water

MDPI

Article

Use of Ensemble-Based Gridded Precipitation Products for Assessing Input Data Uncertainty Prior to Hydrologic Modeling

Scott Pokorny [1,*], Tricia A. Stadnyk [1,2], Rajtantra Lilhare [3], Genevieve Ali [4], Stephen J. Déry [3,5] and Kristina Koenig [6]

1 Department of Civil Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada; tricia.stadnyk@ucalgary.ca
2 Department of Geography, University of Calgary, Calgary, AB T2N 1N4, Canada
3 Natural Resources and Environmental Studies Program, University of Northern British Columbia, Prince George, BC V2N 4Z9, Canada; lilhare@unbc.ca (R.L.); stephen.dery@unbc.ca (S.J.D.)
4 School of Environmental Sciences, University of Guelph, Guelph, ON N1G 2W1, Canada; gali@uoguelph.ca
5 Environmental Science and Engineering Program, University of Northern British Columbia, Prince George, BC V2N 4Z9, Canada
6 Manitoba Hydro, Winnipeg, MB R3C 0G8, Canada; kkoenig@hydro.mb.ca
* Correspondence: umpokors@myumanitoba.ca

Received: 31 August 2020; Accepted: 30 September 2020; Published: 2 October 2020

Abstract: The spatial and temporal performance of an ensemble of five gridded climate datasets (precipitation) (North American Regional Reanalysis, European Centre for Medium-Range Weather Forecasts interim reanalysis, European Union Water and Global Change (WATCH) Watch Forcing data ERA-Interim, Global Forcing Data-Hydro, and The Australian National University spline interpolation) was evaluated towards quantifying gridded precipitation data ensemble uncertainty for hydrologic model input. Performance was evaluated over the Nelson–Churchill Watershed via comparison to two ground-based climate station datasets for year-to-year and season-to-season periods (1981–2010) at three spatial discretizations: distributed, sub-basin aggregation, and full watershed aggregation. All gridded datasets showed spatial performance variations, most notably in year-to-year total precipitation bias. Absolute minimum and maximum realizations were generated and assumed to represent total possible uncertainty bounds of the ensemble. Analyses showed that high magnitude precipitation events were often outside the uncertainty envelope; some increase in spatial aggregation, however, enveloped more observations. Results suggest that hydrologic models can reduce input uncertainty with some spatial aggregation, but begin to lose information as aggregation increases. Uncertainty bounds also revealed periods of elevated uncertainty. Assessing input ensemble bounds can be used to include high and low uncertainty periods in hydrologic model calibration and validation.

Keywords: uncertainty; hydrology; gridded climate data; precipitation

1. Introduction

High-quality precipitation data are essential for an accurate representation of physical environments by hydrological models. In Canada, mountainous and northern regions suffer from data sparsity and paucity issues, with significant gaps in existing long-term records (e.g., [1–4]). This makes observed ground-based climate station data inadequate for data-intensive applications (e.g., forcing hydrologic models), establishing a need for reliable alternatives (e.g., gridded datasets) and estimates of the uncertainty associated with those alternatives. Uncertainty is defined here

as the realistic range of a value or variable [5]. In hydrologic modeling, while uncertainty can be introduced through input data, model structure, parameters, and observed calibration data [6], input precipitation data uncertainty are less often studied; instead, input data sources are selected based on "data quality" criteria (e.g., [3]). To estimate the total uncertainty bounds for hydrologic model output, estimation of input data uncertainty subject to propagation within a hydrologic model is required [7]. Propagation is defined as the transfer of uncertainty from one modeling step to the next, thus allowing an assessment of how cumulative sources of uncertainty impact model output. While some studies have found input uncertainty to be the largest component of total uncertainty (e.g., [8]), there is variation among the approaches presented in the literature when it comes to quantifying that uncertainty. This study addresses several challenges that have been identified in the literature regarding input data uncertainty arising from gridded precipitation products; most notably, challenges related to performance metrics, data product selection, spatial aggregation, temporal period selection, and the definition of ensemble limits.

Gridded precipitation data are used for various applications, and with the numerous products available, assessing performance relative to observed benchmarks becomes a significant issue (e.g., [3,4,9–18]). Performance metrics are used to compare a data product to an observed reference; the most common types of performance metrics are continuous statistics (e.g., percentage bias (PBIAS); [4]), categorical statistics (e.g., equitable threat score (ETS); [19]), extreme indices (e.g., comparison of consecutive wet-days (CWD) between observed and simulated datasets; [20]), and proxy validation by hydrologic models (e.g., [16]), among others. Continuous statistics measure the agreement between simulated and observed timeseries. Categorical statistics measure the agreement between simulated and observed (precipitation) events binned by magnitude. Extreme analysis compares the occurrence and magnitude of extreme events between simulated and observed timeseries, while proxy validation applies input forcing to hydrologic models. In proxy validation, simulated and observed output timeseries for relevant variables are compared using statistics such as the Nash–Sutcliffe efficiency score (NSE score; [21]) or ensemble-based metrics such as reliability, which assess the overlap of modeled uncertainty relative to observed data [22], and inferences are then made about the quality of the input data. The observed data used for comparison are often ground-based climate station data, as they are typically considered the highest-quality observations [2,22]. A comprehensive assessment should consider at least three broad aspects of performance: timing, magnitude, and occurrence of extremes. Performance metrics are often used in the literature to suggest an optimal dataset; yet, performance metric selection is often limited to only two of the three desired performance aspects, therefore limiting the interpretation of results.

Regarding data product selection, it is common in climate change studies to select multiple global climate models (GCMs) to represent uncertainty (e.g., [8,23–26]); where the selected datasets form an ensemble. Studies using only historical data products, however, infrequently use multiple input data products, and thereby misrepresent historical input data uncertainty. The discrepancy in the treatment of precipitation data between climate change and historical studies comes from the widely agreed upon assumption that future climate projections have higher uncertainty than historic climate conditions, which are considered to be known. This, however, does not mean that historical precipitation data are free of uncertainty (e.g., [27]). Furthermore, the inherent future variability of climate change suggests that historical uncertainty envelopes cannot be projected into future periods by assuming stationarity (e.g., [8,28]).

Studies performed within the past decade afford the opportunity to examine historical uncertainty, given the increasing number of gridded climate datasets that have been developed for spatial extents, ranging from Canada-wide (e.g., [18,29]), to North America (e.g., [30]), and global (e.g., [31]). Gridded climate data comparisons generally focus on precipitation (rather than temperature), due to the tendency toward higher uncertainty in the estimation of occurrence, magnitude, and position of storm events (e.g., [4]). Comparisons vary depending on the spatial and temporal extent in each study,

and the performance metrics selected. Uncertainty in both station and gridded data products is often discussed, owing to the paucity and sparsity of ground-based observations.

As uncertainty remains difficult to quantify, many studies do not offer a clear choice of a single best product; instead, concluding to proceed with caution (e.g., [16]). Bukovsky and Karoly [10] compared a suite of gridded precipitation datasets and found that the North American Regional Reanalysis (NARR) performed better over the continental United States of America (USA) compared to outside the USA. Becker et al. [11] further examined NARR and found a systematic bias towards overestimation of light precipitation and an underestimation of extremes. Essou et al. [15] compared the hydrologic performance of NARR with other commonly used gridded products, including ERA-Interim, European Union Water and Global Change (WATCH) Forcing Data Era-Interim (WFDEI), and others: NARR generally produced higher NSE scores through proxy validation, while WFDEI performed best among the global products. Eum et al. [3] compared NARR, Natural Resources Canada's (NRCAN) Australian National University Spline interpolation (ANUSPLIN), and the Canadian Precipitation Analysis (CaPA) at three climate stations in the Athabasca River Basin in western Canada. NARR was found to have a statistical break that occurred in January 2004, coinciding with the discontinuation of Canadian climate station data assimilation [3]. Wong et al. [4] compared a suite of gridded datasets in ecodistricts and found that WFDEI and CaPA performed well, but both varied spatially in their performance. Rapaić et al. [13] compared a large suite of observation-based and reanalysis datasets for the Canadian Arctic and found no dataset was "best"; instead, concluding multiple datasets should be considered [16].

Studies often focus on gridded data performance at a single spatial aggregation (i.e., the spatial averaging of multiple grid points) for varied temporal periods, typically spanning many years. Gridded data are generally aggregated spatially to simplify the comparison with climate station data (e.g., [4]). Aggregation in other studies has generally been dictated by the spatial discretization scheme used by a single hydrological model structure (i.e., distributed, semi-distributed, or lumped [32]). Aggregation is often done by areal averages used to upscale grids [4], inverse distance weighting within a specified region [4], or by bilinear interpolation [20]. Performance assessment is, therefore, limited, because a fixed spatial aggregation leaves a gap in our knowledge, specifically how input uncertainty propagates into hydrologic models of varied spatial structures.

The spatial variability of dataset performance is a common factor complicating dataset intercomparison studies (e.g., [3,4,11,13,15,16,18]). A recent study by Lilhare et al. [33] further highlighted the spatial and temporal variability of gridded datasets in the Lower Nelson River Basin (LNRB) in central Canada. Results presented by Lilhare et al. [33] showed that performance and trends were affected by the amount of spatial aggregation considered. Mekis and Vincent [2] and Wong et al. [4] summarized the limitations associated with meteorological station records, and Choi et al. [12] stated that, due to data sparsity, spatially aggregated comparisons may generate misleading results, due to the varied resolutions of the gridded datasets. Understanding the uncertainty introduced into hydrologic modeling by gridded precipitation data and assessing how best to account for this uncertainty remain major gaps in the literature. The present study addresses the aforementioned gaps, in that, prior to configuring and executing the hydrologic models, we aimed to: (1) characterize model input data uncertainty arising from precipitation dataset selection, temporal period of analysis, and spatial aggregation of a model, and (2) quantify the reliability and limitations of ensemble methods for gridded precipitation data ensembles at spatial aggregations that are representative of common hydrologic model structures.

Our study is organized into seven sections. Section 2 introduces the region of study. Section 3 introduces the observed datasets (Section 3.1) and the gridded datasets (Section 3.2) used in this study. Section 4 presents the study methodology for performance metric selection (Section 4.1), ensemble creation (Section 4.2), and how spatial aggregation was defined (Section 4.3). Section 5 presents our results for gridded dataset (Section 5.1) and ensemble performance (Section 5.2). Section 6 discusses inferences made from results regarding time period selection (Section 6.1), spatial aggregation

(Section 6.2), and overall ensemble reliability (Section 6.3). Conclusions and recommendations are then presented in Section 7.

2. Study Area

The aforementioned objectives were addressed across a mid-latitude watershed which presents important challenges related to input data uncertainty propagation in hydrologic modeling, including its size, paucity of ground-based climate station data, and sensitivity to climate and land-use change. The Nelson–Churchill Watershed (NCW) covers approximately 1.4 million km^2 of the North American landmass, and can be sub-divided into several sub-drainage basins ([34]; Figure 1a).

