township, PNWNAmet, CaPA, and NARR. The method can be divided in three major parts: (a) comparing climate datasets (identified in the method section of the manuscript as "Performance Measure Module"), (b) ranking the gridded datasets based on their performance measures (identified in the method section of the manuscript as "Ranking Module"), and (c) further evaluating climate datasets and their ranking using the VIC hydrological model (identified in the method section of the manuscript as "Proxy validation").

Each part of the method section raises concerns as follows:

Part 1:

In the first part of the methodology, five climate datasets were compared. Three of them (ANUSPLIN, Alberta Township, and PNWNAmet) are climate datasets which are originally generated based on interpolation, and the other two (CaPA and NARR) are generated based on models and satellite technologies. The accuracy of all the datasets is compared to the (observed stations) Adjusted and Homogenized Canadian Climate Data (AHCCD). The main concern is how the authors did this comparison? The study states that "the inverse distance squared weighting method was applied to obtain the values at the AHCCD stations from all the gridded climate datasets. Then, performance measures were calculated by comparing the interpolated values with the data collected at AHCCD stations." This raises major concerns about the method used as follows:

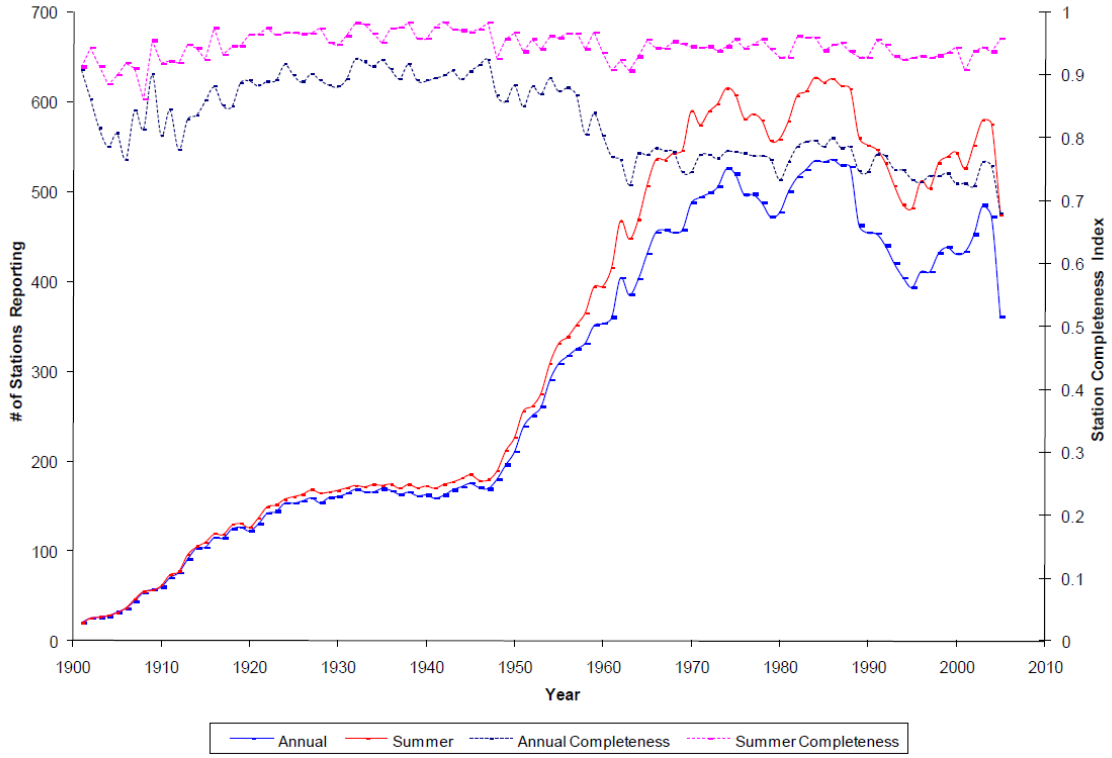
1-1) First and foremost, the ANUSPLIN, Alberta Township, and PNWNAmet climate datasets were originally generated/interpolated based on "the same source of observed data (AHCCD)." If they are slightly different in the interpolated values, this is simply due to:

a. different generation (updated version) of AHCCD were used to interpolate the data (Vincent et al., 2002, 2012; Mekis and Vincent, 2011). This implies that if one dataset illustrates slightly poor performance compared to the others, it doesn't mean it is still the poor choice as they are continuously being updated.

((Reply)) The station-based gridded climate datasets included in this study were not generated based on the adjusted values at the AHCCD stations only, but they employed the raw archive of station data available in a given year to produce gridded climate datasets. While there are only 45 AHCCD stations for precipitation within Alberta, for example, the Alberta Township

dataset employed more than 500 stations (refer to the figure below) to produce the gridded climate data in the 1970s within Alberta. Similarly, for ANUSPLIN, quality-controlled, but unadjusted, station data from the National Climate Data Archive of Environment and Climate Change Canada data (Hutchinson et al., 2009) were interpolated onto the high-resolution grid using thin plate splines. Station density varies over time with changes in station availability, peaking in the 1970s with a general decrease towards the present day (Hutchinson et al., 2009). Thus, the number of stations active across Canada between 1950 and 2011 ranged from 2000 to 3000 for precipitation and 1500 to 3000 for air temperature (Hopkinson et al., 2011).

In other words, the station-based gridded climate datasets have been produced based on different station densities which varied spatially and temporally and by applying different set of rules for inclusion of stations in interpolation. Thus, the number of stations included in each dataset is significantly different apart from differences in the interpolation techniques. Therefore, differences are expected in the interpolated values at a location using different gridded climate datasets.



Number of stations used in the interpolation scheme of the Alberta Township dataset

(Source: https://agriculture.alberta.ca/acis/docs/Methodology-and-data-sources-for-interpolated-data-y2019_m03_d27.pdf)

b. the three climate datasets have been generated based on different interpolation techniques. Therefore, the errors/uncertainties might be associated with the interpolation techniques. In this regard, even if one assumes the three climate datasets were generated using the same version of AHCCD at the time of comparison (which is not the case here), the interpolation method of each individual dataset should have been used to estimate unknown points based on known points. That is the way to evaluate the performance of each climate dataset generated by different interpolation techniques. Instead, authors used their own interpolation method (inverse distance squared weighting method) "to obtain the values at the AHCCD stations", "Then, performance measures were calculated by comparing the interpolated values with the data collected at AHCCD stations." This means the error found in one dataset could be associated with the interpolation techniques used, - not the original datasets. This could be one of the reasons the Alberta Township climate datasets illustrate better accuracy compared to others. Because the Alberta Township climate datasets have been generated based on different versions of the Inverse Distance Weighting method including "the inverse distance squared weighting method" which was used by the authors to do the evaluation.

((Reply)) If all of the climate datasets were generated from the same set of stations data (e.g., only AHCCD), the skill of interpolation techniques can be evaluated as the reviewer commented. However, the three station-based climate datasets have not used the same source of AHCCD stations as commented above in 1-1 a). Due to the limitation of data availability in a given year, each station-based climate dataset investigated in this study employed different numbers of raw station data. For example, ANUSPLIN used the number of stations ranging from 2000 to 3000 for precipitation and from 1500 to 3000 for temperature.

Although ANUSPLIN and PNWNAmet used the same interpolation approach, i.e., thin-plate smoothing spline, it was found in this study that the performance of ANUSPLIN was much better than that of PNWNAmet. The reason of this difference in performance is that ANUSPLIN used all of Canada-wide archive (raw) station data collected by ECCC in a given year while PNWNAmet employed only stations which cover a common period from 1945 to 2002. Therefore, the different number of stations employed in these two climate datasets may induce the different performances in this study. In addition, the Alberta Township dataset has been produced by the archive (raw) station data collected by ECCC, Alberta Environment and Parks (AEP), and Alberta Agriculture and Forestry (AF) over Alberta. It means, additional stations were used in the Alberta Township data for interpolation, so that the accuracy of the dataset was

also improved. In other words, the performance of the station-based climate datasets included in this study considerably depends on the station density employed in interpolation rather than only on the interpolation techniques used.

(1-2) The authors should avoid comparing apples with oranges when the two CaPA and NARR datasets obtained from models and satellites were compared to the ANUSPLIN, Alberta Township, and PNWNAmet datasets obtained from interpolation techniques. This comparison was done based on the observed detests (AHCCD) which was originally used to generate the ANUSPLIN, Alberta Township, and PNWNAmet datasets. Each point of comparison has been initially used as a centre point to generate the ANUSPLIN, Alberta Township, and PNWNAmet datasets, which can result in high correlations between three as well as the AHCCD datasets due the "existing spatial dependency." The point values should have been used for evaluations which are "spatially independent." Otherwise, there is no point in comparing the three interpolated climate datasets with CaPA and NARR which were originally generated to address a poor monitoring network density.

((Reply)) As commented in 1-1, the station-based climate datasets used the archive of raw station data and not the only adjusted values at AHCCD stations. It means the raw but quality controlled observations were used at stations (number of stations used are much greater than only AHCCD stations). Unfortunately, it cannot be guaranteed that the station-based climate datasets are spatially independent with the AHCCD stations as the raw station values at the same location of AHCCD stations might be included in interpolation schemes of each climate dataset. However, it is sure that each station-based climate dataset has been produced using their own spatial structures i.e., different station densities in data generation processes and thus they are unique. On the other hand, CaPA is an amalgamation of rain gauge data, radar data and output from a numerical weather prediction model whereas the NARR data is an amalgamation of NCEP Eta Model (32km/45 layer) output with the Regional Data Assimilation System (RDAS). The archive of raw station data were employed in developing both of these products. As shown in Table 1 (manuscript), the climate datasets used in this study have several inconsistencies with respect to spatial domain, data length, number of climate variables, and spatial resolution. In past, large-scale modelling studies have combined multiple climate datasets to cover the entire study domain or period of record for all the required climate variables, usually without evaluating the performance of different climate datasets for the modelled regions. Thus, the ultimate aim of this study is to suggest a framework that systematically combines multiple climate datasets. In this context, it is meaningful to rank all of the climate datasets and to produce a performance-based hybrid climate dataset to enhance

the performance of numerical models. This study also proved that the hybrid climate dataset provides better representation of historical climatic conditions and thus, enhance the reliability of hydrologic simulations.

Part 2: The gridded datasets have been ranked based on their performance measures. However:

(2-1) We can not necessarily assign a high performance rank to a grid cell just because of being highly correlated with a nearby station - neither due to its distance nor elevation.

((Reply)) The idea is to rank all of available climate datasets (i.e., five datasets included in this study) based on their varying performance spatially. The performance is determined by comparing the interpolated values against the observed values (at several locations within the study area, AHCCD stations). Various performance measures have been used for ranking instead of just using correlation coefficient. Furthermore, as mentioned in (1-1) above, all of five considered climate datasets are different and unique as they have employed different numbers of climate stations (also varied over time) and generation techniques. Two of the datasets (CaPA and NARR) are very different from three station-based datasets as they also employed the output of weather prediction numerical models in addition to assimilation of station based data. When several datasets are available but there are considerable differences between them, it is reasonable to compare them against the observations to determine their accuracy and thus preference for use.

(2-2) The ranking concept may not be still valid considering some of the comments mentioned in part 1.

((Reply)) Based on the response provided above (part 1, 1-1, 1-2; and part 2, 2-1), authors strongly believe that the reviewer now has a better understanding of the five existing gridded climate datasets and how the suggested methodology could help the researchers in identifying the best data for their study area. The suggested methodology will also help in constructing a reliable, gap filled data for larger regional areas which are otherwise affected by the available data domains.

Part 3: Further evaluation of climate datasets and their ranking have been done using the VIC hydrological model.

(3-1) Five VIC models have been calibrated corresponding to each individual climate dataset. How can you justify associating the errors to the climate data rather than to "the calibration parameters and/or the calibration process, and/or the model structure"?

arbitrary adjustment of parameters might have been done to compensate for the errors in the input climate data - which has been done for each VIC model separately.

((Reply)) Authors agrees that in hydrologic simulations, the biases in climate datasets can be compensated or compromised by model parameters that adjust hydrologic processes to observations. That is, a calibrated parameter set may imply biases in a climate dataset. Therefore, this study applied a multiset-parameter hydrologic simulation approach that employs all parameter sets calibrated by the seven climate datasets and the same climate dataset as a forcing input data to assess the sensitivity of the climate dataset to all feasible parameter sets. From the multiset-parameter hydrologic simulations, the bias in a climate dataset can be estimated indirectly by quantifying the variability in hydrologic simulations derived from the feasible calibrated parameter sets under a climate forcing dataset. The results showed the hybrid climate dataset provides a better representation of historical hydrologic simulations compared to the results of individual climate datasets. The authors also clarified this in the revised manuscript as below,

"As mentioned above, however, intrinsic biases exist temporally and spatially in all of the gridded climate datasets, e.g., discrepancies in the amount and spatial distribution of precipitation between the gridded climate datasets and observations. Therefore, the similarity of the gridded climate datasets in terms of magnitude, sequence and spatial distribution of climate events relative to observations is crucial to reproduce historical observed streamflows. In addition to climate forcings, streamflows are mainly affected by geographic characteristics and physical land surface processes (e.g., infiltration and evapotranspiration), which are represented by model parametrization related to infiltration and soil properties (Demaria et al., 2007). In a hydrologic simulation, the biases in climate datasets can be compromised by model parameters that adjust hydrologic processes to observations (Harpold et al., 2017; Kirchner, 2006). That is, a calibrated parameter set may imply biases in a climate dataset. Under the assumption that the calibrated parameter sets are suitable for hydrologic simulations in each sub-basin, this study applied a multiset-parameter hydrologic simulation approach that employs all parameter sets calibrated by the seven climate datasets and the same climate dataset as a forcing input data to assess the sensitivity of the climate dataset to all feasible parameter sets. From the multiset-parameter hydrologic simulations, the bias in a climate dataset can be estimated indirectly by quantifying the variability in hydrologic simulations derived from the feasible calibrated parameter sets under a climate forcing dataset. In other words, the lower variability in the hydrologic simulations indicates higher reliability in the climate forcing dataset."(P16L12-P17L4)

(3-2) It has been mentioned in the manuscript that "The proxy validation also confirmed the superior performance of hybrid climate datasets compared with the other five individual climate datasets investigated in this study." However, the results of the proxy validation (in Table 6) confirm otherwise. Maybe even going one step further, and ask this question whether the two climate datasets; ANUSPLIN, Alberta Township can confirm that there is no need to generate another dataset called "hybrid climate dataset". Overall, I agree the use of various available data sources in hydrological modeling and qualifying them through alternative simulation scenarios prior to calibration of the model parameters (e.g., Faramarzi et al., 2015), but we need way more rigorous method and justification than what are used in this study to introduce 'a reference climate dataset' for a province.

