EMDNA: ~~An~~ **An** Ensemble Meteorological Dataset for North America

Guoqiang Tang[1,~~2~~], Martyn P. Clark[1,~~3,~~2], Simon Michael Papalexiou[2,~~4~~3], Andrew J. Newman[~~5~~4],
Andrew W. Wood[~~5~~4], Dominique Brunet[~~6~~5], Paul H. Whitfield[1,~~2~~]

[1]Centre for Hydrology, U~~U~~niversity of Saskatchewan~~, Coldwater Lab~~, Canmore, Alberta, Canada

[2]Centre for Hydrology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada

[3]Department of ~~Civil, Geological and Environmental Engineering~~Geography and Planning, University of Saskatchewan, Saskatchewan, Canada

[4]Department of Civil, Geological and Environmental Engineering, University of Saskatchewan, Saskatchewan, Canada

[~~5~~4]National Center for Atmospheric Research, Boulder, Colorado

[~~6~~5]Meteorological Research Division, Environment and Climate Change Canada, Toronto, Ontario, Canada

Abstract: Probabilistic methods are ~~very~~ useful to estimate the uncertainty in spatial ~~variability in~~ meteorological ~~conditions~~ fields (e.g., the uncertainty in spatial patterns of precipitation and temperature across large domains). In ensemble probabilistic methods, "equally plausible" ensemble members are used to approximate the probability distribution, hence the uncertainty, of a spatially distributed meteorological variable conditioned on the available information. The ensemble members can be used to evaluate the impact of ~~the~~ uncertainties in spatial meteorological fields for a myriad of applications. This study develops the Ensemble Meteorological Dataset for North America (EMDNA). EMDNA has 100 ensemble members with daily precipitation amount, mean daily temperature, and daily temperature range at 0.1° spatial resolution (approx. 10-km grids) from 1979 to 2018, derived from a fusion of station observations and reanalysis model outputs. The station data used in EMDNA are from a serially complete dataset for North America (SCDNA) that fills gaps in precipitation and temperature measurements using multiple strategies. Outputs from three reanalysis products are regridded, corrected, and merged using the Bayesian Model Averaging. Optimal Interpolation (OI) is used to merge station- and reanalysis-based estimates. EMDNA estimates are generated ~~based~~ using spatiotemporally correlated random fields ~~on~~ to sample from the OI estimates ~~and spatiotemporally correlated random fields~~. Evaluation results show that (1) the merged reanalysis estimates outperform raw reanalysis estimates, particularly in high latitudes and mountainous regions; (2) the OI estimates are more accurate than the reanalysis and station-based regression estimates, with the most notable improvements for precipitation ~~occurring~~ evident in sparsely gauged regions; and (3) EMDNA estimates exhibit good performance according to the diagrams and metrics used for probabilistic evaluation. We ~~also~~ discuss the limitations of the current framework and highlight that ~~persistent efforts are~~further research is needed to ~~further develop probabilistic methods and~~improve ensemble meteorological datasets. Overall, EMDNA is expected to be useful for hydrological and meteorological applications

35   in North America. The ~~whole~~ entire dataset and a teaser dataset (a small subset of EMDNA for easy download and

36   preview) are available at https://doi.org/10.20383/101.0275 (Tang et al., 2020a).

1. Introduction

Precipitation and temperature data are fundamental meteorological ~~inputs~~ variables for a wide variety of geoscientific and ~~operational~~ applications ~~benefitting society~~ (Eischeid et al., 2000; Trenberth et al., 2003; Wu et al., 2014; Yin et al., 2018). Accurately estimating spatial meteorological fields is still challenging despite the availability of many measurement/estimation approaches (e.g., meteorological stations, weather radars, and satellite sensors) and the availability of many atmospheric models (Kirstetter et al., 2015; Sun et al., 2018; Hu et al., 2019; Newman et al., 2019a). There is consequently substantial uncertainty in analyses of spatially distributed meteorological variables.

The uncertainty in spatial meteorological estimates depends on both the measurements available and the climate of the region of study. Whilst meteorological stations provide the most reliable observations at the point scale, spatial meteorological estimates based on station data can be ~~degraded by~~uncertain because of both ~~the~~ spars~~city of~~station networks in remote regions and ~~by~~ because of measurement errors caused by factors such as evaporation/wetting loss and under-catch of precipitation (Sevruk, 1984; Goodison et al., 1998; Nešpor and Sevruk, 1999; Yang et al., 2005; Scaff et al., 2015; Kochendorfer et al., 2018). Interpolating station data to a regular grid can introduce additional uncertainties, especially in regions where there are strong spatial gradients in meteorological fields ~~due to factors such as method choices and topographic variations~~. The accuracy of precipitation estimated from ground radars is affected by factors such as beam blockage, signal attenuation, ground clutter, and uncertainties in the representativeness of radar variables to surface rainfall~~reflectivity rainfall relationships~~ (Dinku et al., 2002; Kirstetter et al., 2015). Moreover, the spatial and temporal coverage of ground radars is limited to large populated areas in most regions of the world. Satellite sensors provide quasi-global estimates of meteorological variables, but their utility can be limited by short sampling periods with insufficient coverage and return frequency, data latency, indirect measurements, imperfect retrieval algorithms, and instrument limitations (Adler et al., 2017; Tang et al., 2016, 2020b). Reanalysis models, which provide long-term global simulations, also contain biases and uncertainties caused by ~~the~~ imperfect model representations of physical processes, observational constraints, and the model resolution~~, and model parameterization~~ (Donat et al., 2014; Parker, 2016).

In recent years, numerous deterministic gridded precipitation and temperature datasets based on observed or simulated data from single or multiple sources have become publicly available ~~to the public~~ (Maurer et al., 2002; Huffman et al., 2007; Mahfouf et al., 2007; Daly et al., 2008; Di Luzio et al., 2008; Haylock et al., 2008; Livneh et al., 2013; Weedon et al., 2014; Fick and Hijmans, 2017; Beck et al., 2019; Ma et al., 2020; Harris et al., 2020). Since the uncertainties vary in space and time, deterministic products do not always agree with each other (Donat et al., 2014; Henn et al., 2018; Sun et al., 2018; Newman et al., 2019a; Tang et al., 2020b). The uncertainties can ~~be~~ propagated to applications such as hydrological modeling and climate analysis (Clark et al., 2006; Hong et al., 2006; Slater and Clark, 2006; Mears et al., 2011; Rodell et al., 2015; Aalto et al., 2016). Proper understanding of the uncertainties can benefit the objective application of meteorological analyses and further improve existing products, yet few gridded datasets provide such uncertainty estimates (Cornes et al., 2018; Frei and Isotta, 2019).

71 Probabilistic datasets ~~now can~~ provide alternatives to deterministic datasets for quantitative precipitation and

72 temperature estimation ~~and have advantages in estimating uncertainties and representing extremes~~ (Kirstetter et al.,

73 2015; Mendoza et al., 2017; Frei and Isotta, 2019). Recently, several ensemble meteorological datasets have become

74 available. For example, Morice et al. (2012) develop the observation-based HadCRUT4 global temperature datasets

75 with 100 members. Caillouet et al. (2019) develop the Spatially COherent Probabilistic Extended Climate dataset

76 (SCOPE Climate) with 25 members in France. Newman et al. (2015, 2019b, 2020) continually extend the probabilistic

77 estimation methodology proposed by Clark and Slater (2006), and produce ensemble precipitation and temperature

78 datasets in the contiguous USA (CONUS), the Hawaii Islands, and Alaska and Yukon, respectively. Moreover, several

79 widely used deterministic datasets now have ensemble versions in view of the advantages of probabilistic estimates.

80 Cornes et al. (2018) developed the ensemble version (100 members) of the Haylock et al. (2008) Europe-wide E-OBS

81 temperature and precipitation datasets. Khedhaouiria et al. (2020) developed the experimental High-Resolution

82 Ensemble Precipitation Analysis (HREPA) for Canada and the northern part of the CONUS with 24 members, which

83 can be regarded as an experimental ensemble version of the Canadian Precipitation Analysis (CaPA; Mahfouf et al.,

84 2007; Fortin et al., 2015).

85 Our objective is to develop an Ensemble Meteorological Dataset for North America (EMDNA) from 1979 to 2018.

86 To improve the quality of estimates in sparsely gauged regions, station data and reanalysis outputs are merged to

87 generate gridded precipitation and temperature estimates. Then, ensemble estimates are produced using the

88 probabilistic method described by Clark and Slater (2006) and Newman et al. (2015, 2019b, 2020). EMDNA has 100

89 members and contains daily precipitation amount, mean daily temperature (Tmean), and daily temperature range

90 (Trange) at 0.1° spatial resolution. Minimum and maximum temperature can be calculated from Tmean and Trange.

91 It is expected that the EMDNA will be useful for a variety of applications in North America.

2. Datasets

93 Station observations are often ~~have~~ subject to temporal discontinuities caused by missing values and short record

94 lengths (Kemp et al., 1983). This study uses station precipitation and minimum/maximum temperature data from the

95 Serially Complete Dataset for North America (SCDNA; (Tang et al., 2020c), which is open-access on Zenodo

96 (https://doi.org/10.5281/zenodo.3735533; Access Date: July 25, 2020). Serially complete datasets ~~could~~ improve the

97 quality of spatial interpolation estimates compared to raw station observations with data gaps (Longman et al., 2020;

98 Tang et al., 2021). Tmean and Trange are calculated from minimum and maximum temperature data. In SCDNA, raw

99 measurements undergo strict quality control checks, and data gaps are filled by combining estimates from multiple

100 strategies (including quantile mapping, spatial interpolation, machine learning, and multi-strategy merging). SCDNA

101 uses reanalysis estimates as the auxiliary data to ensure temporal completeness in sparsely gauged regions. The

102 production of SCDNA has nine steps: (1) matching reanalysis estimates and station data, (2) selecting qualified

103 neighboring stations, (3) building empirical cumulative density functions (CDFs), (4) estimation based on 16 strategies

104 for each day of year, (5) independent validation, (6) merging estimates from the 16 strategies, (7), climatological bias

105 correction, (8) evaluation of SCDNA, and (9) final quality control (Tang et al., 2020c). SCDNA covers the period

106 from 1979 to 2018 and has 24,615 precipitation stations and 19,579 temperature stations. We select precipitation and

107 temperature because they are used in many hydrometeorological studies and are measured by a large number of

108 meteorological stations, while other variables (e.g., humidity and wind speed) are only measured by a much smaller

109 collection of stations.

110 Station-based gridded meteorological estimates usually rely on a certain number of neighboring stations surrounding

111 the target grid cell~~location~~. For most regions in CONUS, the search radius to find 20 or 30 neighboring stations (lower

112 and upper limits for station-based gridded estimates in Sect. 3.1) is smaller than 100 km (Fig. 1). For the regions

113 ~~northern~~ north of ~~to~~ 50°N or ~~southern~~ south of~~to~~ 20°N, however, the search radius required to find 20 or 30 neighboring

114 stations is much larger, and even exceeds 1,000 km in the Arctic Archipelago. The sparse station network at higher

115 latitudes motivates our decision to optimally combine station data with reanalysis products.

116 The reanalysis products used in this study include the fifth generation of European Centre for Medium-Range Weather

117 Forecasts (ECMWF) atmospheric reanalyses of the global climate (ERA5; Hersbach et al., 2020), the Modern-Era

118 Retrospective analysis for Research and Applications, Version 2 (MERRA-2; Gelaro et al., 2017), and the Japanese

119 55-year Reanalysis (JRA-55; Kobayashi et al., 2015). The three widely used products are chosen because of their high

120 spatiotemporal resolutions and suitable time length. The spatial resolutions of ERA5, MERRA-2, and JRA-55 are

121 0.25°×0.25°, 0.5°×0.625°, and ~55 km, respectively. Their start years are 1979, 1980, and 1958, respectively.

122 Therefore, only ERA5 and JRA-55 are used for 1979 throughout this study. Although reanalysis models assimilate

123 observations from various sources, they differ with station measurements in many aspects (Parker, 2016) and often

124 contain large uncertainties as shown by assessment and multi-source merging studies (e.g., Donat et al., 2014; Lader

125 et al., 2016; Beck et al., 2017, 2019; Tang et al., 2020b). ~~We suppose~~~~It is possible~~The dependence ~~between~~of

126 reanalysis estimates ~~and~~on station data may have a negative effect on the merging of reanalysis products (Section 3.2)

127 because ~~station-based weight calculation could be imperfect (e.g., favourable to~~ the reanalysis dataset which

128 assimilates more station data ~~and assigns higher weights to observations~~could be given higher weight~~)~~. The potential

129 dependence, however, is not considered in this study ~~Thereby, the possible dependence between reanalysis estimates~~

130 ~~and station data~~because of the limited understanding of the dependence between reanalysis estimates and station

131 observations~~is not considered when merging them in this study~~. Moreover, none of the reanalysis datasets assimilate

132 precipitation data from stations.

133 The elevation data are sourced from the 3 arc-second resolution Multi-Error-Removed Improved-Terrain digital

134 elevation model (MERIT DEM; Yamazaki et al., 2017).

135

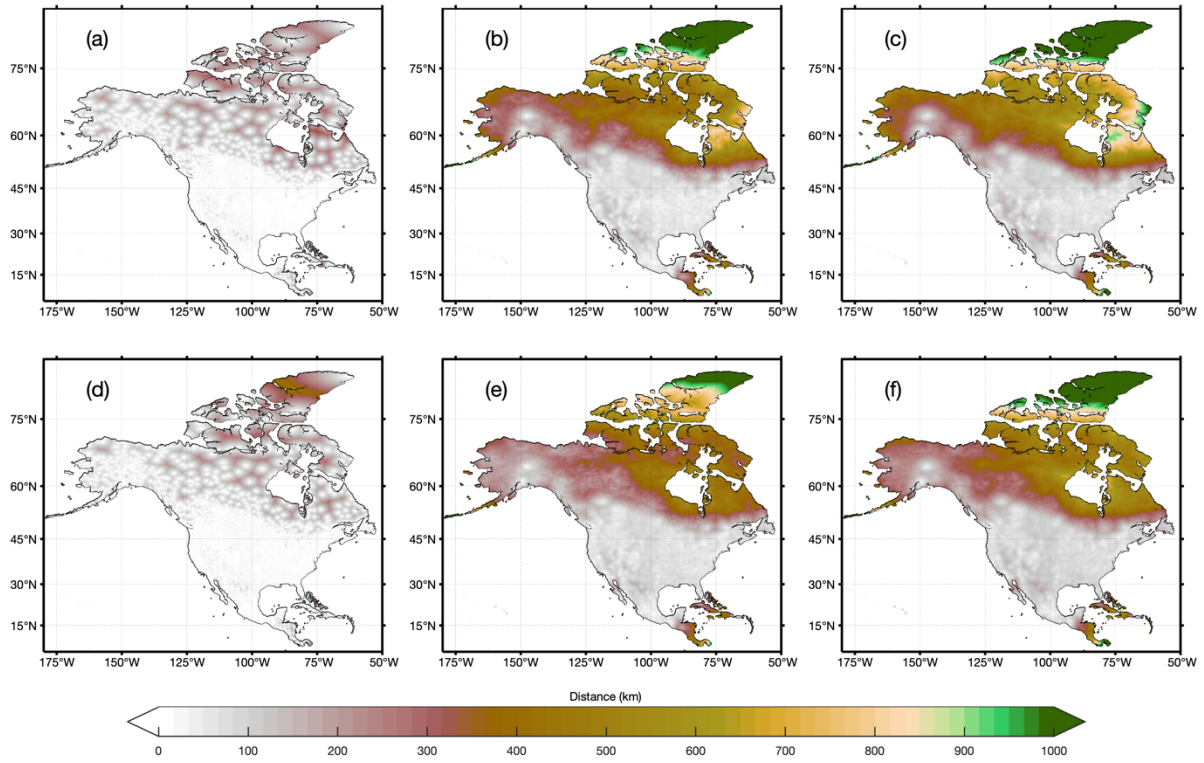



Figure 1. The color of each 0.1° grid indicates the ~~radial~~ radius to find (a) one, (b) 20, and (c) 30 neighboring stations for precipitation (a-c) and temperature (d-f).

3. Methodology

The estimate of a variable at a specific location and time step can be regarded as a random value following a probability distribution. The probability density functions (PDFs) of variables such as ~~the~~ Tmean and Trange~~,~~ can be approximated using the normal distribution. Their value x for ~~the~~ a target location and time step is expressed as:

$$x \sim N(\mu, \sigma^2) \tag{1}$$

where μ is the mean value and σ is the standard deviation. ~~The p~~Probabilistic estimat~~esion~~ of Tmean ~~and~~ or Trange can be realized by sampling from this distribution. In a spatial meteorological dataset, the distribution parameters vary with space and time, and the spatial variability is related to the nature of variables and gridding (interpolation) methods. The performance of gridding methods is critical because accurate estimation of μ can reduce systematic bias and smaller σ means narrower spread.

Precipitation is different from Tmean and Trange because it can be intermittent from local to synoptic scales and its distribution is both highly skewed and bounded at zero. Following Papalexiou (2018) and Newman et al. (2019b), the ~~cumulative density function (CDF)~~ CDF of precipitation can be expressed as below:

150

$$F_X(x) = (1 - p_0)F_{X|X>0}(x) + p_0, \ for \ x \geq 0 \tag{2}$$

151 where $F_X(x)$ is the CDF for $x \geq 0$, $F_{X|X>0}(x)$ is the CDF for $x > 0$, and p_0 is the probability of zero precipitation.

152 The probability of precipitation (PoP) is $1 - p_0$. The CDF $F_{X|X>0}(x)$ is often approximated using the normal

153 distribution after applying suitable transformation functions to observed precipitation. Clark and Slater (2006) perform

154 the normal quantile transformation using an empirical CDF from station observations. Newman et al. (2015) apply a

155 power-law transformation. Newman et al. (2019b) adopts the Box-Cox transformation, that is,

$$x' = \frac{x^\lambda - 1}{\lambda} \tag{3}$$

156 where λ is set to 1/3 following Newman et al. (2019b) and Fortin et al. (2015). Eq. (1) applies to x', enabling the

157 probabilistic estimation of precipitation. Unlike Newman et al. (2019b) that uses transformed precipitation throughout

158 the production, this study only uses Box-Cox transformation when the assumption of normality is necessary (Sect.

159 3.2.4 and 3.3) to reduce the error introduced by the back transformation. The limitations and alternative choices of

160 precipitation transformation are discussed in Sect. 5.2.

161 In summary, seven space- and time-varying parameters (μ and σ for three variables and PoP) should be obtained to

162 realize probabilistic estimation. Our method to develop probabilistic meteorological estimates is summarized in Fig.

163 2a. We apply four main steps to produce EMDNA: (1) station-based regression estimates (Sect. 3.1), (2) the regridding,

164 downscaling, bias correction and merging of three reanalysis products (Sect. 3.2), (3) optimal interpolation-based

165 merging of reanalysis and station-based regression outputs, and the bias correction of the resulting precipitation

166 estimates (Sect. 3.3), and (4) the production of probabilistic estimates in the form of spatial meteorological ensembles

167 (Sect. 3.4).