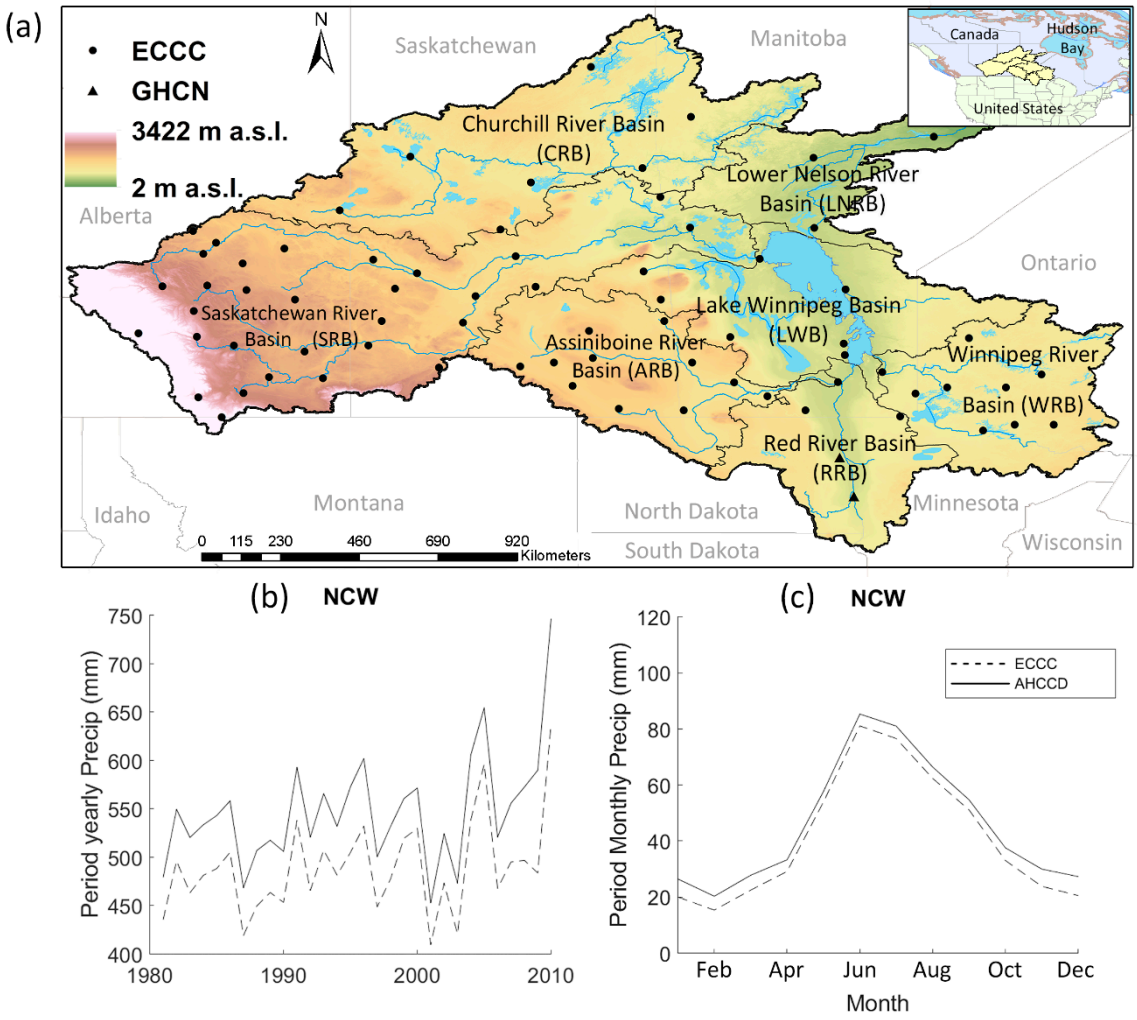

Figure 1. (**a**) Map of the Nelson–Churchill Watershed, including sub-basin delineations and 73 selected ground-based station locations. All Environment and Climate Change Canada (ECCC) stations also have a corresponding Adjusted and Homogenized Climate Change Data (AHCCD) station at the same location. (**b**) Spatially averaged yearly total precipitation timeseries, including all stations shown in (**a**). (**c**) Mean monthly precipitation including all stations shown in (**a**) for 1981–2010.

The NCW spans the USA–Canada border, multiple ecodistricts (see Wong et al. [4] for details), and diverse climatic regions, which generally increase uncertainty among the gridded precipitation data (e.g., [10]). The NCW includes the low-relief prairies across southern Manitoba, Saskatchewan, and Alberta; a mountainous region on the western edge of the Saskatchewan River Basin (SRB); data-sparse, cold climate northern regions (i.e., Sub-Arctic, and the Northwestern boreal forest); extensive water bodies (e.g., Lake Winnipeg, and extensive wetlands); and administrative and

inter-jurisdictional boundaries; each of which introduces unique data and modeling challenges (Figure 1a). Data assimilation challenges are specifically known to occur along the USA–Canada border region, resulting from changes in observation station density [10]. For an in-depth description of the NCW, see Benke and Cushing [35].

3. Precipitation Data

3.1. Observed Ground-Based Climate Station Data

Two sets of ground-based climate (precipitation) station data (referred to hereafter as observed data) were selected for comparison to the gridded precipitation data. The first are near-real-time, daily Environment and Climate Change Canada (ECCC) observations. The second are the daily Adjusted and Homogenized Climate Change Data (AHCCD) [2]. AHCCD use data retrieved from the National Climate Data, consisting of 464 ground-based precipitation timeseries adjusted for common precipitation measurement issues, such as wind undercatch, evaporation, wetting losses, and trace precipitation [36]. The 464 stations were extracted from ECCC data having long, continuous records; multiple stations were combined to extend record lengths when and where possible [2]. Inconsistent methods for the handling of trace precipitation through time and between stations created inhomogeneities in the ECCC data. Therefore, AHCCD utilized varied trace precipitation adjustments based on available metadata to better 'homogenize' these data; however, Mekis and Vincent [2] state that it is possible that inhomogeneities still exist. Mekis and Vincent [2] also state that significant uncertainty remains in these data for extremes, and high spatial and temporal variability in the snow water equivalent adjustment are present; therefore, AHCCD data are not recommended for short-term applications or extreme events, such as blizzards. AHCCD are used by most comparison studies focused on Canada, as they are considered to be the best representation of true observations (e.g., [4]). Hence, in the present study, all comparisons were made to AHCCD unless stated otherwise (i.e., AHCCD were used as station observations in performance metrics).

Two additional stations were used in the USA portion of the Red River Basin (RRB), retrieved from the Global Historical Climatology Network-Daily (GHCN-Daily) database. The GHCN-Daily data quality assurance (QA) procedures include checks for location accuracy, incorrect station identifiers, entries that violate the intended documentation, and other inconsistencies [37]. Data were subjected to 19 tests of data quality, outlined by Durre et al. [38], as well as a secondary tier of quality assurance that evaluate climatic consistency.

For the present study, station selection was based on precipitation data availability, considering both AHCCD and ECCC data. GHCN-Daily data are more similar to ECCC data than to AHCCD data, but no equivalent product for AHCCD is available in the USA portion of the RRB. Therefore, the GHCN-Daily data were included in both the ECCC and AHCCD data comparisons. Selected stations had minimal missing data. A total of 73 stations are selected (Figure 1a): an average of all 73 stations from 1981–2010 across the NCW domain produce the average yearly total precipitation timeseries (Figure 1b) and monthly average annual precipitation (Figure 1c). AHCCD generally has higher precipitation than ECCC, with the largest difference among products often associated with solid precipitation events (Figure 1c) [2]. Lake Winnipeg buoy data were excluded because of different QA processes. Unlike the USA stations, Lake Winnipeg buoy data would not have significantly improved station density or area representation. The final list of selected precipitation gauges is presented in Table S1.

3.2. Gridded Precipitation Datasets

This study is part of the Hudson Bay Systems (BaySys) project [39], a large multi-disciplinary project to partition the environmental effects of river regulation from climate change. The models used to accomplish the goals of the BaySys project require large-scale, gridded climate data for the data-sparse regions of the northern hemisphere, including the Arctic Ocean. Therefore, the current

study focuses on materials and methods relevant to the BaySys project. A total of five gridded datasets were selected based on the following criteria: identification in the literature as high-performing, data availability from 1981 to 2010, and daily temporal resolution or finer. Selected datasets are summarized in Table 1.

Table 1. Main characteristics of the five gridded climate datasets selected for the current study.

Name	Period (Temporal Resolution)	Domain (Spatial Resolution)	Reference	Product Description
The Australian National University spline interpolation (ANUSPLIN)	1950–2013 (daily)	Canada (~0.1°)	[29]	Interpolated ECCC dataset using trivariate thin-plate smoothing spline between latitude, longitude, and elevation. The version updated to cover 1950–2013 was used; a version extending up to 2016 was released after the completion of this study.
North American Regional Reanalysis (NARR)	1979–~present (3 hourly)	North America (~0.32°)	[30]	A reanalysis dataset with many sources of assimilated data, such as the global reanalysis product GR2, gauge observations, and others. NARR stopped assimilating Canadian station data in 2004, which introduced a detectable statistical break [3]. In 2015, the period of April 2009–January 2015 (and thereafter) was updated to address some data processing issues, which improved border effects along the USA–Canada border, particularly focused on southern Ontario.
European Centre for Medium-Range Weather Forecasts interim reanalysis (ERA-Interim; ERA-I)	1979–~present (3-hourly)	Global (0.75°)	[40]	A reanalysis dataset that assimilates a large number of data sources, such as the Integrated Forecast System (IFS) cy31r2, satellite data, and others. ERA-I is a replacement for the previous ERA-40 dataset, featuring 4D-VAR data assimilation among other improvements to the original ERA-40, which stopped in 2002.
European Union Water and Global Change (WATCH) Forcing data ERA-Interim (WFDEI)	1979–2013 (3-hourly)	Global (0.5°)	[41]	An adjusted version of ERA-I using the European Union Water and Global Change (WATCH) Forcing Data (WFD) methodology, which includes various adjustments and bias corrections. These data are a replacement for the original ERA-40-based WFD dataset. The version updated to cover 1979–2013 was used; a version extending up to 2016 was released after the completion of this study.
Global Forcing Data—Hydro (GFD-HYDRO)	1979–~present (3-hourly)	Global (0.5°)	[31]	GFD-Hydro closely mimics the methodology of WFDEI, with updates to current versions of observed data networks. GFD-HYDRO is meant to be a global product similar to WFDEI, but produced at near real-time. Notable differences between WFDEI and GFD-HYDRO exist for precipitation, due to a reduction in undercatch adjustments.

Gridded datasets span a variety of temporal resolutions; therefore, data were aggregated to the largest time step feasible for hydrologic modeling (i.e., daily). For comparison, Wong et al. [4]

chose to upscale and limit data selection to datasets with spatial resolutions of 0.5° and finer, while Rapaić et al. [13] chose to interpolate to the resolution of ECCC's Canadian Gridded temperature and precipitation data series (CANGRD) at 50 km. There is no commonly accepted best practice, as all methods will introduce some degree of uncertainty. Aggregation and interpolation are further discussed in Section 4.3.