((Reply)) The accuracy of the historical gridded climate datasets considerably depends on employed station density which varies with time and region. As commented in (1-1), all of the climate datasets have employed different station densities, methods, and techniques in the processes of data generation, thus they are quite different. Therefore, there is a need to evaluate their performance before application so that an informed decision could be made before their application. Having several products of varying quality may pose serious concerns especially when these are applied without understanding the differences, reliability and accuracies. Further, the performance of a dataset may vary with region and hence requires such assessment for each study area as the results presented here cannot be generalized for the entire data domains. As commented in (1-2), numerical modelers have suffered from the inconsistency of available climate datasets in spatial domain and resolution, data length, and climate variables. In this context, the Athabasca River basin is a good test-bed because the whole domain cannot be covered by the Alberta Township data which was dominantly ranked first. Combining Alberta Township with ANUSPLIN simply for the Athabasca River basin instead of generating the hybrid climate dataset, as commented by the reviewer, we may neglect added values of ANUSPLIN within the domain of Alberta Township. Further, we may neglect added values of other climate datasets available within the basin. Therefore, we suggested the REFerence Reliability Evaluation System (REFRES) that systematically produces a performance-based hybrid climate dataset. For the Clearwater sub-basin, all of five climate datasets contribute to generating the hybrid climate data for precipitation (refer to Table 5), resulting in relatively a larger improvement in hydrologic simulations as shown in Table 6. In addition, the other reviewer also suggested that the usefulness of the hybrid climate dataset can be clearly found

at the whole basin scale instead of a sub-basin scale as the added values may be accumulated at the main stream over the entire ARB. The authors conducted additional analysis to simulate the entire basin and computed NSE values at a few hydrometric stations in the main stream of the Athabasca River (refer to the table below). The results showed that the hybrid climate dataset performs better than the existing gridded climate dataset (in this case ANUSPLIN) as the drainage areas are larger. This is mainly due to the fact that as the watershed area increases, the derived hybrid climate dataset is no longer dominated by a single gridded dataset. We also addressed these results in the reply of (3-2) in AC2.

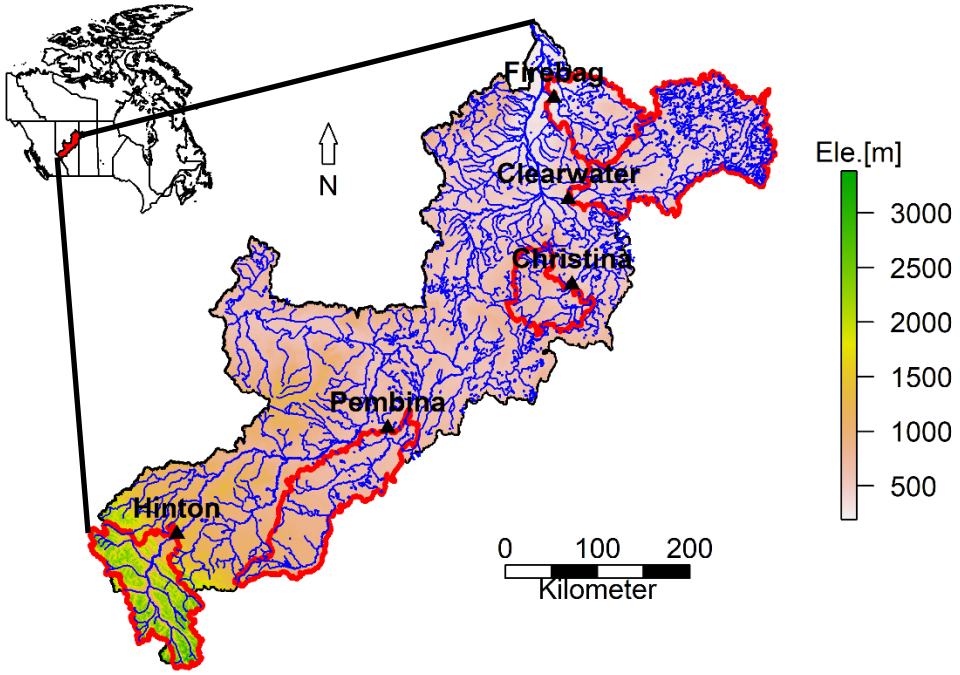
Nash-Sutcliffe Efficiency (NSE) of ANUSPLIN and the hybrid climate datasets at the main stream of the Athabasca River

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>

Authors may consider using coordinate systems for figures, especially Fig. 3 and 8 that can help readers to locate the study area and better investigate its climate.

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



<<Short Comments from Fuad Yassin>>

<General Comments>

This study addresses a relevant topic, particularly in Canada, where there is a huge limitation of reliable high-density observed climate data. Although I find the study very interesting, I have two important general comments that need better clarification.

(1) The first comment is that why other important data sources ignored in this study? If you look at the study of Wong et al. (2017), they demonstrated that GPCC and CRU data are good candidates in Canada compared to NARR. In their study, NARR was found to be the worst data set, and it is not clear why it is accounted in this study, while GPCC and CRU data present unique data globally with long-term and high-temporal resolution data. I believe a better explanation about this is needed, and accounting GPCC and CRU data would provide greater insight for the audience.

((Reply)) Wong et al. (2017) intercompared multiple climate datasets for only precipitation at monthly time step while this study did for precipitation and temperature at daily scale. Both GPCC and CaPA provide daily precipitation at the global and North America domains. However, GPCC has a coarser resolution (1.0° = ~100km) while CaPA provides a higher resolution, at 10km, with a better monitoring network in Canada. Therefore, CaPA has been selected in this study. In addition, CRU has been also excluded as it provides monthly climate datasets. REFRES has a flexible structure to include a new climate dataset when available. For example, the Climate Forecast System Reanalysis (CFSR) dataset will be included in the next version of REFRES.

(2) My second observation is that why only few streamflow stations are used for proxy validation? My understanding is that there are many streamflow stations in the study area, especially around headwaters where huge variability and magnitude of precipitation expected.

((Reply)) Yes, there are other hydrometric stations in the upper reach in the ARB. The five sub-basins were selected for the proxy validation based on three criteria: a) hydrometric record length, b) location defined by upper, middle and lower reaches (Northern River Basin Study, 2002), and c) the number of gridded climate datasets used to generate a hybrid climate dataset for the catchment area of the selected hydrometric station. Based on first criteria, hydrometric stations with a short-period of record and/or severe data gaps were excluded for the proxy validation. There are several stations in the lower watersheds with shorter record length (as they

were installed between 2000 and 2010). In addition several stations are only being operated during open water season (i.e., summer) and do not have any observations during winter seasons and they have also been excluded. For example, the Windfall station (ID: 07AE001) has been excluded because it has hydrometric record only during open-water period. The other reviewer also suggested that the usefulness of the hybrid climate dataset can be clearly found at the whole basin scale instead of a sub-basin scale. Accordingly, we calibrated the VIC model for the whole basin to provide additional results at the main stream of the Athabasca River as shown in the table below. The results showed that the hybrid climate dataset performs better than the existing gridded climate dataset as the drainage areas are larger. We also addressed these results in the reply of (3-2) in AC2 and SC1.

Nash-Sutcliffe Efficiency (NSE) of ANUSPLIN and the hybrid climate datasets at the main stream of the Athabasca River

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

Hybrid climate datasets from a climate data evaluation system and their impacts on hydrologic simulations for the Athabasca River basin in Canada

Hyung-Il Eum[1], ·and Anil Gupta[1,2]

For submission to Hydrology and Earth System Sciences (HESS)

H.-I. Eum (Corresponding author, email: hyung.eum@gov.ab.ca)

[1] Alberta Environment and Parks, Environment Monitoring and Science Division, 3535 Research Road NW, Calgary, Canada, T2L 2K8

[2] Department of Geomatics Engineering, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada

Abstract

A reliable climate dataset is a backbone for modeling the essential processes of the water cycle and predicting future conditions. Although a number of gridded climate datasets are available for North American content, which provides reasonable estimates of climatic conditions in the region, there are inherent inconsistencies in these available climate datasets (e.g., spatial- and temporal-varying data accuracies, meteorological parameters, lengths of records, spatial coverage, temporal resolution, etc). These inconsistencies raise questions as to which datasets are the most suitable for the study area and how to systematically combine these datasets to produce a reliable climate dataset for climate studies and hydrological modeling. This study suggested a framework, called the reference reliability evaluation system (REFRES), that systematically ranks multiple climate datasets to generate a hybrid climate dataset for a region. To demonstrate the usefulness of the proposed framework, REFRES was applied to produce a historical hybrid climate dataset for the Athabasca River basin in Alberta, Canada. A proxy validation was also conducted to prove the applicability of the generated hybrid climate datasets to hydrologic simulations. This study evaluated five climate datasets, including station-based gridded climate datasets (ANUSPLIN, Alberta Township, and PNWNAmet), a multi-source gridded dataset (Canadian Precipitation Analysis - CaPA), and a reanalysis-based dataset (NARR). The results showed that the gridded climate interpolated from station data performed better than multi-source and reanalysis based climate datasets. For the Athabasca River basin, Township and ANUSPLIN were ranked first for precipitation and temperature, respectively. The proxy validation also confirmed the utility of hybrid climate datasets in hydrologic simulations, compared with the other five individual climate datasets investigated in this study. These results indicate that the hybrid climate dataset provides the best representation of historical climatic conditions and thus, enhances the reliability of hydrologic simulations.

Key words: Historical gridded climate data, reference reliability evaluation system, hydrological simulation, Athabasca River basin, proxy validation

Deleted: the

Deleted:

Deleted: length

Deleted:

Deleted: a valid question

Deleted: determines a ranking of

Deleted: mostly

Deleted: superior performance

Deleted: a better

Deleted: enhancing

Deleted: Reference

Deleted: Proxy

1. Introduction

A reliable historical climate dataset is essential in understanding the climatic and hydrological characteristics of a watershed, as it is a crucial forcing input data for simulating key processes of the water and energy cycles in impact models (Deacu et al., 2012; Essou et al., 2016; Wong et al., 2017). Although climate monitoring networks have advanced over the last decades, poor network density still exists, especially in western mountainous and northern parts of Canada. Moreover, climate observations are often spatially interpolated to cover ungauged regions, which may cause unexpected erroneous model predictions as a consequence of the sparse measurements network, especially for mountainous areas affected by orographic effects (Rinke et al., 2004; Wang and Lin, 2015).

As advances in numerical hydrologic and hydrodynamic modeling have increased the capability and reliability in simulating complex natural processes to detect anthropogenic and natural climate changes, a need for temporally- and spatially- reliable climate data has also been grown to accommodate the requirements of input data for numerical models (Shen et al., 2010; Shrestha et al., 2012; Islam and Dery, 2017). For instance, process-based distributed hydrologic models have a grid-based structure that requires input data for each grid cell. However, a simple spatial interpolation of observational station data to all model grid cells may not produce a reliable input forcing dataset for hydrologic models, particularly in a region with a sparse gauging network. A reliable historical climate dataset is also crucial in climate change studies when used for statistical downscaling techniques that employ the relationships between observations and outputs of global (or regional) climate models to produce climate forcing at regional or local scales. Since the resolution of products from a statistical downscaling technique usually corresponds to that of the historical climate dataset (Werner and Cannon, 2016; Eum and Cannon, 2017), the availability of temporally- and spatially- reliable historical climate data is essential for climate-related impact studies (Christensen and Lettenmaier, 2007; Kay et al., 2009; Gutmann et al., 2014; Eum et al., 2016).

A number of high-resolution gridded climate datasets have been developed for various applications such as inter-comparison studies (Eum et al., 2014a; Wong et al., 2017) and hydrologic modeling (Choi et

Deleted: to understand

Deleted: cycle

Deleted: a

Deleted: the

Deleted: produces

Deleted:

al., 2009; Eum et al., 2016). There are various types of gridded climate datasets available for the North American region; 1) station-based interpolated, 2) station-based multiple-source, and 3) reanalysis-based multiple-source (Wong et al., 2017). By interpolation of observational station data, long-term gridded climate datasets have been produced over various domains defined by stations incorporated such as Canada-wide Australia National University's spline (ANUSPLIN, Hutchison et al., 2009), the Alberta Township data (Shen et al., 2001), and the PCIC NorthWest North America meteorological (PNWNAmet) dataset (Werner et al., 2019). The Canadian Precipitation Analysis (CaPA) system, a multiple source-based climate dataset, has been developed to produce near real-time precipitation analyses (6-hr accumulated precipitation) over North America at 15 km resolution which has been further improved to 10km resolution (Lespinas et al., 2015). North American Regional Reanalysis (NARR), one of the reanalysis-based datasets derived from a regional climate model (~32km), has been tested as an alternative climate dataset (Choi et al., 2009; Praskievicz and Bartlein, 2014; Essou et al., 2016; Islam and Dery, 2017).

In most of the large-scale modelling studies, multiple climate data sets were combined to cover the entire modelling domain for all the required climate variables, usually without evaluating the performance of different climate datasets for the modelled regions (Faramarzi et al., 2015; Shrestha et al., 2017; Wong et al., 2017). The lack of performance indicators for available climate datasets may cause inappropriate application of these datasets for various large scale studies, resulting in unreliable outputs, e.g., considerable bias in statistical downscaling studies. Therefore, selecting reliable gridded climate data for a study area is crucial for any hydrological or climate-related studies (Werner and Cannon, 2016; Eum et al., 2014a; 2017). Eum et al. (2014a) intercompared three gridded climate datasets (ANUSPLIN, NARR, and CaPA) for the Athabasca River Basin (ARB) and found that data accuracy varies spatially and temporally over the basin mainly due to the heterogeneity of spatial density of the observational climate network in the basin and limited data assimilation. Wong et al. (2017) also intercompared gridded precipitation datasets derived from different data sources over Canada. Few studies have attempted to incorporate spatially-varied performance measures of various climate datasets to produce a complete long-term historical climate dataset for a study

Deleted: It is an example of a multiple source-based climate dataset.

Deleted:

Deleted: , however

Deleted: an

Deleted: varying

Deleted:

Deleted: measure

Deleted: in producing

region (Faramarzi et al., 2015; Shrestha et al., 2017). In addtion, no systematic framework has been developed yet that could be employed by climatic and hydrologic studies.

Therefore, this study provides a framework, called REFerence Reliability Evaluation System (REFRES), to systematically determine the ranking of multiple climate datasets based on their performance and generate a hybrid climate dataset for a study region by extracting the best candidate (based on the ranking) from multiple climate datasets available in a repository. Several performance measures were identified and calculated by comparing to the Adjusted and Homogenized Canadian Climate Data (AHCCD) over western Canada. Based on the performance measures, the climate datasets were ranked to generate a hybrid climate dataset for the area of interest (target area). A hybrid dataset for two climate variables - precipitation and temperature, key forcing for hydrological modeling, was produced for a period of record that is fully covered by the multiple climate datasets. To validate the applicability of the hybrid climate dataset, a proxy validation approach was employed by comparing simulated streamflows derived from the generated hybrid climate data and other available climate datasets to recorded streamflows at various hydrometric stations in the Athabasca River basin (ARB). Streamflows were simulated using a hydrologic model (Variable Infiltration Capacity, VIC) calibrated and forced by individual climate datasets and the generated hybrid climate dataset. Therefore, the aims of this study are 1) to develop a methodology (i.e., reference reliability evaluation system, REFRES) to compare and rank multiple gridded climate datasets based on the proposed performance measures and to generates the hybrid climate dataset, and 2) to validate the hybrid climate dataset using the proxy validation approach for the Athabasca River basin as a case study to confirm the applicability of hybrid climate dataset to hydrologic simulations.

