Response to comments

The authors thank the reviewers for their constructive comments, which provide the basis to improve the quality of the manuscript and dataset. We address all points in detail and reply to all comments here below. We also updated SCDNA from V1 to V1.1 on Zenodo based on the reviewer's comments. The modifications include adding station source flag, adding original files for location merged stations, and adding a quality control procedure based on the final SCDNA. SCDNA estimates are generally consistent between the two versions, with the total number of stations reduced from 27280 to 27276.

Reviewer 1

General comment

The manuscript presents and advertises a very interesting dataset of temperature and precipitation observation collected over several years in North America. The work is certainly well suited for the readership of ESSD and it is overall very important for the meteorological and climatological community. Furthermore, creation of quality controlled databases is an important contribution to the scientific community in the age of data science. I have a few points to consider before publication, which I recommend, listed below.

1. Measurement instruments: from my background, I am much closer to the instruments themselves (and their peculiarities and issues), as hardware tools. What I missed here was a description of the stations and their instruments. Questions like: which are the instruments deployed in the stations? How is precipitation measured (tipping buckets? buckets? Weighing gauges? Note for example that some instruments may have biases when measuring snowfall while others may not)? How is it temperature measured? How is this different from station to station in your database?

Response: We have added the descriptions of measurement instruments in both the manuscript and dataset documentation. Since a complete introduction to the specifications and the evolution of measurement instruments in North America is not trivial, we only provide a general introduction here, and guide readers to the official sources for more comprehensive knowledge (such as design purpose, instrument structure, accuracy for rain/snow, inter-instrument comparison) in the manuscript and dataset page. As station hardware varies among countries, we successively introduce the overall situations in Canada, U.S., and Mexico as below.

For Canada, the Type-B rain gauge is used since 1970s for most stations by Environment Canada (Devine and Mekis, 2008; Wang et al., 2017). Tipping bucket and weighing gauges are also used in some stations (Metcalfe et al., 1997). For snowfall measurement, Nipher-shielded snow gauges were introduced at nearly 300 synoptic stations in the early 1960s, while most snow observation stations still rely on ruler measurements (https://www.canada.ca/en/environment-climate-change/services/sky-watchers/weather-instruments-tour.html). For temperature, weather observers use as many as 4 different thermometers mounted inside the Stevenson screen. Maximum and minimum thermometers use mercury and alcohol, respectively (https://www.canada.ca/en/environment-climate-change/services/sky-watchers/weather-instruments-tour/thermometers-thermistors.html). However, detailed metadata for an individual station is hard to obtain (e.g., see the detailed analysis of Whitfield (2014) for the station 3053600 in Kananaskis, Alberta).

For the U.S.A., station data are provided by many agencies/programs. The sources are denoted in SCDNA using the source flags provided in the GHCN-D dataset. For stations from the Cooperative Observer Program (COOP), the instruments are summarized in https://www.weather.gov/ilx/coop-equipment. The Standard Rain Gage (SRG) is used, and the method for measuring rainfall and snowfall is summarized in https://www.weather.gov/iwx/coop_8inch. For stations from Community Collaborative Rain, Hail,and Snow (CoCoRaHS), a 4-inch diameter rain gauge is used (https://www.cocorahs.org/Content.aspx?page=rain). For the U.S. Automated Surface Observing System (ASOS), heating Heated Tipping Bucket (HTB) and hygrothermometer are used for most stations, and there is a transition from HTB to All Weather Precipitation Accumulation Gauge (AWPAG) since 2004 (https://www.weather.gov/asos/ASOSImplementation, file:///Users/localuser/Downloads/ASOS_guide_1998.pdf). For NCEI Reference Network Database, a combination of weighing gauge, precipitation detector, and tipping bucket gauge are used, and air temperature is measured using three platinum resistance thermometers housed in fan aspirated solar radiation shields (https://www.ncdc.noaa.gov/crn/instruments.html). For SNOTEL, storage-type gage or tipping bucket is used, and temperature is measured using shielded thermistor (https://www.wcc.nrcs.usda.gov/snotel/snotel_sensors.html, https://www.wcc.nrcs.usda.gov/about/mon_automate.html). For Remote Automatic Weather Station (RAWS), THS-3 temperature and humidity sensor and RG-T tipping bucket rain gauge are used (https://www.fs.fed.us/eacc/library/docs/RAWS_WIMS_Guide.pdf, https://ftsinc.com/fixed-remote-automated-weather-station). For High Plains Regional Climate Center real-time data, tipping bucket or rain gauge is used (https://hprcc.unl.edu/awdn/index.php).