4. Methodology

4.1. Performance Assessment

Analyses of individual datasets were conducted to ensure each product is a reasonable representation of observed data. Months with missing data were excluded from the performance assessment following the World Meteorological Organization standards, in which months are excluded if more than five days total, or three consecutive days, are missing [42].

4.1.1. Continuous Statistics

Three continuous statistics were selected for the evaluation of gridded dataset performance: standard deviation ratio (SDR), percent bias (PBIAS), and Spearman's rank correlation coefficient (Cor) (Equations (1)–(3)):

$$SDR = \frac{\sqrt{\frac{\sum_{i=1}^{N}\left(G_i - \overline{G}\right)^2}{N}}}{\sqrt{\frac{\sum_{i=1}^{N}\left(R_i - \overline{R}\right)^2}{N}}} - 1 \quad (1)$$

$$PBIAS = \frac{\sum_{i=1}^{N}(G_i - R_i)}{\sum_{i=1}^{N}(R_i)} \times 100 \quad (2)$$

$$Cor = 1 - \frac{6\sum_{i=1}^{N}(G_i - R_i)^2}{N(N^2 - 1)} \quad (3)$$

in which N is the number of time steps of observed data and corresponding gridded dataset data, and G and R are gridded and reference timeseries, respectively. *SDR* measures the ratio of the standard deviations of a gridded and station timeseries (one is subtracted from the *SDR* so that positive values are associated with higher standard deviations; values near zero are desired). *PBIAS* measures the tendency of a gridded dataset to over- or under-predict a reference timeseries (smaller values are desired), and *Cor* represents a gridded dataset's ability to correctly reproduce the timing of observed precipitation data (values near one are desired). Spearman's rank correlation coefficient is selected, as it is non-parametric and weights large differences higher than small ones, which was desirable in the current study, as small differences may lie within the uncertainty of the observations.

4.1.2. Categorical Statistics

Categorical statistics measure precipitation events captured within binned ranges, which are used to evaluate event occurrence partitioned by magnitude. This provides more information on uncertainty contributions associated with event magnitude [19,20]. Categorical statistics were a standard metric used in the development and assessment of CaPA [18]. The events captured are measured by a contingency table (Table 2).

Table 2. Contingency table to assess when an event is correctly represented by bin for categorical statistics.

	Observed	
Simulated	Obs = 1	Obs = 0
Sim = 1	Hit (H)	False Positive (F)
Sim = 0	Miss (M)	Correct Negative (C)

A value of 1 represents an event occurring within a bin, and a value of 0 represents an event not occurring within that bin. Bin size selection is adopted from the World Meteorological Organization (WMO) standards: [0, 0.2); [0.2, 1); [1, 2); [2, 5); [5, 10); [10, 25); [25, 50); and [50, inf); all in mm, where square brackets are inclusive and curved brackets are exclusive. Similar to Asong et al. [20] and Lespinas et al. [19], two categorical statistics are used that are also ECCC standard evaluation metrics [18]: the equitable threat score (ETS, Equations (4) and (5)), and the frequency bias (FBIAS, Equation (6)):

$$ETS = \frac{H - H_R}{H + F + M - H_R} \tag{4}$$

$$H_R = \frac{(H + F)(H + M)}{N} \tag{5}$$

$$FBIAS = \frac{H + F}{H + M} - 1 \tag{6}$$

in which N represents the total number of hits (H), false positives (F), misses (M), and correct negatives (C); and H_R is the number of correct forecasts assuming completely random forecasts [19]. One is subtracted from *FBIAS* values to make positive values associated with positive bias. The *FBIAS* does not measure agreement with the observations; *FBIAS* only measures relative frequency [19]. The *ETS* does, however, measure skill, adjusted by the number of correct forecasts if forecasts were random. The assumption of random forecasts may overestimate correct random forecasts, which lowers the *ETS* score [43].

4.1.3. Extreme Indices

Extreme indices measure the occurrence and magnitude of extreme events [28]. Each extreme index was evaluated for seasonal, annual, and full temporal periods. Two precipitation extreme indices were selected: dry spell length (CDD) measuring consecutive days < 1 mm, and wet spell length (CWD) measuring consecutive days ≥ 1 mm. The goal of including extreme indices was to measure precipitation persistence [20]. In addition to extreme indices, one categorical extreme metric was included: the Symmetric Extreme Dependency Index (SEDI) (Equations (7)–(9)):

$$SEDI = \frac{lnF_s - lnH_s + ln(1 - H_s) - ln(1 - F_s)}{lnF_s + lnH_s + ln(1 - H_s) + ln(1 - F_s)} \tag{7}$$

$$H_s = \frac{H}{H + M} \tag{8}$$

$$F_s = \frac{F}{C + F} \tag{9}$$

in which H is the number of hits, F is the number of false positives, C is the number of correct negatives, H_s is the hit rate, and F_s is the false alarm rate. The SEDI metric measures the agreement of the occurrence of extreme precipitation between a gridded dataset and a reference dataset. Following from Asong et al. [20], daily precipitation events above the 75th percentile within each sub-basin were considered extreme events. Station data are reflective of the local precipitation at the location of the climate station, while a grid cell represents a spatial average. The SEDI measures the agreement of

extreme occurrence within a bin, where using a binned metric reduces the reliance on the replication of local precipitation, but the scale difference still impacts extreme value performance.

4.2. Ensemble Creation

The five gridded datasets are treated as an ensemble, which avoids the assumption of a single dataset being "best". To quantify performance limits for the ensemble, upper and lower bounds were generated by selecting the minimum and maximum value of the gridded products in a time step for each grid point. Since each member is considered to be an acceptable representation of the observed environment, the minimum and maximum ensemble members represent the total possible bounds of uncertainty for the selected gridded dataset ensemble at any timestep, location, or aggregation. Maximum and minimum bounds are useful for estimating total possible reliability for the ensemble, as defined by Montanari [44], but likely represent an overestimation of uncertainty beyond what would be considered useful in an operational setting, or for design purposes.

For this study, an ensemble mean is computed using an equal weighting of the five datasets, representing a high likelihood realization. Ideally, all observed values occur within the ensemble upper and lower bounds, therefore, the uncertainty representation (referred to as reliability) is estimated by the number of values within the ensemble range. Precipitation events < 0.2 mm/day were considered as trace precipitation [2]. We note the difference in scale between the point gauges and the areal average grid cells, where a grid cell may capture a small amount of precipitation that is not represented at a point gauge because precipitation did not occur at that exact location (i.e., gauge). Therefore, a gridded dataset and point gauge value for any timestep were assumed to overlap below the trace threshold. Finally, we generate a high and low realization from the second wettest and second driest value for each grid cell for each day. This narrower uncertainty range is included to compare the relative increase in reliability that occurs when the wider minimum and maximum bounds are used.

4.3. Spatial Aggregation

Seasonal, annual, and study period performance analyses were conducted using three spatial aggregations: fully aggregated over the NCW (lumped), aggregated by sub-basin (semi-lumped), and station-based comparison (distributed). Spatial aggregation uses a simple arithmetic mean of points falling inside and along a delineation. Grid cells partially inside a delineation were weighted by their percent overlap. Gridded datasets were aggregated to lumped and semi-lumped aggregations using their original grid resolution. While the difference in grid size will introduce uncertainty into the comparison of gridded datasets, interpolation would have likewise done the same. Gridded datasets were aggregated before upper and lower bounds were generated.

Observed aggregated data were only assumed missing if all stations in a basin were missing data for a particular day. When some, but not all, stations were missing data for a timestep, the spatially aggregated timeseries had lower spatial coverage for that timestep, potentially lowering performance metrics. Spatial aggregation reduces the reliance on storm positioning, hence minimizing the contribution of spatial positioning to uncertainty.

Some hydrologic models ingest meteorological data at the grid point scale; therefore, grid point comparisons were also generated for comparison to spatially aggregated results. Gridded datasets were aggregated to the largest grid (i.e., ERA-Interim at 75 km) by areal average within larger grid cells [4]. For each point gauge, the four nearest grid cells were bilinearly interpolated to the exact location of the gauge [20]. Using the four nearest grid cells reduces representativeness uncertainty [45].

All gridded datasets except ANUSPLIN have data for the full RRB extent; therefore, spatial aggregation in the RRB utilized all data available. This meant that NARR, ERA-I, WFDEI, and GFD-Hydro's RRB spatial averages consider grid points in the USA, while ANUSPLIN's RRB spatial average only considers Canadian grid points. The distributed comparison, however, offers information within the RRB domain that is not dependent on data outside of Canada.

5. Results

5.1. Gridded Dataset Analysis

Period mean monthly plots provided a generalized dataset intercomparison (Figure 2). ERA-Interim was generally the wettest gridded dataset, often wetter than AHCCD, while NARR and ANUSPLIN were often the driest, usually drier than ECCC. The largest mean monthly precipitation differences were in basins with more sparse and non-uniform climate station coverage (Lake Winnipeg Basin (LWB) and LNRB), and basins near the USA-Canada border (Winnipeg River Basin (WRB) and RRB). NARR performed worse in the WRB than in other basins; the 2015 update cited the performance in southern Ontario as a focus for improvement.

