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], Anil Gupta[1,2]

[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

10   *Corresponding to: Hyung-Il Eum (hyung.eum@gov.ab.ca)*

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 the North
15   American content, which provide 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, length of records,   spatial coverage, temporal resolution, etc). These inconsistencies raise a valid question 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
20   hydrological modeling. This study suggested a framework, called reference reliability evaluation system (REFRES), that systematically determines a ranking of 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
25   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 Precipiation 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 mostly ranked first for
30   precipitation and temperature, respectively. The proxy validation also confirmed the superior performance of hybrid climate datasets compared with the other five individual climate datasets investigated in this study. These results indicate that the hybrid climate dataset provides a better representation of historical climatic conditions and thus, enhancing the reliability of hydrologic simulations.

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

5 1. Introduction

A reliable historical climate dataset is essential to understand the climatic and hydrological characteristics of a watershed as it is a crucial forcing for simulating key processes of water and energy cycle 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, a poor network density still exists especially in western mountainous and northern parts of Canada. Moreover, the 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 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 produces 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 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 has been developed to produce near real-time precipitation analyses (6-hr accumulated precipitation) over North America at 15 km

5   resolution which has been further improved to 10km resolution (Lespinas et al., 2015). It is an example of a multiple source-based climate dataset. 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, however, multiple climate data sets were combined to

10   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 an 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

15   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 varying spatial density of the observational climate network in the basin and limited data assimilation. Wong et al. (2017) also intercompared gridded precipitation datasets

20   derived from different data sources over Canada. Few studies have attempted to incorporate spatially varied performance measure of various climate datasets in producing a complete long-term historical climate dataset for a study region (Faramarzi et al., 2015; Shrestha et al., 2017), however, 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

25   (REFRES), to systematically determine the ranking of multiple climate datasets based on their performance and to 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, climate datasets were ranked to generate a

30   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 full period of record overlapped or 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 using individual climate datasets and the generated hybrid climate dataset. Therefore, the aims of this study is 1) to develop a

5   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 superiority of hybrid climate dataset.

10   **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,

15   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, expanding to 464

20   precipitation stations. Vincent et al. (2012) produced the 2nd generation of homogenized daily temperature by adjusting 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).

25   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

30   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 Peasce 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).

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 were 10   included in the generation of hybrid climate dataset for Alberta.

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 15   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 and include 3-climate variables, e.g. daily precipitation, minimum and 20   maximum air temperature, and 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, ranged 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 inter-compare climate products (Eum et al., 2014a; Wong et al., 2017), to evaluate the accuracy of regional climate models (Eum 25   et al., 2012), and hydrologic modeling (Islam and Dery, 2017; Eum et al., 2017; Dibike et al., 2018).

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 30   (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 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).

5   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.,

10   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 climate conditions of the Township center.

15   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.

20   **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

25   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

30   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 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).

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 is 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**ference **R**eliability **E**valuation **S**ystem (REFRES) that consists of three main modules (refer to Figure 2); 1) performance measure module (PMM) to evaluate various performance measures for each climate dataset, 2) 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) data generation module (DGM) to produce a hybrid climate dataset by selecting the most reliable climate dataset based on the ranking provided by RM. These three modules are seamlessly integrated and exchange the required data and information to generate a hybrid climate dataset. The next section provides more details on each module.

Figure 2. Structure of REFRES comprised 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., domain). Therefore, the inverse distance squared weighting method was applied to obtain the values at the

5   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

10   Kolmogorov-Smirnov D statistic (D_{KS}) and standard deviation ratio (σ_{ratio}). Sequence-related performance is assessed by percentage of bias (P_{bias}), root mean square error (RMSE), and temporal correlation coefficient (TCC). Spatial pattern-related performance is evaluated by 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 daily time series of climate variables and PCC is obtained by the mean annual precipitation and temperature

15   of the AHCCD stations over a target domain. Therefore, PCC varies with the user specified target domain.

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

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

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

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

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

where G and O represent gridded and observed climate datasets, 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 calculated for all the selected climate datasets and variables at each AHCCD station. Figure

25   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 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 previous module. For a given target cell, AHCCD stations are selected based on two criteria, distance and elevation. First, 20% (of all AHCCD) stations are selected based on the nearest distance citeria which were then reduced to half 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.

As multiple performance measures are employed in this study, there are situations when a climate dataset may perform well for some measures but not in others. Therefore, a multi-criteria decision-making (MCDM) technique is required to systematically rank all the climate datasets 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 determine 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 ith alternative (climate dataset in this study) and j^{th} criterion (i.e., a performance measure), and a weighted normalized decision matrix, $(t_{ij})_{m \times n}$ is given by

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

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

where, m and n are total number of alternatives and criteria, respectively, n_{ij} is normalized matrix by Eq. (7), and w_j represents weighting on the jth criterion. Under the assumption that all performance measures are fairly 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)

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

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

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

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

Where, t_{bj} and t_{wj} are the best and worst decision matrix 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

5   preference for all climate datasets was determined by the similarity (S_{iw}) to the worst condition in Eq. (15).

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

When $s_{iw} = 1$, 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 evaluated the performance measures (criteria) for individual climate variables, TOPSIS can be applied to

10   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 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

15   measures) for precipitation and temperature individually, while R_{mul} was determined by 4×18 decision matrix (4 climate datasets excluding CaPA that provides only precipitation by 18 performance measures from three variables). To alleviate an 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

20   together to rank the climate datasets for temperature.