For Mexico, the automatic weather station, which is a set of electrical and mechanical devices that perform measurements of meteorological variables automatically (WMO Reference 182) are used by Servicio Meteorológico Nacional. (https://smn.conagua.gob.mx/es/observando-el-tiempo/estaciones-meteorologicas-automaticas-ema-s).

A useful database, the Historical Observing Metadata Repository (HOMR), is maintained by NOAA NCEI (https://www.ncdc.noaa.gov/data-access/land-based-station-data/station-metadata). Users can find detailed information of a station using station ID provided by different station sources, including SCDNA. For example, COOP station USC00244302 measures precipitation using SRG from 2000 to 2018-10-4 and SRG-STN since 2018-10-4. However, instrument information could be missing for many stations outside U.S.

We added a paragraph in Section 2.1: “Many types of precipitation and temperature measurement instruments are used at stations from different sources. For example, the Type-B rain gauge is used by Environment Canada since 1970s for most weather stations (Devine and Mekis, 2008; Wang et al., 2017), while tipping bucket and weighing rain gauges are also used in some stations (Metcalfe et al., 1997). Nipher-shielded snow gauges have been used by some synoptic stations, while ruler measurements are still used by more stations (Mekis and Brown, 2010). Station data in U.S. are from many organizations or programs with different instrument configurations. For instance, the standard rain gauge is used by the Cooperative Observer Program while Snow Telemetry uses storage-type gauges or tipping buckets. A better understanding of instrument specifications and historical changes is important for climate studies (Pielke Sr et al., 2007; Whitfield, 2014; Ma et al., 2019). A detailed summary of station instruments is provided in the documentation of the dataset (https://doi.org/10.5281/zenodo.3953310).”

Reference:

Devine, K. A., & Mekis, E. (2008). Field accuracy of Canadian rain measurements. Atmosphere-ocean, 46(2), 213-227.

Mekis, É., & Brown, R. (2010). Derivation of an adjustment factor map for the estimation of the water equivalent of snowfall from ruler measurements in Canada. Atmosphere-ocean, 48(4), 284-293.

Metcalfe, J. R., B. Routledge, and K. Devine. 1997. Rainfall measurement in Canada: Changing observational methods and archive adjustment procedures. Journal of Climate 10: 92-101.

Pielke Sr, R., Nielsen-Gammon, J., Davey, C., Angel, J., Bliss, O., Doesken, N., ... & Hale, R. (2007). Documentation of uncertainties and biases associated with surface temperature measurement sites for climate change assessment. Bulletin of the American Meteorological Society, 88(6), 913-928.

Whitfield, P. H. 2014. Climate station analysis and fitness for purpose assessment of 3053600 Kananaskis, Alberta. Atmosphere-Ocean 52(5): 363-383.

Wang, X. L., Xu, H., Qian, B., Feng, Y., & Mekis, E. (2017). Adjusted daily rainfall and snowfall data for Canada. Atmosphere-Ocean, 55(3), 155-168.

2. Codes: have you considered adding a little reader with a few capabilities, as additional tool for the interested users?

Response: We have added more detailed descriptions on GitHub (https://github.com/tgq14/GapFill/blob/master/README.md). The functions and their usage of different modules are introduced in Readme.md. Users can utilize the entire or part of the code package with the help of comments contained in scripts.