Deleted:), however

Deleted: to

Deleted: full

Deleted: overlapped or

Deleted: using

Deleted: is

Deleted: superiority

2. Climate data

2.1 Adjusted and Homogenized Canadian Climate Data (AHCCD)

Climate station observations in Canada are available from the national climate data and information archive of Environment and Climate Change Canada (ECCC, http://climate.weather.gc.ca/). Besides the

variable number of observations due to frequent changes in operations including discontinuation of stations, the observations are also subject to various errors from undercatch of solid precipitation, orographic effects, and malfunction of measurements (Mekis and Hogg, 1999; Rinke et al., 2004).

Mekis and Vincent (2011) adjusted daily rainfall and snowfall data, considering wind undercatch, evaporation, and wetting losses corresponding to the types of gauges for 450 stations over Canada. The most recent version released in 2016 provides the adjusted precipitation observations, expanded to 464 precipitation stations. Vincent et al. (2012) produced the 2nd generation of homogenized daily temperature by adjusting the time series at 120 synoptic stations to account for a nation-wide change in observing time and homogenizing discontinuities over 338 temperature (daily minimum and maximum) stations in Canada. The adjusted and homogenized Canadian Climate Data (AHCCD) are available through Environment and Climate Change Canada (http://ec.gc.ca/dccha-ahccd/default.asp?lang=En&n=B1F8423).

Considering that archived raw station data were used to produce the historical gridded climate datasets used in our study, the evaluation of performance at the AHCCD stations is more meaningful because the AHCCD data were adjusted to account for the known measurement issues in the raw station data. For example, the adjusted precipitation data are higher by 5 % to 20 %, varying with topographic characteristics (Mekis and Vincent, 2011). Therefore, the AHCCD dataset is recognized as the best estimate of actual climate variables in Canada, and consequently used in a number of climate-related studies (Asong et al., 2015; Eum et al., 2014a; Shook and Pomeroy, 2012; Wong et al., 2017). As large-scale watersheds in Alberta are crossing the province, e.g., the Peace River and Athabasca River basins, this study evaluated the performance of the historical gridded climate datasets at the AHCCD stations within British Columbia (BC), Alberta (AB), and Saskatchewan (SK) (190 and 129 stations for precipitation and temperature, respectively, in Figure 1). The AHCCD stations have different record lengths. For example, the longest record period is from 1840 to 2016 while the shortest period is from 1967 to 2004. As the data lengths are different at each AHCCD station, we selected a common period between each AHCCD station and climate dataset to estimate performance measures.

Deleted: expanding

Deleted: %

Deleted: Peasce

Figure 1. AHCCD stations within the British Columbia (BC), Alberta (AB), and Saskatchewan (SK) provinces

2.2 Historical gridded climate datasets

In general, the available historical gridded climate dataset can be divided into three categories; 1) station-based, 2) multiple source-based, and 3) reanalysis-based. In this study, five high-resolution gridded climate datasets available for Alberta were selected (Table 1) to evaluate their performance and include in the generation of a hybrid climate dataset for Alberta.

Deleted: were included

Table 1. High-resolution gridded historical climate datasets used in this study

2.2.1 Station-based datasets

Hutchinson et al. (2009) produced a Canada-wide daily climate dataset at 10 km resolution from 1961 to 2003 by the Australia National University's trivariate thin-plate smoothing spline (ANUSPLIN) technique to model the complex spatial patterns (e.g., large variations in ground elevation and station density over Canada) of daily weather data. Hopkinson et al. (2011) updated the existing ANUSPLIN dataset by reducing residuals and extended the daily weather data from 1950 to 2011. Recently, ANUSPLIN data were extended until 2015 for three climate variables, i.e., daily precipitation, minimum and maximum air temperature, which were interpolated with 7,514 surface-based observations (archive data) of Environment Canada. However, the numbers of stations included in interpolation varied year to year, ranging from 2,000 to 3,000 for precipitation and from 1,500 to 3,000 for air temperature. The ANUSPLIN data generated by Natural Resource Canada (NRCan) have been used as the source data to compare climate products (Eum et al., 2014a; Wong et al., 2017), evaluate the accuracy of regional climate models (Eum et al., 2012), and to model hydrologic regimes (Islam and Dery, 2017; Eum et al., 2017; Dibike et al., 2018).

Deleted: and include 3-
Deleted: .g.
Deleted: and
Deleted: ranged
Deleted: inter-
Deleted: to
Deleted: modeling

Similar to the ANUSPLIN dataset, Pacific Climate Impacts Consortium (PCIC) also generated daily precipitation, minimum and maximum air temperature, and wind speed from 1945 to 2012 at 1/16 degree (6~7km) resolution using a thin-plate smoothing spline technique over Northwest North America, called the PCIC North West North America meteorological (PNWNAmet, Werner et al., 2019) dataset (https://data.pacificclimate.org/portal/gridded_observations/map/). While ANUSPLIN utilized a varying number of gauge stations depending on availability of observations in a given year, PNWNAmet set a common period from 1945 to 2012 for all stations included in the interpolation over regularly spaced grid cells within the domain. The PNWNAmet dataset was developed to produce forcing data for an updated version of the Variable Infiltration Capacity model with glaciers (VIC-GL). In addition to precipitation, and minimum and maximum temperature, PNWNAmet includes wind speed, which considerably affects vital hydrologic processes, especially evapotranspiration, sublimation, and snow transport (i.e., snow blowing). Because the AHCCD dataset provides only daily precipitation and temperature, wind speed was excluded in this study.

Alberta Agriculture and Forestry (AF) produced the Alberta Township data (http://agriculture.alberta.ca/acis/township-data-viewer.jsp) from 1961 to 2016 at approximately 10km (Alberta Township grid) resolution using a hybrid inverse distance weighting (IDW) process (Shen et al., 2001) for daily precipitation, minimum and maximum temperature, relative humidity, wind speed, and solar radiation. The archive (raw) station data collected by ECCC, Alberta Environment and Parks (AEP), and AF over Alberta were used in producing the Township dataset. The Township data used various effective radiuses (60 km to 200 km) to ensure a sufficient number of gauge stations in IDW. When there is no station within 200 km, it is assumed that the nearest station represents the climate conditions of the Township center. The domain of Township data covers most of Alberta except the mountainous regions while both ANUSPLIN and PNWNAmet cover all of western Canada (refer to Table 1). Therefore, one of the limitations of the Township dataset is its application to a large watershed spanning Alberta and other neighboring provinces.

2.2.2 Multiple source-based dataset

As an operational system, the Meteorological Service of Canada initiated the Canadian Precipitation Analysis (CaPA) in 2003 to produce superior gridded precipitation data over North America at 10 km resolution (Lespinas et al., 2015), especially for regions with poor observational networks (Mahfouf et al., 2007). CaPA employs an optimum interpolation technique that requires properties of error statistics among observations and a first guess, i.e., background field (Garand and Grassotti, 1995). A short-term forecast of 6-hr accumulated precipitation from the Canadian Meteorological Centre (CMC) regional Global Environmental Multiscale (GEM) model (Côté et al., 1998a; 1998b) is used in CaPA as the background field. The assimilated precipitation from the Canadian weather radar network and 33 US radars near the border are used as additional observations to generate analysis error among multiple sources of observations and the background precipitation. Zhao (2013) tested the applicability of CaPA for hydrologic modelling in the Canadian Prairies and proved its usefulness in data-sparse regions and the winter season. In addition, CaPA has been widely used in agricultural and hydrologic applications (Deacu et al., 2012; NIDIS, 2015). Eum et al. (2014a) further addressed some of the limitations of CaPA, i.e., lack of air temperature which is one of the primary drivers in hydrologic modeling and shorter data length (only from 2002 to 2017), for model calibration and validation. Using 6-hr accumulated precipitation CaPA products, in this study, daily accumulated precipitation was generated over western Canada by adjusting the time zone from Universal Time Coordinated (UTC) to Mountain Time (MT).

Deleted:

Deleted:)

2.2.3 Reanalysis-based dataset

Reanalysis products are another common type of gridded dataset used in climate and hydrologic studies. The North American Regional Reanalysis (NARR) was developed to create a long-term set of dynamically consistent 3-hourly climate data from 1979 to 2003 at a regional scale (0.3°= ~ 32km) for the North America domain (Mesinger et al., 2006). By utilizing advanced land-surface modeling and data

assimilation through the Eta Data Assimilation System (EDAS), NARR improved the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) global reanalysis data. NARR cycled every 3 hours to produce a climate dataset from 1979 to the current year. Choi et al. (2009) tested the applicability of NARR for hydrologic modeling in Manitoba for a region with a poor monitoring network density. However, the NARR dataset after 2004 is not consistent with that of prior years (i.e., 1979 to 2003) because assimilation of precipitation observations was discontinued in 2003 (Eum et al., 2014a). Using the 3-hr NARR climate data, daily precipitation and minimum and maximum temperature were calculated by adjusting the time zone to MT from the original NARR dataset (UTC zone).

3. Methodology

3.1 Reference Reliability Evaluation System (REFRES)

This study suggests a **REF**erence **R**eliability **E**valuation **S**ystem (REFRES) that consists of three main modules (refer to Figure 2): 1) a performance measure module (PMM) to evaluate various performance measures for each climate dataset, 2) a ranking module (RM) to identify the most reliable climate data for a target grid cell using a multi-criteria decision-making technique based on the performance measures provided by PMM, and 3) a data generation module (DGM) to produce a hybrid climate dataset by selecting the most reliable climate dataset based on the ranking provided by the RM (ranking model). These three modules are seamlessly integrated and exchange the required data and information to generate a hybrid climate dataset. The next section provides further details on each module.

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

3.1.1 Performance Measure Module (PMM)

AHCCD is a point (station) dataset while the other climate datasets used in this study (refer to Table 1) are regularly spaced gridded datasets with varying time period, spatial resolution, and coverage (i.e.,

Deleted: is

Deleted: ,

Deleted:);

Deleted: RM.

Deleted: more

Moved (insertion) [1]

Formatted: English (Canada)

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

Formatted: English (United States)

Formatted: Font: Not Bold

domain). Therefore, the inverse distance squared weighting method was applied to obtain the values at the AHCCD stations from all the gridded climate datasets. Then, performance measures were calculated by comparing the interpolated values with the data collected at AHCCD stations. The choice of the performance measures is vital in REFRES, as the ranking of climate datasets entirely depends on included performance measures. In this study, performance measures were selected based on three criteria: 1) distribution, 2) sequencing, and 3) spatial pattern. Distribution-related performance is assessed by the Kolmogorov-Smirnov *D* statistic (D_{KS}) and standard deviation ratio (σ_{ratio}). Sequence-related performance is assessed by the percentage of bias (P_{bias}), root mean square error (RMSE), and temporal correlation coefficient (TCC). Spatial pattern-related performance is evaluated by the pattern correlation coefficient (PCC) as shown in Eq. (1) to Eq. (5). The equations of TCC and PCC are identical but TCC is calculated with the daily time series of climate variables and PCC is obtained by the mean annual precipitation and temperature of the AHCCD stations over a target domain. Therefore, PCC varies with the user specified target domain.

$$D_{KS} = \sup |F_G(x) - F_O(x)| \quad (1)$$

$$\sigma_{\text{ratio}} = \{(\sigma_G/\sigma_O) - 1\} \quad (2)$$

$$P_{\text{bias}} = \frac{\sum_{i=1}^{N}(G_i - O_i)}{\sum_{i=1}^{N} O_i} \times 100 \quad (3)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(G_i - O_i)^2}{N}} \quad (4)$$

$$\text{TCC, PCC} = \frac{\sum_{i=1}^{N}(G_i - \bar{G})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{N}(G_i - \bar{G})^2}\sqrt{\sum_{i=1}^{N}(O_i - \bar{O})^2}} \quad (5)$$

where σ_G and σ_O are the standard deviation of gridded and observed climate datasets, G_i and O_i represent gridded and observed climate datasets at *i*th time step, respectively; *F* is the empirical distribution function of a climate dataset; σ is standard deviation; $\bar{G}$ and $\bar{O}$ represent the mean of gridded and observed climate datasets, respectively and *N* is a total number of data points. These six performance measures were

Deleted: G

Deleted: O

calculated for all the selected climate datasets and variables at each AHCCD station. Figure 2 (blue box in PMM) shows an example of 6 PMs calculated for the precipitation variable using the ANUSPLIN gridded data. Thus, 15 tables (5 climate datasets × 3 variables) were generated by PMM and transferred to the RM.

Deleted: ranking module (RM).

3.1.2 Ranking Module (RM)

The function of the ranking module is to select the appropriate AHCCD stations for a given target grid cell and to rank all the gridded data sets based on the six performance measures calculated in the previous module. For a given target cell, AHCCD stations are selected based on two criteria: distance and elevation. Firstly, 20% (of all AHCCD) stations are selected based on the nearest distance criteria, which were then again reduced by the five nearest stations based on the minimum elevation difference criteria. Then the performance measures are averaged over the selected AHCCD stations to represent the skill of each climate dataset for the given target grid cell.