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 RM. The tool is flexible enough to provide output in various common formats,

25   i.e., NetCDF, ASCII (text), or the specific format of a numerical model. As all of the historical gridded climate datasets has 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

30   on the rank (performance) of each dataset at target grid cells. Thus, 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, when 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. 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 fomat. 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, with 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 confirm the superiority of the derived hybrid climate dataset over other existing climate datasets. 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 for a finer spatial resolution at 1/32° (2~3 km) to better simulate the complex river network in the Lower Athabasca River basin. Five catchment areas listed in Table 2 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. The number of gridded climate datasets was optimized for each catchment area to maximize the evalution of the hydrologic simulations using the hybrid climate data relative to the multiple gridded climate datasets . Hinton is located near the headwaters of ARB which are characterized by mountainous topography and snow- and glacier-ice melt dominated hydrologic regimes. Pembina is one of the major rivers in the middle reach, and 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 number 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. More details on calibration can be found in Eum et al. (2017). Under the assumption of REFRES that all of the existing climate datasets have an equal quality for hydrologic simulations, all of the calibrated parameter sets can be considered as mostly plausible parameter sets for the selected sub-basins. 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. In addition to climate forcings, streamflows are mainly affected by geographic charactersitics 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). Under the assumption that the calibrated parameter sets are suitable for hydrologic simulations in each sub-basin, 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. Therefore, the variability of hydrologic simulations may signify the reliability of the climate dataset used to produce the 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 climate dataset and show low variability in hydrologic simulations, it indicates that the climate dataset may provide reliable historical hydrologic simulations. 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) does not cover the entire catchment areas of these three hydrometric stations.

3. Results

3.1 Precipitation performance measures in Alberta

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 underestimate 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

5   have positive P_{bias}. This indicates that ANUSPLIN and Township may underestimate extreme precipitation and other climate datasets (especially multiple sources and reanalysis data) overestimate extreme precipitation. These results are consistent with findings in Eum et al. (2014a) that CaPA and NARR overestimate extreme precipitation events by overly reflecting the orographic effects on precipitation in wetern Alberta.

10   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. TCC between ANUSPLIN and Township was the highest among climate datasets except for the observations (i.e., OBS) even though they incorporated different interpolation

15   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, reanalysis-based climate dataset, showed higher TCC with CaPA than with other datasets as it is assimilated with multiple sources of observations.

20   Maps of each performance measure are shown in Figure 5. It is evident from the spatial variability that the ANUSPLIN and Township datasets outperformed other datasets in D_{KS} throughout Alberta. In the mountains region of southwest Alberta, most of the climate datasets performed poorly in P_{bias}, σ_{ratio}, RMSE, and PCC, resulting mainly from sparse observation network and inconsistent observations near the Canada-US border. PNWNAmet highly overestimates the mean annual precipitation in the mountainous area (e.g.,

25   > 300 mm/year for station ID 3050519), which may considerably affect simulated streamflows originating in mountainous headwaters and further downstream.

3.2 Air temperature performance measures in Alberta

The performance measures for air temperature averaged over 37 AHCCD stations in Alberta are

30   presented in Table 4. As CaPA provides only precipitation, it was excluded in the assessment for temperature. 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 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, 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 temperature while a cold bias in extreme warm temperature.

3.3 Ranking of climate datasets in the ARB

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 the 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 (Case A and Case B) and multi-variables (Case C), i.e., precipitation and (minimum and maximum) temperature in this study.

For precipitation, the Alberta township dataset was ranked first in most of grid cells within the basin (78%), 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 the Township dataset overwhelmingly outperformed other climate datasets for precipitation.

For temperature, ANUSPLIN was ranked first (in 62% grid cells), followed by Township (31%) and PNWNAmet (7%). 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. 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 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.

5   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 and 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), on the contrary, ANUSPLIN was ranked mostly first and Township was a distinct second choice in the majority of grid cells, followed by

10   PNWNAmet and NARR. For Multi-variables (Case C), Township and ANUSPLIN were the first and second choice in the majority of grid cells in the ARB.

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 superior performance and consequently recommended

15   for future climate-related studies. A proxy validation approach was applied using both the generated hybrid climate datasets to determine the superiority of one dataset over the other.

3.4 Proxy validation of generated hybrid climate datasets

In addition to the five gridded climate datasets, the two hybrid climate datasets were implemented for

20   proxy validation using the VIC model. Since CaPA 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 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

25   hydrometric stations, most of the climate datasets performed well with the exception of NARR in the Pembina catchment. The two hybrid climate datasets performed slightly better 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 VIC simulations. The NSE ranges were obtained from multiple VIC simulations with each climate dataset (as climate forcing) for all the plausible model parameter sets which were calibrated with seven climate datasets, individually. The values above each boxplot represent

30   the averaged value of NSEs over the multiple hydrologic simulations. A narrower range of NSE values represents 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 a 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, which might be a reason that 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 in further assessment of the precision and accuracy even though all 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 and suitable as a forcing dataset. For Clearwater, Hybrid(R_{ind}) is the top performer, followed by Hybrid(R_{mul}), ANUSPLIN, PNWNAmet, and NARR. Clearwater has 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 used the same temperature data from NARR. For Firebag, Hybrid(R_{ind}) also showed the 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 a 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.

4. Discussion

Among the station-based gridded climate datasets, 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 enormously overestimated precipitation over mountainous area, which considerably affects simulated low flows at Hinton in the ARB. Figure 12 shows observed and simulated hydrographs from six climate datasets (excluding Township) at Hinton, the headwater region in the ARB. It clearly shows that PNWNAmet highly overestimated the low flows during winter, which is caused by overestimated precipitation in the drainage area of Hinton. As with PNWNAmet, NARR also overestimates the low flows during winter, which is induced by 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).

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 300~400 stations in the 1960s while 400~500 since 1970. A change-point analysis of these datasets may provide some useful information to

5   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, it may also have the

10   inconsistencies issue 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, and 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

15   western Canada, and is low in the cold climate of mountainous and northern areas. Therefore, the ranking could further be improved with a more uniform density of AHCCD stations over western Canada.