168

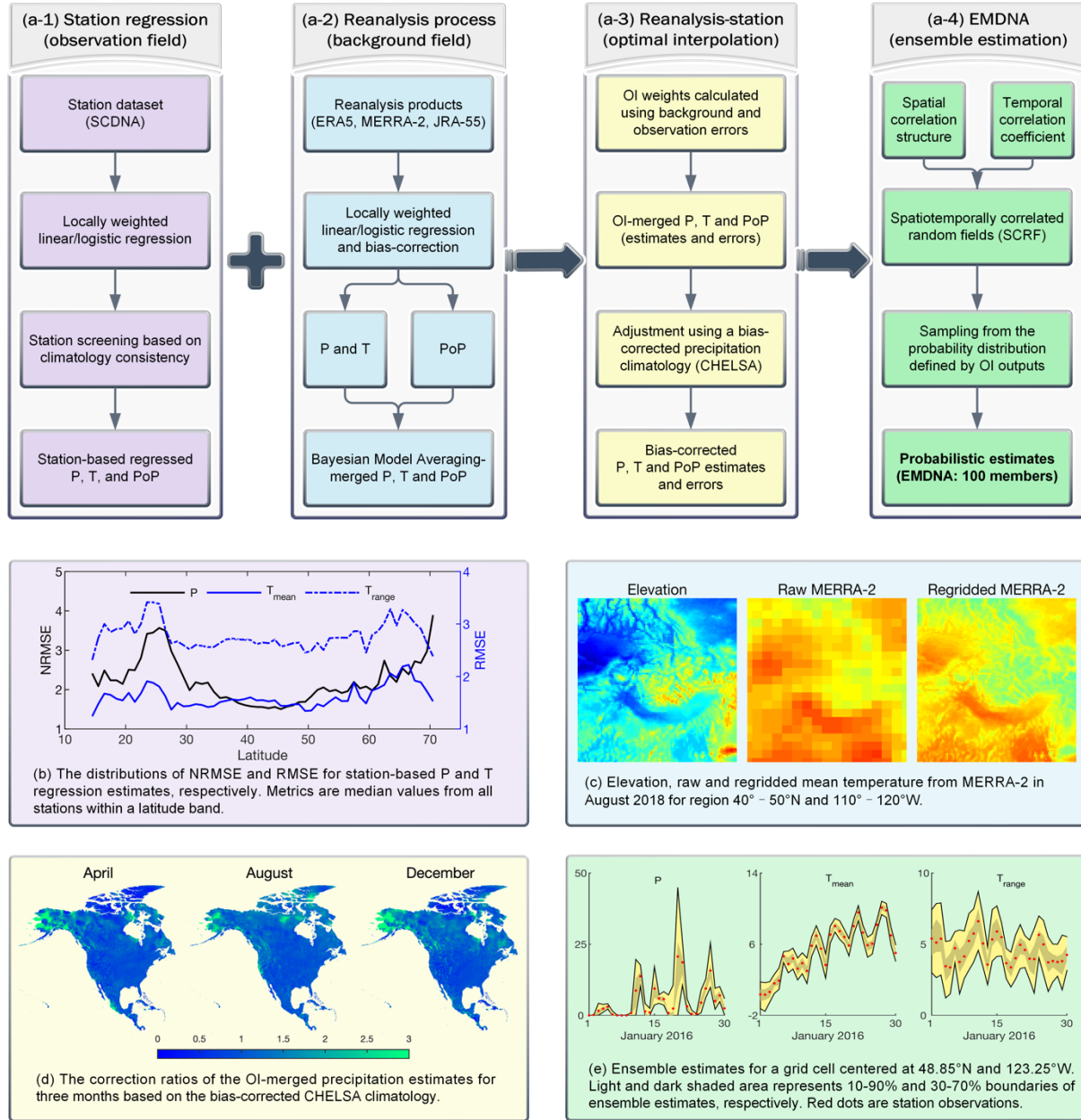



Figure 2. (a) The flowchart outlining the main steps for producing EMDNA. P represents precipitation and T represents temperature. (b-e) demonstrate output examples from (a-1 to -4), respectively. (b) Latitudinal distribution of the root mean square error (RMSE) for temperature and normalized RMSE (NRMSE) for precipitation (Sect. 3.1). (c) Example showing the mean temperature of MERRA-2 before and after regridding (Sect. 3.2). (d) The correction ratios calculated using precipitation climatology from the bias-corrected CHELSA (Sect. 3.3). (e) Example of the ensemble-based distributions of precipitation and temperature estimates from EMDNA (Sect. 3.4).

3.1 Regression estimates from station data

Clark and Slater (2006) and Newman et al. (2015, 2019b) use locally weighted linear regression and logistic regression to obtain gridded ~~precipitation and~~ temperature and gridded precipitation estimates which are used as parameters in Eq. (1). However, for high-latitude regions in North America where stations are scarce (Fig.1), such gridded estimates based only on station data could contain large uncertainties (Fig. 2b) due to the long distances needed to assemble a sufficient sample of stations to form the regressions. This study uses optimal interpolation (OI) to merge data from stations and reanalysis models. In this section, we only obtain regression estimates and their errors at the locations of stations, which are used as inputs to OI in Sect. 3.3.

3.1.1 Locally weighted linear regression

Daily precipitation amount, Tmean and Trange are estimated for all stations based on the locally weighted linear regression (also known as the geographically weighted regression). Let x_o be the station observation for variable X (precipitation, Tmean, and Trange), the regression estimate $\hat{x}$ for the target point and time step is obtained as below:

$$x_o = \hat{x} + \varepsilon = \beta_0 + \sum_{i=1}^{n} A_i \beta_i + \varepsilon \tag{4}$$

where A_i is the ith time-invariant topographic attribute (or predictor variables), β_0 and β_i are regression coefficients estimated using ordinary least squares, and ε is the residual (or error term). The topographic attributes are latitude, longitude, and elevation for Tmean and Trange. For precipitation, two more topographic attributes (west-east and south-north slopes) are used to account for windward and leeward slope precipitation differences. An isotropic Gaussian low-pass filter is used to smooth DEM before calculating slopes, which can reduce the influence of noise in a high-resolution DEM on the large-scale topographic effect of precipitation (Newman et al., 2015). Ideally the scale of this smoothing reflects the scale at which terrain most directly influences precipitation or temperature spatial patterns; in this case the filter bandwidth is 180 km.

For a target station point, $\hat{x}$ is obtained based on data from neighboring stations. Newman et al. (2015, 2019b) used 30 neighboring stations, without controlling for maximum station distance. The very low station density in high-latitude regions makes this configuration infeasible, hence this study adopts a relatively flexible criterion for selecting neighboring stations: (1) finding at most 30 stations within a fixed search radius (400 km), and (2) if fewer than 20 stations are found, extending the search radius until 20 stations are found. The ~~least number~~lower threshold is set to 20 to ensure that linear/logistic regression is robust. To incorporate local dependence, a tricube weighting function is used to calculate the weight $w_{i,j}$ between the target station i and the neighboring station j.

$$w_{i,j} = \left[1 - \left(\frac{d_{i,j}}{d_{max}}\right)^3\right]^3 \tag{5}$$

where $d_{i,j}$ is the distance between i and j, and d_{max} depends on the maximum distance ($d_{i,j}^{max}$) between i and all its neighboring stations. If $d_{i,j}^{max}$ is smaller than 100 km, d_{max} is set to 100 km; otherwise, d_{max} is set to $d_{i,j}^{max} + 1$ km

204   (Newman et al., 2015, 2019b). The cubic weight function is smoother compared to functions such as exponential

205   functions and inverse distance functions, indicating that $w_{i,j}$ degrades with distance in a relatively slow way which

206   generally leads to smooth spatial variations of variables. The comparison of different weight functions could be a

207   direction for future studyresearch direction. Regression coefficients are estimated by weighted least squares method

208   (described in inAppendix A).

209   We found that a small number of observations stations show a climatology that is notably statistically different from

210   surrounding stations, which could cause an adverse effect on gridded estimates, particularly in sparsely gauged regions.

211   Strategies are designed to identify and exclude such stations are summarized in (Appendix B).

3.1.2   Locally weighted logistic regression

213   PoP is estimated using the locally weighted logistic regression by fitting binary precipitation occurrence to spatial

214   topographic attributes:

$$PoP = \frac{1}{1 + \exp(-\beta_0 + \sum_{i=1}^{n} A_i \beta_i)} \tag{6}$$

215   The topographic attributes (A_i) are the same as those used by precipitation regression. Appendix A describes the

216   method to estimate rRegression coefficients are estimated in Appendix A.

217   The errors of precipitation, temperature, and PoP estimates for all stations are calculated as the difference between

218   regression estimates and station observations using the leave-one-out cross-validation procedure (also known as the

219   jackknife procedure). The leave-one-out evaluation could be affected by the distributions of stations in some cases.

220   For example, two stations with very close distance may both show very high accuracy in the leave-one-out evaluation,

221   which is actually (this is a problem for all station-based evaluation methods).

3.2 Regridding, correction, and merging of reanalysis datasets

223   The three reanalysis datasets (ERA5, MERRA-2, and JRA-55) have different spatial resolutions and contain

224   systematic biases. In this section, we discuss steps taken to (1) re-grid all reanalysis datasets to the resolution of

225   EMDNA (0.1°);: (2) perform a correction to remove the systematic bias in original estimates;, and (3) merge the three

226   reanalysis datasets to produce a background field that improves over any individual reanalysis dataset, in support of

227   the reanalysis-station merging described in Sect. 3.3.

3.2.1   Regridding of reanalysis datasets

229   Precipitation, Tmean, and Trange are regridded to 0.1° using locally weighted regression (Fig. 2c). Latitude, longitude,

230   and elevation are used as predictor variables for simplicity. Precipitation or temperature lapse rates are implicitly

231   considered by involving elevation in the regression. Raw reanalysis data from a 5 × 5 space window (i.e., 25 coarse-

232   resolution grids) centered by the 0.1° target grid are used to perform the regression. Each grid is represented using its

233   center point. This regridding method has been proven effective in previous studies (Xu et al., 2015; Duan and Li, 2016;

234   Lu et al., 2020). Reanalysis estimates are also regressed to the locations of all stations to facilitate evaluation and

235   weight estimation in the following steps, which can avoid the scale mismatch caused by using point-scale observations

236   to evaluate 0.1° gridded estimates (Tang et al., 2018a).

237   We also tested other regridding methods such as the nearest neighbor, bilinear interpolation, and temperature lapse

238   rate-based downscaling (Tang et al., 2018b). Results (not shown) indicated that their performance is generally inferior

239   to the locally weighted regression with respect to several accuracy metrics.

240   **3.2.2   Probability of precipitation estimation**

241   Reanalysis precipitation can exhibit large biases in the number of wet days because the models often generate many

242   light precipitation events. To overcome this limitation, we designed two methods for determining the occurrence of

243   reanalysis precipitation. The first is to use positive thresholds to determine precipitation occurrence. The threshold

244   was estimated in two ways, namely by forcing reanalysis precipitation (1) to have the same number of wet days with

245   station data, or (2) to achieve the highest critical success index (CSI). Gridded thresholds can be obtained through

246   interpolation and used to discriminate between precipitation events or non-events. However, this method can only

247   obtain binary occurrence instead of continuous PoP between zero and one. The second method is based on univariate

248   logistic regression. The amount of reanalysis precipitation is used as the predictor and the binary occurrence from

249   station data is used as the predictand. The logistic regression is implemented for each reanalysis product in the same

250   way as Sect. 3.1.2. The comparison between the threshold-based method and the logistic regression-based method

251   shows the latter achieves higher accuracy. Therefore, we adopt the univariate logistic regression to estimate PoP for

252   each reanalysis product in this study. The possible bias caused by station measurements is not considered.

253   **3.2.3   Bias correction of reanalysis datasets**

254   Considering reanalysis products ~~usually~~ contain systematic biases (Clark and Hay, 2004; Mooney et al., 2011; Beck

255   et al., 2017; Tang et al., 2018b, 2020b), the linear scaling method (also known as multiplicative/additive correction

256   factor; Teutschbein and Seibert, 2012) is used to correct reanalysis precipitation, Tmean, and Trange estimates.

257   Reanalysis PoP is not corrected because station information has been incorporated in the logistic regression. Let x_r

258   be the reanalysis estimate for variable X, the corrected estimate for a target grid/point i is calculated as:

$$x^*_{r,i} = \begin{cases} x_{r,i} + \dfrac{\sum_{j=1}^{m} w_{i,j}\left(\bar{x}_{o,j} - \bar{x}_{r,j}\right)}{\sum_{j=1}^{m} w_{i,j}} & \text{additive correction} \\[3em] x_{r,i}\dfrac{\sum_{j=1}^{m} w_{i,j}\dfrac{\bar{x}_{o,j}}{\bar{x}_{r,j}}}{\sum_{j=1}^{m} w_{i,j}} & \text{multiplicative correction} \end{cases} \qquad (7)$$

259   where $x^*_{r,i}$ is the corrected reanalysis estimate, $w_{i,j}$ is the distance-based weight (Eq. (5)), and $\bar{x}_{o,j}$ and $\bar{x}_{r,j}$ are the

260   climatological mean for each month (e.g., all January from 1979 to 2018) from station observations and reanalysis

261   estimates for the *j*th neighboring station, respectively. The additive correction is used for Tmean and Trange, and the

262   multiplicative correction is used for precipitation. The number of neighboring stations (*m*) is set to 10, which is smaller

263   than that used for linear or logistic regression (Sect. 3.1) but should be enough for bias correction. The upper bound

264   of $\frac{\bar{x}_{o,j}}{\bar{x}_{r,j}}$ is set to 10 to avoid over-correction in some cases (Hempel et al., 2013).

265   Linear scaling can ~~also~~ be performed at monthly (Arias-Hidalgo et al., 2013; Herrnegger et al., 2018; Willkofer et al.,

266   2018) or daily (Vila et al., 2009; Habib et al., 2014) time scales by replacing $\bar{x}_{o,j}$ and $\bar{x}_{r,j}$ by the monthly mean (e.g.,

267   January in one year) or daily values. We compared the performance of corrections at different scales and found that

268   monthly- or daily-scale corrections acquire more accurate estimates than the climatological correction. The

269   climatological correction was adopted because (1) it preserves the absolute/relative trends better than daily or monthly

270   corrections, and (2) the OI merging (Sect. 3.3) adjusts daily variability of estimates, which compensates for the

271   limitation of climatological correction and makes daily/monthly-scale correction unnecessary.

272   Quantile mapping is another widely used correction method (Wood et al., 2004; Cannon et al., 2015). We compared

273   quantile mapping and linear scaling and found that they are similar in statistical accuracy, while quantile mapping

274   achieves better probability distributions with much smaller Hellinger distance (Hellinger, 1909) which is a metric used

275   to quantify the similarity between estimated and observed probability distributions. Nevertheless, quantile mapping

276   could result in spatial smoothing of precipitation and temperature, particularly in high-latitude regions where stations

277   are few. For example, Ellesmere Island, the northernmost island of the Canadian Arctic Archipelago, usually shows

278   lower temperature in inland regions ~~due to orographic uplift of air parcels/clouds~~. However, quantile mapping will

279   erase this gradient because reanalysis grids for this island are corrected based on ~~almost the same reference~~

280   ~~stations~~stations on the coast. To ensure the authenticity of spatial distributions, quantile mapping is not used in this

281   study.

3.2.4   Merging of reanalysis datasets

283   The three reanalysis products are merged using the Bayesian Model Averaging (BMA, Hoeting et al., 1999), which

284   has proved to be effective in fusing multi-source datasets (Chen et al., 2015; Ma et al., 2018a, b). According to the

285   law of total probability, the PDF of the BMA estimate can be written as:

$$p(E) = \sum_{r=1}^{3} p(E|x_r^*, x_o) \cdot p(x_r^*|x_o) \tag{8}$$

286   where E is the ensemble estimate, x_r^* (*r*=1, 2, 3) is the bias-corrected estimate from three reanalysis products,

287    $p(E|x_r^*, x_o)$ is the predicted PDF based only on a specific reanalysis product, and $p(x_r^*|x_o)$ is the posterior probability

288   of reanalysis products given the station observation x_o. The posterior probability $p(x_r^*|x_o)$ can be identified as the

289   fractional BMA weight w_r with $\sum_{r=1}^{3} w_r = 1$. BMA prediction can be written as the weighted sum of individual

290   reanalysis products.

For Tmean and Trange, $p(E|x_r^*, x_o)$ can be regarded as the normal distribution $g(E|\theta_r)$ defined by the parameter $\theta_r = \{\mu_r, \sigma_r^2\}$, where μ_r is the mean and σ_r^2 is the variance (Duan and Phillips, 2010). For precipitation, if we apply Box-Cox transformation (Eq. (3)) to positive events (>0) and exclude zero events, its distribution is approximately normal, and $p(E|x_r^*, x_o)$ can be represented using $g(E|\theta_r)$. Therefore, Eq. (8) can be written as:

$$p(E) = \sum_{r=1}^{3} w_r \cdot g(E|\theta_r) \tag{9}$$

There are different approaches to infer w_r and θ_r (Schepen and Wang, 2015). This study uses the log-likelihood function to estimate the parameters (Duan and Phillips, 2010; Chen et al., 2015; Ma et al., 2018b). The Expectation-Maximization algorithm (Raftery et al., 2005) can be applied to estimate parameters by maximizing the likelihood function. BMA weights are obtained for all stations and each month. Gridded weights are obtained using the inverse distance weighting interpolation.

Merging multiple datasets could affect the probability distributions and extreme characteristics of original datasets. This is not a major concern because the merged reanalysis data are further adjusted by station data in OI merging (Sect. 3.3), a later step in the EMDNA process. Also, the probabilistic estimation of ensemble members (Sect. 3.4) has a large effect on estimates of extreme events.

Gridded errors of BMA-merged estimates are necessary to enable optimal interpolation (Sect. 3.3). The error estimation is realized using a two-layer cross-validation (Appendix C).

3.3 Optimal Interpolation-based merging of reanalysis and station data

3.3.1 Optimal Interpolation

OI has proven to be effective in merging multiple datasets (Sinclair and Pegram, 2005; Xie and Xiong, 2011) and has been applied in operational products such as CaPA (Mahfouf et al., 2007; Fortin et al., 2015) and the China Merged Precipitation Analysis (CMPA, Shen et al., 2014, 2018). Let x_A be the OI analysis estimate. The OI analysis estimate ($x_{A,i}$) for a target grid/point i and time step is obtained by adding an increment to the first guess of the background ($x_{B,i}$). The increment is a weighted sum of the difference between observation and background values at neighboring stations.

$$x_{A,i} = x_{B,i} + \sum_{j=1}^{m} w_j(x_{O,j} - x_{B,j}) \tag{10}$$

where $x_{O,j}$, $x_{B,j}$, and w_j are the observed value (subscript O), background value (subscript B), and weight for the jth neighboring station. Let x_T be the true value, the errors of observed and background values are $\varepsilon_{O,j} = x_{O,j} - x_{T,j}$ and $\varepsilon_{B,j} = x_{B,j} - x_{T,j}$ (or $\varepsilon_{B,i} = x_{B,i} - x_{T,i}$), respectively. Assuming that (1) the observation and background errors are unbiased with an expectation of zero and (2) there is no correlation between background and observation errors, the weights that minimize the variance of the analysis errors can be obtained by solving:

$$w(R + B) = b \qquad (11)$$

where $\mathbf{w}$ is the vector of w_j ($j = 1,2,...,$ m), $\mathbf{R}$ and $\mathbf{B}$ are $m \times m$ covariance matrices of $\varepsilon_{O,j}$ and $\varepsilon_{B,j}$, respectively, and $\mathbf{b}$ is the $m \times 1$ vector of covariance between $\varepsilon_{B,i}$ and $\varepsilon_{B,j}$. The background provided by reanalysis models assimilates observations in the production and is corrected in a way using station data (described in Sect. 3.2.3), which may affect the soundness of the second assumption. The effect of this slight violation, however, is rather small according to our results and previous studies (Xie and Xiong, 2011; Shen et al., 2014b, 2018).