Minor/Details

1. P2: as trivial as it can be, it is worth to define the term "station".

Response: We added the definition. The revised sentence in P2 is "Many methods have been developed to estimate missing observations and reconstruct time series of meteorological stations that provide point-scale regular observations of atmospheric conditions".

2. P3, L96: Why exactly the variables of Tmin, Tmax, and precipitation have been chosen? Is it a matter of (lack of) availability of other measurements? (humidity, wind, etc). I just suggest to clarify.

Response: We selected the three variables for two reasons. First, as you have indicated, precipitation, Tmin and Tmax are the most common variables provided by meteorological stations, while other variables such as wind or humidity are less common. Second, most previous studies focus on precipitation and temperature, while other variables attract less attention. Thus, whether our methodology will work for other variables needs further investigation. We added explanation in the first paragraph in P3: "The three variables are selected because (1) most stations measure precipitation and temperature, while other variables, such as humidity and wind speed are measured at fewer stations, and (2) precipitation and temperature data are fundamental inputs for hydrological modeling."

We also added discussion on involving other variables in future work in Section 5.4.

3. Is precipitation the daily amount? I probably missed this information.

Response: Yes, it is. We added explanation in the first paragraph in Section 2.1: "In this dataset, precipitation is the daily amount."

Reviewer 2

This study develops a very useful dataset (SCDNA) of serially complete precipitation and temperature in North America. The dataset will benefit researchers in various fields with the long-term and gap-filled station data collected from multiple sources. The sophisticated framework for imputing missing values is well designed, which can be potentially applied in other regions of the world for the production of regional or even global serially complete datasets. From my perspective, the paper can be published on ESSD after the minor revisions, and I also have a few comments as below.

1. The differences between SCDNA and MSWEP show distinct differences along the boundaries of CONUS and Canada. Can you provide more detailed explanation about how observation time inconsistency causes this problem?

Response: MSWEP merges data from satellite products, reanalysis models and ground observations. Station data in different regions could have different observation time. To match station and reanalysis/satellite data, MSWEP calculates daily grid- and gauge-based time series, with the grid-based time series shifted by offsets of −36, −33, −30, …, +30, +33, and +36 h. Then, the temporal offset with the highest correlation is used to calculate 24-h accumulation of daily precipitation (Beck et al., 2019). Therefore, the final MSWEP estimates do not necessarily correspond to the raw observation of stations. For CONUS and Canada, the temporal offset is different and thus the mismatch between MSWEP and original station data is different.

We added an explanation in the third paragraph in Section 4.4: "Fig. 15 shows notable differences between MSWEP and SCDNA at the Canada-USA border and the USA-Mexico border. This is because MSWEP infers gauge reporting time by searching for the highest correlation between gauge data and the temporally shifted reanalysis/satellite estimates (Beck et al., 2019). The estimated temporal shift could vary with countries, which results in distinct differences of station-based evaluation results along national boundaries."

Reference:

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I., ... & Adler, R. F. (2019). MSWEP V2 global 3-hourly 0.1 precipitation: methodology and quantitative assessment. Bulletin of the American Meteorological Society, 100(3), 473-500.

2. The paper said "Outputs from three reanalysis products (ERA5, JRA-55, and MERRA-2) provided auxiliary information to estimate station records and were also used as an assessment benchmark. ". Can you give more explanation why you selected reanalysis products for benchmark?

Response: We choose the three products because (1) they are produced by representative reanalysis models from organizations in U.S., Europe, and Japan, and (2) they or their predecessors (ERA-Interim, JRA-25, and MERRA) have are been widely used by previous studies (e.g., Sun et al., 2018). The three reanalysis products are used as benchmark because they are widely used as the source of long-term precipitation and temperature data and have been applied to support infilling and reconstruction in this study.

We added an explanation in Section 2.2: "The three products are chosen because they are representative products from different international organizations and they or their predecessor (ERA-Interim, JRA-25, and MERRA) have are been widely used by researchers.".

Reference:

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K. L. (2018). A review of global precipitation data sets: Data sources, estimation, and intercomparisons. Reviews of Geophysics, 56(1), 79-107.