Much literature 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-

20   elevation regions (Hinton in this study) while it did not perform so well in the middle and lower reaches, i.e., lower-elevation watersheds. Especially, NARR performed poorly in the Pembina sub-basin, a region where hydrologic simulations are highly sensitive to model parameters other than climate forcings (Eum et al., 2014a). 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. Since the validation

25   period in this study is only 7 years from 2010 to 2016, however, 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 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

30   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 could be 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

5   the order of preference for combining various climate datasets at each grid cell. The proxy validation approach also confirmed the superiority of a generated hybrid climate dataset over other data sets especially in hydrologic simulations.

5. Summary and concluding remarks

10   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 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-

15   (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

20   -   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 southwest Alberta, including mountainous areas, due to sparse observation network, orographic effects, topographic complexity, and inconsistency
25   in observation between Canada and 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.

30   -   Hybrid(R_{ind}) especially outperformed other climate datasets in Clearwater sub-basin where the highest number of climate datasets were combined in generating Hybrid(R_{ind}) for precipitation. This clearly indicates that the hybrid climate dataset generated by REFRES may lead to more reliable hydrologic simulations and thus 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 representative of the historical climatic conditions. Therefore, the hybrid climate dataset should be used as
5 a reference dataset for statistical downscaling and hydrologic model forcing, resulting in more reliable high-resolution climatic and hydrologic projections.

Acknowledgements

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

References

Asong, Z. E., Khaliq, M. N. and Wheater, H. S.: Regionalization of precipitation characteristics in the
15 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
20 change impacts on the hydrology and water resources of the Colorado River Basin, Hydrology and
Earth System Sciences Discussions, 11(4), 1417–1434, 2007.

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.

25 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–
30 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

5   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.

10   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:

15   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, 28(14), 4250–4271, doi:10.1002/hyp.10236, 2014a.

Eum, H.-I., Yonas, D. and Prowse, T.: Uncertainty in modelling the hydrologic responses of a large

20   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.

25   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.

30   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.

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.

5 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.

Hwang, C.L. and Yoon, K.: Multiple attribute decision making: methods and applicasions. Springer, New 10 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 15 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.

20 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.

25 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, 30 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,

5   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.

10   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

15   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.

Shen, Y., Xiong, A., Wang, Y. and Xie, P.: Performance of high-resolution satellite precipitation products

20   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

25   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

30   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

5   *: 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 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	Climate dataset				
	ANUSPLIN	Township	PNWNAmet	NARR	CaPA
(A) Precipitation	2985 (13%)	17515 (78%)	691 (3%)	499 (2%)	682 (3%)
(B) Temperature (Min & Max Temp.)	13809 (62%)	6924 (31%)	1639 (7%)	0 (0%)	-
(C) Multi-variables	8049 (36%)	14323 (64%)	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.
ANU SPLIN	0.88	0.83	0.61	0.64	0.52	0.46	0.76	0.54	0.61	0.49
Town ship	-	-	0.62	0.66	0.54	0.49	-	-	-	-
PNWNA met	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