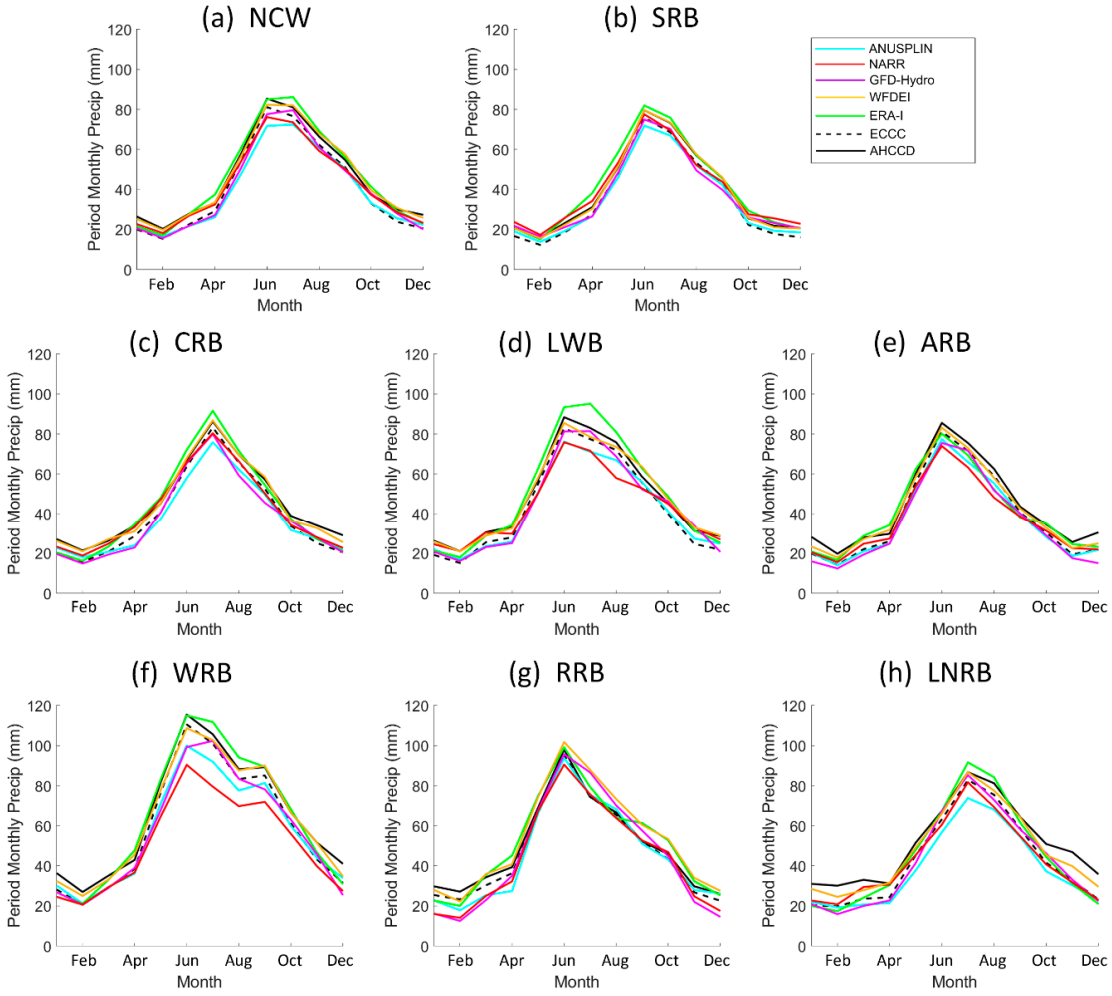

Figure 2. Mean monthly precipitation for 1981–2010, spatially averaged over each of the sub-basins. Aggregations include (**a**) the lumped Nelson–Churchill Watershed (NCW), and the semi-lumped sub-basin aggregations that include: (**b**) the Saskatchewan River Basin (SRB), (**c**) the Churchill River Basin (CRB), (**d**) the Lake Winnipeg Basin (LWB), (**e**) the Assiniboine River Basin (ARB), (**f**) the Winnipeg River Basin (WRB), (**g**) the Red River Basin (RRB), and (**h**) the Lower Nelson River Basin (LNRB).

The performance of each dataset, temporally partitioned into annual periods (Figure 3) is less consistent, indicating variable performance between years. Hydrologic models generally assess performance on a continuous daily or monthly time scale; therefore, Figure 3 reflects general calibration methodologies for hydrologic modeling. Correlations (Figure 3a) were generally higher in larger basins with better climate station coverage, such as the NCW or the SRB; and lower when climate station

coverage was low (e.g., RRB and LWB). Years that showed lower correlations across all precipitation datasets were generally the result of missing data among some of the AHCCD stations included in the aggregations (Table S2). Basin averages were only considered "missing" if all stations were missing for a given timestep, thereby reducing data coverage when some stations had missing data and lowering resulting correlations. The highest correlations were often from ANUSPLIN, which is expected, because ANSUPLIN is an interpolated product, or the ensemble mean. The year that NARR stopped ingesting Canadian climate station data (2004) showed notably worse performance than other years, caused by the structural changes imposed by the data ingestion step. All reported correlation coefficient values are statistically significant at the 99% confidence level (Figure 3a).

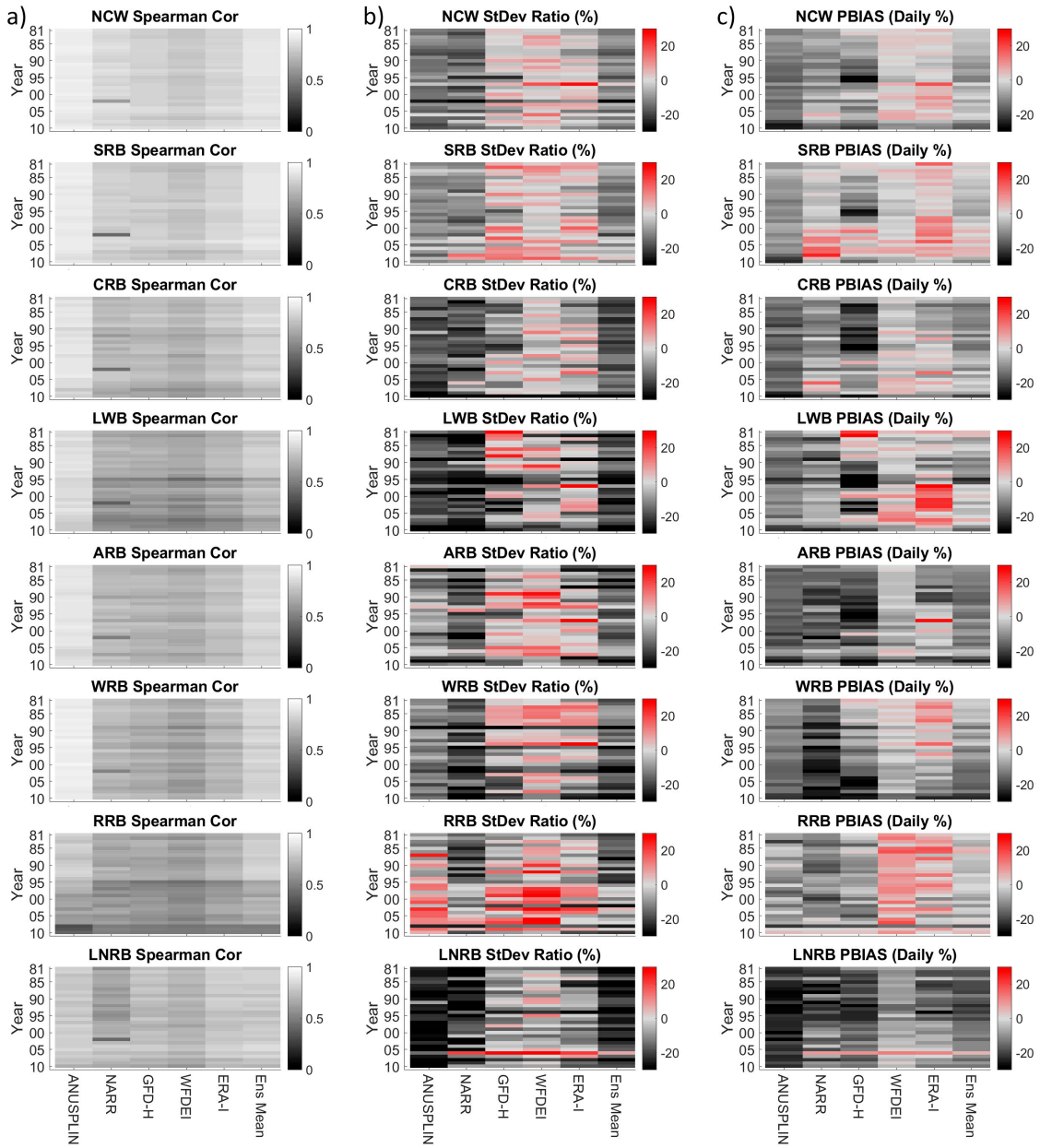

Figure 3. Daily precipitation spatially aggregated annual continuous statistics with reference to the AHCCD observed data set in each sub-basin (1981–2010). (**a**) Daily Spearman correlation, (**b**) daily ratio of standard deviations, and (**c**) daily PBIAS.

Standard deviation ratios were generally negative for ANUSPLIN, NARR, and the ensemble mean; suggesting under dispersion of the precipitation datasets. PBIAS (Figure 3c) highlighted temporal

inconsistencies obscured by the longer-term averaging period used in Figure 2. While Figure 2 suggested the existence of a strong wet bias for ERA-Interim in the LWB, Figure 3 revealed that this wet bias was mainly influenced by the 1997–2004 period. Similar temporal inconsistencies existed for each dataset as performance varied from year to year. The ensemble mean often resulted in the smallest PBIAS. ANUSPLIN showed high correlations, negative PBIAS, and lower variability than observed data. Together, this suggests that ANUSPLIN matches the observed reference timing well, but consistently underestimates precipitation. NARR showed similar PBIAS and ratio of standard deviations to ANUSPLIN in the LNRB, but with lower correlation, suggesting a timing issue as well as underestimation. Positive PBIAS was generally present when positive ratios of standard deviation were present. ANUSPLIN was expected to show negative PBIAS since it was generated from ECCC station data, and AHCCD generally added water to the ECCC data (i.e., was wetter). Similarly, reanalysis datasets that ingest Canadian climate station data use ECCC data (not AHCCD); therefore, under-estimation and negative PBIAS were anticipated.

Performance was generally worse for the distributed comparisons (Supplementary materials: Figure S1), in which yearly correlations ranged, on average, from 0.31 to 0.67 for all climate datasets except ANUSPLIN, whose correlations were generally above 0.7. Since ANUSPLIN was interpolated from ECCC stations, the distributed comparison was expected to perform well, since the two datasets are not independent. It is important to note that observed data were also ingested by the reanalysis products, meaning they, too, are not fully independent. PBIAS and the ratio of standard deviations showed similar amounts of disagreement and temporal variability to those in Figure 3c for each dataset (Figure S2), respectively, but were more often negative. Negative values were expected considering point gauges were compared to grid cells at notably different scales (Figure S3). The performance was generally worst in summer or winter and best in spring or autumn for all three spatial aggregations. Comparisons with ECCC data were similar, except for PBIAS more often suggesting a gridded dataset wet bias. This result was expected, as AHCCD adjustments generally added precipitation.

Categorical statistics, similar to continuous statistics, reflected better performance for lumped or semi-lumped aggregations (Figures S4–S7). ETS scores were generally highest for the 0–0.2 mm bin, the 5–10 mm bin, and the 10–25 mm bin; performance was generally worst for the 0.2–1 mm bin and the 1–2 mm bin. FBIAS scores were more often negative for events below 0.2 mm and those above 2 mm; events between 1 and 10 mm generally had the smallest FBIAS scores, suggesting that the gridded datasets did not have a tendency to consistently over- or underestimate events of those magnitudes. There were fewer high-magnitude events in larger sub-basins; the NCW, SRB, and Churchill River Basin (CRB) often had no events larger than 25 mm. This was expected since (larger) areal averages tend to dampen high precipitation events, decreasing their frequency with respect to station-based records. The distributed comparison ETS scores suggested events above 5 mm to be well represented (Figure S8) and FBIAS values generally decreased, suggesting more often underestimation as precipitation volume increased (Figure S9).