3. The period from 1979 to 2018 is total 40 years. Numbers of stations with only at least 8-year records are shown in Table 1. Why only 8-year period records are showed? Are only stations with at least 8-year precipitation or Tmin and Tmax records between 1979 to 2018 utilized to evaluate the performance? Is there some difference between 8-year records and total records for evaluation?

Response: For the first question, we only show 8-year records because according to our sensitivity analysis, eight years are enough to ensure gap filling is generally reliable (Figure S1). Using a higher period threshold can improve the quality of the final dataset but will reduce the number of stations.

For the second question, yes, only stations with at least eight-year records are used for evaluation to be consistent with inputs.

For the third question, our evaluation is based on 30% samples of each station. For example, if a station has 8-year/40-year observations, the validation samples are about 2.4-year/12-year. Therefore, the evaluation period length could be different for different stations. According to our results (Figures 6 and 12), the spatial distributions of accuracy metrics and contribution ratios are smooth, indicating that the difference between 8-year records and total records for evaluation is not evident. We added explanation in Step-5 in Section 3.3.3: "Although the evaluation samples are different among stations, the results are reliable and stable as shown in the results section."

4. Precipitation and minimum/maximum temperature are very widely used in hydrometeorological studies. I think probably this is why the three variables are chosen. Considering meteorological stations can usually measure more variables which also suffer from missing values, expanding this work to other variables would be very interesting for future studies. I suggest that the authors add some discussion about the applicability of your method to other variables.

Response: Thank you for this suggestion. Expanding this work to other variables will be an interesting study. We added discussion in Section 5.4: "Furthermore, other variables such as wind and humidity observed by stations also suffer from the same problems faced by precipitation and temperature. Future studies should explore whether the current methodology is applicable to other variables. A SCD covering more variables would be useful for research in various fields."

SCDNA: a serially complete precipitation and temperature dataset for North America from 1979 to 2018

Guoqiang Tang[1,2], Martyn P. Clark[1,2], Andrew J. Newman[3], Andrew W. Wood[3], Simon Michael Papalexiou[2,4], Vincent Vionnet[5], and Paul H. Whitfield[1,2]

[1]University of Saskatchewan Coldwater Lab, Canmore, Alberta, Canada
[2]Centre for Hydrology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
[3]National Center for Atmospheric Research, Boulder, Colorado
[4]Department of Civil, Geological and Environmental Engineering, University of Saskatchewan, Saskatchewan, Canada
[5]Environmental Numerical Research Prediction, Environment and Climate Change Canada, Dorval, Quebec, Canada

Correspondence to: Guoqiang Tang (guoqiang.tang@usask.ca)

Abstract: Station-based serially complete datasets (SCDs) of precipitation and temperature observations are important for hydrometeorological studies. Motivated by the lack of serially-complete station observations for North America, this study seeks to develop a SCD from 1979 to 2018 from station data. The new SCD for North America (SCDNA) includes daily precipitation, minimum temperature (T_{min}), and maximum temperature (T_{max}) data for ~~27280~~27276 stations. Raw meteorological station data were obtained from the Global Historical Climate Network Daily (GHCN-D), the Global Surface Summary of the Day (GSOD), Environment and Climate Change Canada (ECCC), and a compiled station database in Mexico. Stations with at least 8-year records were selected, which underwent location correction and were subjected to strict quality control. Outputs from three reanalysis products (ERA5, JRA-55, and MERRA-2) provided auxiliary information to estimate station records~~ and were also used as an assessment benchmark~~. Infilling during the observation period and reconstruction beyond the observation period were accomplished by combining estimates from 16 strategies (variants of quantile mapping, spatial interpolation, and machine learning). A sensitivity experiment was conducted by assuming 30% observations of stations were missing – this enabled independent validation and provided a reference for reconstruction. Quantile mapping and mean-value corrections were applied to the final estimates. The median Kling-Gupta efficiency (KGE') values of the final SCDNA for all stations are 0.90, 0.98, and 0.99 for precipitation, T_{min} and T_{max}, respectively. The SCDNA is closer to station observations than four benchmark gridded product, and can be used in applications that require either quality-controlled meteorological station observations or reconstructed long-term estimates for analysis and modelling. The dataset is available at https://doi.org/10.5281/zenodo.3735533 ~~https://doi.org/10.5281/zenodo.3735534~~ (Tang et al., 2020).