Deleted: ,
Deleted: First
Deleted: citeria
Deleted: to half
Deleted:

As multiple performance measures are employed in this study, there are situations when a climate dataset may perform well for some measures but not for others. Therefore, a multi-criteria decision-making (MCDM) technique is required to systematically rank all of the climate datasets while considering multiple performance measures. This study applied a multi-criteria decision-making technique called the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS, Hwang and Yoon 1981) to systematically determine the order of preference for all climate datasets at each target grid cell. TOPSIS calculates the geometric distance between alternatives and an ideal solution defined by the best performance on each criterion from the alternatives, and then determines the best and worst alternatives based on the distance. TOPSIS has been successfully applied to watershed management for multi-criteria problems (Jun et al., 2013; Lee et al., 2013). TOPSIS starts with the averaged performance measures, $(x_{ij})_{m\times n}$ for the i^{th} alternative (climate dataset in this study) and j^{th} criterion (i.e., a performance measure). A weighted normalized decision matrix, $(t_{ij})_{m\times n}$ is given by

Deleted: in
Deleted: ,
Deleted:),
Deleted: determine
Deleted:), and a

$$(t_{ij})_{m\times n} = (w_j n_{ij})_{m\times n} \quad i = 1,2,\cdots,m; \quad j = 1,2,\cdots,n \tag{6}$$

$$n_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}^2} \tag{7}$$

where, m and n are the total number of alternatives and criteria, respectively, n_{ij} is normalized matrix by Eq. (7), and w_j represents weighting on the j^{th} criterion. Under the assumption that all performance measures are important, this study used an equal weighting. Then, Euclidean distances (d_{ib} and d_{iw}) of climate datasets from the best (A_b) and worst (A_w) conditions were calculated respectively by Eq. (8) to Eq. (11)

$$A_w = \{\langle \max(t_{ij}|i = 1,2,\cdots,m)\,|j \in J_-\rangle, \langle \min(t_{ij}|i = 1,2,\cdots,m)\,|j \in J_+\rangle\} \equiv \{t_{wj}|j = 1,2,\cdots,n\} \tag{8}$$

$$A_b = \{\langle \min(t_{ij}|i = 1,2,\cdots,m)\,|j \in J_-\rangle, \langle \max(t_{ij}|i = 1,2,\cdots,m)\,|j \in J_+\rangle\} \equiv \{t_{bj}|j = 1,2,\cdots,n\} \tag{9}$$

$$d_{iw} = \sqrt{\sum_{j=1}^{n}(t_{ij} - t_{wj})^2} \quad i = 1,2,\cdots,m \tag{10}$$

$$d_{ib} = \sqrt{\sum_{j=1}^{n}(t_{ij} - t_{bj})^2} \quad i = 1,2,\cdots,m \tag{11}$$

Where, t_{bj} and t_{wj} are the best and worst decision matrices determined by Eq. (8) and (9), respectively, and J_+ and J_- represent criteria that have a positive and a negative impact on performance. For example, TCC and PCC are in J_+ while D_{KS}, σ_{ratio}, P_{bias}, and RMSE are in J_-. Using the Euclidean distances, the order of preference for all climate datasets was determined by the similarity (S_{iw}) to the worst condition in Eq. (15).

$$s_{iw} = \frac{d_{iw}}{d_{iw}+d_{ib}}, 0 \le s_{iw} \le 1, \quad i = 1,2,\cdots,m \tag{15}$$

s_{iw} = 1 when the alternative is equal to the best condition (A_b) and s_{iw} = 0 if the alternative is equal to the worst condition (A_w). In other words, a higher s_{iw} represents higher preference among alternatives. As we evaluate the performance measures (criteria) for individual climate variables, TOPSIS can be applied to decide the preference of climate datasets considering the performance measures for either individual or multiple variables. In this study, TOPSIS provides two types of ranking information by using performance measures from i) individual climate variable and ii) all climate variables. That is, one is the ranking for

Deleted: fairly

Deleted: matrix

Deleted: When

Deleted: ,

Deleted:),

Deleted: evaluated

precipitation and temperature separately (R_{ind}) and the other is the ranking for multiple variables (R_{mul}). For example, in this study, R_{ind} was determined by a 5×6 decision matrix (5 climate datasets and 6 performance measures) for precipitation and temperature individually, while R_{mul} was determined by a 4×18 decision matrix (4 climate datasets excluding CaPA that provides only precipitation by 18 performance measures from three variables). To alleviate the erroneous output that minimum temperature is higher than maximum temperature on a certain day when producing the hybrid climate dataset by the ranking of temperature values individually, the performance measures of both minimum and maximum temperature are employed together to rank the climate datasets for temperature.

Deleted: an

3.1.3 Data Generation Module (DGM)

DGM extracts the most reliable climate data for a user-specified target region based on the ranking information obtained from the RM. The tool is flexible enough to provide output in various common formats, i.e., NetCDF, ASCII (text) or in the specific format of a numerical model. As all of the historical gridded climate datasets have been tested and employed in numerous climatic and hydrologic studies, an assumption was made in generating the hybrid climate dataset that all of the climate datasets are equally qualified for inclusion but the final selection can be determined by the proven superiority evaluated through the performance measures. Under this assumption, the available datasets can be combined systematically based on the rank (performance) of each dataset at target grid cells. As each climate dataset has different data periods shown in Table 1, the first ranked dataset cannot fully cover a whole target period to be extracted from a set of climate data candidates. DGM provides a systematic procedure to identify the most reliable dataset for a target region and extracts the data from the inventory of climate datasets considering the ranking and availability of each dataset for a desired period. For instance, if CaPA and ANUSPLIN ranked first and second for precipitation and the desired period is 1950 to 2016, DGM starts searching for the availability of precipitation in 1950. As CaPA is only available between 2002 to 2016, DGM reorders the rank to select ANUSPLIN as the best climate dataset available in 1950. In this way, a hybrid dataset

Deleted:

Deleted:),

Deleted: has

Deleted: Thus,

Deleted: when

over the period 1950 to 2016 is generated by extracting from ANUSPLIN from 1950 to 2001 and CaPA from 2002 to 2016 in this particular case. Once the best climate datasets are extracted over all the target grid cells (study domain), the hybrid climate dataset is produced in a user-defined format. This study generated the hybrid climate datasets in the form of the VIC forcing input format to be directly employed into the hydrologic model.

3.2 Proxy validation

Although the AHCCD dataset has been adjusted to provide better estimates of actual precipitation and temperature, it contains statistical artifacts that include inevitable errors from sequential data processes that can be propagated in the derived hybrid climate dataset. Given that the AHCCD stations, the reference dataset for the performance measures, are not regularly distributed and have especially poor density in the northern parts of the study area (refer to Figure 1), it is questionable if the hybrid climate dataset can represent a historical climate better than the individual gridded climate dataset. Utilizing a proxy validation approach (Klyszejko, 2007), this study applied streamflow records to validate the utility of the derived hybrid climate dataset over other existing climate datasets in hydrologic simulations. In this study, the proxy validation was conducted using an existing hydrologic model (Eum et al., 2017), Variable Infiltration Capacity (VIC, Liang et al., 1994), for the Athabasca River basin (ARB). The VIC model was further refined at 1/32° (2~3 km) for a finer spatial resolution and to better simulate the complex river network in the Lower Athabasca River basin. Five of the catchment areas listed in Table 2 were selected for the proxy validation based on three criteria: i) hydrometric record length, ii) location defined by upper, middle and lower reaches (Northern River Basin Study, 2002), and iii) the number of gridded climate datasets used to generate a hybrid climate dataset for the catchment area of the selected hydrometric station. 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. Hinton is located near the headwaters of ARB, which are characterized by mountainous topography and

Deleted:

Deleted: ¶

Deleted: , with

Deleted: confirm

Deleted: superiority

Deleted: at 1/32° (2~3 km)

Deleted: a

Deleted: b

Deleted: c

Deleted: The

Deleted: was optimized for each catchment area

Deleted: maximize the evalution of the hydrologic simulations using

Deleted: multiple

Deleted:

snow- and glacier-ice melt dominated hydrologic regimes. Pembina is one of the major rivers in the middle reach. The other three stations (Christina, Clearwater above Christina and Firebag) are located in the lower reach, which is a water-limited (dry) region due to a higher amount of evapotranspiration (Eum et al., 2014b). The sub-basins of Hinton, Firebag, and Clearwater include a partial area outside of the Township data domain, thus inducing a higher or lower number of climate datasets in the derived hybrid dataset. A total of seven climate datasets (five individual and two hybrid climate datasets from the R_{ind} and R_{mul}) are available to calibrate the VIC hydrologic model parameter set related to soil properties and routing. 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. More details on calibration can be found in Eum et al. (2017). Under the assumption of REFRES that all of the existing climate datasets are of equal quality for hydrologic simulations, all of the calibrated parameter sets can be considered as mostly plausible parameter sets for the selected sub-basins. However, as mentioned above, intrinsic biases exist temporally and spatially in all of the gridded climate datasets, e.g., discrepancies in the amount and spatial distribution of precipitation between the gridded climate datasets and observations. Therefore, the similarity of the gridded climate datasets in terms of magnitude, sequence, and spatial distribution of climate events relative to observations is crucial to reproduce historically observed streamflows. In addition to climate forcings, streamflows are mainly affected by geographic characteristics and physical land surface processes (e.g., infiltration and evapotranspiration), which are represented by model parametrization related to infiltration and soil properties (Demaria et al., 2007). In a hydrologic simulation, the biases in climate datasets can be compromised by model parameters that adjust hydrologic processes to observations (Harpold et al., 2017; Kirchner, 2006). That is, a calibrated parameter set may imply biases in a climate dataset. Under the assumption that the calibrated parameter sets are suitable for hydrologic simulations in each sub-basin, this study applied a multiset-parameter hydrologic simulation approach that employs all parameter sets calibrated by the seven climate datasets and the same climate dataset as a forcing input data to assess the

Deleted: , and

Deleted: number

Deleted: have an

Deleted: As

Deleted: , however

Deleted: charactersitics

Deleted: the similarity of the gridded

Deleted: in terms of magnitude, sequence

Deleted: spatial distribution of

Deleted: events relative

Deleted: observations is crucial to reproduce historical observed streamflows. Therefore, the variability of hydrologic simulations may signify the reliability

sensitivity of the climate dataset to all feasible parameter sets. From the multiset-parameter hydrologic simulations, the bias in a climate dataset can be estimated indirectly by quantifying the variability in hydrologic simulations derived from the feasible calibrated parameter sets under a climate forcing dataset. In other words, lower variability in the hydrologic simulations indicates higher reliability in the climate forcing dataset. The suitability of the hybrid climate dataset for improving historical hydrologic simulations was also tested by directly comparing the performances of calibration and validation for each climate dataset. Proxy validations were carried out by conducting 49 hydrologic simulations (7 climate forcing × 7 parameter sets) for the Pembina and Christina catchment areas, whereas only 36 simulation runs were possible for Hinton, Firebag, and Clearwater sub-basins, as one of the gridded data sets (i.e., Township) did not cover the entire catchment areas of these three hydrometric stations.

Deleted: used to produce

Deleted: multiple calibrated parameter sets, i.e., the lower variability indicates higher reliability. For instance, all of hydrologic simulations from the calibrated parameter sets perform well for a

Deleted: and show low

Deleted: , it indicates that the

Deleted: may provide reliable historical

Deleted: does

4. Results

Deleted: ¶
3

4.1 Precipitation performance measures in Alberta

Deleted: 3

Although the performance measures were calculated for 190 AHCCD stations in western Canada, the target area of this study is in Alberta, where only 45 stations are located. Therefore, the results for the 45 AHCCD stations are given in this study. Table 3 shows spatially averaged performance measures for precipitation. The Township data outperformed other climate datasets for all performance measures except P_{bias}. ANUSPLIN is the second best climate dataset for Alberta. All climate datasets underestimate the standard deviation of observed daily precipitation (i.e., negative σ_{ratio}), especially PNWNAmet and CaPA which underestimated by 34 % and 39 %, respectively. Interestingly, two station-based gridded climate datasets, ANUSPLIN and Township, show negative P_{bias} while PNWNAmet, CaPA, and NARR datasets have positive P_{bias}. This indicates that ANUSPLIN and Township may underestimate extreme precipitation, as they employed the raw station data instead of the adjusted precipitation data which is higher than the raw station data by 5%-20%. In contrast, other climate datasets (especially multiple sources and reanalysis data) overestimate extreme precipitation. These results are consistent with findings in Eum et al. (2014a) that

Deleted:

Deleted: underestimate

Deleted: and

CaPA and NARR overestimate extreme precipitation events by overly reflecting the orographic effects on precipitation in western Alberta.

Figure 4 shows the temporal correlation coefficient (TCC) data averaged over the AHCCD stations in Alberta to investigate the similarity between historical precipitation datasets employed in this study. As expected, station-based climate datasets (i.e., ANUSPLIN, PNWNAmet, and Township) showed better TCCs than CaPA and NARR. The TCC between ANUSPLIN and Township was the highest among climate datasets except for the observations (i.e., OBS), even though they incorporated different interpolation techniques. PNWNAmet showed the highest TCC with ANUSPLIN because they both are based on thin plate spline interpolation. TCCs between CaPA and other climate datasets are similar, as CaPA is produced from multiple sources such as GEM's outputs and weather radar networks of Canada and US. NARR, the reanalysis-based climate dataset, showed higher TCC with CaPA than with other datasets, as it is assimilated with multiple sources of observations.

Maps of each performance measure are shown in Figure 5. It is evident from the spatial variability that the ANUSPLIN and Township datasets outperformed the other datasets in D_{KS} throughout Alberta. In the mountainous region of southwest Alberta, most of the climate datasets performed poorly in P_{bias}, σ_{ratio}, RMSE, and PCC, resulting mainly from the sparse observation network and inconsistent observations near the Canada-US border. PNWNAmet highly overestimates the mean annual precipitation in the mountainous area (e.g., 300 mm/year higher than that observed at station ID 3050519), which may considerably affect simulated streamflows originating in mountainous headwaters and further downstream.

4.2 Air temperature performance measures in Alberta

The performance measures for air temperature averaged over 37 AHCCD stations in Alberta are presented in Table 4. As CaPA provides only precipitation, it was excluded in the assessment for temperature. All of the performance measures for temperature are better than those for precipitation except P_{bias}. NARR is highly biased as it underestimates minimum and maximum temperatures, which might be an attribute of

Deleted: wetern

Deleted:)

Deleted: mountains

Deleted: >

Deleted: for

Deleted: 3

Deleted: Performance

discontinuation of observation assimilation since 2003 (Eum et al., 2014a). ANUSPLIN and Township showed an almost perfect linear relationship (TCC) with the observations (i.e., > 0.97 for all of the climate datasets). The performance measures for maximum temperature are better than those for minimum temperature as maximum temperature is dominated by mainly large-scale heat waves while minimum temperature is affected by local physical processes, e.g., topography and surface conditions (Eum et al., 2012). NARR showed less skill in capturing these local effects due to the coarse spatial resolution (~32km) compared to other station-based climate datasets. As with precipitation, the maps of performance measures for minimum and maximum temperature presented in Figure 6 and Figure 7 showed that data from the mountainous areas performed poorly in most of the performance measures. NARR showed positive and negative P_{bias} for minimum and maximum temperature, respectively, in the mountainous region, indicating that NARR has a warm bias in extreme cold temperatures and a cold bias in extreme warm temperatures.

Deleted: temperature while

Deleted: temperature

4.3 Ranking of climate datasets in the ARB

Deleted: 3

The geospatial information (i.e., latitude, longitude, and elevation) of 22,372 grid cells within the ARB was extracted from the Canadian digital elevation data provided by Natural Resources Canada (refer to https://open.canada.ca/data/dataset/7f245e4d-76c2-4caa-951a-45d1d2051333). Using this information, the RM in REFRES ranked the five climate datasets by TOPSIS for each grid cell. Table 5 presents the first-ranked number of grid cells and their percentage for each climate dataset according to the performance measures of individual variables (Case A and Case B) and multi-variables (Case C), i.e., precipitation and (minimum and maximum) temperature in this study.

Deleted: the

For precipitation, the Alberta township dataset was ranked first in most of the grid cells within the basin (78%) for the whole ARB, followed by ANUSPLIN (13%), PNWNAmet (3%), CaPA (3%), and NARR (2%). However, the Township data domain covers only 83% of the ARB within Alberta; the remaining 17% of the watershed area that lies on the outside the province is not covered (Figure 8). The Township dataset was ranked first for almost 95% of grid cells within its domain, indicating that the

Deleted: %),

Township dataset overwhelmingly outperformed other climate datasets for precipitation. Township was dominantly ranked first for the subbasins (Pembina and Christina) within the Township domain.