Results from the analysis of the yearly extremes are presented in Figure 4. Consecutive dry and wet days were generally well represented by all gridded products, with some notable temporal variations in performance, suggesting that persistence patterns were often captured by the gridded datasets. SEDI values (Figure 4c) were highest for ANUSPLIN, followed by ERA-Interim. Since ANUSPLIN was interpolated from ECCC station data, its high performance was expected. Many of the events evaluated by SEDI occurred in summer, where the higher resolution of NARR would be expected to resolve convective storms better than ERA-Interim. Spatial aggregation of NARR also averaged more grid points, which would reduce the aggregated performance for higher precipitation events. Excluding ANUSPLIN, the effects of aggregation are notable as a gradient from higher ERA-Interim performance to lower NARR performance. The ensemble average generally outperformed the gridded datasets, with the exception of ANUSPLIN. Distributed comparisons (Figures S10–S12) showed that CDD values were generally lower than AHCCD, and CWD values were generally higher than AHCCD. Grids at 75 km resolution would be expected to average precipitation that was not observed at a

point gauge. Similarly, a grid cell would be expected to often dampen the magnitude of convective precipitation events by averaging with areas in a grid cell where no precipitation occurred. Therefore, persistence patterns would be more difficult to capture at point locations than when aggregated to basins. The same reasoning applies to SEDI values for point locations, which were generally lower than the aggregated comparisons. The SEDI values were generally above 0.6 for NARR, GFD-HYDRO, WFDEI, and ERA-I, and generally near 0.7 and 0.8 for the ensemble mean and ANUSPLIN, respectively. Comparison with ECCC showed fewer wet days than AHCCD, suggesting better extreme value representation by the gridded datasets.

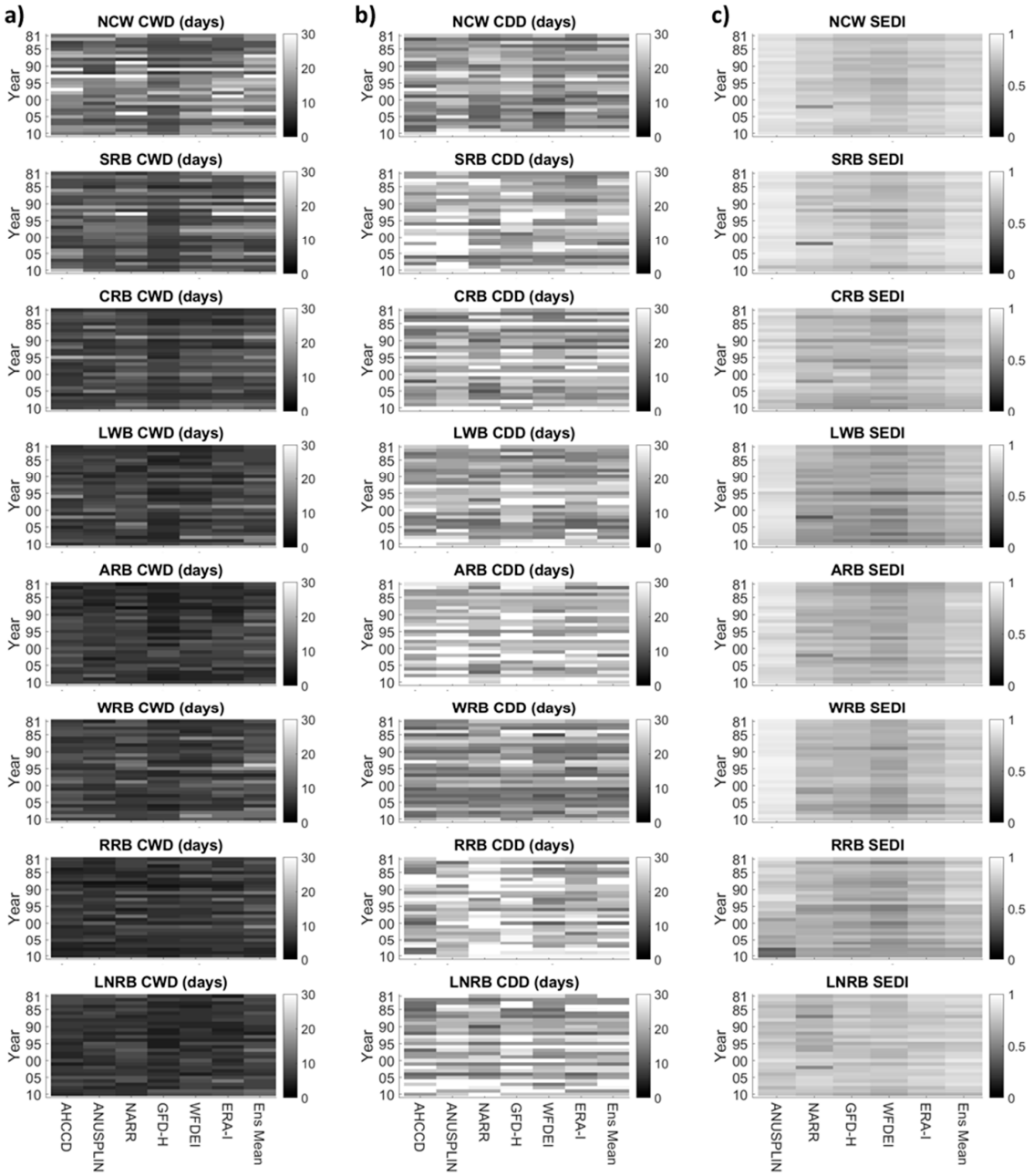

Figure 4. Spatially aggregated daily precipitation extreme indices (yearly periods) in each sub-basin (1981–2010). (**a**) CWD, (**b**) CDD, and (**c**) SEDI.

5.2. Ensemble Analysis

By assuming that the minimum to maximum ensemble range is a total possible uncertainty envelope for the ensemble (Figure 5), we find that wet conditions, low spatial coverage of climate stations, and the presence of large water bodies [46] increased uncertainty (Figure 5d,f,h). The upper and lower bounds represented a performance limit for the ensemble and can be used to identify periods of elevated uncertainty. Upper and lower bounds correlated well with the ensemble mean: when the bounds diverge, the underlying datasets disagree on the occurrence of precipitation, and when bounds converge, the underlying datasets agree on the occurrence of precipitation. Convergence suggests lower uncertainty in precipitation estimation. The ensemble mean was often similar to AHCCD, but approached ECCC for smaller aggregations. This relationship varied seasonally, with the ensemble mean generally approaching AHCCD in winter, spring, and autumn for large basin aggregations, but was similarly closer to ECCC during summer. The envelope created by the ensemble minimum and maximum realizations was widest in summer and narrowest in winter. When considered at annual timesteps, the high and low ensemble realizations appeared sufficient to envelope most of the observed precipitation events.

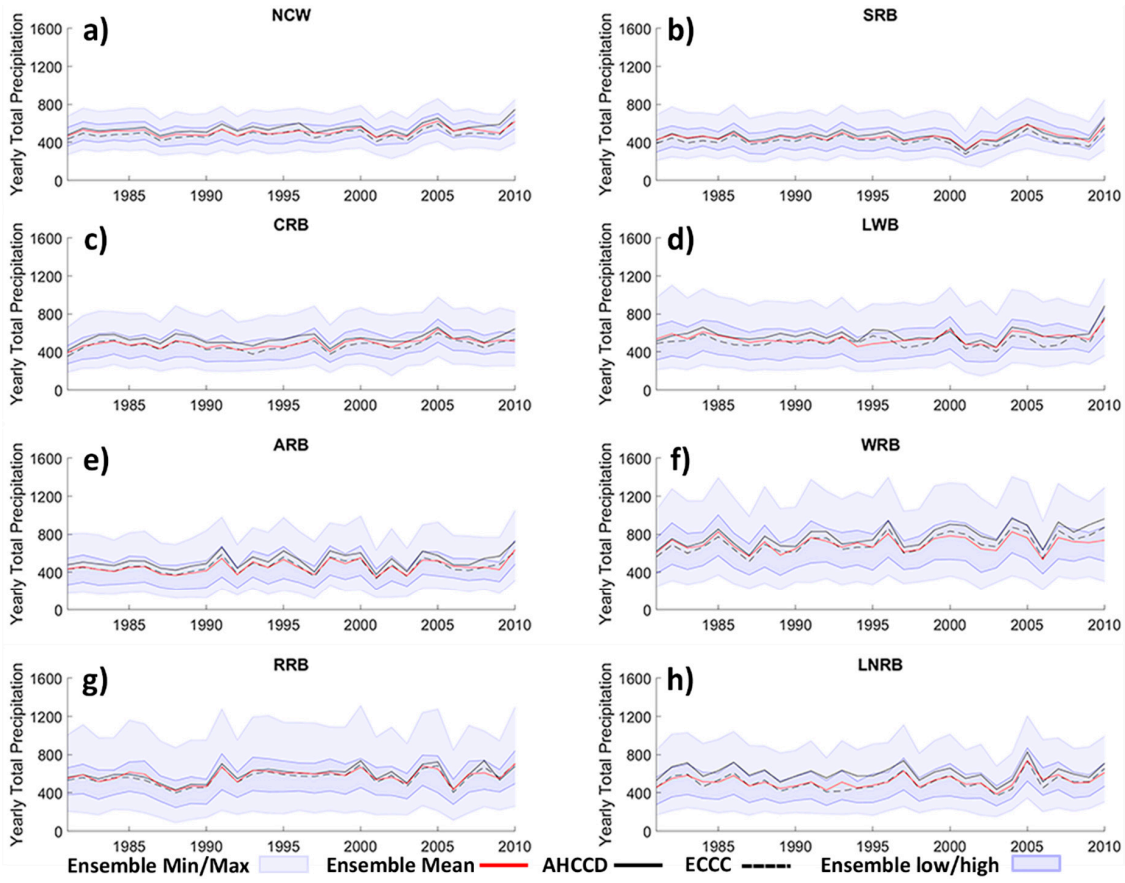

Figure 5. Basin-averaged annual total precipitation timeseries showing the ensemble minimum, mean, and maximum; the high and low ensemble realizations generated from taking the second wettest and second driest value for each timestep and grid; as well as ECCC and AHCCD for each sub-basin in the NCW (1981–2010): (**a**) NCW, (**b**) SRB, (**c**) CRB, (**d**) LWB, (**e**) ARB, (**f**) WRB, (**g**) RRB, (**h**) LNRB.

Spatial aggregation was influential when assessing events captured by the ensemble (Figure 6). Figure 6a indicates the lumped annual average ensemble mean precipitation for the NCW to be 512 mm year^{-1}. Figure 6b indicates that 31% of events fell outside the ensemble envelope (Figure 6b),

with 26% of those being below the ensemble minimum. The volume difference of events falling below (above) the ensemble minimum (maximum) increased as precipitation event magnitude increased.