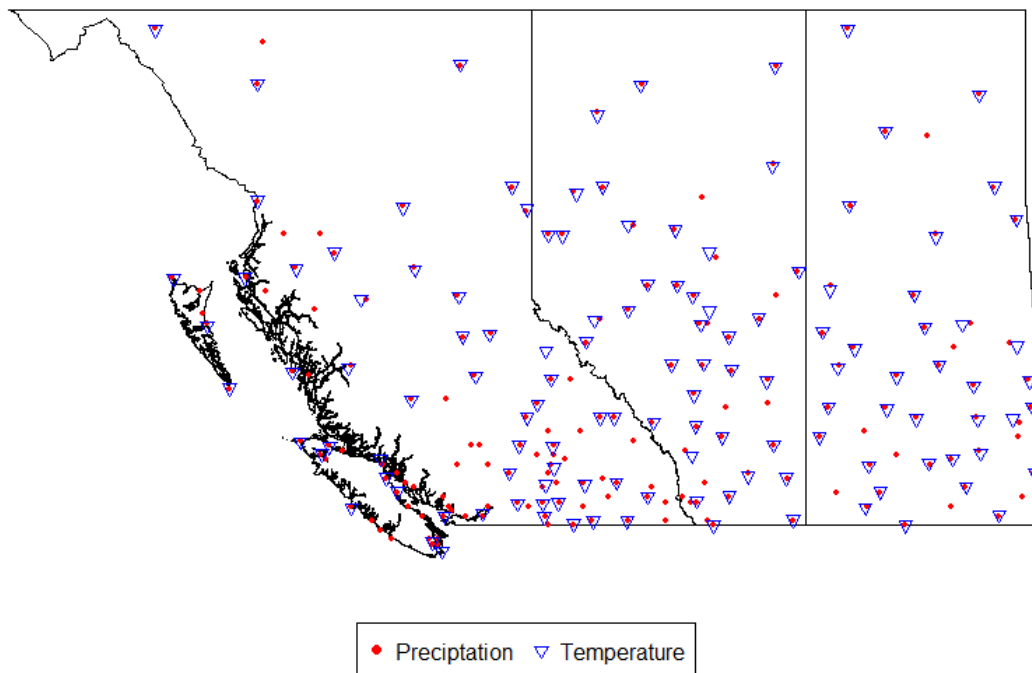



Figure 1. AHCCD stations within the BC, AB, and SK provinces

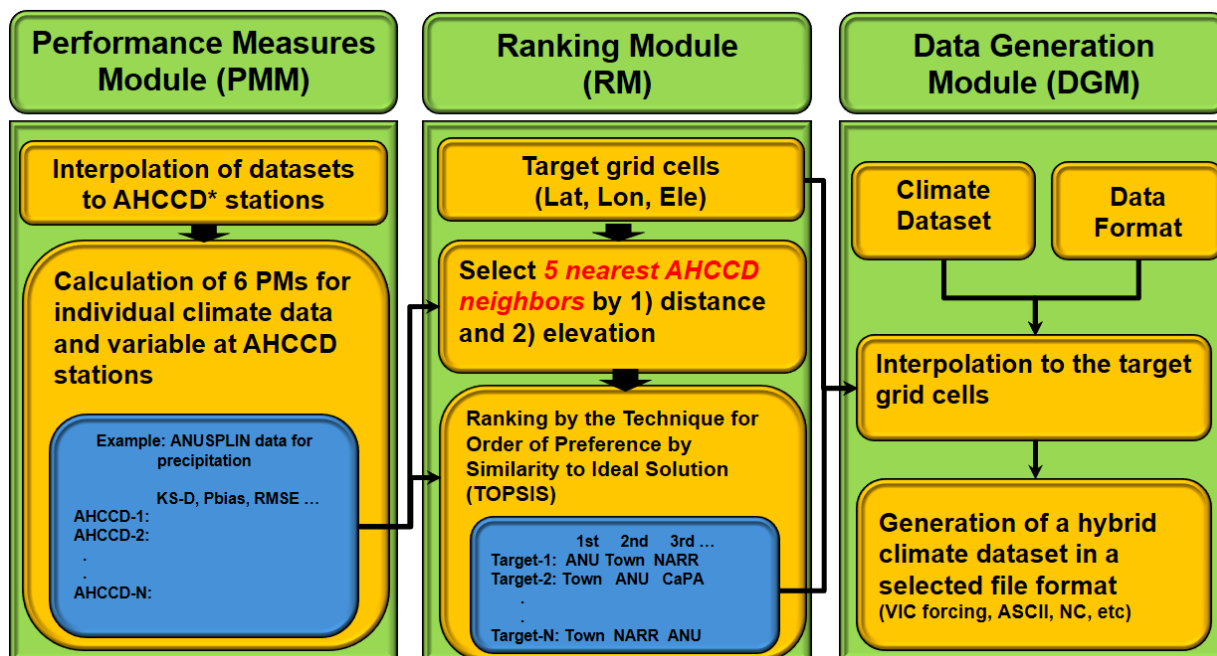



Figure 2. Structure of REFRES comprised of three modules; 1) Performance Measure Module (PMM), 2)

Ranking Module (RM), and 3) Data Generation Module (DGM)

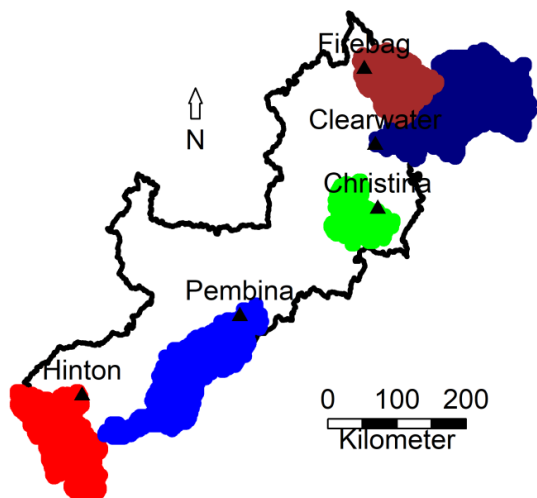



Figure 3. Five sub-basins selected in the Athabasca River basin for the proxy validation