Different approaches can be used to implement OI. For example, Fortin et al. (2015) used raw station observations as x_O, and assumed that the background error is a function of error variance and correlation length, and the observation error is a function of error variance. The variances and correlation length are obtained by fitting a theoretical variogram using station observations. Xie and Xiong (2011) and Shen et al. (2014) use station-based gridded estimates as x_O, and assume that the background error variance is a function of precipitation intensity, the cross-correlation of background errors is a function of distance, and the observation error variance is a function of precipitation intensity and gauge density. The parameters of those functions are estimated based on station data in densely gauged regions.

In this study, we adopt a novel design that calculates weights based on error estimation, a feature that is enabled by the probabilistic nature of the observational dataset. Regression estimates and their errors at station points (Sect. 3.1) are used as x_O and ε_O, respectively. BMA-merged reanalysis estimates and their errors (Sect. 3.2) are used as x_B and ε_B, respectively. We do not use gridded regression estimates because (1) $x_{O,j} - x_{B,j}$ will show weak variation if neighboring stations are replaced by neighboring grids, and (2) estimates of weights $\mathbf{w}$ could be unrealistic because of the spatial smoothing of interpolated regression errors. The advantages of this design are (1) weights and inputs closely match each other and (2) weights in sparsely gauged regions are not determined by parameters fitted in densely gauged regions. In regions with few stations, the errors of regression estimates could be larger than reanalysis estimates, resulting in a smaller contribution from regression estimates and a larger contribution from reanalysis estimates, which is the complementary effect we expect by involving reanalysis datasets in EMDNA.

The Box-Cox transformation is applied to precipitation estimates. Then, precipitation, PoP, Tmean, and Trange estimates provided by OI are used as μ and PoP required for generating meteorological ensembles.

3.3.2   Error of OI-merged estimates

Variance is a necessary parameter to enable ensemble estimation. The variance σ^2 is represented using the mean squared error of OI estimates in this study. First, the error of OI analysis estimates ($\varepsilon_A = x_A - x_o$) is obtained for all stations using the leave-one-out strategy. Then, the σ_i^2 for the ith grid is obtained as a weighted sum of squared errors from neighboring stations:

$$\sigma_i^2 = \frac{\sum_{j=1}^{m} w_{i,j} (\varepsilon_{A,j})^2}{\sum_{j=1}^{m} w_{i,j}} \qquad (12)$$

348 where $\varepsilon_{A,j}$ is the difference between the station observation and OI estimate at the jth neighboring station, and $w_{i,j}$ is

349 the weight (Eq. (5)).

3.3.3 Correction of precipitation under-catch

351 Considering station precipitation data usually contain measurement errors such as wind-induced under-catch

352 particularly in high-latitude and mountainous regions, OI-merged precipitation is further adjusted using thee

353 Precipitation Bias Correction (PBCOR) datasetthe bias-corrected precipitation climatology produced by Beck et al.

354 (2020). The This PBCOR climatology infers the long-term precipitation (without rain-snow separation) using a

355 Budyko curve and streamflow observations collected from seven national and international sources, among which the

356 Global Runoff Data Centre (GRDC), the U.S. Geological Survey (USGS), and the Water Survey of Canada

357 Hydrometric Data (HYDAT) are data sources in North America. The streamflow stations are scarce in high latitude

358 regions and absent in Greenland.-. Three corrected datasets are provided, including WorldClim, version 2 (WorldClim

359 V2; Fick and Hijmans, 2017), the Climate Hazards Group Precipitation Climatology, version 1 (CHPclim V1; Funk

360 et al., 2015) and Climatologies at High Resolution for the Earth's Land Surface Areas, version 1.2 (CHELSA V1.2;

361 Karger et al., 2017). The water balance-based method of Beck et al. (2020) considers all measurement errors (e.g.,

362 under-catch and wetting/evaporation loss) as a whole and under-catch is the major error source in many regions. Note

363 that the rain gauge catch error includes both under-catch an(Hempel et al., 2013)d over-catch. The potential over-

364 catch could be caused by splash of rain or blow snow collected on the wind shield (Folland, 1988; Zhang et al., 2019).

365 Since over-catch is less common compared to under-catch and the PBCOR dataset does not consider over-catch, the

366 bias correction in this study only addresses the under-catch problem. Moreover, the water balance estimates of

367 precipitation under-catch do not consider non-contributing areas of river basins (e.g., endorheic sub-catchments),

368 which are common in the Canadian Prairies and the northern Great Plains in the USA.

369 Although the three datasets show similar precipitation distributions after bias correction, CHELSA V1.2 is used

370 because its period (1979–2013) is most similar to our study period (1979–2018). The correction of OI-merged

371 precipitation is performed in two steps: (1) the ratio between bias-corrected CHELSA V1.2 and OI-merged long-term

372 monthly precipitation is calculated at the 0.1° resolution during 1979–2013, and (2) daily OI-merged precipitation

373 estimates during 1979–2018 are scaled using the corresponding monthly ratio map. The bias correction notably

374 increases precipitation in northern Canada and Alaska (Fig. 2d) where precipitation -under-catch of precipitation is

375 often largesignificant due to the large proportion of snowfall. The uncertainties of gridded estimates are typically

376 larger in high-latitude sparsely gauged regions and topographically elevated regions, which is partly related to the

377 increased proportion of snowfall and hence larger gauge catch errors.

3.4 Ensemble generation

3.4.1 Spatiotemporally correlated random fields

Spatially correlated random fields (SCRFs) are used to sample from the probability distributions of precipitation and temperature. The SCRFs are produced using the following three steps. First, the spatial correlation structure is generated based on an exponential correlation function:

$$c_{i,j} = \exp\left(-\frac{d_{i,j}}{C_{len}}\right) \tag{13}$$

where $d_{i,j}$ is the distance between grids i and j, and C_{len} is the spatial correlation length determined for each climatological month based on regression using station data for precipitation, Tmean, and Trange, separately. The spatial correlation structure is generated using the conditional distribution approach. Every point is conditioned on previously generated points which are determined using a nested simulation strategy to improve the calculation efficiency (Clark and Slater, 2006).

Second, the spatially correlated random field ($\mathbf{R}_t$) for the tth time step is generated by sampling from the normal distribution with the mean value and standard deviation depending on the random numbers of previously generated grids (Clark and Slater, 2006).

Third, the SCRF is generated by incorporating spatial and temporal correlation relationships. Let ρ_{TM} and ρ_{TR} be the lag-1 auto-correlation for Tmean and Trange, respectively, ρ_{CR} be the cross-correlation between Trange and precipitation, $\mathbf{R}_{t-1,TM}$, $\mathbf{R}_{t-1,TR}$ and $\mathbf{R}_{t-1,PR}$ be the SCRF for the $(t-1)$th time step for Tmean, Trange, and precipitation, respectively, the SCRF for tth time step following (Newman et al., 2015) is written as:

$$\begin{cases} \mathbf{R}_{t,TM} = \rho_{TM}\mathbf{R}_{t-1,TM} + \sqrt{1-\rho_{TM}^2}\,\mathbf{R}_{t-1,TM} \\[2mm] \mathbf{R}_{t,TR} = \rho_{TR}\mathbf{R}_{t-1,TR} + \sqrt{1-\rho_{TR}^2}\,\mathbf{R}_{t-1,TR} \\[2mm] \mathbf{R}_{t,PR} = \rho_{CR}\mathbf{R}_{t-1,TR} + \sqrt{1-\rho_{CR}^2}\,\mathbf{R}_{t-1,PR} \end{cases} \tag{14}$$

3.4.2 Probabilistic estimation

Probabilistic estimates are produced using the probability distribution $N(\mu, \sigma^2)$ in Eq. (1) and $\mathbf{R}$ in Eq. (14). For Tmean and Trange, the SCRF ($\mathbf{R}_{TM}$ and $\mathbf{R}_{TR}$) is directly used as the standard normal deviate (R_X). The estimate (x_e) for the ensemble member e is written as:

$$x_e = \mu + \mathrm{R}_X \cdot \sigma \tag{15}$$

399 For precipitation, an additional step is to judge whether an event occurs or not according to OI-merged PoP and the
400 estimated probability from the SCRF. Let $F_N(x)$ be the CDF of the standard normal distribution, $F_N(R_{PR})$ is the
401 cumulative probability corresponding to the random number R_{PR}. If $F_N(R_{PR})$ is larger than p_0, the scaled cumulative
402 probability of precipitation (p_{cs}) is calculated as:

$$p_{cs} = \frac{F_N(R_{PR}) - p_0}{1 - p_0} \tag{16}$$

403 The probabilistic estimate for precipitation can be expressed as:

$$x_e = \begin{cases} 0 & if & F_N(R_{PR}) \le p_0 \\ \mu + F_N^{-1}(p_{cs}) \cdot \sigma & if & F_N(R_{PR}) > p_0 \end{cases} \tag{17}$$

404 **3.5 Evaluation of probabilistic estimates**

405 Independent stations that are not used in SCDNA are used to evaluate EMDNA because the leave-one-out strategy is
406 too time-consuming ~~for~~ to evaluat~~cing~~ probabilistic estimates. GHCN-D stations with precipitation or temperature
407 records less than eight years are extracted because SCDNA restricts attention to stations with at least eight-year records.
408 In total, 15,018 precipitation stations and 2,455 temperature stations are available for independent testing.

409 The Brier skill score (BSS; Brier, 1950) is used to evaluate probabilistic precipitation estimates. The continuous ranked
410 probability skill score (CRPSS) is used to evaluate probabilistic temperature estimates. Their definitions are described
411 in Appendix D.

412 Furthermore, the reliability and discrimination diagrams are used to assess the behavior of probabilistic precipitation
413 estimates. The reliability diagram shows the conditional probability of an observed event (precipitation above a
414 threshold) given the probability of probabilistic precipitation estimates. In a reliability diagram, a perfect match has
415 all points located on the 1-1 line. The discrimination diagram shows the PDF of probabilistic precipitation estimates
416 for different observed categories. For precipitation, two categories are defined: events or non-events, i.e., observed
417 precipitation above or below a threshold. The difference between PDF curves of events or non-events represents the
418 degree of discrimination. Larger discrimination is preferred. The PDF for non-event/event should be maximized at the
419 probability of zero/one.

420 **4. Results**

421 **4.1 Comparison between raw and merged reanalysis estimates**

422 The three raw reanalysis estimates are regridded, corrected for bias, and merged. In this section, we directly compare
423 raw and BMA-merged estimates. The evaluation is performed for all stations using the two-layer cross-validation
424 strategy. The correlation coefficient (CC), ~~and~~ root mean square error (RMSE), and normalized RMSE (NRMSE) are

425   used as evaluation metrics. RMSE is sued for Tmean, and NRMSE is used for precipitation and Trange to remove

426   patterns caused by climatology.

427   For precipitation, the three reanalysis products show the highest CC in CONUS and the lowest CC in Mexico (Fig. 3).

428   The slight spatial discontinuity of CC along the Canada-USA border and the USA-Mexico border (Fig. 3 and 6) is

429   caused by the inconsistent reporting time of stations. Daily precipitation from reanalysis products is accumulated from

430   0 to 24 UTC, while stations from different countries or regions usually have different UTC accumulation periods

431   (Beck et al., 2019; Tang et al., 2020a). The distributions of RMSE agrees with those of precipitation amounts with

432   higher values in the southern corner and west coast of North America and western CONUSNRMSE is higher in central

433   CONUS and Mexico compared to other regions. Overall, ERA5 outperforms MERRA-2 followed by JRA-55.

434   BMA-merged precipitation estimates show higher accuracy than all reanalysis products (Fig. 3). For ERA5 and JRA-

435   55, tThe improvement of CC and NRMSE is the most evident in the Rocky Mountains, while for MERRA-2, the

436   largest improvement is also obvious occurs in central CONUS. ERA5 is the closest to BMA estimates concerning CC

437   and NRMSE. The improvement of BMA estimates against ERA5 is more prominent in the high-latitude regions.

438   Specifically, the mean CC increases by 0.05 and 0.07 in regions southern and northern to 55°N, respectively. The

439   corresponding decrease of mean NRMSE is 0.150.72 and 0.210.89 mm/d, respectively.

440

441

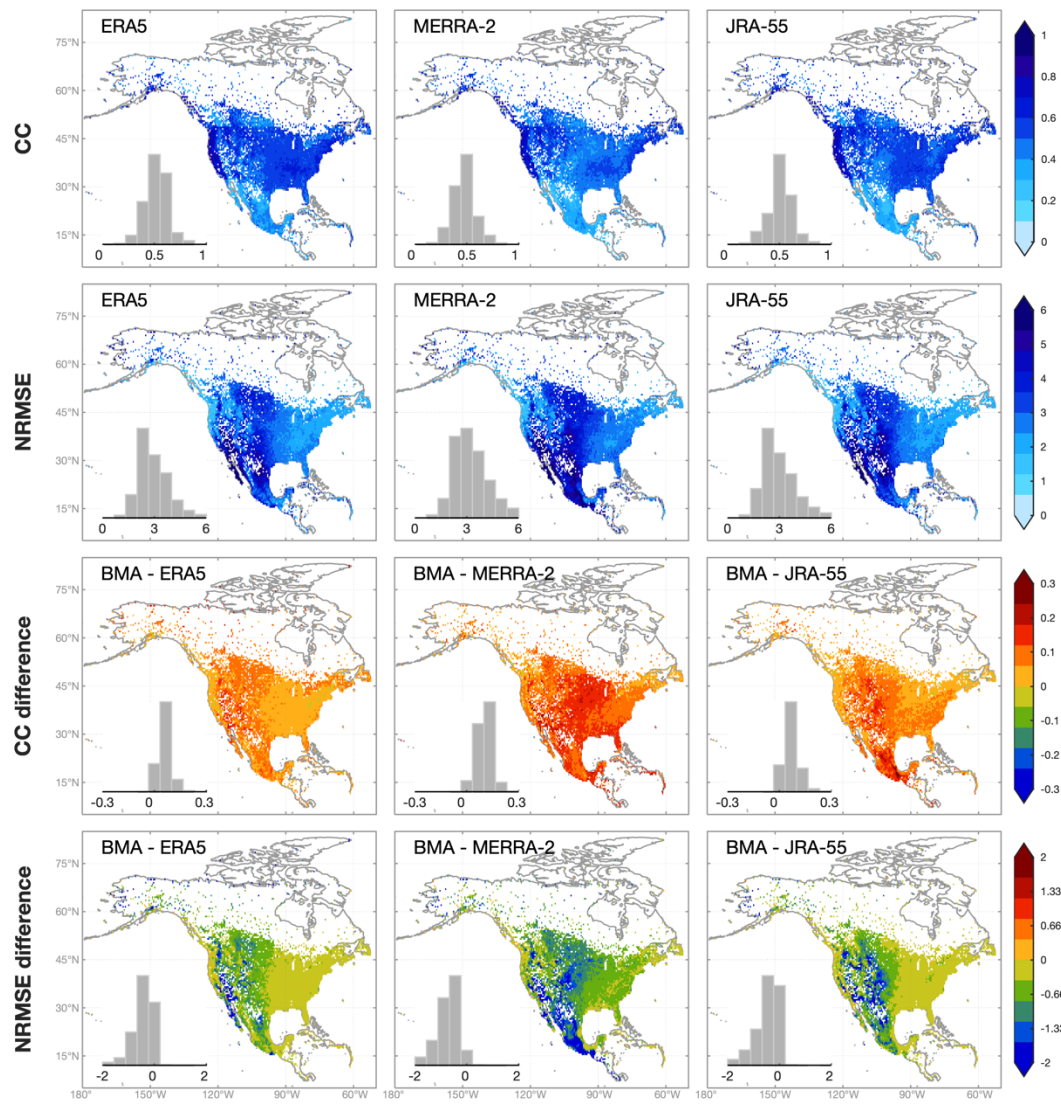
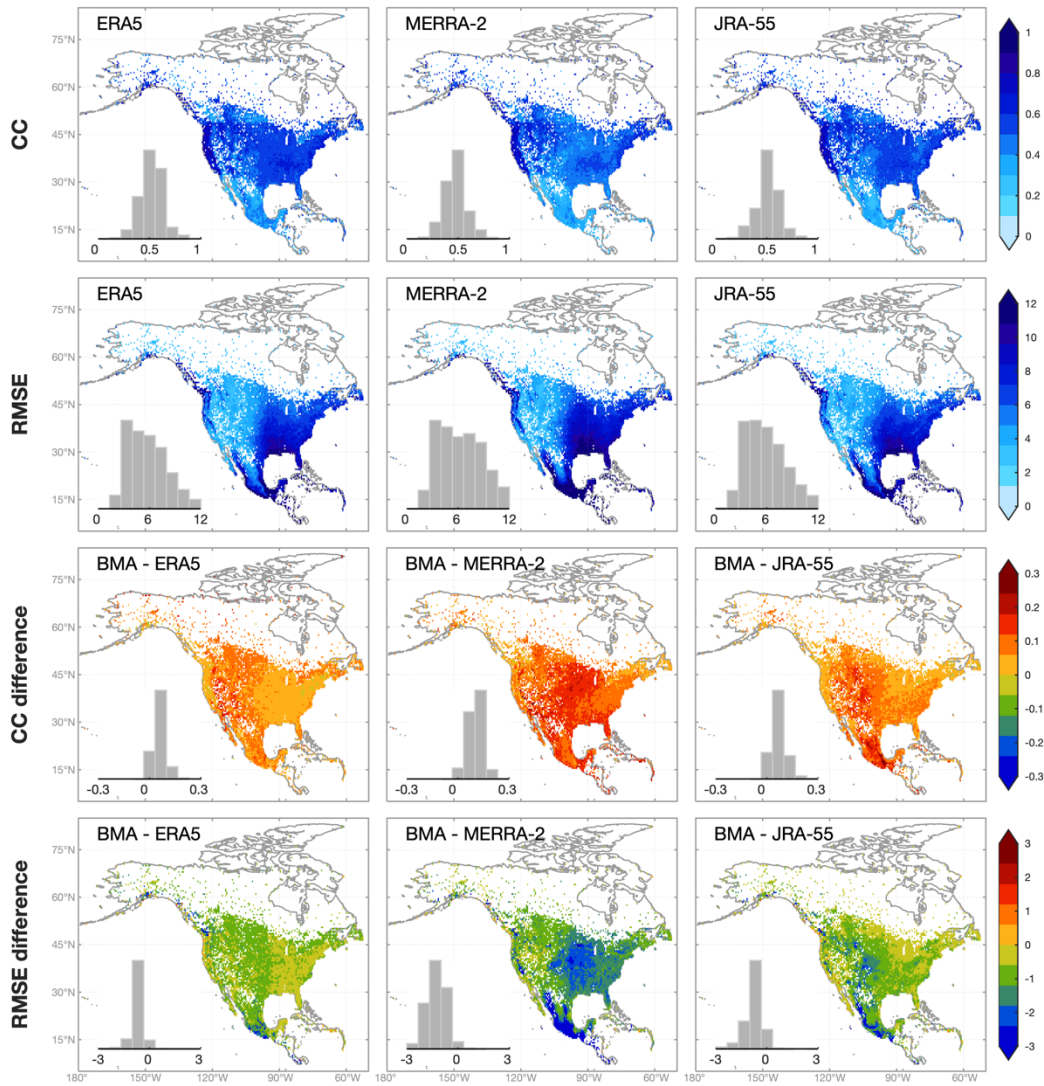



Figure 3. The spatial distributions and histograms of CC (the first row) and NRMSE (the second row) based on raw reanalysis precipitation estimates (ERA5, MERRA-2, and JRA-55). The improvement of BMA-merged estimates against raw reanalysis estimates is shown in the third and fourth rows. The maps are at the ~~The grid resolution is~~ 0.5° resolution, and the value of ~~For~~ each 0.5° grid point~~, the~~ is the median metric of ~~median value of~~ all stations located within the grid ~~is shown~~.