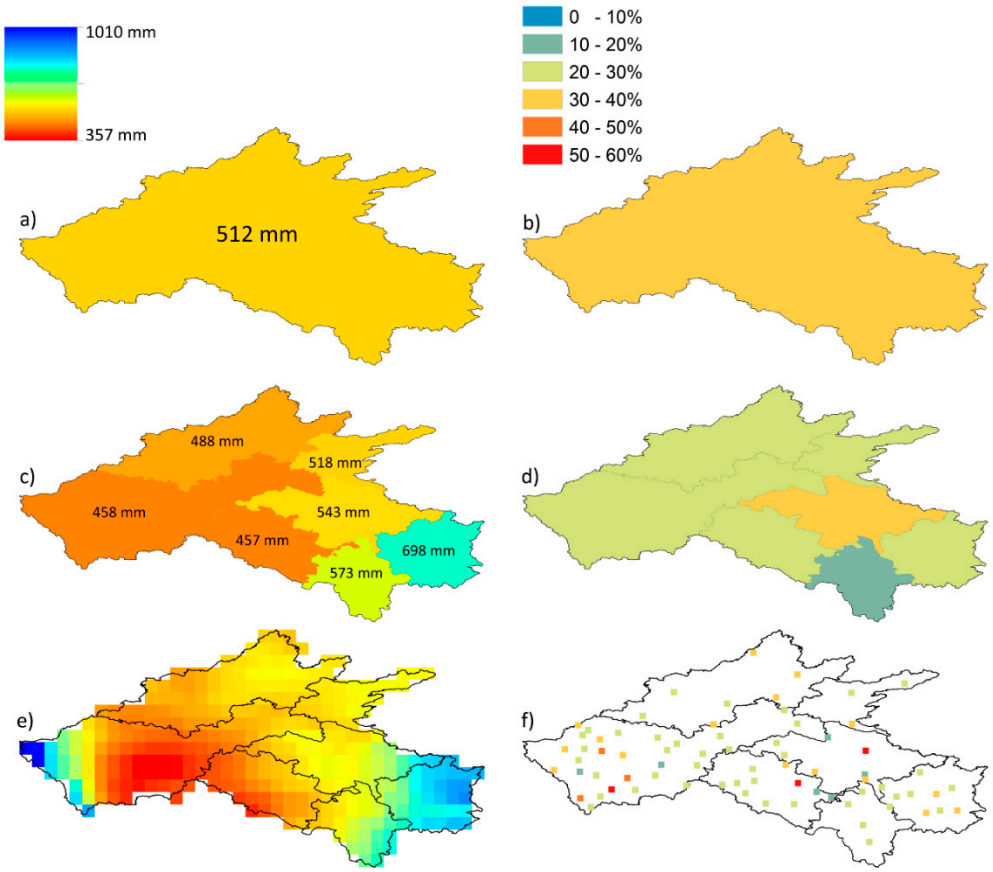

Figure 6. Period mean annual sum precipitation (1981–2010) according to different spatial aggregations: fully lumped (**a**), semi-lumped (**c**), and distributed (**e**). The % of non-zero AHCCD events outside the ensemble range when using different spatial aggregation schemes: fully lumped (**b**), semi-lumped (**d**), and distributed (**f**).

The semi-lumped aggregation varied more in annual average precipitation and the number of events falling outside the ensemble bounds (Figure 6c,d). The worst performing basin was the LWB, missing 30% of precipitation events. This was an expectedly poor performance given the low spatial density of climate stations within this basin. The best performing basin was the RRB; however, datasets diverged due to border effects and discontinuous ANUSPLIN data. The wider ensemble bounds suggested higher ensemble uncertainty for the RRB, but subsequently led to a low percentage (19%) of events beyond the ensemble envelope. The high/low ensemble realizations enveloped approximately half as many observations as the min/max ensemble (e.g., 64% for the NCW), which supports our finding that anomalies such as border effects that increase uncertainty are likely responsible for the higher reliability observed within the RRB. Border effects were also present in the WRB and the SRB, but were also anticipated to be less impactful than in the RRB, owing to their smaller percent land area in the USA. The percentage of events falling below the ensemble's lower bound ranged from 16% to 40% (Figure 6d). The volume difference of events below or above the ensemble minimum or maximum increased as precipitation event magnitude increased.

Distributed comparisons were found to be the worst, ranging from 16% to 52% of events falling beyond the ensemble bounds (Figure 6f), although they provided the highest resolution of spatial variation in precipitation (Figure 6e). Events were below the ensemble minimum; an average of 25% across all stations. Similar to the aggregated comparison, the volume difference of events below or

above the ensemble minimum or maximum increased as precipitation event magnitude increased. Reliability with respect to ECCC data was better, showing that 28% of precipitation events missed in the lumped aggregation, with a range of 15% to 26% among the semi-lumped basins, and 12% to 50% among the distributed comparison, and an average of 36% of events being below the ensemble minimum. Since the AHCCD process added water, the ECCC improvement was sourced from fewer events above the ensemble bounds. Finally, the low/high uncertainty realizations resulted in reliabilities ranging from 31% to 63%, which was consistently worse among all stations with respect to AHCCD.

6. Discussion

6.1. Uncertainty from Temporal Period of Analysis

Performance metrics calculated for the full 1981–2010 period agree with results from Eum et al. [3] and Wong et al. [4], as high- and low-performance years average out but obscure temporally dependent information [13]. The comparison of individual gridded datasets (Figure 2) does not indicate temporally dependent performance evolution [3]. Events that affect wet/dry years, such as El Niño-Southern Oscillation (ENSO) years, are smoothed out over multi-year temporal periods of evaluation [47,48]. Therefore, yearly and seasonal periods better represent the temporal variation of input data uncertainty for hydrologic modeling purposes (Figure 3).

Results presented in Figure 5 were aggregated to annual precipitation totals for the ease of presentation. A similar plot at the daily, weekly, or monthly timescale can be used to identify periods of dataset convergence (low uncertainty) and divergence (elevated uncertainty). Since each dataset was evaluated, they are assumed to be reasonable approximations of the observed environment. Therefore, when they diverge, it is reasonable to assume that they diverge because climatic conditions were uncertain, not because the products perform poorly. This concept is consistent with recommendations for the ERA5 uncertainty envelope [49]. Applying this concept to an ensemble of products overcomes some of the limitations with the ERA5 uncertainty product, which only accounts for some uncertainty in the observational data ingested into the reanalysis model. The present ensemble includes some representation of structural uncertainty, parameter uncertainty, and input uncertainty for climate models, suggesting that it is a more robust measure of periods of high or low data uncertainty.

A simple arithmetic mean is used to generate a high likelihood ensemble realization, however, other methods such as linear weighting and Bayesian model averaging have produced promising results in the literature (e.g., [50–52]). Ensemble methods assign weights to the ensemble members. This suggests that events beyond the bounds of an ensemble will not be well represented by ensemble methods. The high and low ensemble realizations further show that methods seeking to constrain uncertainty may misrepresent a higher proportion of observed events. Therefore, the maximum and minimum bounds of an ensemble, as presented in this work, can be used to identify the upper (lower) limit of ensemble performance, or performance at the extremes, or tails, of the distribution. Examining ensemble bounds offers an assessment of the quality of the ensemble for a target application at a target temporal resolution (e.g., drought or flood studies). It is, however, important to reinforce that the uncertainty bounds are not of equal likelihood. Rather, the ensemble range should be considered only from a likelihood perspective. Ingesting the minimum or maximum realization is equivalent to ingesting the lowest likelihood realizations from the ensemble. The communication of wide uncertainty bounds without inferring a lack of confidence in the results is a common topic in the literature (e.g., [53–55]). The selection of wider uncertainty bounds preserves the ability to sample low likelihood events [56], which tend to be of particular interest for climate change studies where non-stationarity can increase (decrease) the likelihood of event occurrence. Ideally, we would ingest a near-infinite number of samples from the uncertainty range, weighted by their likelihood, which would include low likelihood realizations. We caution, however, that if wider uncertainty bounds are rejected, the sampling of such events is no longer possible, even as computational power makes such sampling rigor plausible.

It is common to select hydrologic model calibration and validation periods to include both wet and dry periods [57]. Hydrologic models are likely not capable of performing well too far beyond the climatic conditions they are originally calibrated to reflect. It is, therefore, reasonable to assume that a hydrologic model calibrated using only periods of low uncertainty may be less robust in periods of high uncertainty. Ensemble convergence appears to be a viable way to integrate input uncertainty conditions into the selection of calibration and validation periods.

6.2. Uncertainty from Spatial Aggregation

Gridded datasets were weakest in their representation of extreme events (Figure 4), which generally agrees with the literature [4,11–13,16]. It was expected that a spatial average would not capture (well) the extremes recorded at a point gauge, which are only representative of local precipitation [28,58–60]. There were more grid points than observed stations, which led to differing degrees of spatial averaging between the gridded and observed datasets, suggesting added value in considering multiple spatial aggregations (e.g., [59,60]). The most extreme events within any given year generally always occurred in summer, due to convective storms and frontal systems [61]. A gridded dataset may reasonably estimate storm event magnitude, but not the storm's spatial positioning. This made the distributed comparison dependent on both magnitude and positioning, while a spatially aggregated basin was mainly dependent on magnitude resulting in lower uncertainty [62]. Since both the station data and the gridded products were aggregated when compared at lumped or semi-lumped aggregations, the scale differences were less noticeable. This also suggests that spatially aggregated hydrologic models may ingest less input data uncertainty during summer, when the spatial positioning of convective storms is most relevant. It should be noted, however, that this may lead to high magnitude, local precipitation events being damped due to areal averaging, which can introduce higher output uncertainty in some modeling applications, such as flood forecasting.

6.3. Ensemble Reliability

The minimum and maximum ensemble realizations were clearly beyond the uncertain range of observed precipitation data in this study, but highlight the possible range of the ensemble, and therefore its limitations. The choice to generate the ensemble minimum and maximum realizations was equivalent, by design, to maximizing reliability without regard for sharpness (e.g., [63,64]). The high/low ensemble realizations represent a simple attempt to reduce the uncertainty range, but generate notably lower reliability. Sharpness is a measure of ensemble spread: an ensemble with narrow bounds results in a high performing sharpness value. The high/low ensemble realizations outperform the ensemble minimum and maximum realizations in terms of their sharpness. With no temporal aggregation (i.e., by considering individual days), the reliability of the high/low ensemble bounds was lower than expected (Figure 6). This suggests that any attempt to generate a single best ensemble realization would misrepresent precipitation events beyond the ensemble range, and similarly, under-represent input data uncertainty ingested by the hydrologic models. Generating a narrower envelope that preserves reliability could be seen as constraining uncertainty in a valuable way for operations, in that the representation of observed uncertainty is preserved without producing unrealistically wide ensemble bounds for well-simulated events. However, it is important to consider that observed data are also not free of uncertainty [36]. In addition, generating an ensemble range that is too narrow risks losing information about low likelihood events that may increase in likelihood with climate change.