	OBS*	ANUSPLIN	PNWNAme	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

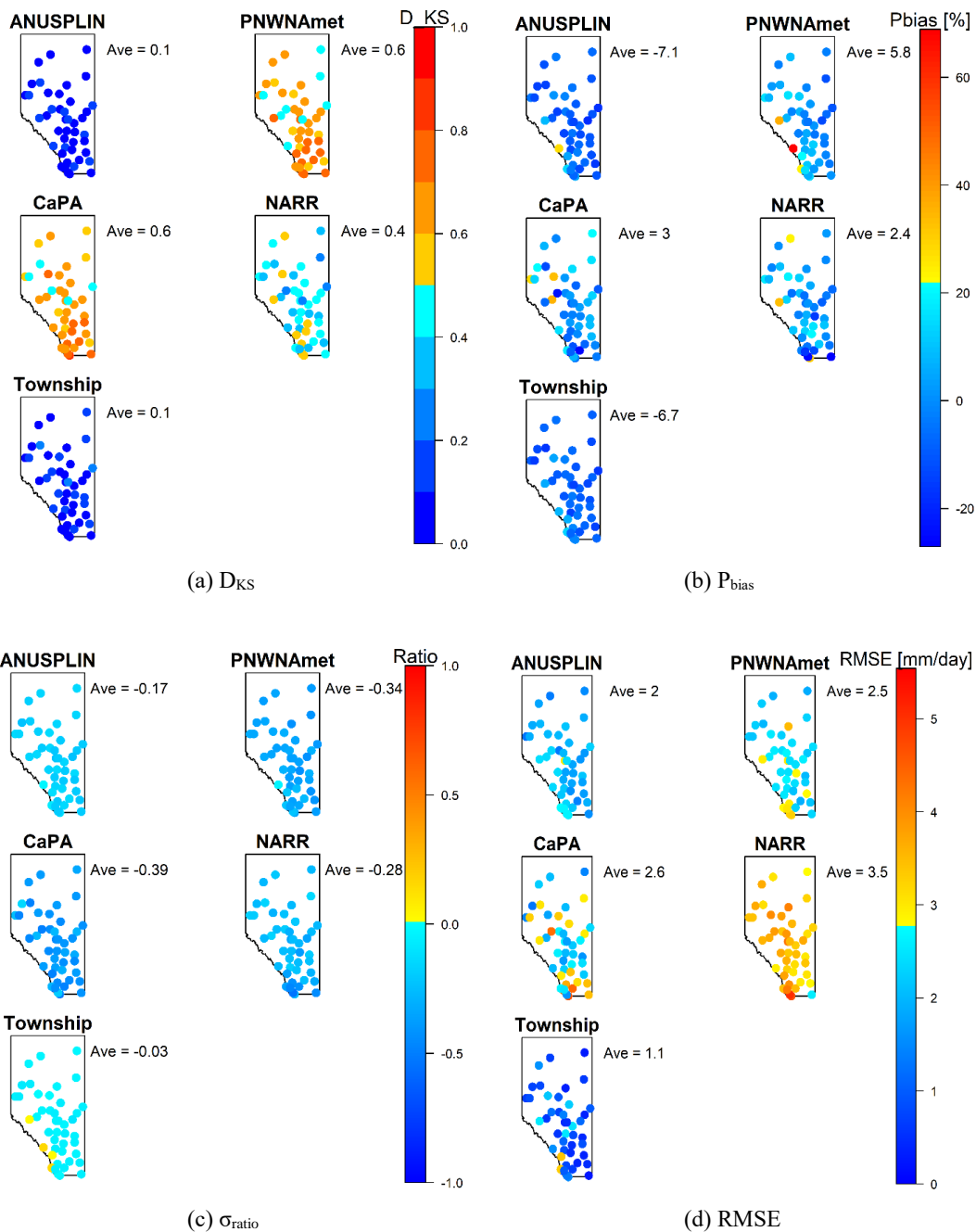



Figure 5. Maps of performance measures for AHCCD precipitation stations in Alberta