The CC of reanalysis Tmean estimates is close to one in most regions of North America (Fig. 4) and still above 0.9 in Mexico where the CC is the lowest. According to RMSE, Tmean estimates have the largest error in western North America because coarse-resolution raw reanalysis estimates cannot reproduce the variability of temperature caused by elevation variations. The rank of three reanalysis products for Tmean is the same as that for precipitation with ERA5 being the best one. BMA estimates show higher CC than reanalysis products particularly in Mexico, while the improvement of RMSE is the most notable in the Rocky Mountains. For a few stations, the RMSE of BMA estimates

453   is slightly worse than raw reanalysis estimates (Fig. 4) because the downscaling of reanalysis temperature could

454   occasionally magnify the error in low-altitude regions (Tang et al., 2018b).

455   For Trange, BMA estimates show much larger improvement than Tmean, while the differences of CC and NRMSE

456   are relatively evenly distributed (Fig. 5). The improvement of BMA estimates against JRA-55 estimates is especially

457   large. In general, BMA is effective in improving the accuracy of reanalysis precipitation and temperature estimates.

458

459

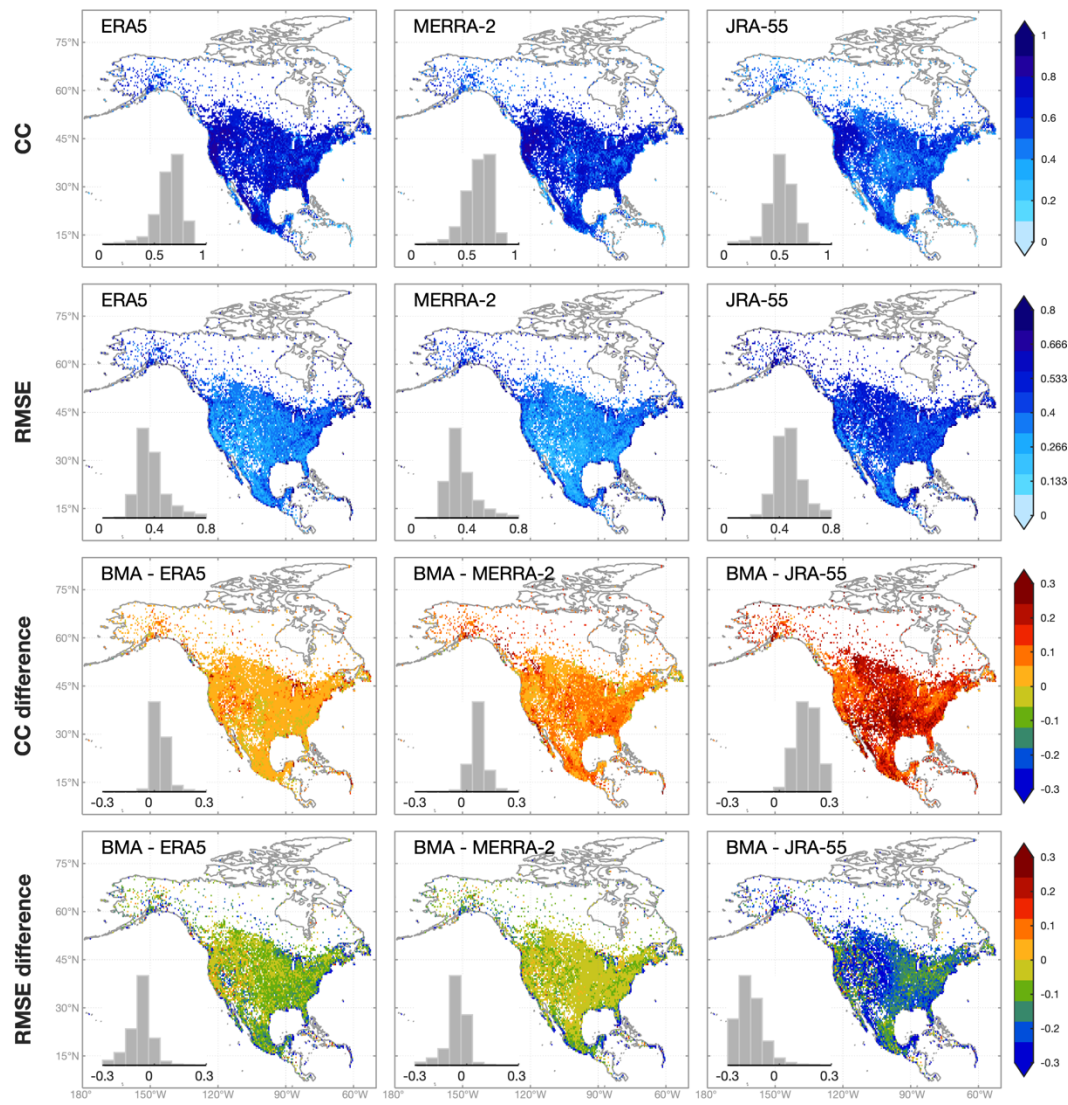
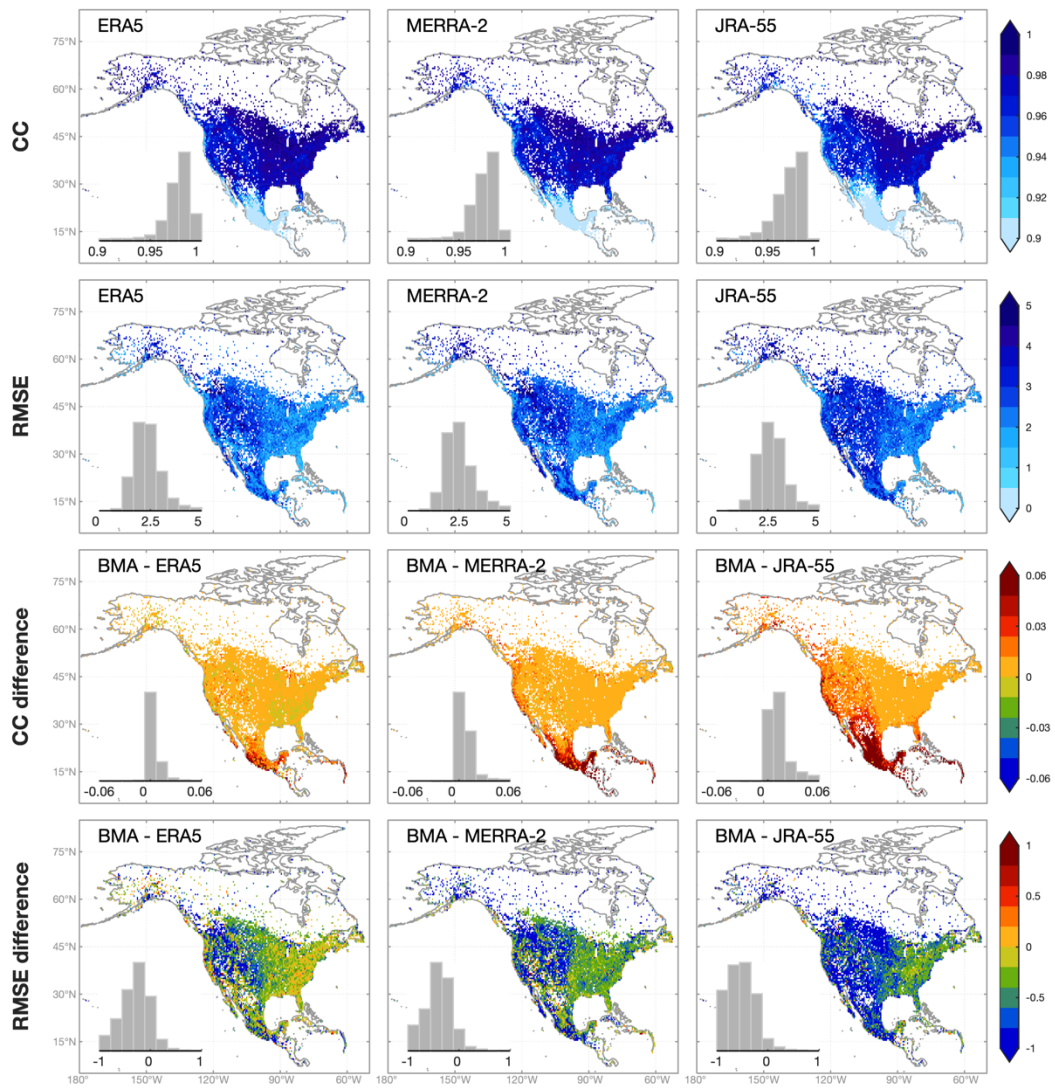



Figure 4. Same with Figure 3, but for mean temperature.

461

462

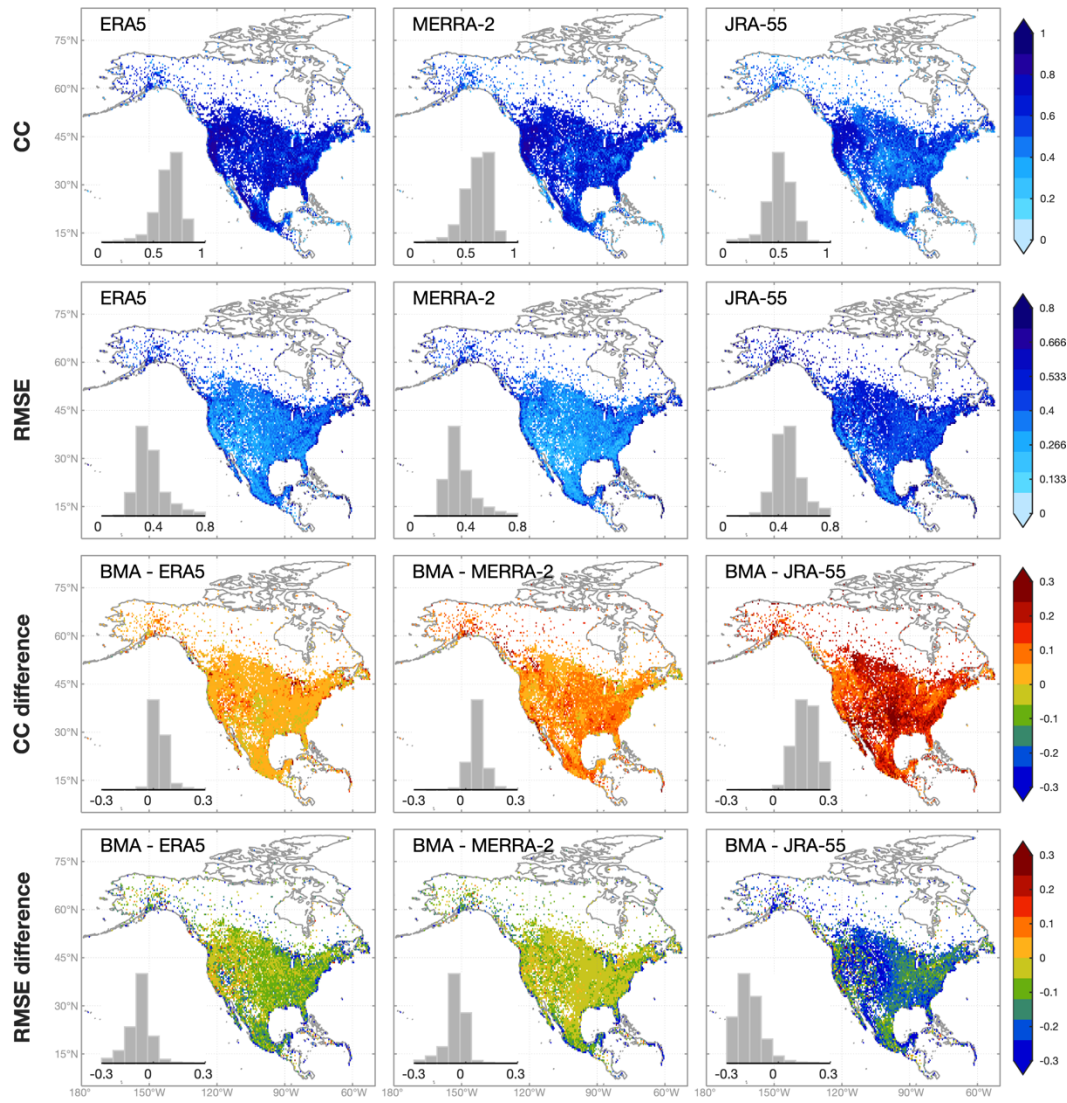
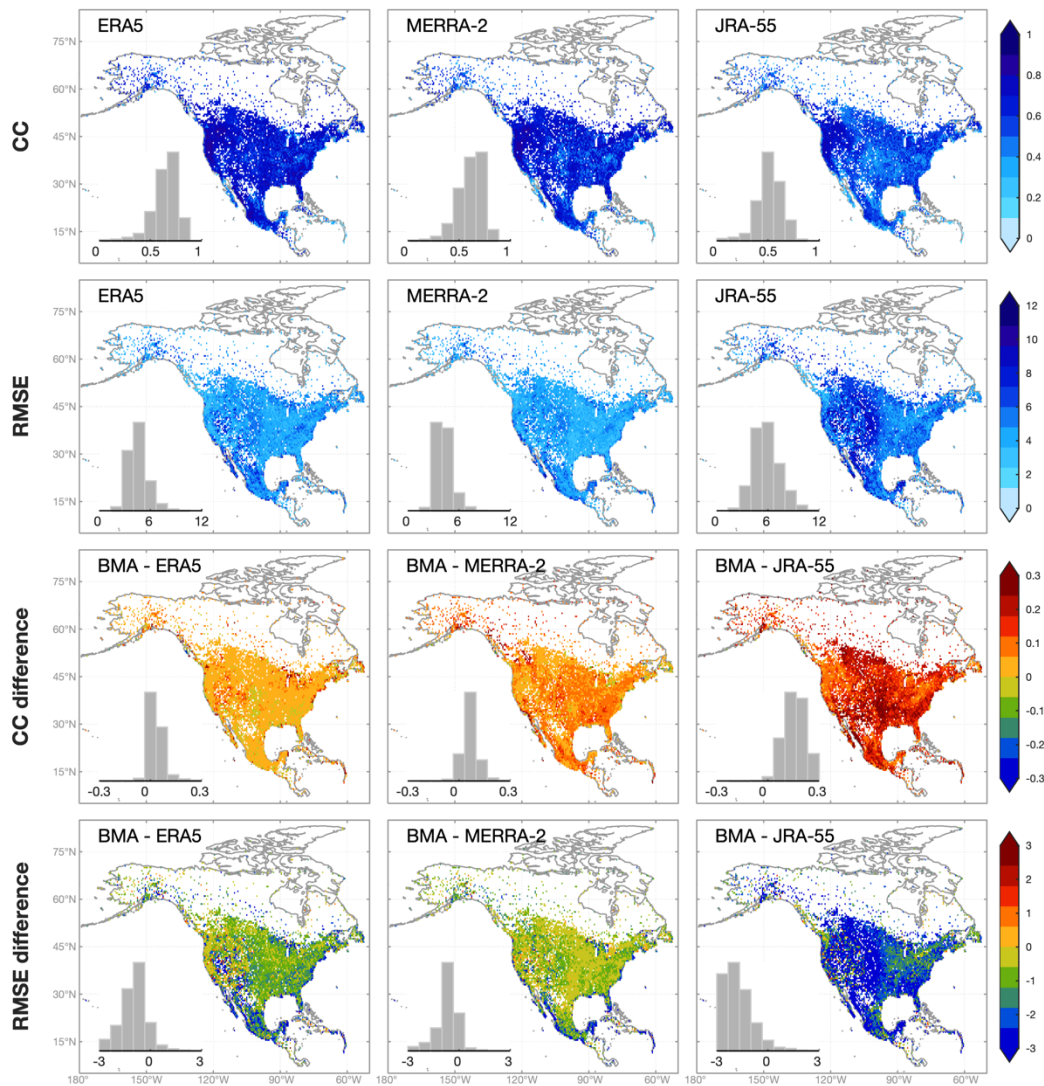



Figure 5. Same with Figure 3, but for daily temperature range.

4.2 The performance of optimal interpolation

Optimal interpolation is used to combine station-based estimates with reanalysis estimates. The performance of OI-merged precipitation and temperature estimates is compared to the background (BMA-merged reanalysis estimates; Fig. 6) and observation (station-based regression estimates; Fig. 7) inputs. To better show the spatial variations of the improvement of OI estimates, RMSE for precipitation and Trange is normalized using the mean value (termed as NRMSE), while Tmean is evaluated using RMSE.

Overall, OI estimates are more accurate than merged reanalysis or station regression estimates for all variables across North America. Comparing OI estimates to reanalysis estimates, for precipitation, Tmean, and Trange, the mean CC is improved by 0.24, 0.02, and 0.15, respectively, and the mean RMSE is reduced by 1.88 mm/d, 0.52°C, and 0.87°C,

473 respectively. The improvement of OI estimates against station estimates is smaller with the mean CC increasing by
474 0.06, 0.01 and 0.05, and the mean RMSE decreasing by 0.56 mm/d, 0.18℃, and 0.29℃ for precipitation, Tmean, and
475 Trange, respectively.

476 OI can utilize the complementarity between station and reanalysis estimates. For example, according to CC, the
477 improvement of OI estimates against reanalysis estimates is larger in the eastern than the western CONUS, while the
478 improvement against station estimates is larger in western than eastern CONUS. This means that although station
479 estimates generally show higher accuracy than reanalysis estimates, station estimates face more severe quality
480 degradation in mountainous regions. Moreover, the latitudinal curves of CC and NRMSE in Fig. 6 and 7 indicate that
481 the improvement of OI estimates against reanalysis estimates decreases as the latitude increases from southern
482 CONUS to northern Canada, while the improvement against station estimates shows a reverse trend.

483 For Tmean, the CC improvement for OI estimates is the largest in Mexico and decreases from low to high latitudes,
484 while based on RMSE, the improvement increases with latitude. For Trange, the latitudinal variation exhibits a similar
485 pattern with precipitation for regions north of 50°N, with larger/smaller improvement in higher latitudes against
486 station/reanalysis estimates. For regions south of 50°N, the improvement of CC and NRMSE against station estimates
487 shows different trends.

488 The latitudinal variations in Fig. 6 and 7 are related to station densities (Fig. 8). Station-based estimates often have
489 lower accuracy in regions with scarce stations (i.e., high-latitude North America), while reanalysis estimates could
490 have less dependence on station densities due to the compensation of physically-based models. Therefore, For
491 precipitation, the improvement of OI estimates against regression estimates increases with the distance according to
492 both CC and NRMSE, while the improvement against reanalysis estimates shows an inverse trend (Fig. 8). The shaded
493 area figure within Fig. 8 shows that most stations can find the 20 neighboring stations within the search radius of 20-
494 100 km. However, as the distance increases beyond 200 km, the number of stations becomes very small while the
495 number of grids is still large. For Tmean, the trend with distance is not obvious probably because it is usually easier
496 to interpolate Tmean observations due to its strong linkage with elevation and latitude. For Trange, the improvement
497 against reanalysis and station estimates both increases with the distance. The results show that OI merging is
498 particularly useful in sparsely gauged regions.