For temperature, ANUSPLIN was ranked first (in 62% grid cells) for the whole ARB, followed by Township (31%) and PNWNAmet (7%). In the upper and middle reaches, i.e., Hinton and Pembina, PNWNAmet and Township were mostly ranked first, respectively, while ANUPLIN outperformed other climate datasets for the subbasins in the lower reach. When considering the performance measures for multiple variables simultaneously, the Township dataset was ranked first, followed by ANUSPLIN for 64% and 36% of the grid cells for the whole ARB. Figure 9 shows maps of the first-ranked climate datasets for each case in Table 5, i.e., individual variable (Case A and B) and multi-variables (Case C). Due to the limited spatial coverage of the Township dataset, other climate datasets were ranked first in the headwaters of the ARB and the area of the river basin in Saskatchewan. For instance, ANUSPLIN and PNWNAmet were ranked first in the headwaters, while no specific climate dataset dominated in Saskatchewan for precipitation (refer to Figure 9A). For temperature, ANUSPLIN outperformed in the northern part (middle and lower reaches of the ARB) due to outstanding performance of the P_{bias} performance measure for minimum temperature as shown in Table 4 and Figure 6(b). For multi-variables, Township was mostly ranked first within its domain and ANUSPLIN was ranked first outside the Township dataset domain and also for a small part of lower reach area in the ARB.

Figure 10 shows the percentage of each climate dataset at each rank for the three cases (e.g. A, B, and C in Table 5). For precipitation (Case A), Township overwhelmed other climate datasets. The second alternative was ANUSPLIN in the majority of grid cells in the ARB. PNWNAmet, NARR and CaPA were mostly ranked 3rd, 4th and 5th, respectively. For temperature (Case B), ANUSPLIN was ranked mostly first and Township was a distinct second choice in the majority of grid cells, followed by PNWNAmet and NARR. For multi-variables (Case C), Township and ANUSPLIN were the first and second choices in the majority of grid cells in the ARB, respectively.

Deleted:),

Deleted: and the

Deleted: on the contrary,

Deleted: Multi

Deleted: choice

As two different hybrid climate datasets were generated using the ranking information from single- and multi-variable approaches, i.e., Hybrid (R_{ind}) and Hybrid (R_{mul}), further investigation is required to identify which hybrid climate dataset may provide better performance and consequently will be recommended for future climate-related studies. A proxy validation approach was applied using both generated hybrid climate datasets to validate the utility of one dataset over the other.

4.4 Proxy validation of generated hybrid climate datasets

In addition to the five gridded climate datasets, the two hybrid climate datasets were implemented for proxy validation using the VIC model. In contrast to the station-based climate datasets, both CaPA and NARR were produced from climate models and multiple sources of observations, consequently showing a higher correlation with each other as shown in Figure 4. Since CaPA also provides only precipitation, this study combined precipitation of CaPA with the NARR temperature to prepare the CaPA climate forcing dataset for the proxy validation. Table 6 presents the Nash-Sutcliffe Efficiency (NSE) for the calibration and validation periods at the selected hydrometric stations (Hinton, Pembina, Christina, Clearwater, and Firebag) in the ARB to assess the suitability of each climate dataset as a climate forcing input data for hydrologic simulations. Over the five hydrometric stations, most of the climate datasets performed well with the exception of NARR in the Pembina catchment. Most of NSE values in calibration for Christina and Firebag were above 0.50, which is the threshold of satisfactory performance in hydrologic models as suggested by Moriasi et al. (2007). However, model performance is not satisfactory but acceptable for Christina and Firebag during the validation period. The two hybrid climate datasets performed well, with comparably good and better NSE values than other climate datasets, especially at Pembina, Clearwater, and Firebag, located in the middle and lower reaches. Figure 11 presents the boxplots of NSEs obtained through the multiset-parameter VIC simulations. The NSE ranges were obtained from multiple VIC simulations, with each climate dataset used as climate forcing for all the plausible model parameter sets, which were calibrated with seven climate datasets, individually. The values above each boxplot represent the averaged

Deleted: superior
Deleted: the
Deleted: determine
Deleted: superiority
Deleted: 3
Deleted: Since CaPA
Deleted: datasets
Deleted: slightly
Deleted: (
Deleted:)

value of the NSEs over the multiset-parameter hydrologic simulations. A narrower range of NSE values represents a higher precision for a climate dataset and a higher averaged NSE value means higher accuracy. Therefore, a climate dataset showing both a higher averaged NSE and a narrow range of NSEs indicates that it is a relatively more appropriate and reliable climate forcing dataset for hydrologic simulations.

At Hinton, all of the climate datasets showed satisfactory NSE values for accuracy, while ANUSPLIN, Hybrid(R_{ind}), and Hybrid(R_{mul}) showed better precision. 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. Hybrid(R_{mul}) and ANUSPLIN showed the highest accuracy as forcing data, followed by Hybrid(R_{ind}), PNWNAmet, and NARR. In the Pembina and Christina catchments, the Hybrid(R_{ind}), Hybrid(R_{mul}), and Township datasets had the highest precision and accuracy. NARR produced negative NSEs at Pembina, indicating it is not reliable or suitable as a forcing dataset. For Clearwater, Hybrid(R_{ind}) is the top performer, followed by Hybrid(R_{mul}), ANUSPLIN, PNWNAmet, and NARR. Clearwater had the highest number of climate datasets combined in the hybrid climate dataset within the basin for precipitation as shown in Figure 9. Interestingly, the precision of NARR is similar to that of CaPA because they shared the temperature data from NARR. For Firebag, Hybrid(R_{ind}) also showed top performance in both precision and accuracy, followed by Hybrid(R_{mul}), ANUSPLIN, PNWNAmet, and NARR. Overall, Hybrid(R_{ind}) showed the best accuracy and precision at all hydrometric stations, indicating that it has the potential not only to improve historical hydrologic simulations but also to be used as reference data for statistical downscaling of climate change projections in the province.

Deleted: multiple

Deleted: a

Deleted: , which

Deleted: that

Deleted: in

Deleted: and

Deleted: has

Deleted: used

Deleted: same

Deleted: the

Deleted: a

Deleted: 4

5. Discussion

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. 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 the 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

Deleted: Among the station-based gridded climate datasets

Deleted: enormously

Deleted: six

Deleted: (excluding Township)

Deleted: , the headwater region in the ARB.

Deleted: flows during winter

Deleted: Hinton.

Deleted: overestimates

Deleted: during winter

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.

Among the station-based gridded climate datasets, ANUSPLIN and Township employed a different number of stations depending on their periods of record. Therefore, there is an inconsistency in these climate datasets over time. For example, the Township dataset employed only 300~400 stations in the 1960s, but has increased to 400~500 since 1970. A change-point analysis of these datasets may provide some useful information to end-users with respect to when and where changes occurred, which will help in establishing spatial and temporal accuracies of these datasets (Eum et al. 2014a). Further, PNWNAmet employed the same number of stations over time to avoid the above mentioned inconsistency, but this study found that it induced overestimation of precipitation in data-poor regions such as mountainous regions in Alberta. As the hybrid climate datasets are generated from the multiple historical gridded datasets, they may also have the same inconsistencies identified in other datasets. The proxy validation, however, demonstrated that the generated hybrid climate datasets can improve the performance of hydrologic simulations.

This study identified the preference order of all gridded climate datasets based on the performance measures evaluated at the AHCCD stations, therefore the ranking somewhat relies on the spatial distribution of the AHCCD stations. As shown in Figure 1, the density of AHCCD stations varies across western Canada, and is low in the cold climates of mountainous and northern areas. Therefore, the ranking could further be improved with a more uniform density of AHCCD stations over western Canada.

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

Deleted: while

Deleted: it

Deleted: issue

Deleted: and

Deleted: climate

Deleted: Much literature

Deleted: Especially,

Deleted: other than climate forcings (Eum et al., 2014a).

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 hydrologic simulations as any parameter sets did not produce a fair NSE value from NARR in Pembina. CaPA was more suitable than NARR for the selected sub-basins in this study, which indicates that CaPA might be a better alternative in low station-density regions such as the ARB. However, since the validation period in this study is only 7 years from 2010 to 2016, a longer data period is necessary to validate the suitability of CaPA as indicated in Eum et al. (2014a) and Wong et al. (2017).

In the proxy validation, Hybrid(R_{ind}) performed well in the Clearwater sub-basin where the highest number of climate datasets were combined in the generated hybrid climate datasets. The Township dataset, which mostly ranked first within its spatial domain, partially covers the drainage area of Clearwater, so that the generated hybrid climate dataset, Hybrid(R_{ind}), is composed of many climate datasets in this sub-basin. In a traditional approach to hydrological modelling for Clearwater, either the Township dataset might be completely excluded (as it does not cover the entire Clearwater watershed), or potentially combined with other gridded climate datasets to cover the entire watershed. However, combining different climate datasets to construct the climate forcing for a larger region requires an evaluation of the datasets to identify the order of preference for such aggregation when multiple choices are available. Therefore, this study suggested the REFRES methodology to systematically compare all-available climate datasets for a region to produce a hybrid climate dataset that covers a desired period of record and spatial domain by considering the order of preference for combining various climate datasets at each grid cell. The proxy validation approach also confirmed the utility of a generated hybrid climate dataset over other data sets, especially in hydrologic simulations.

6. Summary and concluding remarks

This study suggested a framework called reference reliability evaluation system (REFRES) to systematically generate a performance-based hybrid climate dataset from multiple climate datasets for a

Formatted: Font: 12 pt, Font color: Auto

Deleted: Since

Deleted: however,

Deleted: .(

Deleted: .(

Deleted:)

Deleted: could be

Deleted: superiority

Deleted: 5

region. The hybrid dataset was found to more reliable for hydrological modelling. The REFRES is composed of three modules; 1) performance measures, 2) ranking, and 3) data generation. The suggested framework was applied to the ARB as a test-bed and generated two hybrid climate datasets from single- (R_{ind}) and multi-variable (R_{mul}) approaches by evaluating the performance of five available gridded climate datasets: station-based gridded climate datasets (i.e. ANUSPLIN, Alberta Township, and PNWNAmet), a multi-source dataset (CaPA), and a reanalysis-based dataset (NARR). A hydrologic modelling-based proxy validation approach was applied to demonstrate the applicability of the hybrid climate dataset generated for the five sub-basins in the ARB. The results showed that

- Among the five climate datasets, the station-based climate datasets performed better than multi-source- and reanalysis-based datasets. The Township dataset, in particular, outperformed other climate datasets in the selected performance measures over northern Alberta.
- Most of the climate datasets performed poorly in the mountainous areas of southwest Alberta, due to a sparse observation network, orographic effects, topographic complexity, and inconsistencies in observation between Canada and the US.
- As a result of REFRES' application for the ARB, the Township and ANUSPLIN datasets are mostly ranked the highest among the five climate datasets for precipitation and temperature, respectively.
- In the proxy validation, two hybrid climate datasets, Hybrid(R_{ind}) and Hybrid(R_{mul}), performed better in terms of precision and accuracy as forcing data for hydrologic simulations.
- Hybrid(R_{ind}) especially outperformed other climate datasets in the Clearwater sub-basin where the highest number of climate datasets were combined in generating Hybrid(R_{ind}) for precipitation. This indicates that the hybrid climate dataset generated by REFRES may lead to more reliable hydrologic simulations, resulting in improved hydrologic predictions.

This study provided the preference order of climate datasets available in Alberta, which may be useful for modelers and decision-makers as to which climate dataset is the most suitable for their studies and projects. Furthermore, this study demonstrated that the hybrid climate dataset produced by REFRES is more

Deleted:
Deleted:
Deleted: including mountainous areas,
Deleted: inconsistency
Deleted: clearly
Deleted: and thus

representative of historical climatic conditions. Therefore, the hybrid climate dataset is recommended to be used as a reference dataset for statistical downscaling and hydrologic model forcing, resulting in more reliable high-resolution climatic and hydrologic projections.

Deleted: the
Deleted: should

Acknowledgements

The authors would like to thank the Natural Resources Canada, Alberta Agriculture and Forest, the Pacific Climate Impacts Consortium (PCIC), Environment and Climate Change Canada, and NOAA/OAR/ESRL PSD for providing the historical gridded climate datasets.

Deleted: National

References

Asong, Z. E., Khaliq, M. N. and Wheater, H. S.: Regionalization of precipitation characteristics in the Canadian Prairie Provinces using large-scale atmospheric covariates and geophysical attributes, Stochastic Environmental Research and Risk Assessment, 29(3), 875–892, 2015.

Choi, W., Kim, S. J., Rasmussen, P. F. and Moore, A. R.: Use of the North American Regional Reanalysis for hydrological modeling in Manitoba, Can. Water Resour. J., 34, 13–36, 2009.

Christensen, N. S. and Lettenmaier, D. P.: A multimodel ensemble approach to assessment of climate change impacts on the hydrology and water resources of the Colorado River Basin, Hydrology and Earth System Sciences, 11(4), 1417–1434, 2007.

Deleted: Discussions

Côté, J., Desmarais, J.-G., Gravel, S., Méthot, A., Patoine, A., Roch, M. and Staniforth, A.: The operational CMC–MRB global environmental multiscale (GEM) model. Part II: Results, Monthly Weather Review, 126(6), 1397–1418, 1998a.

Côté, J., Gravel, S., Méthot, A., Patoine, A., Roch, M. and Staniforth, A.: The operational CMC–MRB global environmental multiscale (GEM) model. Part I: Design considerations and formulation, Monthly Weather Review, 126(6), 1373–1395, 1998b.

Deacu, D., Fortin, V., Klyszejko, E., Spence, C. and Blanken, P. D.: Predicting the Net Basin Supply to the Great Lakes with a Hydrometeorological Model, Journal of Hydrometeorology, 13(6), 1739–1759, doi:10.1175/JHM-D-11-0151.1, 2012.

Demaria, E.M, Nijssen, B., Wagener, T.: Monte Carlo sensitivity analysis of land surface parameters using the variable infiltration capacity model, Journal of Geophysical Research, 112, D11113, 2007

Dibike, Y., Eum, H.-I. and Prowse, T.: Modelling the Athabasca watershed snow response to a changing climate, Journal of Hydrology: Regional Studies, 15, 134–148, doi:10.1016/j.ejrh.2018.01.003, 2018.

Essou, G. R. C., Sabarly, F., Lucas-Picher, P., Brissette, F. and Poulin, A.: Can Precipitation and Temperature from Meteorological Reanalyses Be Used for Hydrological Modeling?, Journal of Hydrometeorology, 17(7), 1929–1950, doi:10.1175/JHM-D-15-0138.1, 2016.

Eum, H.-I. and Cannon, A. J.: Intercomparison of projected changes in climate extremes for South Korea: application of trend preserving statistical downscaling methods to the CMIP5 ensemble, International Journal of Climatology, 37(8), 3381–3397, doi:10.1002/joc.4924, 2017.

Eum, H.-I., Dibike, Y. and Prowse, T.: Climate-induced alteration of hydrologic indicators in the Athabasca River Basin, Alberta, Canada, Journal of Hydrology, 544, 327–342, doi:10.1016/j.jhydrol.2016.11.034, 2017.

Eum, H.-I., Dibike, Y. and Prowse, T.: Comparative evaluation of the effects of climate and land-cover changes on hydrologic responses of the Muskeg River, Alberta, Canada, Journal of Hydrology: Regional Studies, 8, 198–221, doi:10.1016/j.ejrh.2016.10.003, 2016.

Eum, H.-I., Dibike, Y., Prowse, T. and Bonsal, B.: Inter-comparison of high-resolution gridded climate data sets and their implication on hydrological model simulation over the Athabasca Watershed, Canada, Hydrological Processes, 28(14), 4250–4271, doi:10.1002/hyp.10236, 2014a.