(e) TCC   (f) Mean annual precipitation

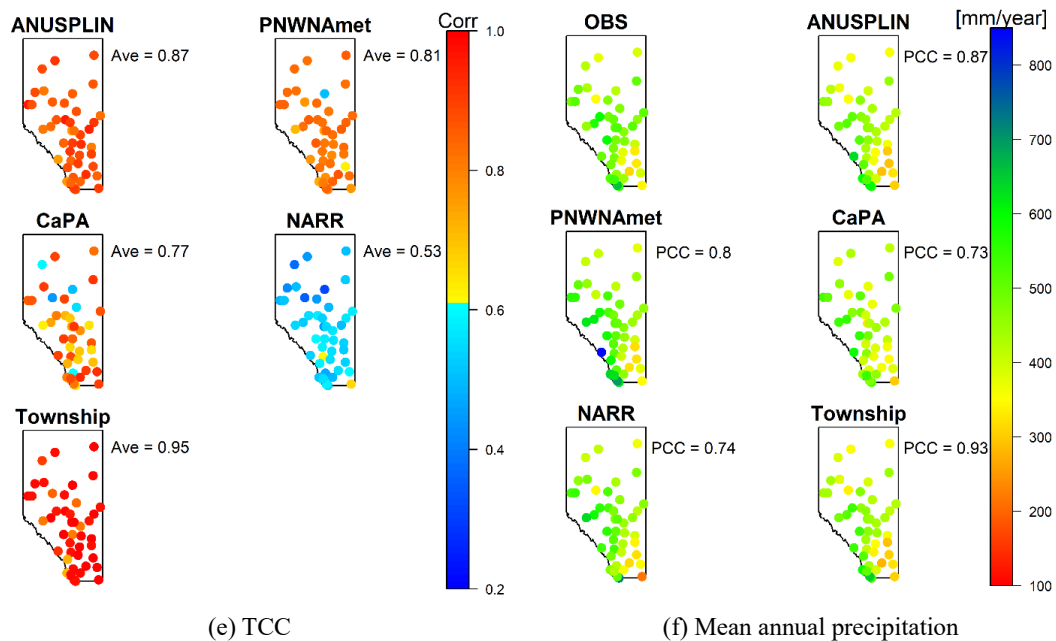



Figure 5. Continued

(a) D$_{KS}$

(b) P$_{bias}$

(c) σ$_{ratio}$

(d) RMSE

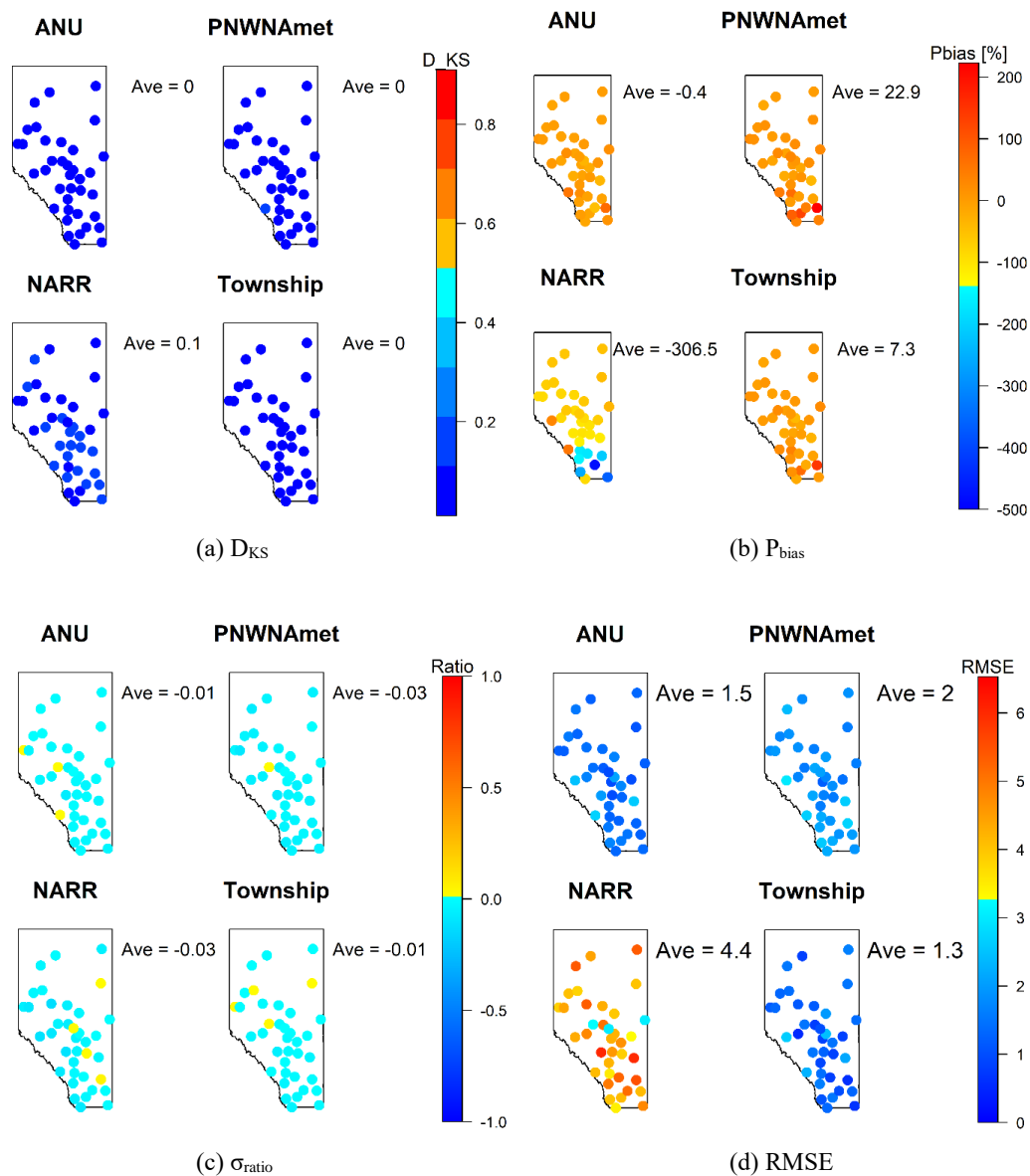



Figure 6. Maps of performance measures for minimum temperature over the AHCCD stations in Alberta

(e) TCC   (f) Mean annual minimum temperature

Figure 6. Continued

(e) TCC
(f) Mean annual maximum temperature

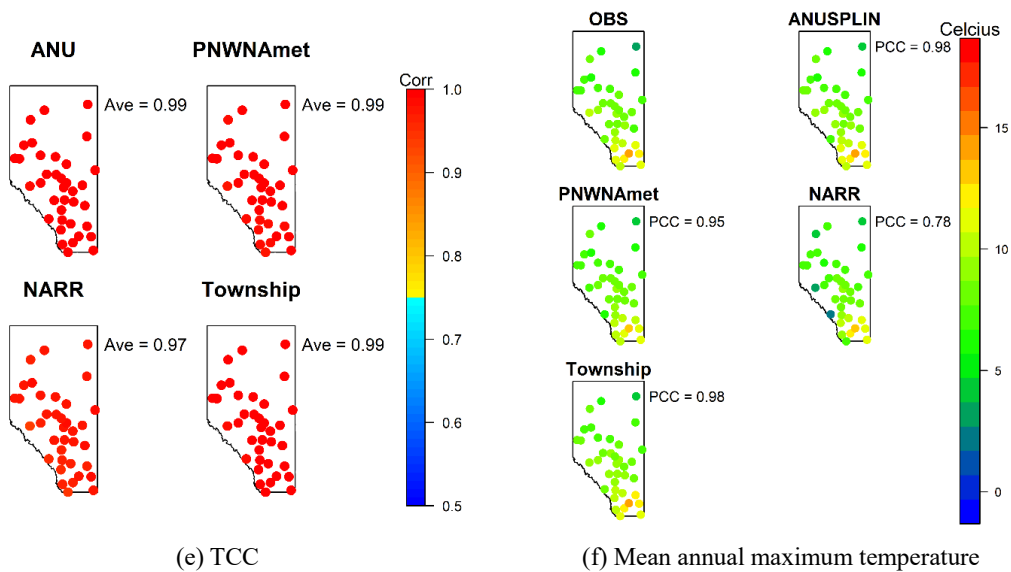



Figure 7. Continued

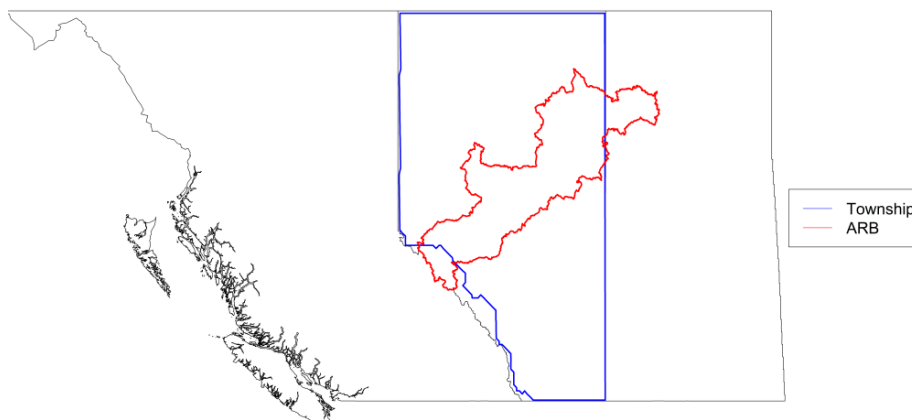



Figure 8. Domain of the Township dataset (blue line) and the boundary of the Athabasca River basin (red line)