499

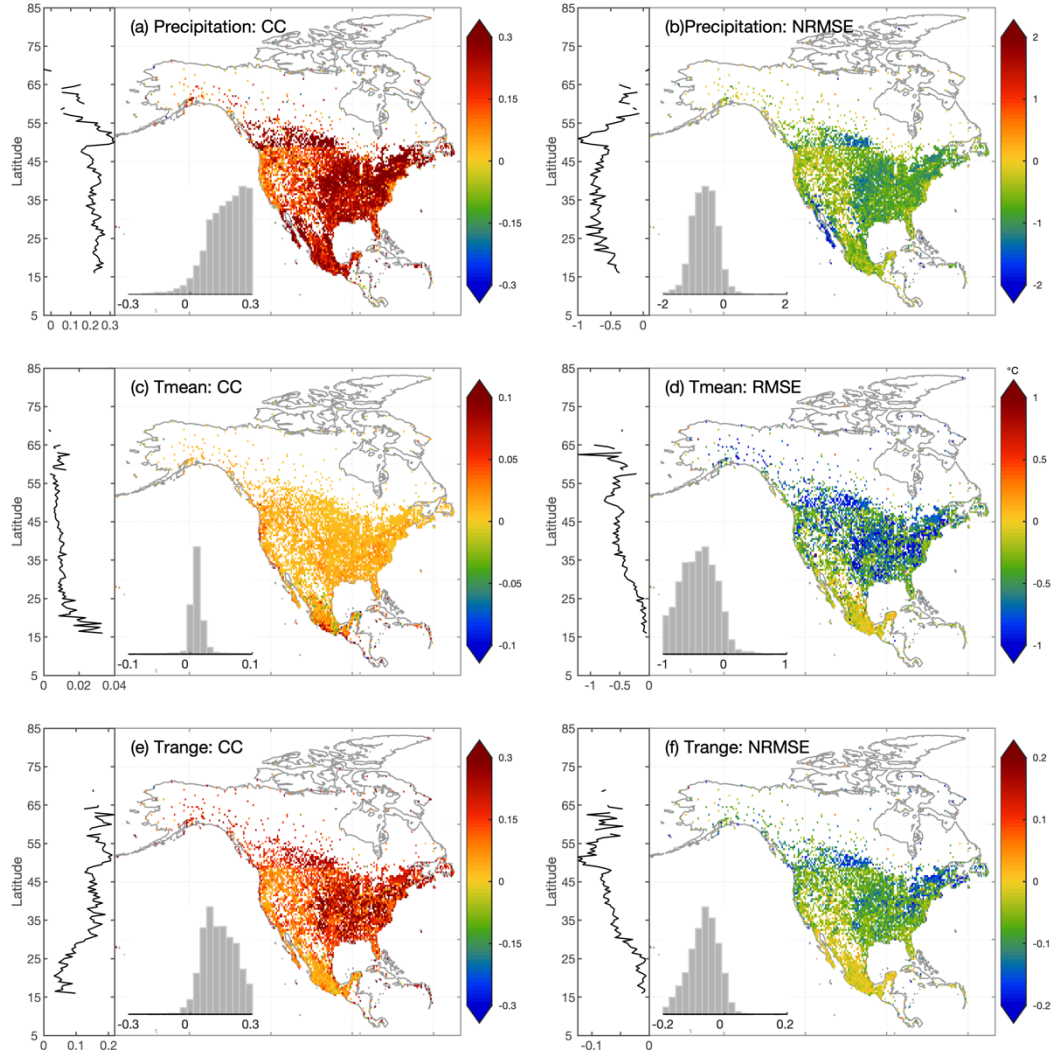



Figure 6. The differences of (a) CC and (b) NRMSE (normalized RMSE) between OI-merged precipitation estimates and BMA-merged reanalysis precipitation estimates. The latitudinal distributions of metrics are attached on the left side, showing the median value for 0.5° latitude bands. (c-d) are the same with (a-b) but for mean temperature and RMSE is not normalized. (e-f) are the same with (a-b) but for daily temperature range.

504

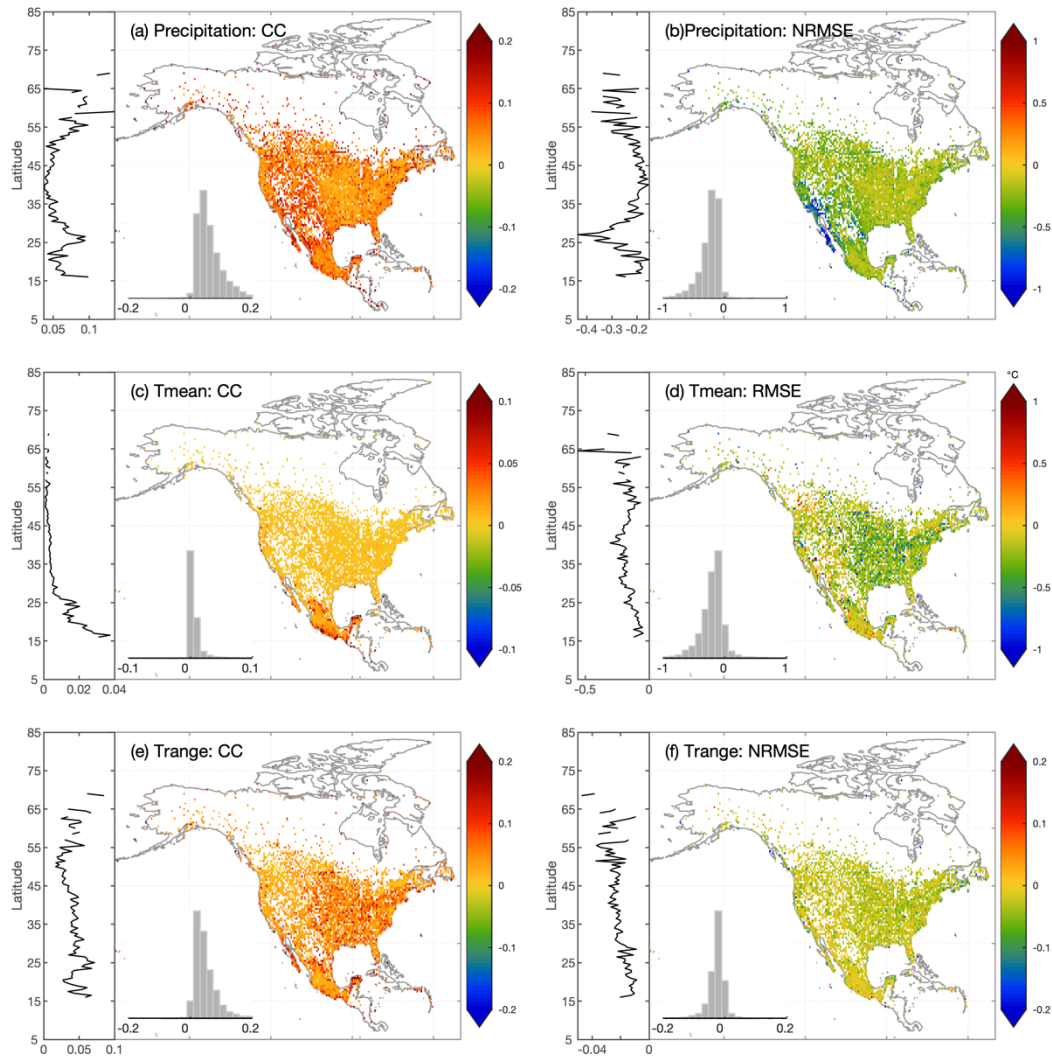



Figure 7. Similar with Figure 6, but the differences are between OI-merged precipitation estimates and station-based

506   regression precipitation estimates.

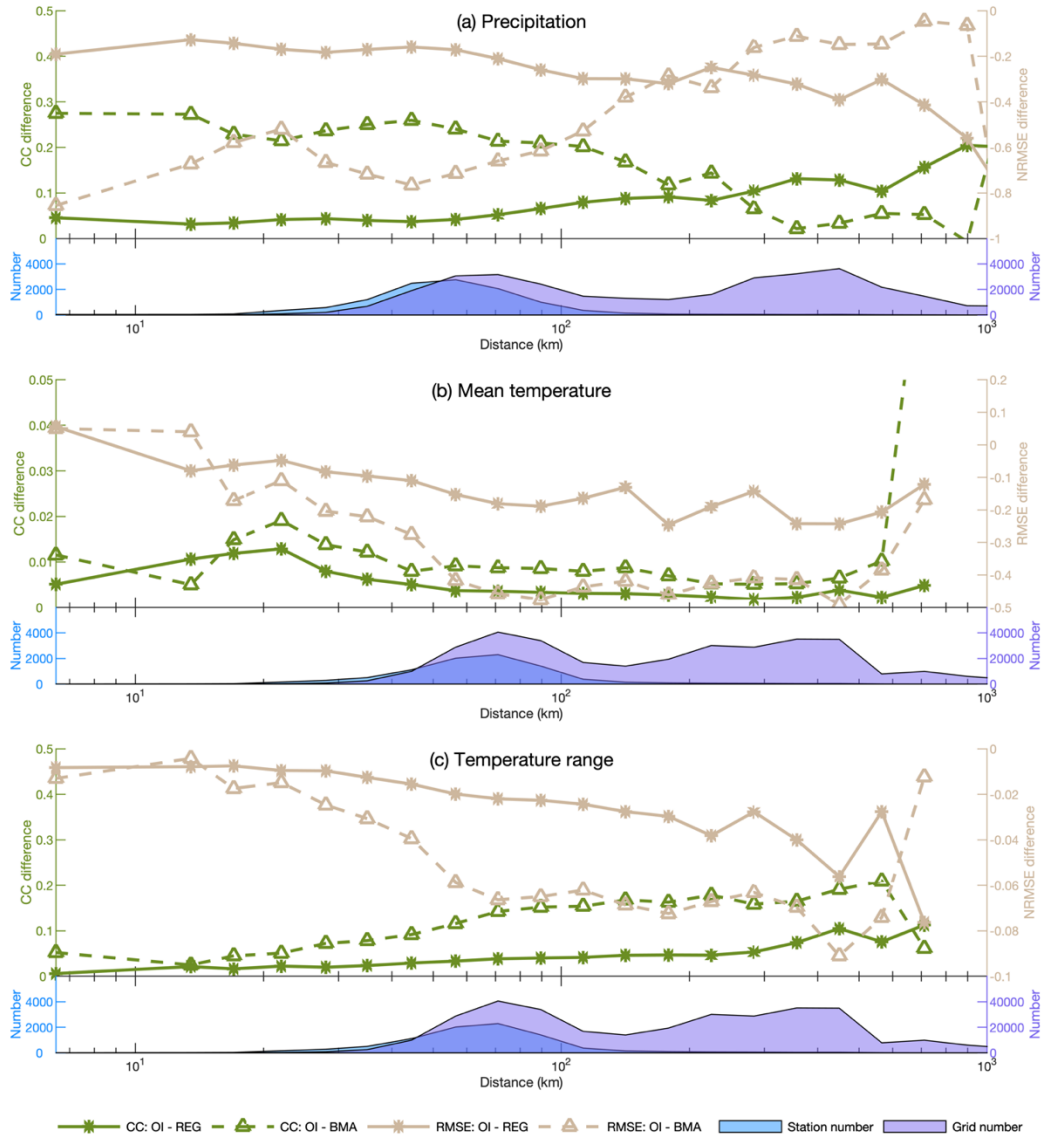



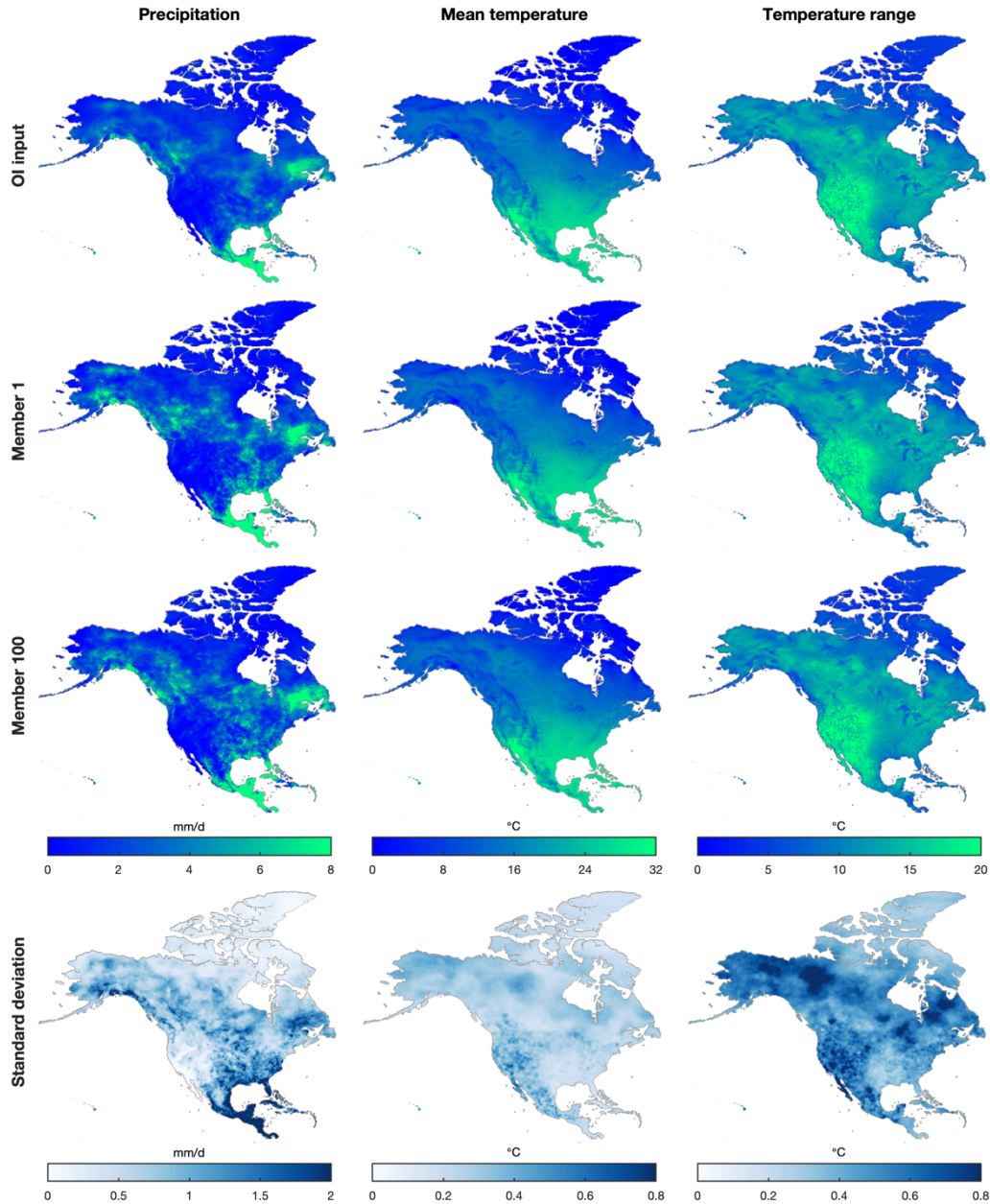
Figure 8. ~~The improvement of OI-based station-reanalysis merged estimates against station-based regression (REG) and BMA-merged reanalysis (BMA). The logarithmic X-axis shows the distance between the target station/grid and its 20th distant neighboring station. A larger distance represents a lower station density. The shaded area chart shows the numbers of stations and grid points corresponding to the same distance, which is the same for mean temperature and temperature range.~~

4.3 Evaluation of probabilistic estimates

The distributions of the OI and ensemble precipitation, Tmean, and Trange estimates in June 2016 are shown in Fig. 8~~9~~. Compared with OI precipitation estimates, ensemble precipitation estimates show generally consistent but less smooth distributions because of the relatively short spatial correlation length in the warm season. For Tmean and

517 Trange, OI and ensemble estimates show very similar spatial distributions. Precipitation shows the largest standard
518 deviation, while Tmean shows the smallest, because the standard deviation is determined by the errors of OI estimates.

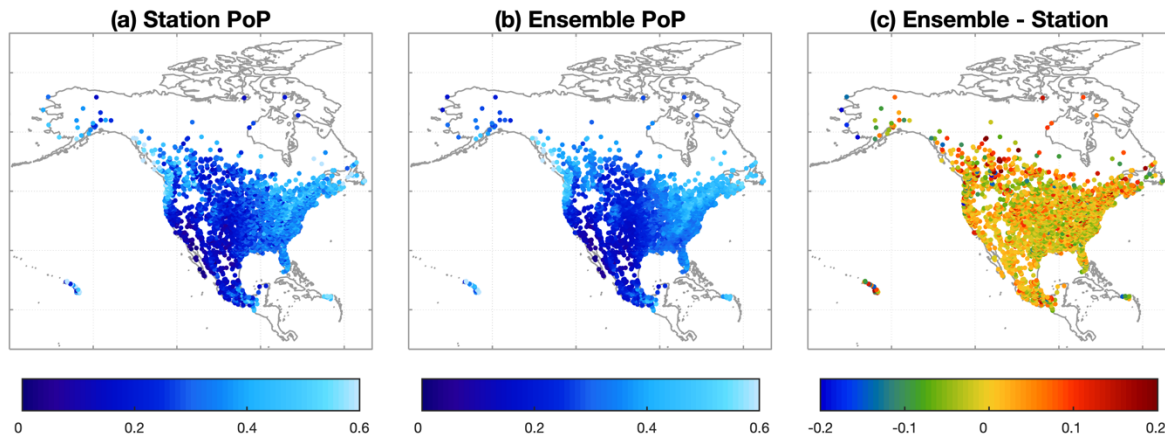
519 The PoP from station observations and ensemble estimates is compared based on stations with at least 5-year-long
520 records from 1979 to 2018 (Fig. 9̶10). The comparison cannot represent climatological PoP (Newman et al., 2019b)
521 due to short time length of independent stations (Sect. 3.5). Overall, EMDNA estimates show similar PoP distributions
522 with station observations. The PoP in Canada is slightly overestimated because (1) the quality of EMDNA is lower in
523 regions with fewer stations and (2) point-scale station observations could underestimate the PoP at a larger scale (e.g.,
524 0.1° grids) as shown by Tang et al. (2018a).

525

526 Figure 89. The distributions of average values from precipitation (the first column), mean daily temperature (the

527 second column), and daily temperature range (the third column) averaged over the period 1-30 June 2016. The first to

528 third rows represent estimates from OI-merged inputs, ensemble member 1, and ensemble member 100. The fourth

529 row represents the standard deviation of all the 100 members for one month (June 2016).

530

531 Figure 910. The probability of precipitation (PoP) from (a) station observations and (b) concurrent EMDNA ensemble

532 estimates with their differences shown in (c). Stations with at least 5-year-long records from 1979 to 2018 are involved

533 in the comparison.

534 The discrimination diagram (Fig. 101) shows that ensemble precipitation assigns the highest occurrence frequency at

535 the lowest estimated probability for non-precipitation events, and the performance becomes better as the threshold

536 increases from 0 to 50 mm. For precipitation events, ensemble estimates show the highest frequency at the highest

537 estimated probability for the thresholds of 0, 10, and 25 mm, while as the threshold increases, the frequency curve

538 becomes skewed to the lower estimated probability. This problem is also seen in Clark and Slater (2006) and Newman

539 et al. (2015). Ensemble precipitation shows good reliability for all precipitation thresholds with the points located at

540 or close to the 1-1 line (Fig. 101). At low and high estimated probabilities of occurrence, ensemble precipitation shows

541 slight wet bias. The reliability performance does not show clear dependence with thresholds.