Spatially aggregated reliabilities were often better than the reliabilities at stations within those aggregations; however, the NCW had the lowest reliability of the spatially aggregated comparisons. This suggests that hydrologic model spatial structures could potentially be optimized to balance the loss of information, through both choice of spatial aggregation (Figure 6a,c,e) and the reductions in input uncertainty. If increasing spatial aggregation narrows the ensemble bounds, which can be interpreted as reducing uncertainty, too much narrowing may actually reduce the reliability (Figure 6b,d). Some aggregation and reduction of uncertainty are beneficial, but also reduce the

information on the spatial positioning of precipitation. Optimizing the degree of spatial aggregation is out of scope for this study, but would be interesting future work.

It should be noted that there exist more formal methods for estimating precipitation data uncertainty in the literature. As an example, Newman et al. [64] present a framework for estimating climate data uncertainty from station data. Their methodology was applied to precipitation and temperature data. Data density was generally high in their study, and therefore may not perform well in regions like northern Manitoba, or the Arctic regions of relevance to the BaySys project. As shown in the present study, reliability and dataset convergence can be tested with any climate variable, with any data density. In this way, reliability and convergence do not replace formal estimates of uncertainty. An estimate of dataset convergence could be applied to the dataset produced by Newman et al. [65] to determine periods of elevated uncertainty. Additionally, the concepts of reliability and convergence can be applied across scales to aid in the evaluation of spatial aggregation decisions for hydrologic modeling.

7. Conclusions

Many gridded precipitation datasets have been developed and included in comparison studies. These studies often struggled to suggest a generalized best product, instead suggesting the use of ensembles or an ensemble method-derived realization. There is a need for a better understanding of the uncertainty associated with gridded precipitation data ensembles, particularly for applications with varied spatial and temporal resolutions. This is an important subject, given the rising popularity of new reanalysis forcing products and hydrologic ensembles, which include diverse methods and spatial scales for data ingestion. This study compared dataset performance at multiple spatial and temporal aggregations to explore the effect of the various spatial aggregation choices made in recent literature studies. Based on the analysis conducted, the major conclusions can be summarized as follows (Table 3):

Table 3. Summary of findings and suggestions.

General Findings	• Dataset performance is dependent on performance metric • Dataset performance varies spatially and temporally
Specific Findings	• The amount of spatial and temporal aggregation impacts dataset performance, uncertainty, and reliability • Dataset convergence (divergence) can be used to assess periods of low (high) uncertainty to include input uncertainty conditions into the selection of calibration/validation periods • Observations that fall outside the minimum/maximum range of the ensemble will likely never be well represented by ensemble methods
Suggestions	• Multiple ensemble members should be used to account for gridded dataset uncertainty • Dataset assessment should be conducted across a range of spatial and temporal scales relevant to those of a target hydrologic model(s) • A reliability analysis should be conducted to ensure sufficient overlap of gridded datasets with observations preceding hydrologic modeling • An assessment of convergence/divergence should be done preceding hydrologic modeling to include periods of low/high precipitation uncertainty into calibration/validation periods

The findings of this study are not meant as an exhaustive search for viable gridded precipitation datasets for inclusion in an ensemble, but rather to present a simple procedure to assess the limitations associated with an ensemble. Therefore, the procedure presented in the present study identifies the performance limit of a gridded precipitation data ensemble, and how spatial and temporal aggregation methods affect that limit. The results reported in this study are likely not unique to the NCW; therefore,

future research should include an examination of the effect that each gridded dataset in an ensemble has on the ensemble reliability.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4441/12/10/2751/s1, and are referenced in the main body of the manuscript. Figure S1: Daily precipitation spatially aggregated annual correlation with reference to the AHCCD observed data set in each gauge (1981–2010). Figure S2: Daily precipitation spatially aggregated annual PBIAS with reference to the AHCCD observed data set in each gauge (1981–2010). Figure S3: Daily precipitation spatially aggregated annual SEDI score with reference to the AHCCD observed data set in each gauge (1981–2010). Figure S4: Daily precipitation spatially aggregated annual ETS score with reference to the AHCCD observed data set in each sub-basin (1981–2010). Scores for 0–5 mm bins are presented. Figure S5: Daily precipitation spatially aggregated annual ETS score with reference to the AHCCD observed data set in each sub-basin (1981–2010). Scores for 5-inf mm bins are presented. Figure S6: Daily precipitation spatially aggregated annual FBIAS score with reference to the AHCCD observed data set in each sub-basin (1981–2010). Scores for 0–5 mm bins are presented. Figure S7: Daily precipitation spatially aggregated annual ETS score with reference to the AHCCD observed data set in each sub-basin (1981–2010). Scores for 5-inf mm bins are presented. Figure S8: Daily precipitation spatially aggregated annual ETS score with reference to the AHCCD observed data set in each gauge (period mean for 1981–2010). Figure S9: Daily precipitation spatially aggregated annual FBIS score with reference to the AHCCD observed data set in each gauge (period mean for 1981–2010). Figure S10: Daily precipitation spatially aggregated annual CDD score with reference to the AHCCD observed data set in each gauge (1981–2010). Figure S11: Daily precipitation spatially aggregated annual CWD, with AHCCD observed data set presented for reference, in each gauge (1981–2010). Figure S12: Daily precipitation spatially aggregated annual SEDI score with reference to the AHCCD observed data set in each gauge (1981–2010). Table S1: Final list of ground-based climate stations with mean yearly precipitation and the standard deviation of yearly precipitation. Stations are ordered from highest to lowest latitude. Table S2: Station coverage for each year in each basin as the percentage of stations with data available in a basin. Percentages below 100% represent the presence of one or more days with some stations missing data (AHCCD).

Author Contributions: Conceptualization, S.P., T.A.S., R.L., G.A. and S.J.D.; methodology, S.P., T.A.S., R.L., G.A. and S.J.D.; validation, S.P., T.A.S., R.L., G.A. and S.J.D.; formal analysis, S.P. and R.L.; investigation, S.P. and T.A.S.; resources, S.P. and T.A.S.; data curation, S.P.; writing—original draft preparation, S.P., T.A.S., R.L., G.A. and S.J.D.; writing—review and editing, S.P., T.A.S., R.L., G.A., S.J.D. and K.K.; visualization, S.P.; supervision, T.A.S., G.A., S.J.D. and K.K.; project administration, T.A.S.; funding acquisition, T.A.S. All authors contributed significantly to the editing and revising of the manuscript. Initial investigation and study development was conducted by authors: S.P., T.A.S., R.L., G.A. and S.J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded by the BaySys project and partners Manitoba Hydro, and ArcticNet and well as partnered universities under BaySys grant 280926454.

Acknowledgments: Thanks to the University of Manitoba, Manitoba Hydro, AcrticNet (in-kind), and partners funding through the Natural Sciences and Engineering Research Council of Canada through the funding of the BaySys project. Thanks to developers of the gridded datasets used in this study. Thanks to NRCAN for providing access to the ANUSPLIN dataset and the AHCCD dataset, and SMHI for providing access to the Hydro-GFD dataset. We also thank the reviewers for their time and feedback that assisted in improving this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Price, D.T.; McKenney, D.W.; Nalder, I.A.; Hutchinson, M.F.; Kesteven, J.L. A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. *Agric. For. Meteorol.* **2000**, *101*, 81–94. [CrossRef]
2. Mekis, E.; Vincent, L.A. An overview of the second generation adjusted daily precipitation dataset for trend analysis in Canada. *Atmosphere-Ocean* **2011**, *49*, 163–177. [CrossRef]
3. Eum, H.-I.; Dibike, Y.; Prowse, T.; Bonsal, B. Inter-comparison of high-resolution gridded climate data sets and their implication on hydrological model simulation over the Athabasca Watershed, Canada. *Hydrol. Process.* **2014**, *28*, 4250–4271. [CrossRef]
4. Wong, J.S.; Razavi, S.; Bonsal, B.R.; Wheater, H.S.; Asong, Z.E. Inter-comparison of daily precipitation products for large-scale hydro-climatic applications over Canada. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 2163–2185. [CrossRef]
5. ISO. *Guide to the Expression of Uncertainty in Measurement*; ISO: Geneva, Switzerland, 1995; pp. 4–5.
6. Matott, L.S.; Babendreier, J.E.; Purucker, S.T. Evaluating uncertainty in integrated environmental models: A review of concepts and tools. *Water Resour. Res.* **2009**, *45*, W06421. [CrossRef]

7. Ajami, N.K.; Duan, Q.; Sorooshian, S. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resour. Res.* **2007**, *43*, W01403. [CrossRef]
8. Dams, J.; Nossent, J.; Senbeta, T.B.; Willems, P.; Batelaan, O. Multi-model approach to assess the impact of climate change on runoff. *J. Hydrol.* **2015**, *529*, 1601–1616. [CrossRef]
9. Pavelsky, T.M.; Smith, L.C. Intercomparison of four global precipitation data sets and their correlation with increased Eurasian river discharge to the Arctic Ocean. *J. Geophys. Res. Space Phys.* **2006**, *111*, D21112. [CrossRef]
10. Bukovsky, M.S.; Karoly, D.J. A brief evaluation of precipitation from the North American regional reanalysis. *J. Hydrometeorol.* **2007**, *8*, 837–846. [CrossRef]
11. Becker, E.J.; Berbery, E.H.; Higgins, R.W. Understanding the characteristics of daily precipitation over the United States using the North American regional reanalysis. *J. Clim.* **2009**, *22*, 6268–6286. [CrossRef]
12. Choi, W.; Kim, S.J.; Rasmussen, P.F.; Moore, A.R. Use of the North American regional reanalysis for hydrological modelling in Manitoba. *Can. Water Resour. J.* **2009**, *34*, 17–36. [CrossRef]
13. Rapaić, M.; Brown, R.; Markovic, M.; Chaumont, D. An evaluation of temperature and precipitation surface-based and reanalysis datasets for the Canadian arctic, 1950–2010. *Atmosphere-Ocean* **2015**, *53*, 283–303. [CrossRef]
14. Kluver, D.; Mote, T.L.; Leathers, D.; Henderson, G.R.; Chan, W.; Robinson, D.A. Creation and validation of a comprehensive 1° by 1° daily gridded North American dataset for 1900–2009: Snowfall. *J. Atmos. Ocean. Technol.* **2016**, *33*, 857–871. [CrossRef]
15. Essou, G.R.C.; Sabarly, F.; Lucas-Picher, P.; Brissette, F.; Poulin, A. Can precipitation and temperature from meteorological reanalyses be used for hydrological modeling? *J. Hydrometeorol.* **2016**, *17*, 1929–1950. [CrossRef]
16. Gbambie, A.S.B.; Poulin, A.; Boucher, M.-A.; Arsenault, R. Added value of alternative information in interpolated precipitation datasets for hydrology. *J. Hydrometeorol.* **2017**, *18*, 247–264. [CrossRef]
17. Boluwade, A.; Zhao, K.-Y.; Stadnyk, T.; Rasmussen, P. Towards validation of the Canadian Precipitation Analysis (CaPA) for hydrologic modeling applications in the Canadian prairies. *J. Hydrol.* **2018**, *556*, 1244–1255. [CrossRef]
18. Fortin, V.; Roy, G.; Stadnyk, T.; Koenig, K.; Gasset, N.; Mahidjiba, A. Ten years of science based on the Canadian precipitation analysis: A CaPA system overview and literature review. *Atmosphere-Ocean* **2018**, *56*, 1–19. [CrossRef]
19. Lespinas, F.; Fortin, V.; Roy, G.; Rasmussen, P.F.; Stadnyk, T.A. Performance evaluation of the Canadian Precipitation Analysis (CaPA). *J. Hydrometeorol.* **2015**, *16*, 2045–2064. [CrossRef]
20. Asong, Z.E.; Razavi, S.; Wheater, H.S.; Wong, J.S. Evaluation of Integrated Multisatellite Retrievals for GPM (IMERG) over southern Canada against ground precipitation observations: A preliminary assessment. *J. Hydrometeorol.* **2017**, *18*, 1033–1050. [CrossRef]
21. Nash, J.; Sutcliffe, J. River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290. [CrossRef]
22. Vincent, L.A.; Wang, X.L.; Milewska, E.J.; Wan, H.; Yang, F.; Swail, V. A second generation of homogenized Canadian monthly surface air temperature for climate trend analysis. *J. Geophys. Res. Space Phys.* **2012**, *117*, D18110. [CrossRef]
23. Masson, D.; Knutti, R. Climate model genealogy. *Geophys. Res. Lett.* **2011**, *38*, L08703. [CrossRef]
24. Knutti, R.; Masson, D.; Gettelman, A. Climate model genealogy: Generation CMIP5 and how we got there. *Geophys. Res. Lett.* **2013**, *40*, 1194–1199. [CrossRef]
25. Sanderson, B.; Knutti, R.; Caldwell, P. A representative democracy to reduce interdependency in a multimodel ensemble. *J. Clim.* **2015**, *28*, 5171–5194. [CrossRef]
26. LeDuc, M.; Laprise, R.; De Elía, R.; Šeparović, L. Is institutional democracy a good proxy for model independence? *J. Clim.* **2016**, *29*, 8301–8316. [CrossRef]
27. Steinschneider, S.; Wi, S.; Brown, C. The integrated effects of climate and hydrologic uncertainty on future flood risk assessments. *Hydrol. Process.* **2014**, *29*, 2823–2839. [CrossRef]
28. Sillmann, J.; Kharin, V.V.; Zhang, X.; Zwiers, F.; Bronaugh, D. Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *J. Geophys. Res. Atmos.* **2013**, *118*, 1716–1733. [CrossRef]

