<<Reviewer 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 x 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.

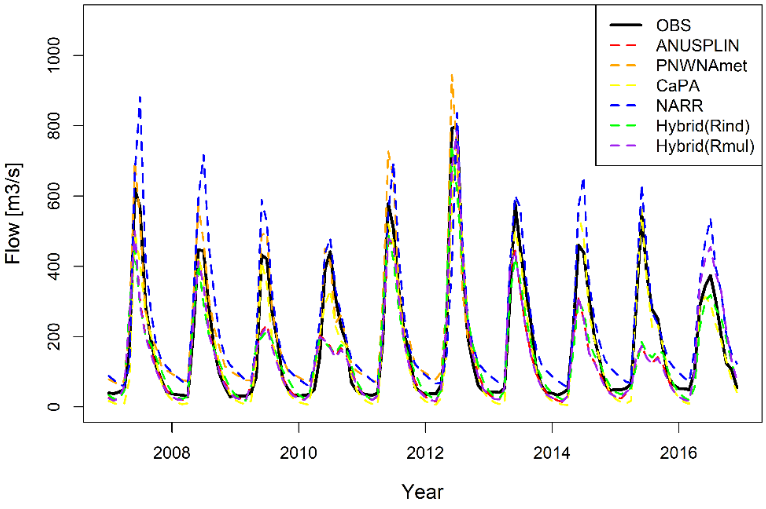
((Reply)) In case of Pembina watershed with NAAR data set: The NSE value for calibration period (1985 to 1997) is 0.5 while it is -0.14 for the validation period (1998 -2016). There are some reasons of such a poor performance of NARR in most of the watersheds including Pembina. Since 2003, assimilation 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
$0.75 \leq \text{NSE} \leq 1.00$	$0.65 \leq \text{NSE} \leq 0.75$	$0.50 < \text{NSE} \leq 0.65$	$\text{NSE} \leq 0.50$

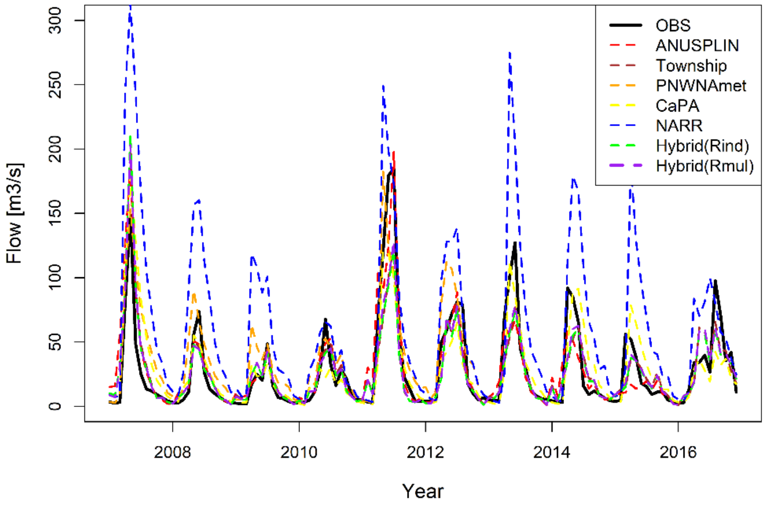
The goodness-of-fit statistics table shows modelling is satisfactory when NSE > 0.5. Table 6 presents Nash-Sutcliffe Efficiency (NSE) for the calibration and validation periods at the selected hydrometric stations (Hinton, Pembina, Christina, Clearwater, and Firebag) in the ARB to assess the suitability 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