542

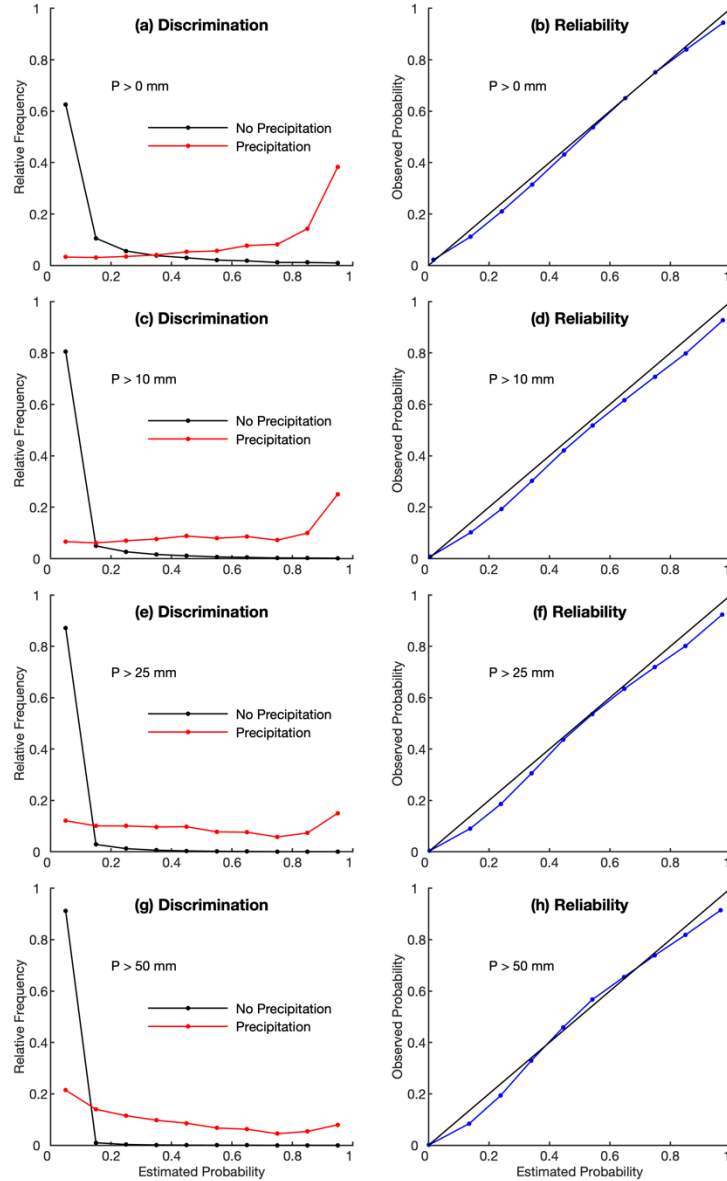



Figure 10̶1̶. The discrimination and reliability diagrams based on ensemble precipitation estimates. Four rain/no rain thresholds (0, 10, 25, 50 mm) are used.

The BSS for precipitation and CRPSS for Tmean and Trange are shown in Fig. 11̶2̶. In most cases, ensemble precipitation shows the highest frequency when BSS is above 0.5. As the precipitation threshold increases, the BSS values decrease. The median BSS values are 0.62, 0.54, and 0.46 for the thresholds of 0, 10, and 20 mm/d, respectively. We note that a small number of cases show BSS values smaller than zero, indicating that the ensemble estimated probability is worse than climatological probability. A low BSS value usually occurs in regions where precipitation is hard to estimate (e.g., Rocky Mountains) resulting in inaccurate parameters of Eq. (1).

551 The BSS for all thresholds shows a clear increasing trend from 1979 to 2018 (Fig. 1~~1~~2b) because the observed

552 precipitation samples from SCDNA increase during this period (Fig. 2 in Tang et al. (2020b)). The increasing trend

553 of BSS is particularly prominent from 2003 to 2009, during which precipitation samples in the USA experience the

554 greatest increase (Tang et al., 2020a). The results show that although infilled station data contribute to higher station

555 densities, observation samples still have a significant effect on gridded data estimation.

556 Tmean shows high CRPSS for most cases with the frequency peak occurring at ~0.8. The CRPSS of Trange is much

557 lower with the peak occurring at ~0.6. The median CRPSS for Tmean and Trange is 0.74 and 0.51, respectively.

558 Trange shows lower CRPSS probably because the bias direction (i.e., overestimation or underestimation) of daily

559 minimum and maximum temperature could be different, resulting in the larger bias of Trange than Tmean. Analyses

560 show that among stations with negative CRPSS, most are located in Mexico due to the degraded quality of temperature

561 estimates (Sect. 4.1 and 4.2). The long-term variation of CRPSS is not shown because independent temperature

562 stations are insufficient to support validation between 1986 and 2010.

563

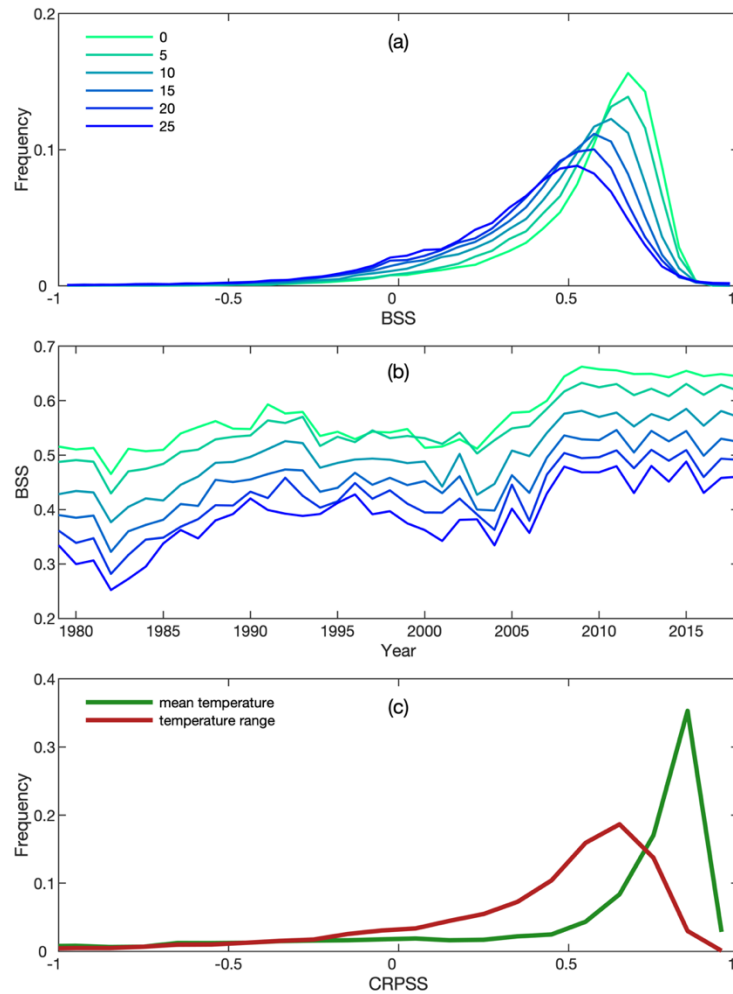



Figure 12. (a) The frequency distributions of the Brier Skill Score (BSS) for precipitation corresponding to rain/no rain thresholds from 0 to 25 mm/d. (b) The distributions of BSS for precipitation from 1979 to 2018. For each year, the median value of all stations is used. (c) The frequency distributions of the continuous ranked probability skill score (CRPSS) for daily mean temperature and daily temperature range.

5. Discussion

This study presents the framework for producing an ensemble precipitation and temperature dataset over North America. Although we have tested multiple choices of methods (Sect. 3) and overall the product shows good performance (Sect. 4), the methodology still has limitations that need to be improved through continued efforts.

5.1 Implementation of OI

OI is used to merge reanalysis outputs and station data. To implement OI-based merging, a critical step is to estimate the weights. Previous studies usually adopt empirical error or variogram functions and fit the parameters using station observations (e.g., CaPA (Fortin et al., 2015) and CMPA (Shen et al., 2018)); then the parameters are constant for the whole study area in the actual application.

In this study, we proposed a novel design, which uses station-based regression estimates as the observation fielld and calculates weights by directly solving the weight functions based on observation and background errors. Compared with methods that use station data as the observation field, our method is characterized by inferior estimation of the observation field but realistic estimation of weights. The close linkage between the observation field and the weights could benefit OI estimates but comparing different OI implementations is still meaningful and necessary considering that OI has been widely used and is the core algorithm of some operational products.

Furthermore, regression estimates show worse performance in regions with fewer stations. More advanced interpolation methods that can utilize climatology information and comprehensively consider topographic and atmospheric conditions (Daly et al., 2008; Newman et al., 2019b; Newman and Clark, 2020) should be examined in future studies.

5.2 Probabilistic estimation

Power transformations (e.g., Box-Cox and root/cubic square) with fixed parameters have proven to be useful in precipitation estimation and dataset production (Fortin et al., 2015, 2018; Cornes et al., 2018; Khedhaouiria et al., 2020; Newman et al., 2020). The Box-Cox transformation with a constant parameter is applied following Fortin et al. (2015) and Newman et al. (2019b, 2020). A fixed parameter, however, cannot ensure that transformed precipitation is normally distributed everywhere as is desirable.

We tested a series of additional parametric and non-parametric transformations based on power functions, logarithmic functions, or a mix of both, and optimized the parametric transformation functions (including Box-Cox) for every grid

595 by minimizing the objective function which is the sum of squared L-skewness and L-kurtosis (Papalexiou and

596 Koutsoyiannis, 2013). Theoretically, compared to a Box-Cox transformation with a fixed parameter, the optimized

597 functions can obtain precipitation series closer to the normal distribution which should benefit probabilistic estimation,

598 while the evaluation results show that the Box-Cox transformation with a fixed parameter is better at probabilistic

599 estimation than optimized functions. We suggest there are three reasons for this: (1) the standard deviation in Eq. (1)

600 is obtained by interpolating OI errors (Sect. 3.2.2) from neighboring stations, whereas the optimized transformation

601 parameters could be different at those stations, (2) zero precipitation is excluded during optimization to avoid invalid

602 transformation or optimization, which reduces the number of stations for every time step and thus degrades the quality

603 of the spatial interpolation, and (3) the errors caused by back transformation could be large if the optimized

604 transformation is too powerful. More efforts are needed to resolve this problem.

605 There are other potential directions for improvement. For example, SCRF is generated from Gaussian distributions,

606 while other choices such as copulas functions (Papalexiou and Serinaldi, 2020) show potential in probabilistic

607 estimation. The spatial correlation length is constant for the whole study area following Newman et al. (2015, 2019b),

608 which may introduce uncertainties for a large domain. Overall, studies related to the production of ensemble

609 meteorological datasets are still insufficient, particularly for large areas. More studies are needed to clarify the critical

610 issues in large-scale probabilistic estimation and explore the effect of parameter/method choices on probabilistic

611 estimates.

612 ## 5.3 Alternate data sources

613 The quality of source data (station observations and reanalysis models) primarily determines the quality of output

614 datasets. The density of stations has a critical effect on the accuracy of the observation field and probabilistic estimates.

615 While SCDNA collects data from multiple datasets, efforts are ongoing to expand the database by involving station

616 sources such as provincial station networks in Canada.

617 For reanalysis products, ERA5, MERRA-2, and JRA-55 are regridded using locally weighted linear regression to meet

618 the target resolution. There are some choices for future improvement, such as (1) adopting/developing better

619 downscaling methods or (2) utilizing outputs from high-resolution re-analysis products or forecasting models such as

620 ERA5-Land (Muñoz-Sabater et al., 2021) or the Arctic System Reanalysis (Bromwich et al., 2018)Weather Research

621 and Forecasting (WRF) model. For the latter one, a comprehensive assessment of available products is necessary

622 before substituting the three reanalysis products used by EMDNA. Moreover, including other data sources such as

623 satellite (e.g. (Huffman et al., 2015)GPM IMERG) and weather radar estimates is also an opportunity for regions with

624 adequate sample coverage.

625 ## 5.4 Precipitation under-catch

626 Although station precipitation observations are used as the reference in this study, these values are subject to

627 measurement errors such as wetting loss, wind-induced under-catch, and trace precipitation. Station temperature

628 measurements also contain errors due to microclimate and sensor design(Lin and Hubbard, 2004), which is generally

629 small and not discussed here. Under-catch of precipitation is particularly severe in high latitudes and mountains due
630 to the stronger wind and frequent snowfall (Sevruk, 1984; Goodison et al., 1998; Nešpor and Sevruk, 1999; Yang et
631 al., 2005; Scaff et al., 2015; Kochendorfer et al., 2018). For example, underestimation of precipitation could be larger
632 than 100% in Alaska (Yang et al., 1998). Bias correction of station precipitation data should consider many factors
633 such as gauge types, precipitation phase, and environmental conditions, which would be very complicated when a
634 large number of sparsely distributed stations are involved over the whole of North America.

635 The under-catch correction used in this study relies on bias-corrected precipitation climatology produced by Beck et
636 al. (2020), which infers the long-term precipitation using a Budyko curve and streamflow observations. The bias-
637 corrected precipitation climatology, however, is less accurate in northern Canada where streamflow stations are few
638 (Beck et al., 2020). TIn addition, the streamflow data used by the bias-corrected climatology also contain uncertainties
639 (Hamilton and Moore, 2012; Kiang et al., 2018) related to factors such as streamflow derivation methods (e.g., rate
640 curves) and measurement instruments. In addition, this correction method aims to constrain the total precipitation
641 amount and cannot distinguish between rainfall and snowfall which show different gauge catch performance. Data
642 users can realize rain-snow classification using approaches such as temperature threshold-based methods and
643 reanalysis model-based snowfall proportion. Moreover, as mentioned earlier, the water balance estimates of
644 precipitation under-catch do not consider non-contributing areas of river basins (e.g., endorheic sub-catchments),
645 which are common in the Canadian Prairies and the northern Great Plains in the USA. Whilst various under-catch
646 correction methods (e.g., Fuchs et al., 2001; Beck et al., 2020; Newman et al., 2020) exist, further studies are needed
647 to compare these solutions considering their effectiveness and availability of input data in a large domain.

6. Data availability

649 The EMDNA dataset is available at https://doi.org/10.20383/101.0275 (Tang et al., 2020a) in netCDF format.
650 Individual ensemble member, ensemble mean, and ensemble spread of precipitation, Tmean, and Trange are provided.
651 The total data size is 3.35 TB. Since the 100 members are equally plausible, users can download fewer members if the
652 storage space and processing time are limited.

653 The deterministic OI estimates of precipitation, PoP, Tmean, and Trange produced in this study are also available in
654 netCDF format. The high-quality OI estimates merge reanalysis and station data, which can be useful to applications
655 that do not need ensemble forcings. The data sizes are 3.35 TB for the probabilistic part and 40.84 GB for the
656 deterministic part, respectively.

657 The ensemble mean of the 100 members for Tmean and Trange is similar to deterministic OI estimates. For
658 precipitation, the ensemble mean is slightly higher than deterministic OI estimates due to the back transformation. We
659 recommend that users select the deterministic dataset instead of the ensemble mean if their applications do not involve
660 uncertainty characterization.

661 The total data size is 3.35 TB. The total data size is 40.84 GB.

662   A teaser dataset of probabilistic estimates is provided to facilitate easy preview of EMDNA without downloading the
663   entire dataset. The teaser dataset covers the region from -116.8° to -115.2°W, and 50.7° to 51.9°N, the time from 2014
664   to 2015, and the ensemble members from 1 to 25. The total data size is smaller than 30 MB. See Appendix E for a
665   brief introduction.

7. Summary and Conclusions

667   Ensemble meteorological datasets are of great value to hydrological and meteorological studies. Given the lack of a
668   historical ensemble dataset for the entire North America, this study develops EMDNA by integrating multi-source
669   information to overcome the limitation of sparse stations in high-latitude regions. EMDNA contains precipitation,
670   Tmean, and Trange estimates at 0.1° spatial resolution and daily temporal resolution from 1979 to 2018 with 100
671   members. Multiple methodological choices are examined when determining critical steps in the production of
672   EMDNA. The ultimate framework composes of four main steps: (1) generating station-based interpolation estimates
673   from SCDNA using locally weighted linear/logistic regression, (2) regridding, correction, and merging of reanalysis
674   products (ERA5, MERRA-2, and JRA-55), (3) merging station-reanalysis estimates using OI based on a novel method
675   of OI weight calculation, and correcting precipitation under-catch using the PBCOR dataset , and (4) generating
676   ensemble estimates by sampling from the estimated probability distributions with the perturbations provided by SCRF.

677   The performance of each step is comprehensively evaluated using multiple methods. The results show that the design
678   of the framework is effective. In short, we find that (1) station-based interpolation estimates are less accurate in regions
679   with sparse stations (e.g., high latitudes) and complex terrain; (2) BMA-merged reanalysis estimates show notable
680   improvement against raw reanalysis estimates, particularly for precipitation and Trange and over high-latitude regions;
681   (3) OI achieves more accurate estimates than interpolation and reanalysis estimates from (1) and (2), respectively, and
682   the complementary effect between reanalysis and interpolation estimates contributes to the large improvement of OI
683   estimates in sparsely gauged regions; and (4) ensemble precipitation estimates show good discrimination and
684   reliability performance for all thresholds, and the BSS values for ensemble precipitation and CRPSS values for
685   ensemble Tmean and Trange are high in most cases. BSS values of ensemble precipitation increase from 1979 to 2018
686   due to the increase of the number of stations.

687   Overall, EMDNA (version 1) will be useful for many applications in North America such as regional or continental
688   hydrological modeling. Meanwhile, we recognize that the current framework is not perfect and have provided
689   suggestions on the future directions for large-scale ensemble estimation of meteorological variables. (Sampson et al.,
690   2020)Continuing efforts from the community are needed to promote the development of probabilistic estimation
691   methods and datasets. Development of datasets at higher resolutions (e.g., 1 km and hourly) is also an important
692   direction to enable more sophisticated hydrometeorological studies (e.g., Sampson et al., 2020).

693

694   **Author contributions:** GT and MC designed the framework of this study. GT collected data, performed the analyses

695   and wrote the paper. MC, SP, AN and AW contributed to the design of the methodology and result evaluation. SP,

696   DB and PW contributed to the evaluation of methodology and results. All authors contributed to data analysis,

697   discussions about the methods and results, and paper improvement.

698   **Competing interests:** The authors declare that they have no conflict of interest.

699   **Acknowledgment:** The study is funded by the Global Water Futures (GWF) program in Canada. The authors

700   appreciate the extensive efforts from the developers of the ground and reanalysis datasets to make their products

701   available. The authors also thank Federated Research Data Repository (FRDR; https://www.frdr-dfdr.ca; Access Date:

702   September 29, 2020) for publishing our dataset as open access to users.

703

Appendix A. Regression coefficients

705   The coefficients for locally weight linear regression are estimated using weighted least square. Given a station i with

706   m neighboring stations, let $\mathbf{A} = [1, A_1, \ldots, A_n]$ be the $m \times n + 1$ attribute matrix, let $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ be the station

707   observations from neighboring stations, and let $\mathbf{w}_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,m})$ be the weight vector with distance-based

708   weights computed from Eq. (5). The regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_n)$ for Eq. (4) are estimated from the

709   weighted normal equation as

$$\boldsymbol{\beta} = (\mathbf{A}^\mathrm{T}\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^\mathrm{T}\mathbf{W}\mathbf{x}, \tag{A1}$$

710   where the $m \times m$ weight matrix $\mathbf{W} = \mathbf{I}_m\mathbf{w}_i$ is a diagonal matrix obtained by multiplying the $m \times m$ identity matrix

711    $\mathbf{I}_m$ with the weight vector $\mathbf{w}_i$.

712   The regression coefficients for logistic regression (Eq. 6) are estimated iteratively as:

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} + (\mathbf{A}^\mathrm{T}\mathbf{W}\mathbf{V}\mathbf{A})^{-1}\mathbf{A}^\mathrm{T}\mathbf{W}(\mathbf{P}_0 - \boldsymbol{\pi}) \tag{A2}$$

$$\boldsymbol{\pi} = \frac{1}{1 + \exp(-\mathbf{A}\boldsymbol{\beta}^{old})} \tag{A3}$$

$$\mathbf{V} = \mathbf{I}_m\boldsymbol{\pi}(1 - \boldsymbol{\pi}) \tag{A4}$$

713   where $\mathbf{P}_0$ is a vector of binary precipitation occurrence for neighboring stations, $\boldsymbol{\pi}$ is the vector of estimated PoP for

714   neighboring stations, and $\mathbf{V}$ is the diagonal variance matrix for PoP. The regression coefficients $\boldsymbol{\beta}^{old}$ are initialized as

715   a vector of ones.