Eum, H.-I., Dibike, Y. and Prowse, T.: Uncertainty in modelling the hydrologic responses of a large watershed: a case study of the Athabasca River basin, Canada, Hydrological Processes, 28(14), 4272–4293, doi:10.1002/hyp.10230, 2014b.

Eum, H.-I., Gachon, P., Laprise, R. and Ouarda, T.: Evaluation of regional climate model simulations versus gridded observed and regional reanalysis products using a combined weighting scheme, Climate Dynamics, 38(7-8), 1433–1457, doi:10.1007/s00382-011-1149-3, 2012.

Faramarzi, M., Srinivasan, R., Iravani, M., Bladon, K. D., Abbaspour, K. C., Zehnder, A. J. B. and Goss, G. G.: Setting up a hydrological model of Alberta: Data discrimination analyses prior to calibration, Environmental Modelling & Software, 74, 48–65, doi:10.1016/j.envsoft.2015.09.006, 2015.

Garand, L. and Grassotti, C.: Toward an objective analysis of rainfall rate combining observations and short-term forecast model estimates, Journal of Applied Meteorology, 34, 1962–1977, 1995.

Gutmann, E., Pruitt, T., Clark, M. P., Brekke, L., Arnold, J. R., Raff, D. A. and Rasmussen, R. M.: An intercomparison of statistical downscaling methods used for water resource assessments in the United States, Water Resources Research, 50(9), 7167–7186, doi:10.1002/2014WR015559, 2014.

Harpold, A. A., Kaplan, M. L., Klos, P. Z., Link, T., McNamara, J. P., Rajagopal, S., Schumer, R. and Steele, C. M.: Rain or snow: hydrologic processes, observations, prediction, and research needs, Hydrology and Earth System Sciences, 21(1), 1–22, doi:10.5194/hess-21-1-2017, 2017.

Hopkinson, R. F., McKenney, D. W., Milewska, E. J., Hutchinson, M. F., Papadopol, P. and Vincent, L. A.: Impact of Aligning Climatological Day on Gridding Daily Maximum–Minimum Temperature and Precipitation over Canada, Journal of Applied Meteorology and Climatology, 50(8), 1654–1665, doi:10.1175/2011JAMC2684.1, 2011.

Hutchinson, M. F., McKenney, D. W., Lawrence, K., Pedlar, J. H., Hopkinson, R. F., Milewska, E. and Papadopol, P.: Development and Testing of Canada-Wide Interpolated Spatial Models of Daily Minimum–Maximum Temperature and Precipitation for 1961–2003, Journal of Applied Meteorology and Climatology, 48(4), 725–741, doi:10.1175/2008JAMC1979.1, 2009.

Deleted: Yonas, D

Hwang, C.L. and Yoon, K.: Multiple attribute decision making: methods and applicasions. Springer, New York, 1981.

Islam, S. U. and Déry, S. J.: Evaluating uncertainties in modelling the snow hydrology of the Fraser River Basin, British Columbia, Canada, Hydrology and Earth System Sciences, 21(3), 1827–1847, doi:10.5194/hess-21-1827-2017, 2017.

Jun, K. S., Chung, E.-S., Kim, Y.-G. and Kim, Y.: A fuzzy multi-criteria approach to flood risk vulnerability in South Korea by considering climate change impacts, Expert Systems with Applications, 40(4), 1003–1013, 2013.

Kay, A. L., Davies, H. N., Bell, V. A. and Jones, R. G.: Comparison of uncertainty sources for climate change impacts: flood frequency in England, Climatic Change, 92(1-2), 41–63, doi:10.1007/s10584-008-9471-4, 2009.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology: GETTING THE RIGHT ANSWERS FOR THE RIGHT REASONS, Water Resources Research, 42(3), doi:10.1029/2005WR004362, 2006.

Klyszejko, E. S.: Hydrologic Validation of Real-Time Weather Radar VPR Correction Methods, University of Waterloo., 2007.

Lee, G., Jun, K.-S. and Chung, E.-S.: Integrated multi-criteria flood vulnerability approach using fuzzy TOPSIS and Delphi technique, Natural Hazards and Earth System Science, 13(5), 1293–1312, doi:10.5194/nhess-13-1293-2013, 2013.

Lespinas, F., Fortin, V., Roy, G., Rasmussen, P. and Stadnyk, T.: Performance Evaluation of the Canadian Precipitation Analysis (CaPA), Journal of Hydrometeorology, 16(5), 2045–2064, doi:10.1175/JHM-D-14-0191.1, 2015.

Liang, X., Lettenmaier, D. P., Wood, E. F. and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation model, Joural of Geophysical Reserch, 99(D7), 14,415–14,428, doi:10.1029/94JD00483, 1994.

Mahfouf, J.-F., Brasnett, B. and Gagnon, S.: A Canadian Precipitation Analysis (CaPA) Project: Description and Preliminary Results, ATMOSPHERE-OCEAN, 45(1), 1–17, doi:10.3137/ao.v450101, 2007.

Mekis, E. and Hogg, W. D.: Rehabilitation and analysis of Canadian daily precipitation time series, Atmosphere-Ocean, 37(1), 53–85, doi:10.1080/07055900.1999.9649621, 1999.

Mekis, É. and Vincent, L. A.: An Overview of the Second Generation Adjusted Daily Precipitation Dataset for Trend Analysis in Canada, Atmosphere-Ocean, 49(2), 163–177, doi:10.1080/07055900.2011.583910, 2011.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E. H., Ek, M. B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D. and Shi, W.: North American Regional Reanalysis, Bulletin of the American Meteorological Society, 87(3), 343–360, doi:10.1175/BAMS-87-3-343, 2006.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D. and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Transactions of the ASABE, 50(3), 885–900, 2007.

NIDIS, U.S. Drought Portal. NOAA, 2015 [Available online at http://www.drought.gov.]

Praskievicz, S. and Bartlein, P.: Hydrologic modeling using elevationally adjusted NARR and NARCCAP regional climate-model simulations: Tucannon River, Washington, Journal of Hydrology, 517, 803–814, doi:10.1016/j.jhydrol.2014.06.017, 2014.

Rinke, A., Marbaix, P. and Dethloff, K.: Internal variability in Arctic regional climate simulations: case study for the SHEBA year, Climate research, 27(3), 197–209, doi:doi:10.3354/cr027197, 2004.

Shen, S. S., Dzikowski, P., Li, G. and Griffith, D.: Interpolation of 1961–97 daily temperature and precipitation data onto Alberta polygons of ecodistrict and soil landscapes of Canada, Journal of applied meteorology, 40(12), 2162–2177, 2001.

Formatted

Shen, Y., Xiong, A., Wang, Y. and Xie, P.: Performance of high-resolution satellite precipitation products over China, Journal of Geophysical Research, 115(D2), doi:10.1029/2009JD012097, 2010.

Shrestha, R. R., Cannon, A. J., Schnorbus, M. A. and Zwiers, F. W.: Projecting future nonstationary extreme streamflow for the Fraser River, Canada, Climatic Change, 145(3-4), 289–303, doi:10.1007/s10584-017-2098-6, 2017.

Shrestha, R. R., Schnorbus, M. A., Werner, A. T. and Berland, A. J.: Modelling spatial and temporal variability of hydrologic impacts of climate change in the Fraser River basin, British Columbia, Canada, Hydrological Processes, 26(12), 1840–1860, 2012.

Shook, K. and Pomeroy, J.: Changes in the hydrological character of rainfall on the Canadian prairies, Hydrological Processes, 26(12), 1752–1766, 2012.

Vincent, L. A., Wang, X. L., Milewska, E. J., Wan, H., Yang, F. and Swail, V.: A second generation of homogenized Canadian monthly surface air temperature for climate trend analysis: HOMOGENIZED CANADIAN TEMPERATURE, Journal of Geophysical Research: Atmospheres, 117(D18), n/a–n/a, doi:10.1029/2012JD017859, 2012.

Wang, X. L. and Lin, A.: An algorithm for integrating satellite precipitation estimates with in situ precipitation data on a pentad time scale: BLENDED PENTAD PRECIPITATION DATA, Journal of Geophysical Research: Atmospheres, 120(9), 3728–3744, doi:10.1002/2014JD022788, 2015.

Werner, A. T. and Cannon, A. J.: Hydrologic extremes – an intercomparison of multiple gridded statistical downscaling methods, Hydrology and Earth System Sciences, 20(4), 1483–1508, doi:10.5194/hess-20-1483-2016, 2016.

Werner, A., Schnorbus, M., Shrestha, R., Cannon, A., Zwiers, F., Dayon, G., and Anslow, F.: A long-term, temporally consistent, gridded daily meteorological dataset for northwestern North America, Scientific Data, 6, 180299, 2019.

Wong, J. S., Razavi, S., Bonsal, B. R., Wheater, H. S. and Asong, Z. E.: Inter-comparison of daily precipitation products for large-scale hydro-climatic applications over Canada, Hydrology and Earth System Sciences, 21(4), 2163–2185, doi:10.5194/hess-21-2163-2017, 2017.

Zhao, K.: Validation of the Canadian Precipitation Analysis (CaPA) for hydrological modelling in the Canadian Prairies, University of Manitoba (Canada)., 2013.

Table 1. High-resolution gridded historical climate datasets used in this study

Dataset	Full name	Variable	Type	Period	Resolution	Domain	Institution
ANUSPLIN	Australia National University Spline	PRCP, TMX, TMN	Station-based	1950-2015	10 km, Daily	Canada	Natural Resource Canada (NRCan)
Township	Alberta Township	PRCP, TMX, TMN, Tave, WS, RH, SR	Station-based	1961-2016	10km, Daily	Alberta	Alberta Agriculture and Forestry
PNWNAmet	PCIC NorthWest North America meteorological dataset	PRCP, TMX, TMN, WS	Station-based	1945-2012	1/16 degree (6~7 km), Daily	Western Canada (BC, AB, SK) and Alaska	Pacific Climate Impacts Consortium
CaPA	Canadian Precipitation Analysis	PRCP	Multiple source-based	2002-2017	10 km, 6-hr	North America	Canadian Meteorological Centre
NARR	North American Regional Reanalysis	PRCP, Tair, WS, RH, SR, GH, etc*	Reanalysis-based	1979-2017	32km, 3-hr	North America	National Oceanic and Atmospheric Administration (NOAA)

PRCP: precipitation, TMX: maximum temperature, TMN: minimum temperature, Tave: average temperature, Tair: air temperature, WS: wind speed, RH: relative humidity, SR: solar radiation, GH: Geopotential Height

*: Refer to https://www.esrl.noaa.gov/psd/data/gridded/data.narr.monolevel.html for details

Table 2. Characteristics of hydrometric stations selected in this study

Station name	Station ID	Record length	Drainage (km^2)	Reach
Hinton	07AD002	1961-2016	9,760	Upper
Pembina	07BC002	1957-2016	13,100	Middle
Christina	S29 (07CE002)	1982-2016	4,836	Lower
Clearwater above Christina	S42 (07CD005)	1966-2016	18,061	Lower
Firebag	S27 (07DC001)	1971-2016	5,980	Lower

Table 3. Performance measures averaged over AHCCD stations in Alberta for precipitation

Performance measure	Climate Dataset				
	ANUSPLIN	PNWNAmet	CaPA	NARR	Township
D_{KS}	0.09	0.62	0.60	0.42	0.09
σ_{ratio}	-0.17	-0.34	-0.39	-0.28	-0.03
P_{bias}	-7.05	5.80	3.02	2.43	-6.73
RMSE	2.02	2.50	2.59	3.53	1.07
TCC	0.87	0.81	0.77	0.53	0.95
PCC	0.87	0.80	0.73	0.74	0.93

Table 4. Performance measures averaged over the AHCCD stations in Alberta for minimum and maximum temperature

Performance measure	Climate Dataset							
	ANUSPLIN		PNWNAmet		NARR		Township	
	Tmin	Tmax	Tmin	Tmax	Tmin	Tmax	Tmin	Tmax
D_{KS}	0.03	0.02	0.05	0.04	0.12	0.08	0.03	0.02
σ_{ratio}	-0.01	-0.01	-0.03	-0.03	-0.03	-0.03	-0.01	-0.02
P_{bias}	-0.43	-0.28	22.90	-3.89	-306.52	-14.09	7.33	-0.86
RMSE	1.48	1.25	1.97	1.82	4.40	3.47	1.31	0.97
TCC	0.99	0.99	0.98	0.99	0.96	0.97	0.99	0.99
PCC	0.91	0.98	0.87	0.95	0.71	0.78	0.93	0.98

Table 5. First ranked number of grid cells in the five sub-basins and the whole Athabasca River Basin (ARB) and their percentages 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 (98%)	40 (2%)	0 (0%)	0 (0%)	-
	Firebag	924 (97%)	28 (3%)	0 (0%)	0 (0%)	-
(C) Multi-variables	ARB	8049 (36%)	14323 (64%)	0 (0%)	0 (0%)	-

Deleted: percentage

Formatted: Line spacing: single
Inserted Cells
Formatted: Line spacing: single
Formatted Table
Formatted: Line spacing: single
Merged Cells
Formatted: Line spacing: single
Formatted: Line spacing: single
Formatted: Line spacing: single
Inserted Cells
Formatted: Line spacing: single
Inserted Cells

	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%)	-

Table 6. Nash-Sutcliffe Efficiency (NSE) for the calibration and validation periods at five sub-basins in ARB for the climate datasets investigated in this study

Climate forcing	Hinton		Pembina		Christina		Clearwater		Firebag	
	Cal.	Val.	Cal.	Val.	Cal.	Val.	Cal.	Val.	Cal.	Val.
ANUSPLIN	0.88	0.83	0.61	0.64	0.52	0.46	0.76	0.54	0.61	0.49
Township	-	-	0.62	0.66	0.54	0.49	-	-	-	-
PNWNAmet	0.82	0.81	0.53	0.54	0.40	0.35	0.73	0.59	0.65	0.48
CaPA	0.89	0.90	0.53	0.61	0.55	0.44	0.74	0.74	0.51	0.53
NARR	0.84	0.79	0.50	-0.14	0.39	0.34	0.75	0.42	0.44	0.32
Hybrid (R_{ind})	0.82	0.78	0.61	0.66	0.55	0.49	0.78	0.67	0.60	0.52
Hybrid (R_{mul})	0.89	0.83	0.61	0.65	0.54	0.48	0.77	0.53	0.59	0.47