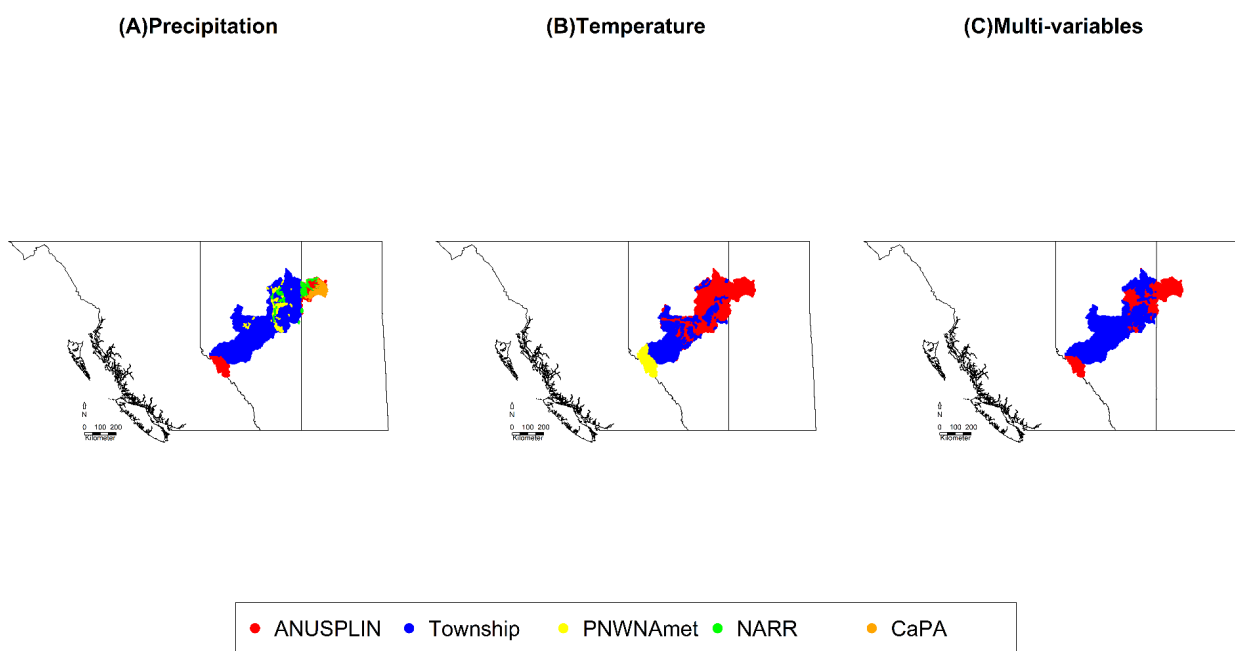
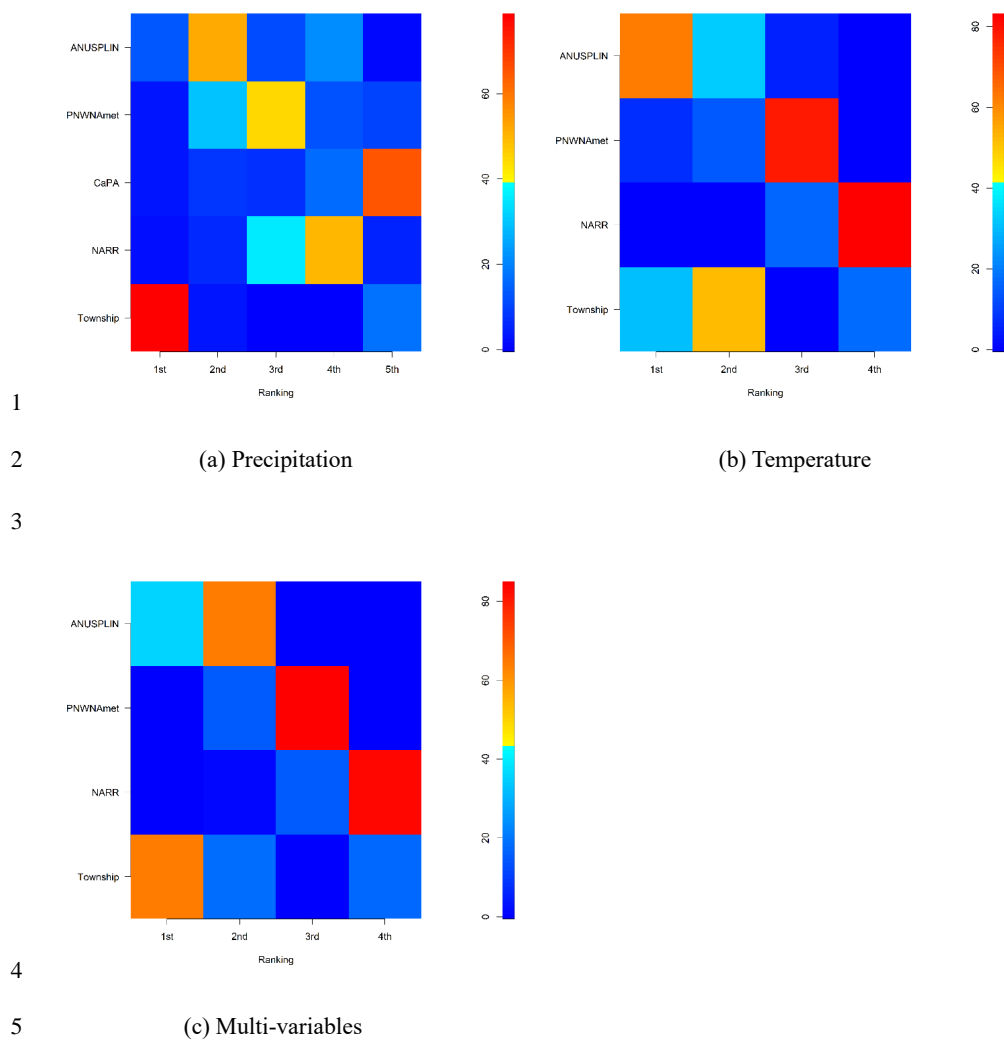
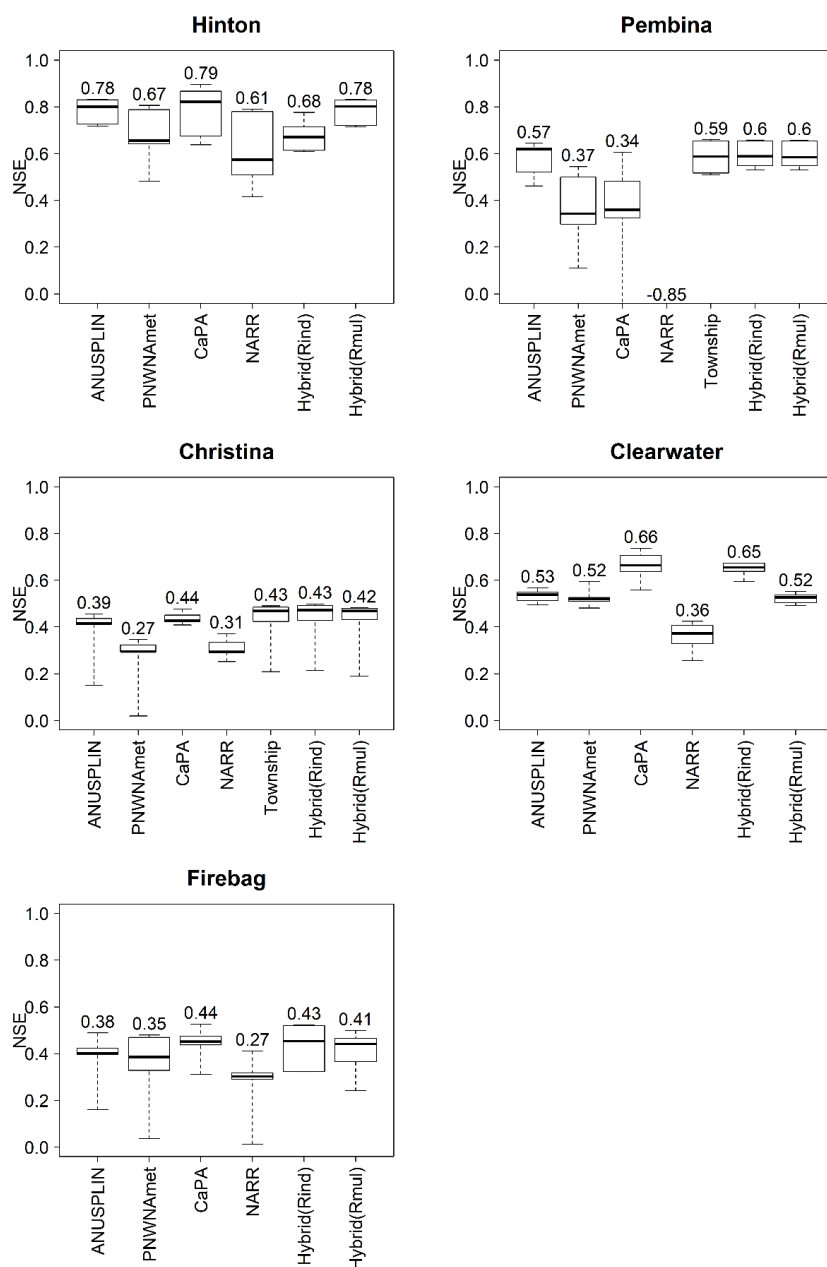