716

Appendix B. Anomalous stations

718 To exclude climatologically anomalous stations, for temperature (Tmean or Trange), we calculate: (1) the absolute
719 difference of the climatological mean between the target station and the average value of its 10 neighboring stations
720 (referred as Diff-1), and (2) the absolute difference of the climatological mean between station observation and
721 regression estimates (referred as Diff-2). A temperature station will be excluded if its Diff-1 is larger than the 95%
722 percentile and its Diff-2 larger than the 99% percentile of all stations simultaneously. The threshold of percentiles for
723 Diff-1 is lower to better identify some climatologically anomalous stations.

724 For precipitation, the ratio (Ratio-1 and Ratio-2) is obtained in the same way with the Diff-1 and Diff-2 of temperature.
725 A two-tailed check is used for precipitation compared with the one-tailed check for temperature. A precipitation station
726 will be excluded if its Ratio-1 is larger (or smaller) than the 99.9% (1%) percentile and its Ratio-2 larger (or smaller)
727 than the 99.9% (1%) percentile simultaneously. This check has more tolerance for heavy precipitation but tries to
728 exclude more extremely dry stations.

729 As a result, ~1.5% precipitation and temperature stations are rejected, after which algorithms described in Sect. 3.1.1
730 and 3.1.2 are re-run. Stations can be anomalous because they are badly operated or simply because they are unique in
731 terms of topography or climate. The usage of Diff-2 or Ratio-2 is helpful to avoid excluding unique stations, but for
732 cases where the regression is ineffective, the unique stations can still be wrongly excluded. Although the effect on
733 final estimates could be rather small, better strategies could be used in future studies.

734

Appendix C. Error of BMA-merged reanalysis estimates

736 The errors of BMA-merged estimates are first estimated for all stations and then interpolated to grids. Considering
737 station observations cannot be used to evaluate merged estimates once they are used in bias correction or BMA weight
738 estimation, a two-layer cross-validation strategy is designed. In the first layer, we treat i as the target station and find
739 its m ($j_1 = 1, 2, …, m; i \notin j_1$) neighboring stations. In the second layer, we treat each j_1 as a target station, and (1)
740 find m ($j_2 = 1, 2, …, m; i \notin j_2$) neighboring stations for each j_1, (2) calculate linear scaling correction factors for all
741 j_2, (3) estimate the correction factor for the target j_1 by interpolating factors at all j_2 stations using inverse distance
742 weighting, (4) correct estimates at j_1 using the correction factor, (5) calculate BMA weights of three reanalysis
743 products for all j_1 stations, (6) interpolate BMA weights from all j_1 stations to the target station i and merge the three
744 reanalysis products for i, and (7) calculate the difference between merged reanalysis estimates and station observations
745 for i. This two-layer design may seem convoluted but is necessary to ensure that the error estimation is realistic. j_1
746 and j_2 could be partly overlapped due to their close locations but should not cause a large effect on the error estimation

747   for i because data for i are only used in (7) in this design. The station-based errors are interpolated to all grids using

748   inverse distance weighting.

Appendix D. Metrics for probabilistic evaluation

750   BSS is calculated based on the Brier Score (BS):

$$\text{BSS} = 1 - \frac{\text{BS}}{\text{BS}_{clim}} \qquad\qquad \text{D1}$$

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}(\text{PoP}_{ens} - \text{PoP}_{obs})^2 \qquad\qquad \text{D2}$$

751   where PoP_{ens} is the estimated probability of ensemble precipitation, PoP_{obs} is the observed binary precipitation

752   occurrence, n is the sample number, and BS_{clim} is the climatological BS by assigning the climatological probability

753   to all samples. When the two series match the value of BSS will be equal to one.

754   CRPSS is calculated based on the continuous ranked probability skill score (CRPS; Hersbach, 2000):

$$\text{CRPSS} = 1 - \frac{\text{CRPS}}{\text{CRPS}_{clim}} \qquad\qquad \text{D3}$$

$$\text{CRPS} = \int_{-\infty}^{\infty}(F(x) - H(x \geq x_o))^2 dx \qquad\qquad \text{D4}$$

755   where $F(x)$ is the CDF of the ensemble temperature estimate x, x_o is the observed temperature, $H(x \geq x_o)$ is the

756   Heaviside step function with the value being one if the condition $x \geq x_o$ is satisfied and zero if not satisfied, and

757   CRPS_{clim} is the climatological CPRS. CRPS measures the distance between the CDF of probabilistic estimates and

758   observations. For a perfect match, the value of CRPSS would be one.

Appendix E. Teaser dataset

760   The teaser dataset is a subset of EMDNA probabilistic estimates for a small region (-116.8° to -115.2°W, 50.7° to

761   51.9°N) and a short period (2014 to 2015) with only 25 ensemble members. Users can easily download and preview

762   the teaser dataset (<30 MB) before downloading the entire EMDNA dataset (~3 TB or ~40 GB) as shown in Sect. 6.

763   The region covers the Bow River basin above Banff, Canada, which is located in the Canadian Rockies (Figure A1).

764   The spread of ensemble members in this region could be large due to the complex topography and limited stations.

765

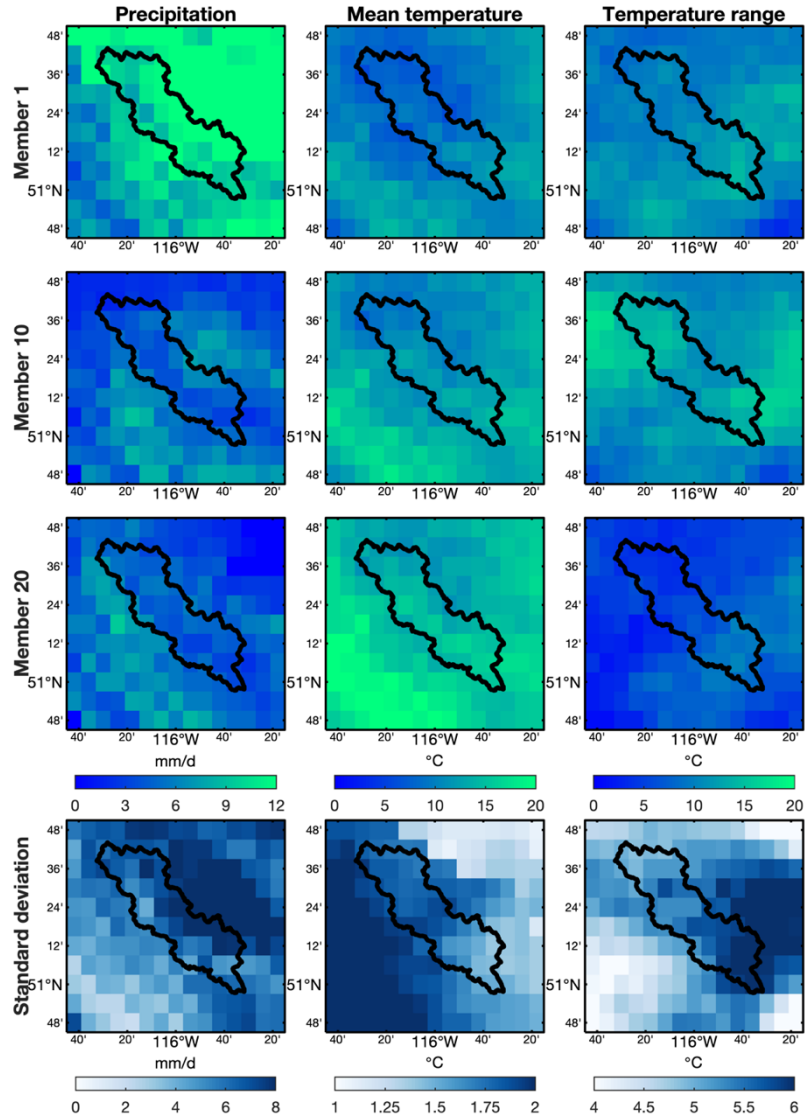



Figure A1. The distributions of daily precipitation (the first column), mean daily temperature (the second column), and daily temperature range (the third column) on 29 June 2015. The first to third rows represent ensemble members 1, 10, and 20, respectively. The fourth row represents the standard deviation of 25 members for this day. The black line outlines the Bow River basin above Banff, Canada.

770

References

772

Aalto, J., Pirinen, P., and Jylhä, K.: New gridded daily climatology of Finland: Permutation-based uncertainty estimates and temporal trends in climate, J. Geophys. Res. Atmospheres, 121, 3807–3823, https://doi.org/10.1002/2015JD024651, 2016.

Adler, R. F., Gu, G. J., Sapiano, M., Wang, J. J., and Huffman, G. J.: Global Precipitation: Means, Variations and Trends During the Satellite Era (1979-2014), Surv. Geophys., 38, 679–699, https://doi.org/10.1007/s10712-017-9416-4, 2017.

Arias-Hidalgo, M., Bhattacharya, B., Mynett, A. E., and van Griensven, A.: Experiences in using the TMPA-3B42R satellite data to complement rain gauge measurements in the Ecuadorian coastal foothills, Hydrol. Earth Syst. Sci., 17, 2905–2915, https://doi.org/10.5194/hess-17-2905-2013, 2013.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F.: Global-scale evaluation of 23 precipitation datasets using gauge observations and hydrological modeling, Hydrol. Earth Syst. Sci., 21, 6201–6217, https://doi.org/10.5194/hess-2017-508, 2017.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., van Dijk, A. I. J. M., McVicar, T. R., and Adler, R. F.: MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment, Bull. Am. Meteorol. Soc., 100, 473–500, https://doi.org/10.1175/BAMS-D-17-0138.1, 2019.

Beck, H. E., Wood, E. F., McVicar, T. R., Zambrano-Bigiarini, M., Alvarez-Garreton, C., Baez-Villanueva, O. M., Sheffield, J., and Karger, D. N.: Bias Correction of Global High-Resolution Precipitation Climatologies Using Streamflow Observations from 9372 Catchments, J. Clim., 33, 1299–1315, https://doi.org/10.1175/JCLI-D-19-0332.1, 2020.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Mon. Weather Rev., 78, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2, 1950.

Bromwich, D. H., Wilson, A. B., Bai, L., Liu, Z., Barlage, M., Shih, C.-F., Maldonado, S., Hines, K. M., Wang, S.-H., and Woollen, J.: The Arctic system reanalysis, version 2, Bull. Am. Meteorol. Soc., 99, 805–828, 2018.

Caillouet, L., Vidal, J.-P., Sauquet, E., Graff, B., and Soubeyroux, J.-M.: SCOPE Climate: a 142-year daily high-resolution ensemble meteorological reconstruction dataset over France, Earth Syst. Sci. Data, 11, 241–260, https://doi.org/10.5194/essd-11-241-2019, 2019.

Cannon, A. J., Sobie, S. R., and Murdock, T. Q.: Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes?, J. Clim., 28, 6938–6959, https://doi.org/10.1175/JCLI-D-14-00754.1, 2015.

Chen, Y., Yuan, W., Xia, J., Fisher, J. B., Dong, W., Zhang, X., Liang, S., Ye, A., Cai, W., and Feng, J.: Using Bayesian model averaging to estimate terrestrial evapotranspiration in China, J. Hydrol., 528, 537–549, https://doi.org/10.1016/j.jhydrol.2015.06.059, 2015.

Clark, M. P. and Hay, L. E.: Use of medium-range numerical weather prediction model output to produce forecasts of streamflow, J. Hydrometeorol., 5, 15–32, 2004.

Clark, M. P. and Slater, A. G.: Probabilistic Quantitative Precipitation Estimation in Complex Terrain, J. Hydrometeorol., 7, 3–22, https://doi.org/10.1175/JHM474.1, 2006.

Clark, M. P., Slater, A. G., Barrett, A. P., Hay, L. E., McCabe, G. J., Rajagopalan, B., and Leavesley, G. H.: Assimilation of snow covered area information into hydrologic and land-surface models, Adv. Water Resour., 29, 1209–1221, https://doi.org/10.1016/j.advwatres.2005.10.001, 2006.

Cornes, R. C., Schrier, G. van der, Besselaar, E. J. M. van den, and Jones, P. D.: An ensemble version of the E-OBS temperature and precipitation data sets, J. Geophys. Res. Atmospheres, 123, 9391–9409, https://doi.org/10.1029/2017JD028200, 2018.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., Curtis, J., and Pasteris, P. P.: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States, Int. J. Climatol., 28, 2031–2064, https://doi.org/10.1002/joc.1688, 2008.

Di Luzio, M., Johnson, G. L., Daly, C., Eischeid, J. K., and Arnold, J. G.: Constructing Retrospective Gridded Daily Precipitation and Temperature Datasets for the Conterminous United States, J. Appl. Meteorol. Climatol., 47, 475–497, https://doi.org/10.1175/2007JAMC1356.1, 2008.

Dinku, T., Anagnostou, E. N., and Borga, M.: Improving radar-based estimation of rainfall over complex terrain, J. Appl. Meteorol., 41, 1163–1178, 2002.

Donat, M. G., Sillmann, J., Wild, S., Alexander, L. V., Lippmann, T., and Zwiers, F. W.: Consistency of Temperature and Precipitation Extremes across Various Global Gridded In Situ and Reanalysis Datasets, J. Clim., 27, 5019–5035, https://doi.org/10.1175/JCLI-D-13-00405.1, 2014.

Duan, Q. and Phillips, T. J.: Bayesian estimation of local signal and noise in multimodel simulations of climate change, J. Geophys. Res. Atmospheres, 115, https://doi.org/10.1029/2009JD013654, 2010.

Duan, S.-B. and Li, Z.-L.: Spatial Downscaling of MODIS Land Surface Temperatures Using Geographically Weighted Regression: Case Study in Northern China, IEEE Trans. Geosci. Remote Sens., 54, 6458–6469, https://doi.org/10.1109/TGRS.2016.2585198, 2016.

Eischeid, J. K., Pasteris, P. A., Diaz, H. F., Plantico, M. S., and Lott, N. J.: Creating a Serially Complete, National Daily Time Series of Temperature and Precipitation for the Western United States, J. Appl. Meteorol., 39, 1580–1591, https://doi.org/10.1175/1520-0450(2000)039<1580:CASCND>2.0.CO;2, 2000.

Fick, S. E. and Hijmans, R. J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, Int. J. Climatol., 37, 4302–4315, 2017.

Folland, C. K.: Numerical models of the raingauge exposure problem, field experiments and an improved collector design, Q. J. R. Meteorol. Soc., 114, 1485–1516, https://doi.org/10.1002/qj.49711448407, 1988.

Fortin, V., Roy, G., Donaldson, N., and Mahidjiba, A.: Assimilation of radar quantitative precipitation estimations in the Canadian Precipitation Analysis (CaPA), J. Hydrol., 531, 296–307, https://doi.org/10.1016/j.jhydrol.2015.08.003, 2015.

Fortin, V., Roy, G., Stadnyk, T., Koenig, K., Gasset, N., and Mahidjiba, A.: Ten Years of Science Based on the Canadian Precipitation Analysis: A CaPA System Overview and Literature Review, Atmosphere-Ocean, 56, 178–196, https://doi.org/10.1080/07055900.2018.1474728, 2018.

Frei, C. and Isotta, F. A.: Ensemble Spatial Precipitation Analysis From Rain Gauge Data: Methodology and Application in the European Alps, J. Geophys. Res. Atmospheres, 124, 5757–5778, https://doi.org/10.1029/2018JD030004, 2019.

Fuchs, T., Rapp, J., Rubel, F., and Rudolf, B.: Correction of synoptic precipitation observations due to systematic measuring errors with special regard to precipitation phases, Phys. Chem. Earth Part B Hydrol. Oceans Atmosphere, 26, 689–693, 2001.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen, J.: The climate hazards infrared precipitation with stations--a new environmental record for monitoring extremes, Sci. Data, 2, 150066, https://doi.org/10.1038/sdata.2015.66, 2015.

853 Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich,
854 M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A.
855 M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W.,
856 Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era Retrospective Analysis for Research
857 and Applications, Version 2 (MERRA-2), J. Clim., 30, 5419–5454, https://doi.org/10.1175/jcli-d-16-0758.1, 2017.

858 Goodison, B. E., Louie, P. Y., and Yang, D.: WMO solid precipitation measurement intercomparison, 1998.

859 Habib, E., Haile, A. T., Sazib, N., Zhang, Y., and Rientjes, T.: Effect of Bias Correction of Satellite-Rainfall Estimates
860 on Runoff Simulations at the Source of the Upper Blue Nile, Remote Sens., 6, 6688–6708,
861 https://doi.org/10.3390/rs6076688, 2014.

862 Hamilton, A. S. and Moore, R. D.: Quantifying Uncertainty in Streamflow Records, Can. Water Resour. J. Rev. Can.
863 Ressour. Hydr., 37, 3–21, https://doi.org/10.4296/cwrj3701865, 2012.

864 Harris, I., Osborn, T. J., Jones, P., and Lister, D.: Version 4 of the CRU TS monthly high-resolution gridded
865 multivariate climate dataset, Sci. Data, 7, 109, https://doi.org/10.1038/s41597-020-0453-3, 2020.

866 Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M.: A European daily high-
867 resolution gridded data set of surface temperature and precipitation for 1950–2006, J. Geophys. Res. Atmospheres,
868 113, https://doi.org/10.1029/2008JD010201, 2008.

869 Hellinger, E.: Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen., J. Für Reine
870 Angew. Math. Crelles J., 1909, 210–271, 1909.

871 Hempel, S., Frieler, K., Warszawski, L., Schewe, J., and Piontek, F.: A trend-preserving bias correction – the
872 ISI-MIP approach, Earth Syst. Dyn., 4, 219–236, https://doi.org/10.5194/esd-4-219-2013, 2013.

873 Henn, B., Newman, A. J., Livneh, B., Daly, C., and Lundquist, J. D.: An assessment of differences in gridded
874 precipitation datasets in complex terrain, J. Hydrol., 556, 1205–1219, https://doi.org/10.1016/j.jhydrol.2017.03.008,
875 2018.

876 Herrnegger, M., Senoner, T., and Nachtnebel, H.-P.: Adjustment of spatio-temporal precipitation patterns in a high
877 Alpine environment, J. Hydrol., 556, 913–921, https://doi.org/10.1016/j.jhydrol.2016.04.068, 2018.

878 Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather
879 Forecast., 15, 559–570, 2000.

880 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R.,
881 Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J.,
882 Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes,
883 M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P.,
884 Lupu, C., Radnoti, G., Rosnay, P. de, Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global
885 reanalysis, Q. J. R. Meteorol. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

886 Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T.: Bayesian Model Averaging: A Tutorial, Stat. Sci.,
887 14, 382–401, 1999.

888 Hong, Y., Hsu, K., Moradkhani, H., and Sorooshian, S.: Uncertainty quantification of satellite precipitation estimation
889 and Monte Carlo assessment of the error propagation into hydrologic response, Water Resour. Res., 42,
890 https://doi.org/10.1029/2005wr004398, 2006.

891 Hu, Q., Li, Z., Wang, L., Huang, Y., Wang, Y., and Li, L.: Rainfall Spatial Estimations: A Review from Spatial
892 Interpolation to Multi-Source Data Merging, Water, 11, 579, https://doi.org/10.3390/w11030579, 2019.

893 Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., Hong, Y., Bowman, K. P., and Stocker,
894 E. F.: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor
895 Precipitation Estimates at Fine Scales, J. Hydrometeorol., 8, 38–55, https://doi.org/10.1175/jhm560.1, 2007.