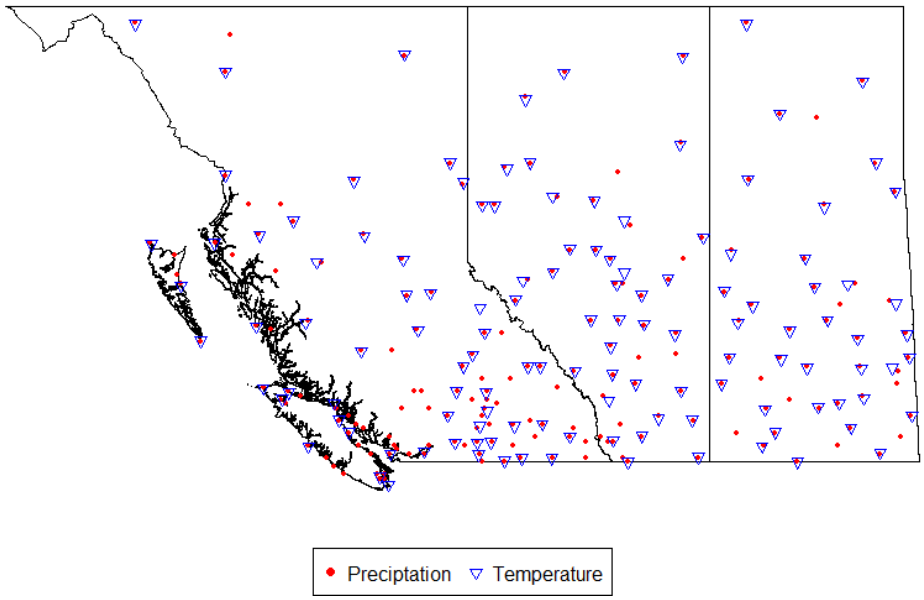


Figure 1. AHCCD stations within the BC, AB, and SK provinces

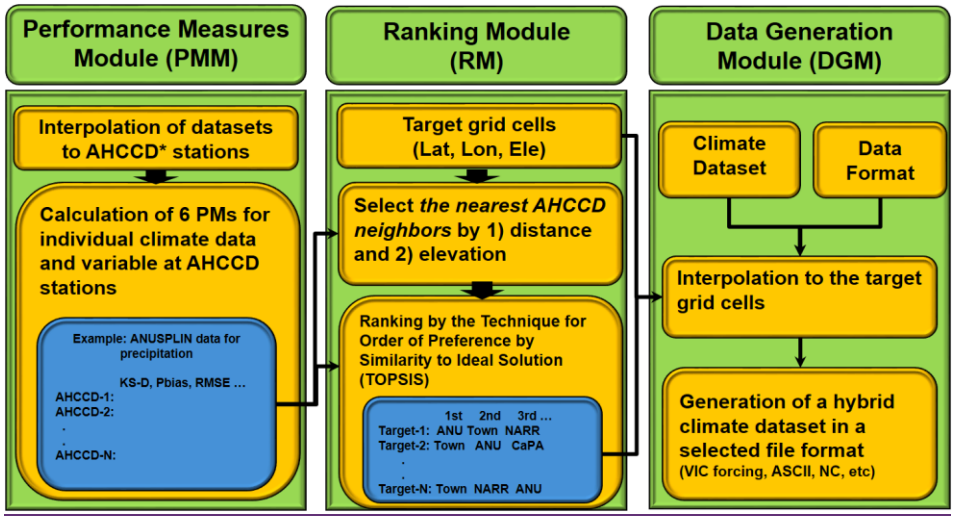


Figure 2. Structure of REFRES comprised of three modules; 1) Performance Measure Module (PMM), 2) Ranking Module (RM), and 3) Data Generation Module (DGM)

Moved (insertion) [2]

Formatted: English (United States)

Formatted: Font: Not Bold

Formatted: Centered

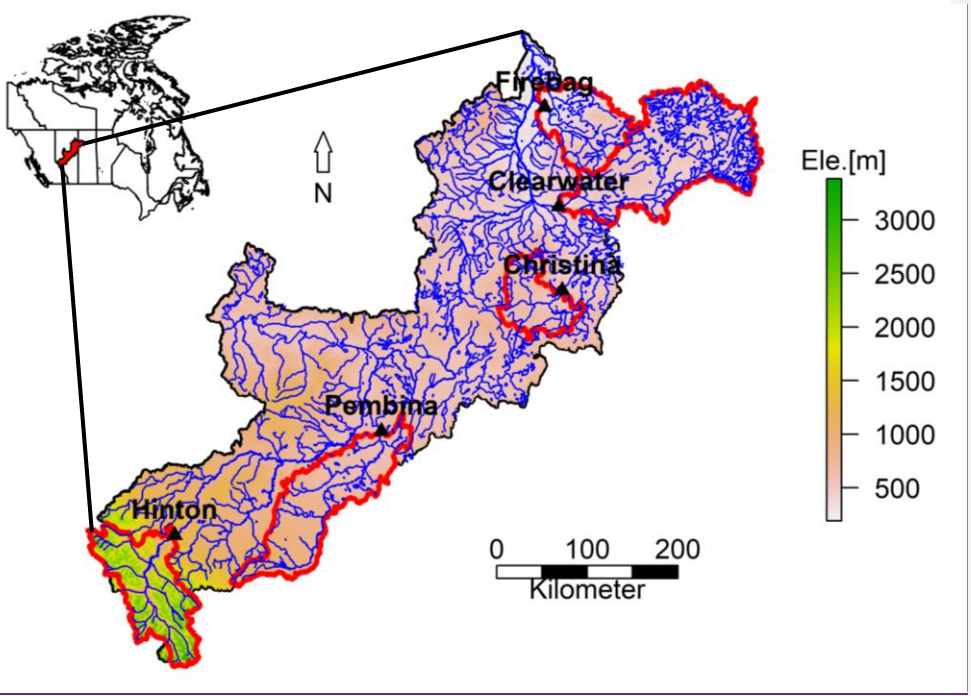


Figure 3. Geographical information on the five sub-basins (red line) selected in the Athabasca River basin for the proxy validation

Moved up [1]: Figure 2.

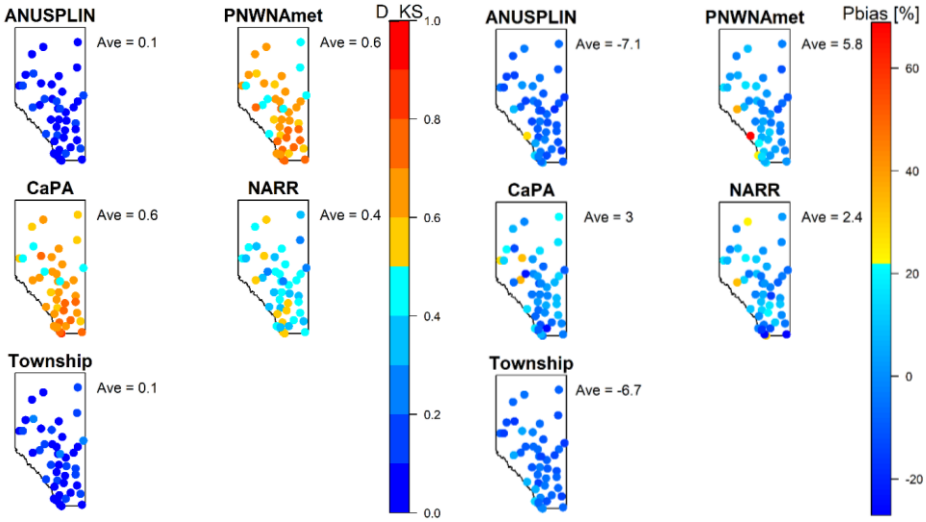
Formatted: English (Canada)

Deleted: Five

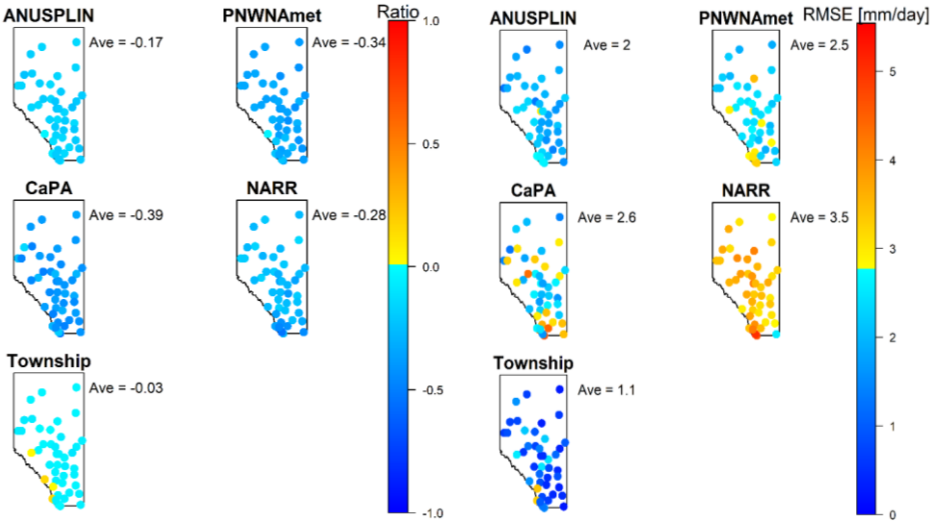
	OBS*	ANUSPLIN	PNWNAmet	CaPA	NARR	Township
OBS*	1	0.87	0.81	0.77	0.53	0.95
ANUSPLIN	0.87	1	0.84	0.81	0.61	0.86
PNWNAmet	0.81	0.84	1	0.81	0.65	0.78
CaPA	0.77	0.81	0.81	1	0.76	0.81
NARR	0.53	0.61	0.65	0.76	1	0.55
Township	0.95	0.86	0.78	0.81	0.55	1

Figure 4. Temporal Correlation Coefficient (TCC) between historical precipitation data.

*: AHCCD data

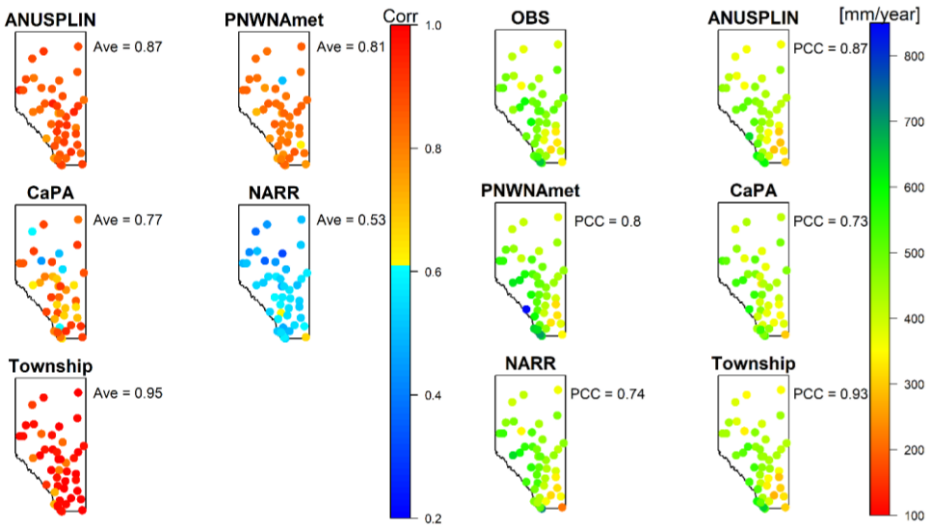


(a) D_{KS} (b) P_{bias}



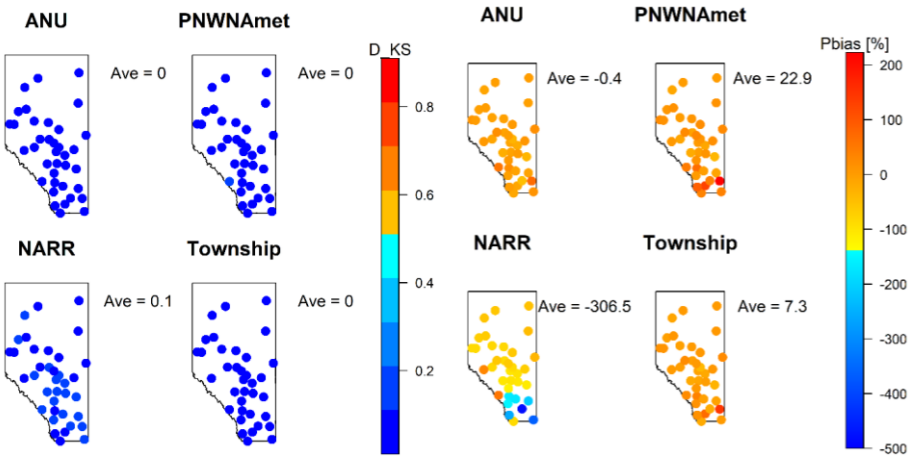
(c) σ_{ratio} (d) RMSE

Figure 5. Maps of performance measures for AHCCD precipitation stations in Alberta

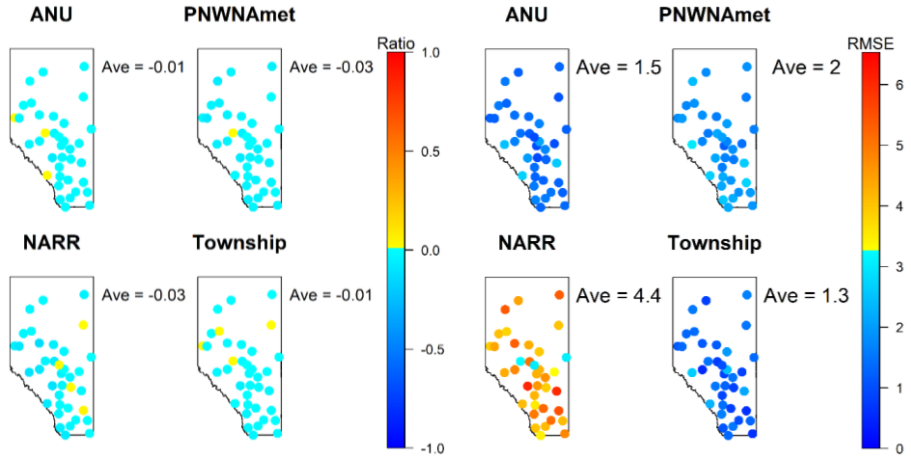


(e) TCC (f) Mean annual precipitation

Figure 5. Continued

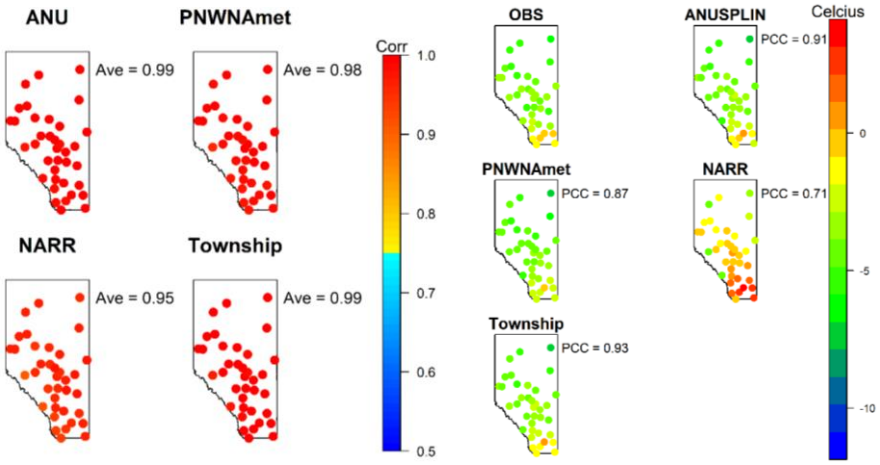


(a) D_{KS} (b) P_{bias}



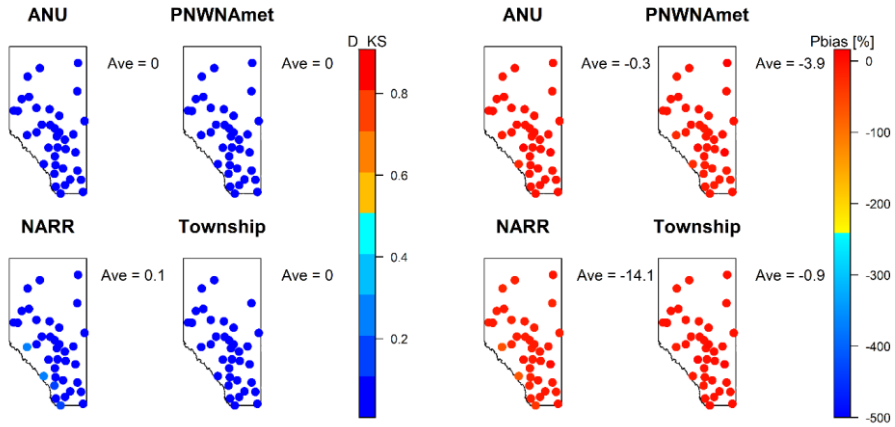
(c) σ_{ratio} (d) RMSE

Figure 6. Maps of performance measures for minimum temperature over the AHCCD stations in Alberta