29. Hutchinson, M.F.; McKenney, D.W.; Lawrence, K.; Pedlar, J.H.; Hopkinson, R.F.; Milewska, E.; Papadopol, P. Development and testing of Canada-Wide interpolated spatial models of daily minimum–maximum temperature and precipitation for 1961–2003. *J. Appl. Meteorol. Clim.* **2009**, *48*, 725–741. [CrossRef]
30. Mesinger, F.; DiMego, G.; Kalnay, E.; Mitchell, K.; Shafran, P.C.; Ebisuzaki, W.; Jovic, D.; Woollen, J.; Rogers, E.; Berbery, E.H.; et al. North American regional reanalysis. *Bull. Am. Meteorol. Soc.* **2006**, *87*, 343–360. [CrossRef]
31. Berg, P.; Donnelly, C.; Gustafsson, D. Near-real-time adjusted reanalysis forcing data for hydrology. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 989–1000. [CrossRef]
32. Khakbaz, B.; Imam, B.; Hsu, K.; Sorooshian, S. From lumped to distributed via semi-distributed: Calibration strategies for semi-distributed hydrologic models. *J. Hydrol.* **2012**, *418*, 61–77. [CrossRef]
33. Lilhare, R.; Déry, S.J.; Pokorny, S.; Stadnyk, T.A.; Koenig, K.A. Intercomparison of multiple hydroclimatic datasets across the lower nelson river basin, Manitoba, Canada. *Atmosphere-Ocean* **2019**, *57*, 1–17. [CrossRef]
34. Government of Canada. Natural Resources Canada. Available online: https://www.nrcan.gc.ca/home (accessed on 5 July 2018).
35. Benke, A.C.; Cushing, C.E. *Rivers of North America*; Elsevier: Burlington, MA, USA, 2005; pp. 853–888.
36. McMillan, H.; Krueger, T.; Freer, J. Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrol. Process.* **2012**, *26*, 4078–4111. [CrossRef]
37. Menne, M.J.; Durre, I.; Vose, R.S.; Gleason, B.E.; Houston, T.G. An overview of the global historical climatology network-daily database. *J. Atmos. Ocean. Technol.* **2012**, *29*, 897–910. [CrossRef]
38. Durre, I.; Menne, M.J.; Gleason, B.E.; Houston, T.G.; Vose, R.S. Comprehensive automated quality assurance of daily surface observations. *J. Appl. Meteorol. Clim.* **2010**, *49*, 1615–1633. [CrossRef]
39. Barber, D.G. BaySys—Contributions of Climate Change and Hydroelectric Regulation to the Variability and Change of Freshwater-Marine Coupling in the Hudson Bay System. Available online: http://umanitoba.ca/faculties/environment/departments/ceos/media/BaySys_PROJECT_DESCRIPTION.pdf (accessed on 15 January 2014).
40. Dee, D.P.; Uppala, S.M.; Simmons, A.J.; Berrisford, P.; Poli, P.; Kobayashi, S.; Andrae, U.; Balmaseda, M.A.; Balsamo, G.; Bauer, P.; et al. The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **2011**, *137*, 553–597. [CrossRef]
41. Weedon, G.P.; Balsamo, G.; Bellouin, N.; Gomes, S.; Best, M.; Viterbo, P. The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resour. Res.* **2014**, *50*, 7505–7514. [CrossRef]
42. World Meteorological Organization. *Calculation of Monthly and Annual 30-Year Standard Normals*; WCDP-No. 10, WMO-TD/No. 341; World Meteorological Organization: Geneva, Switzerland, 1989.
43. Zhu, Y.; Luo, Y. Precipitation calibration based on the frequency-matching method. *Weather Forecast.* **2015**, *30*, 1109–1124. [CrossRef]
44. Montanari, A. Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations. *Water Resour. Res.* **2005**, *41*, W08406. [CrossRef]
45. Tustison, B.; Harris, D.; Foufoula-Georgiou, E. Scale issues in verification of precipitation forecasts. *J. Geophys. Res. Space Phys.* **2001**, *106*, 11775–11784. [CrossRef]
46. Rouse, W.R. Impacts of hudson bay on the terrestrial climate of the Hudson bay lowlands. *Arct. Alp. Res.* **1991**, *23*, 24. [CrossRef]
47. Shabbar, A.; Khandekar, M. The impact of el Niño-Southern oscillation on the temperature field over Canada: Research note. *Atmosphere-Ocean* **1996**, *34*, 401–416. [CrossRef]
48. Trenberth, K.E. The definition of El Niño. *Bull. Amer. Meteor. Soc.* **1997**, *78*, 2771–2778. [CrossRef]
49. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 1999–2049. [CrossRef]
50. Wilks, D.S. Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorol. Appl.* **2006**, *13*, 243. [CrossRef]
51. Yang, C.; Yan, Z.; Shao, Y. Probabilistic precipitation forecasting based on ensemble output using generalized additive models and Bayesian model averaging. *Acta Meteorol. Sin.* **2012**, *26*, 1–12. [CrossRef]
52. Yao, Y.; Liang, S.; Xie, X.; Cheng, J.; Jia, K.; Li, Y.; Liu, R. Estimation of the terrestrial water budget over northern China by merging multiple datasets. *J. Hydrol.* **2014**, *519*, 50–68. [CrossRef]

53. DeMeritt, D.; Cloke, H.; Pappenberger, F.; Pozo, J.T.-D.; Bartholmes, J.C.; Ramos, M. Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environ. Hazards* **2007**, *7*, 115–127. [CrossRef]
54. Pappenberger, F.; Stephens, E.; Thielen, J.; Salamon, P.; DeMeritt, D.; Van Andel, S.J.; Wetterhall, F.; Alfieri, L. Visualizing probabilistic flood forecast information: Expert preferences and perceptions of best practice in uncertainty communication. *Hydrol. Process.* **2012**, *27*, 132–146. [CrossRef]
55. Pokorny, S.; Stadnyk, T.A.; Ali, G.; Lilhare, R.; Déry, S.J.; Koenig, K. Cumulative effects of uncertainty on simulated streamflow in a hydrologic modeling environment. *Elem. Sci. Anth.* **2020**. [CrossRef]
56. Westerberg, I.K.; Sikorska-Senoner, A.E.; Viviroli, D.; Vis, M.J.P.; Seibert, J. Hydrological model calibration with uncertain discharge data. *Hydrol. Sci. J.* **2020**, 1–16. [CrossRef]
57. Vaze, J.; Post, D.; Chiew, F.; Perraud, J.-M.; Viney, N.R.; Teng, J. Climate non-stationarity—Validity of calibrated rainfall–runoff models for use in climate change studies. *J. Hydrol.* **2010**, *394*, 447–457. [CrossRef]
58. Sun, X.; Barros, A. An Evaluation of the statistics of rainfall extremes in rain gauge observations, and satellite-based and reanalysis products using universal multifractals. *J. Hydrometeorol.* **2010**, *11*, 388–404. [CrossRef]
59. Fischer, E.M.; Beyerle, U.; Knutti, R. Robust spatially aggregated projections of climate extremes. *Nat. Clim. Chang.* **2013**, *3*, 1033–1038. [CrossRef]
60. Pendergrass, A.G.; Knutti, R.; Lehner, F.; Deser, C.; Sanderson, B. Precipitation variability increases in a warmer climate. *Sci. Rep.* **2017**, *7*, 17966. [CrossRef] [PubMed]
61. Dingman, S.L. *Physical Hydrology*, 3rd ed.; Waveland Press: Long Grove, IL, USA, 2015; pp. 47–203.
62. Carpenter, T.M.; Georgakakos, K.P. Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales. *J. Hydrol.* **2006**, *329*, 174–185. [CrossRef]
63. Shafii, M.; Tolson, B.A.; Matott, L.S. Addressing subjective decision-making inherent in GLUE-based multi-criteria rainfall–runoff model calibration. *J. Hydrol.* **2015**, *523*, 693–705. [CrossRef]
64. Zhou, R.; Li, Y.; Lu, D.; Liu, H.; Zhou, H. An optimization based sampling approach for multiple metrics uncertainty analysis using generalized likelihood uncertainty estimation. *J. Hydrol.* **2016**, *540*, 274–286. [CrossRef]
65. Newman, A.J.; Clark, M.P.; Craig, J.; Nijssen, B.; Wood, A.W.; Gutmann, E.; Mizukami, N.; Brekke, L.; Arnold, J.R. Gridded ensemble precipitation and temperature estimates for the contiguous United States. *J. Hydrometeorol.* **2015**, *16*, 2481–2500. [CrossRef]