896 Karger, D. N., Conrad, O., Bohner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P.,
897 and Kessler, M.: Climatologies at high resolution for the earth's land surface areas, Sci. Data, 4, 170122,
898 https://doi.org/10.1038/sdata.2017.122, 2017.

899 Kemp, W. P., Burnell, D. G., Everson, D. O., and Thomson, A. J.: Estimating Missing Daily Maximum and Minimum
900 Temperatures, J. Clim. Appl. Meteorol., 22, 1587–1593, https://doi.org/10.1175/1520-
901 0450(1983)022<1587:EMDMAM>2.0.CO;2, 1983.

902 Khedhaouiria, D., Bélair, S., Fortin, V., Roy, G., and Lespinas, F.: High Resolution (2.5km) Ensemble Precipitation
903 Analysis across Canada, J. Hydrometeorol., https://doi.org/10.1175/JHM-D-19-0282.1, 2020.

904 Kiang, J. E., Gazoorian, C., McMillan, H., Coxon, G., Coz, J. L., Westerberg, I. K., Belleville, A., Sevrez, D., Sikorska,
905 A. E., Petersen-Øverleir, A., Reitan, T., Freer, J., Renard, B., Mansanarez, V., and Mason, R.: A Comparison of
906 Methods for Streamflow Uncertainty Estimation, Water Resour. Res., 54, 7149–7176,
907 https://doi.org/10.1029/2018WR022708, 2018.

908 Kirstetter, P.-E., Gourley, J. J., Hong, Y., Zhang, J., Moazamigoodarzi, S., Langston, C., and Arthur, A.: Probabilistic
909 precipitation rate estimates with ground-based radar networks, Water Resour. Res., 51, 1422–1442,
910 https://doi.org/10.1002/2014WR015672, 2015.

911 Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo,
912 H., Miyaoka, K., and Takahashi, K.: The JRA-55 Reanalysis: General Specifications and Basic Characteristics, J.
913 Meteorol. Soc. Jpn. Ser II, 93, 5–48, https://doi.org/10.2151/jmsj.2015-001, 2015.

914 Kochendorfer, J., Nitu, R., Wolff, M., Mekis, E., Rasmussen, R., Baker, B., Earle, M. E., Reverdin, A., Wong, K.,
915 Smith, C. D., Yang, D., Roulet, Y.-A., Meyers, T., Buisan, S., Isaksen, K., Brækkan, R., Landolt, S., and Jachcik, A.:
916 Testing and development of transfer functions for weighing precipitation gauges in WMO-SPICE, Hydrol. Earth Syst.
917 Sci., 22, 1437–1452, https://doi.org/10.5194/hess-22-1437-2018, 2018.

918 Lader, R., Bhatt, U. S., Walsh, J. E., Rupp, T. S., and Bieniek, P. A.: Two-Meter Temperature and Precipitation from
919 Atmospheric Reanalysis Evaluated for Alaska, J. Appl. Meteorol. Climatol., 55, 901–922,
920 https://doi.org/10.1175/JAMC-D-15-0162.1, 2016.

921 Livneh, B., Rosenberg, E. A., Lin, C., Nijssen, B., Mishra, V., Andreadis, K. M., Maurer, E. P., and Lettenmaier, D.
922 P.: A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States:
923 Update and Extensions, J. Clim., 26, 9384–9392, https://doi.org/10.1175/JCLI-D-12-00508.1, 2013.

924 Longman, R. J., Newman, A. J., Giambelluca, T. W., and Lucas, M.: Characterizing the Uncertainty and Assessing
925 the Value of Gap-Filled Daily Rainfall Data in Hawaii, J. Appl. Meteorol. Climatol., 59, 1261–1276,
926 https://doi.org/10.1175/JAMC-D-20-0007.1, 2020.

927 Lu, X., Tang, G., Wang, X., Liu, Y., Wei, M., and Zhang, Y.: The Development of a Two-Step Merging and
928 Downscaling Method for Satellite Precipitation Products, Remote Sens., 12, 398, 2020.

929 Ma, Y., Yang, Y., Han, Z., Tang, G., Maguire, L., Chu, Z., and Hong, Y.: Comprehensive evaluation of Ensemble
930 Multi-Satellite Precipitation Dataset using the Dynamic Bayesian Model Averaging scheme over the Tibetan plateau,
931 J. Hydrol., 556, 634–644, https://doi.org/10.1016/j.jhydrol.2017.11.050, 2018a.

932 Ma, Y., Hong, Y., Chen, Y., Yang, Y., Tang, G., Yao, Y., Long, D., Li, C., Han, Z., and Liu, R.: Performance of
933 Optimally Merged Multisatellite Precipitation Products Using the Dynamic Bayesian Model Averaging Scheme Over
934 the Tibetan Plateau, J. Geophys. Res. Atmospheres, 123, 814–834, https://doi.org/10.1002/2017jd026648, 2018b.

935 Ma, Z., Xu, J., Zhu, S., Yang, J., Tang, G., Yang, Y., Shi, Z., and Hong, Y.: AIMERG: a new Asian precipitation
936 dataset (0.1°/half-hourly, 2000–2015) by calibrating the GPM-era IMERG at a daily scale using APHRODITE, Earth
937 Syst. Sci. Data, 12, 1525–1544, https://doi.org/10.5194/essd-12-1525-2020, 2020.

938 Mahfouf, J.-F., Brasnett, B., and Gagnon, S.: A Canadian precipitation analysis (CaPA) project: Description and
939 preliminary results, Atmosphere-Ocean, 45, 1–17, https://doi.org/10.3137/ao.v450101, 2007.

940 Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., and Nijssen, B.: A Long-Term Hydrologically Based
941 Dataset of Land Surface Fluxes and States for the Conterminous United States, J. Clim., 15, 15, 2002.

942 Mears, C. A., Wentz, F. J., Thorne, P., and Bernie, D.: Assessing uncertainty in estimates of atmospheric temperature
943 changes from MSU and AMSU using a Monte-Carlo estimation technique, J. Geophys. Res. Atmospheres, 116,
944 https://doi.org/10.1029/2010JD014954, 2011.

945 Mendoza, P. A., Wood, A. W., Clark, E., Rothwell, E., Clark, M. P., Nijssen, B., Brekke, L. D., and Arnold, J. R.: An
946 intercomparison of approaches for improving operational seasonal streamflow forecasts, Hydrol. Earth Syst. Sci., 21,
947 3915–3935, 2017.

948 Mooney, P. A., Mulligan, F. J., and Fealy, R.: Comparison of ERA-40, ERA-Interim and NCEP/NCAR reanalysis
949 data with observed surface air temperatures over Ireland, Int. J. Climatol., 31, 545–557,
950 https://doi.org/10.1002/joc.2098, 2011.

951 Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D.: Quantifying uncertainties in global and regional
952 temperature change using an ensemble of observational estimates: The HadCRUT4 data set, J. Geophys. Res.
953 Atmospheres, 117, https://doi.org/10.1029/2011JD017187, 2012.

954 Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga,
955 M., Harrigan, S., and Hersbach, H.: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications,
956 Earth Syst. Sci. Data Discuss., 1–50, 2021.

957 Nešpor, V. and Sevruk, B.: Estimation of Wind-Induced Error of Rainfall Gauge Measurements Using a Numerical
958 Simulation, J. Atmospheric Ocean. Technol., 16, 450–464, https://doi.org/10.1175/1520-
959 0426(1999)016<0450:EOWIEO>2.0.CO;2, 1999.

960 Newman, A. J. and Clark, M. P.: TIER version 1.0: an open-source Topographically InformEd Regression (TIER)
961 model to estimate spatial meteorological fields, Geosci. Model Dev., 13, 1827–1843, https://doi.org/10.5194/gmd-13-
962 1827-2020, 2020.

963 Newman, A. J., Clark, M. P., Craig, J., Nijssen, B., Wood, A., Gutmann, E., Mizukami, N., Brekke, L., and Arnold,
964 J. R.: Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous United States, J. Hydrometeorol.,
965 16, 2481–2500, https://doi.org/10.1175/JHM-D-15-0026.1, 2015.

966 Newman, A. J., Clark, M. P., Longman, R. J., and Giambelluca, T. W.: Methodological Intercomparisons of Station-
967 Based Gridded Meteorological Products: Utility, Limitations, and Paths Forward, J. Hydrometeorol., 20, 531–547,
968 https://doi.org/10.1175/JHM-D-18-0114.1, 2019a.

969 Newman, A. J., Clark, M. P., Longman, R. J., Gilleland, E., Giambelluca, T. W., and Arnold, J. R.: Use of Daily
970 Station Observations to Produce High-Resolution Gridded Probabilistic Precipitation and Temperature Time Series
971 for the Hawaiian Islands, J. Hydrometeorol., 20, 509–529, https://doi.org/10.1175/JHM-D-18-0113.1, 2019b.

972 Newman, A. J., Clark, M. P., Wood, A. W., and Arnold, J. R.: Probabilistic Spatial Meteorological Estimates for
973 Alaska and the Yukon, J. Geophys. Res. Atmospheres, 2020.

974 Papalexiou, S. M.: Unified theory for stochastic modelling of hydroclimatic processes: Preserving marginal
975 distributions, correlation structures, and intermittency, Adv. Water Resour., 115, 234–252, 2018.

976 Papalexiou, S. M. and Koutsoyiannis, D.: Battle of extreme value distributions: A global survey on extreme daily
977 rainfall, Water Resour. Res., 49, 187–201, https://doi.org/10.1029/2012WR012557, 2013.

978 Papalexiou, S. M. and Serinaldi, F.: Random Fields Simplified: Preserving Marginal Distributions, Correlations, and
979 Intermittency, With Applications From Rainfall to Humidity, Water Resour. Res., 56, e2019WR026331,
980 https://doi.org/10.1029/2019WR026331, 2020.

981 Parker, W. S.: Reanalyses and Observations: What's the Difference?, Bull. Am. Meteorol. Soc., 97, 1565–1572,
982 https://doi.org/10.1175/BAMS-D-14-00226.1, 2016.

983 Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to Calibrate
984 Forecast Ensembles, Mon. Weather Rev., 133, 1155–1174, https://doi.org/10.1175/MWR2906.1, 2005.

985 Rodell, M., Beaudoing, H. K., L'Ecuyer, T. S., Olson, W. S., Famiglietti, J. S., Houser, P. R., Adler, R., Bosilovich,
986 M. G., Clayson, C. A., Chambers, D., Clark, E., Fetzer, E. J., Gao, X., Gu, G., Hilburn, K., Huffman, G. J., Lettenmaier,
987 D. P., Liu, W. T., Robertson, F. R., Schlosser, C. A., Sheffield, J., and Wood, E. F.: The Observed State of the Water
988 Cycle in the Early Twenty-First Century, J. Clim., 28, 8289–8318, https://doi.org/10.1175/JCLI-D-14-00555.1, 2015.

989 Sampson, A. A., Wright, D. B., Stewart, R. D., and LoBue, A. C.: The role of rainfall temporal and spatial averaging
990 in seasonal simulations of the terrestrial water balance, Hydrol. Process., 34, 2531–2542,
991 https://doi.org/10.1002/hyp.13745, 2020.

992 Scaff, L., Yang, D., Li, Y., and Mekis, E.: Inconsistency in precipitation measurements across the Alaska–Yukon
993 border, The Cryosphere, 9, 2417–2428, https://doi.org/10.5194/tc-9-2417-2015, 2015.

994 Schepen, A. and Wang, Q. J.: Model averaging methods to merge operational statistical and dynamic seasonal
995 streamflow forecasts in Australia, Water Resour. Res., 51, 1797–1812, https://doi.org/10.1002/2014WR016163, 2015.

996 Sevruk, B.: International comparison of national precipitation gauges with a reference pit gauge., WMO Instrum. Obs.
997 Methods Rep. No 17, 111, 1984.

998 Shen, Y., Zhao, P., Pan, Y., and Yu, J. J.: A high spatiotemporal gauge-satellite merged precipitation analysis over
999 China, J. Geophys. Res.-Atmospheres, 119, 3063–3075, https://doi.org/10.1002/2013jd020686, 2014a.

1000 Shen, Y., Zhao, P., Pan, Y., and Yu, J.: A high spatiotemporal gauge-satellite merged precipitation analysis over China,
1001 J. Geophys. Res. Atmospheres, 119, 3063–3075, https://doi.org/10.1002/2013JD020686, 2014b.

1002 Shen, Y., Hong, Z., Pan, Y., Yu, J., and Maguire, L.: China's 1 km Merged Gauge, Radar and Satellite Experimental
1003 Precipitation Dataset, Remote Sens., 10, 264, https://doi.org/10.3390/rs10020264, 2018.

1004 Sinclair, S. and Pegram, G.: Combining radar and rain gauge rainfall estimates using conditional merging,
1005 Atmospheric Sci. Lett., 6, 19–22, https://doi.org/10.1002/asl.85, 2005.

1006 Slater, A. G. and Clark, M. P.: Snow Data Assimilation via an Ensemble Kalman Filter, J. Hydrometeorol., 7, 478–
1007 493, https://doi.org/10.1175/JHM505.1, 2006.

1008 Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K.-L.: A Review of Global Precipitation Data Sets:
1009 Data Sources, Estimation, and Intercomparisons, Rev. Geophys., https://doi.org/10.1002/2017rg000574, 2018.

1010 Tang, G., Zeng, Z., Long, D., Guo, X., Yong, B., Zhang, W., and Hong, Y.: Statistical and Hydrological Comparisons
1011 between TRMM and GPM Level-3 Products over a Midlatitude Basin: Is Day-1 IMERG a Good Successor for TMPA
1012 3B42V7?, J. Hydrometeorol., 17, 121–137, https://doi.org/10.1175/jhm-d-15-0059.1, 2016.

1013   Tang, G., Behrangi, A., Long, D., Li, C., and Hong, Y.: Accounting for spatiotemporal errors of gauges: A critical
1014   step to evaluate gridded precipitation products, J. Hydrol., 559, 294–306,
1015   https://doi.org/10.1016/j.jhydrol.2018.02.057, 2018a.

1016   Tang, G., Behrangi, A., Ma, Z., Long, D., and Hong, Y.: Downscaling of ERA-Interim Temperature in the Contiguous
1017   United States and Its Implications for Rain–Snow Partitioning, J. Hydrometeorol., 19, 1215–1233,
1018   https://doi.org/10.1175/jhm-d-18-0041.1, 2018b.

1019   Tang, G., Clark, M. P., Papalexiou, S. M., Newman, A. J., Wood, A. W., Brunet, D., and Whitfield, P. H.: EMDNA:
1020   Ensemble Meteorological Dataset for North America [Dataset], FRDR, https://doi.org/10.20383/101.0275, 2020a.

1021   Tang, G., Clark, M. P., Papalexiou, S. M., Ma, Z., and Hong, Y.: Have satellite precipitation products improved over
1022   last two decades? A comprehensive comparison of GPM IMERG with nine satellite and reanalysis datasets, Remote
1023   Sens. Environ., 240, 111697, https://doi.org/10.1016/j.rse.2020.111697, 2020b.

1024   Tang, G., Clark, M. P., Newman, A. J., Wood, A. W., Papalexiou, S. M., Vionnet, V., and Whitfield, P. H.: SCDNA:
1025   a serially complete precipitation and temperature dataset for North America from 1979 to 2018, Earth Syst. Sci. Data,
1026   12, 2381–2409, https://doi.org/10.5194/essd-12-2381-2020, 2020c.

1027   Tang, G., Clark, M. P., and Papalexiou, S. M.: The use of serially complete station data to improve the temporal
1028   continuity of gridded precipitation and temperature estimates, J. Hydrometeorol., 2021.

1029   Teutschbein, C. and Seibert, J.: Bias correction of regional climate model simulations for hydrological climate-change
1030   impact studies: Review and evaluation of different methods, J. Hydrol., 456–457, 12–29,
1031   https://doi.org/10.1016/j.jhydrol.2012.05.052, 2012.

1032   Trenberth, K. E., Dai, A., Rasmussen, R. M., and Parsons, D. B.: The Changing Character of Precipitation, Bull. Am.
1033   Meteorol. Soc., 84, 1205–1218, https://doi.org/10.1175/BAMS-84-9-1205, 2003.

1034   Vila, D. A., de Goncalves, L. G. G., Toll, D. L., and Rozante, J. R.: Statistical Evaluation of Combined Daily Gauge
1035   Observations and Rainfall Satellite Estimates over Continental South America, J. Hydrometeorol., 10, 533–543,
1036   https://doi.org/10.1175/2008JHM1048.1, 2009.

1037   Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P.: The WFDEI meteorological forcing
1038   data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resour. Res., 50, 7505–
1039   7514, https://doi.org/10.1002/2014wr015638, 2014.

1040   Willkofer, F., Schmid, F.-J., Komischke, H., Korck, J., Braun, M., and Ludwig, R.: The impact of bias correcting
1041   regional climate model results on hydrological indicators for Bavarian catchments, J. Hydrol. Reg. Stud., 19, 25–41,
1042   https://doi.org/10.1016/j.ejrh.2018.06.010, 2018.

1043   Wood, A. W., Leung, L. R., Sridhar, V., and Lettenmaier, D. P.: Hydrologic Implications of Dynamical and Statistical
1044   Approaches to Downscaling Climate Model Outputs, Clim. Change, 62, 189–216,
1045   https://doi.org/10.1023/B:CLIM.0000013685.99609.9e, 2004.

1046   Wu, H., Adler, R. F., Tian, Y., Huffman, G. J., Li, H., and Wang, J.: Real-time global flood estimation using satellite-
1047   based precipitation and a coupled land surface and routing model, Water Resour. Res., 50, 2693–2717,
1048   https://doi.org/10.1002/2013wr014710, 2014.

1049   Xie, P. and Xiong, A.-Y.: A conceptual model for constructing high-resolution gauge-satellite merged precipitation
1050   analyses, J. Geophys. Res. Atmospheres, 116, https://doi.org/10.1029/2011JD016118, 2011.

1051   Xu, S., Wu, C., Wang, L., Gonsamo, A., Shen, Y., and Niu, Z.: A new satellite-based monthly precipitation
1052   downscaling algorithm with non-stationary relationship between precipitation and land surface characteristics, Remote
1053   Sens. Environ., 162, 119–140, https://doi.org/10.1016/j.rse.2015.02.024, 2015.

1054 Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., and
1055 Bates, P. D.: A high-accuracy map of global terrain elevations, Geophys. Res. Lett., 44, 5844–5853,
1056 https://doi.org/10.1002/2017GL072874, 2017.

1057 Yang, D., Goodison, B. E., Ishida, S., and Benson, C. S.: Adjustment of daily precipitation data at 10 climate stations
1058 in Alaska: Application of World Meteorological Organization intercomparison results, Water Resour. Res., 34, 241–
1059 256, https://doi.org/10.1029/97WR02681, 1998.

1060 Yang, D., Kane, D., Zhang, Z., Legates, D., and Goodison, B.: Bias corrections of long-term (1973-2004) daily
1061 precipitation data over the northern regions, Geophys. Res. Lett., 32, n/a-n/a, https://doi.org/10.1029/2005gl024057,
1062 2005.

1063 Yin, J., Gentine, P., Zhou, S., Sullivan, S. C., Wang, R., Zhang, Y., and Guo, S.: Large increase in global storm runoff
1064 extremes driven by climate and anthropogenic changes, Nat. Commun., 9, 4389, https://doi.org/10.1038/s41467-018-
1065 06765-2, 2018.

1066 Zhang, Y., Ren, Y., Ren, G., and Wang, G.: Bias Correction of Gauge Data and its Effect on Precipitation Climatology
1067 over Mainland China, J. Appl. Meteorol. Climatol., 58, 2177–2196, https://doi.org/10.1175/JAMC-D-19-0049.1, 2019.

1068