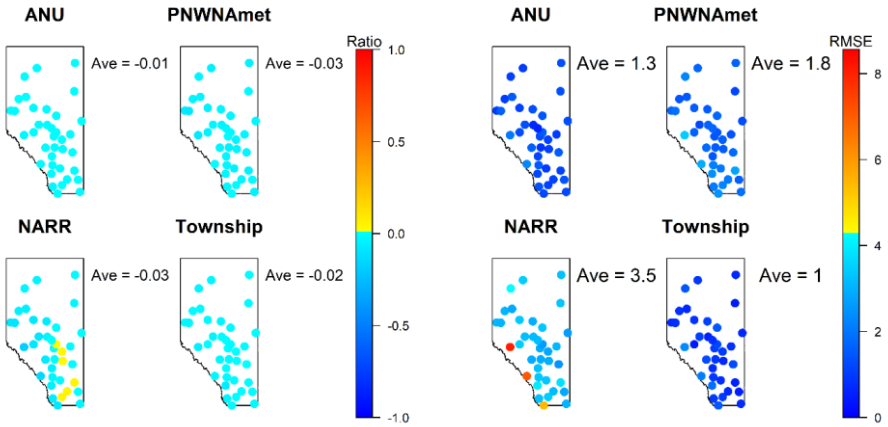


(e) TCC (f) Mean annual minimum temperature

Figure 6. Continued

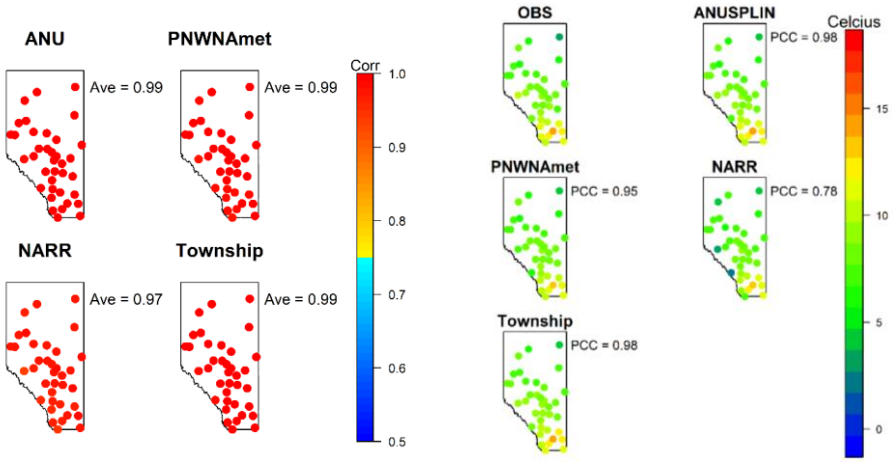


(a) D_{KS} (b) P_{bias}



(c) σ_{ratio} (d) RMSE

Figure 7. Maps of performance measures for maximum temperature over the AHCCD stations in Alberta



(e) TCC (f) Mean annual maximum temperature

Figure 7. Continued

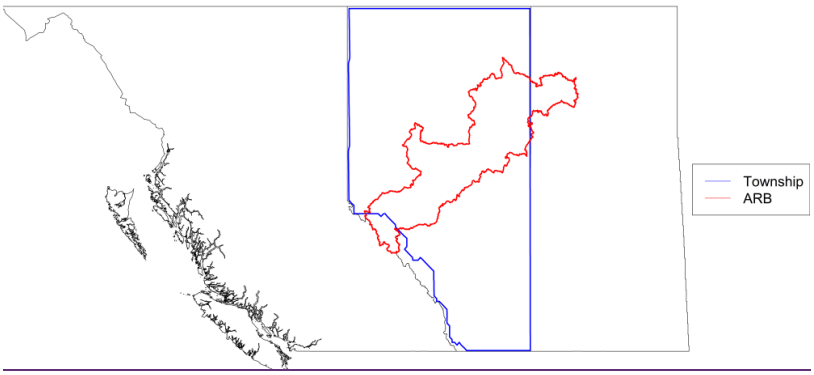


Figure 8. Domain of the Township dataset (blue line) and the boundary of the Athabasca River basin (red line)

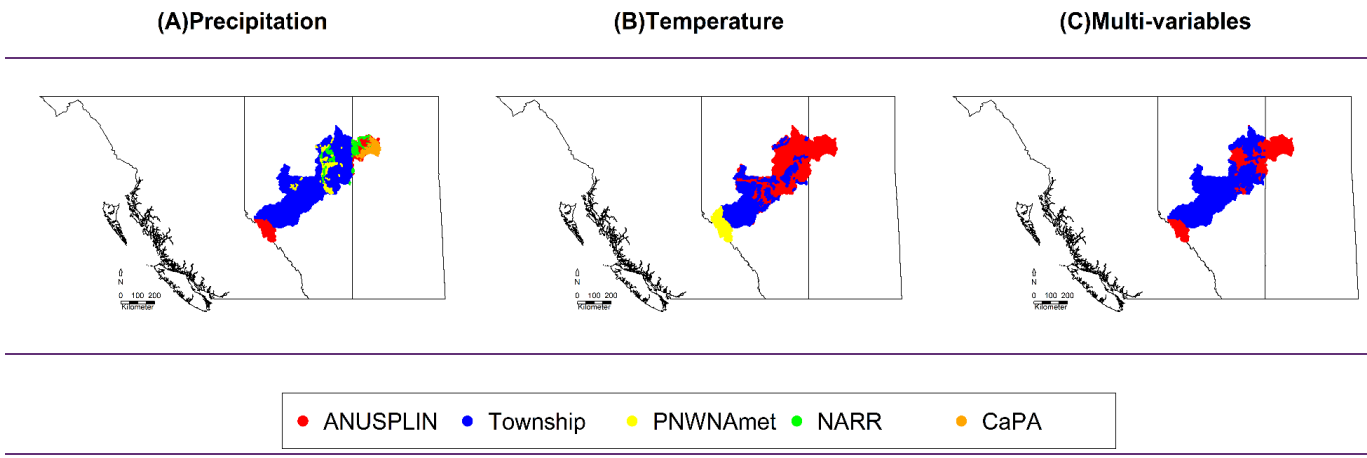
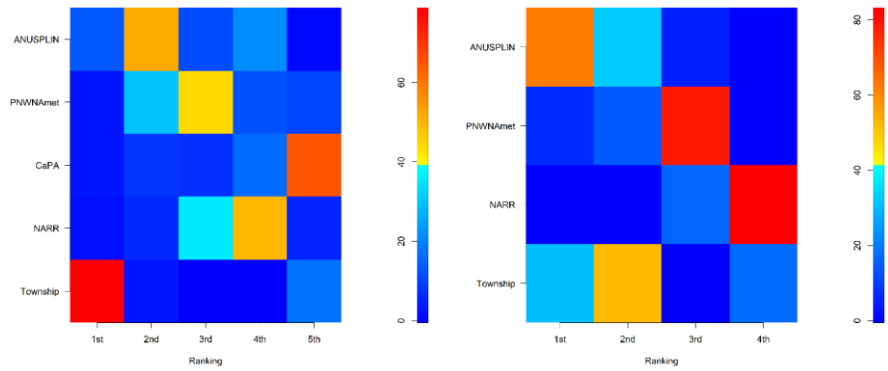
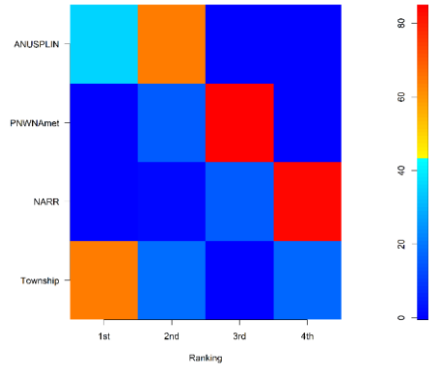


Figure 9. Maps of the first-ranked climate datasets in ARB for the individual variable (A and B) and multi-variables (C)



(a) Precipitation

(b) Temperature



(c) Multi-variables

Figure 10. Percentage of climate datasets on each rank for R_{ind} and R_{mul}

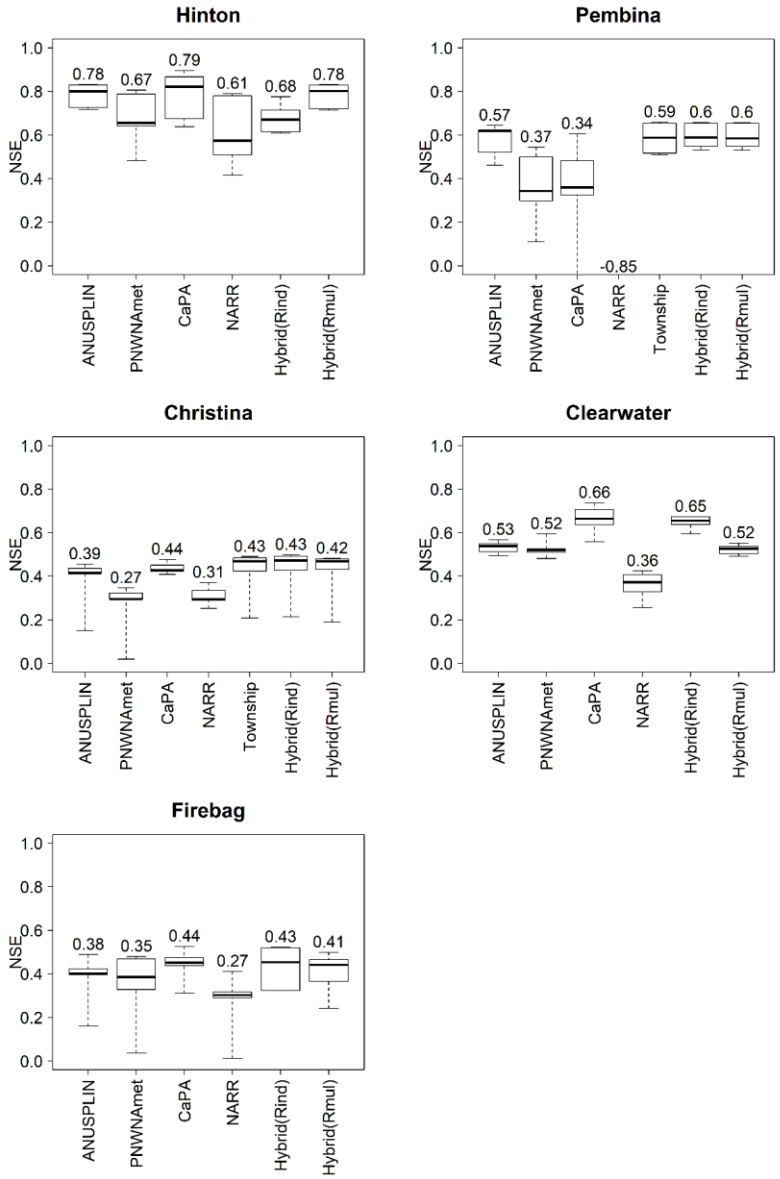
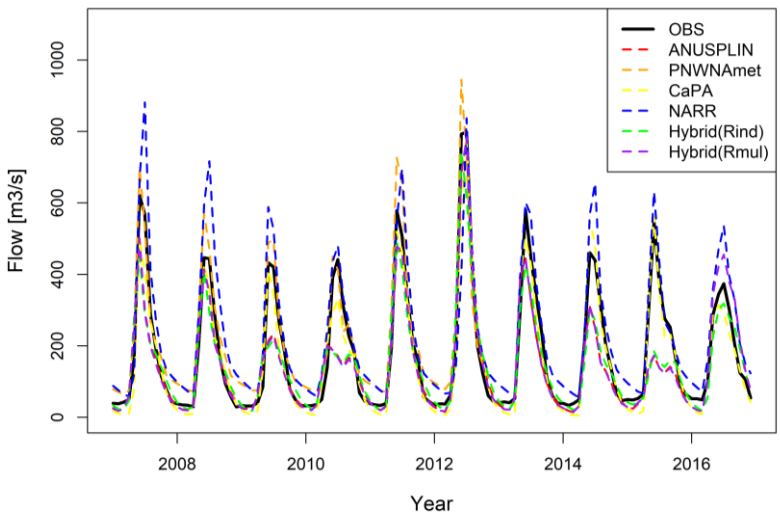
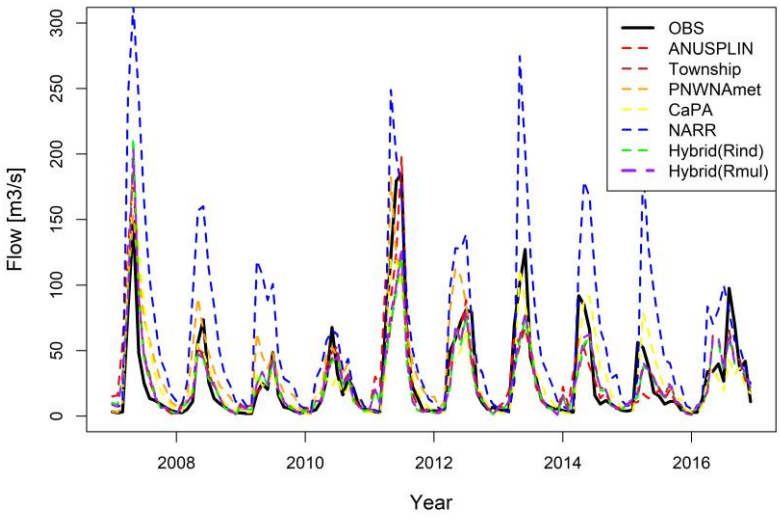


Figure 11. Boxplots of the NSEs of the proxy validation at the five sub-basins in ARB. The values above each boxplot represent the average over NSEs of the proxy validation.

Deleted: NSE



(a) Hinton



(a) Pembina

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

Deleted: Daily

Deleted: six