Figure 9. Maps of the first-ranked climate datasets in ARB for the individual variable (A and B) and multi-variables (C)

2   (a) Precipitation   (b) Temperature

5   (c) Multi-variables

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

2   Figure 11. Boxplots of NSE of the proxy validation at the five sub-basins in ARB. The values above

3   each boxplot represent the average over NSEs of the proxy validation.

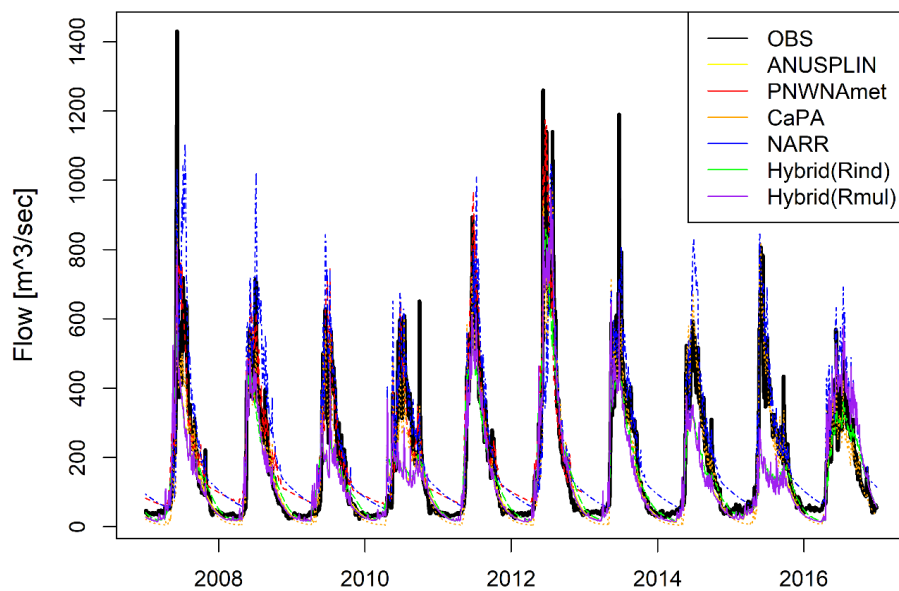



Figure 12. Daily observed and simulated hydrographs from six gridded climate datasets at Hinton