Key words: serially complete dataset; precipitation; temperature; North America

1 Introduction

Station-based serially complete datasets (SCDs, see Table A1 for all acronyms) are important for meteorological, climatological and hydrological studies (Kanda et al., 2018; Ramos-Calzado et al., 2008), such as ~~the production of~~producing retrospective gridded products (Di Luzio et al., 2008; Kenawy et al., 2013; Newman et al., 2019; Serrano-Notivoli et al., 2019), trend analyse~~i~~s (Knowles et al., 2006; Anderson et al., 2009; Papalexiou and Montanari, 2019), and climatologic index calculation (Alexander et al., 2006; Papalexiou et al., 2018). These SCDs are useful because station-based observational datasets often contain missing values due to factors such as observer absence, instrument~~al~~ failures and interrupted communication (Hasanpour Kashani and Dinpashoh, 2012). Moreover, station observations failing quality control tests such as outlier and homogeneity checks may not be reliable (Menne et al., 2012), and many stations are only maintained over a relatively short period of time or portions of the year, resulting in data gaps that could affect the analysis of climate variability or long-term trends (Rubin, 1976; Stooksbury et al., 1999). Serial completeness is also a critical requirement for real-time station-based applications, which regularly contend with missing data values due to latencies in station reporting, quality control and processing (Tang et al., 2009).

Many methods have been developed to estimate missing observations and reconstruct time series of meteorological stations that provide point-scale regular observations of atmospheric conditions.~~;~~ T~~t~~hey can be ~~grouped~~ classified ~~in~~ as self-contained infilling, spatial interpolation, quantile mapping ~~(QM)~~, and machine learning methods.

1. Self-contained infilling only uses records ~~of~~ from the target station to estimate its own missing values. Typical methods include interpolation based on data from previous and subsequent days or replacing missing values by long-term mean (Kemp et al., 1983; Pappas et al., 2014). Self-contained infilling, however, only performs well for variables with high temporal autocorrelation such as temperature and is problematic for daily precipitation (Simolo et al., 2010; Teegavarapu and Chandramouli, 2005), and in covering lengthy data gaps.

2. Spatial interpolation uses neighboring stations (identified on spatial distance or statistical similarity) to estimate data at the target station. Spatial interpolation methods~~,~~ ~~which~~ can be divided into two types: the first uses information only from neighboring stations; and common methods include linear interpolation and inverse distance weighting (IDW; Shepard, 1968). The second method needs information from both neighboring and target stations. Typical examples include the revised normal ratio (NR; Young, 1992) and the single best estimator (Eischeid et al., 1995, 2000), both of which use correlation coefficients (CCs) between target and neighboring stations to estimate merging weights. This second type of spatial interpolation also includes more sophisticated methods (e.g., multiple linear regression, optimal interpolation, and kriging) that build a functional relationship between neighboring and target stations (Simolo et al., 2010). Previous studies have shown that multiple linear regression based on the least absolute deviation criteria (MLAD) performs better than many interpolation methods such as IDW, NR, and optimal interpolation in infilling/reconstruction (Eischeid et al., 2000; Kanda et al., 2018).

3. Quantile mapping (QM) is widely used to correct biases ~~of~~ in meteorological data (Maraun, 2013; Cannon et al., 2015) and performs well in estimating missing station data (Simolo et al., 2010; Newman et al., 2015, 2019; Devi et al., 2019). In QM-based estimation, the cumulative distribution functions (CDFs) of observations from neighboring and target stations are derived, and the record at the target station is estimated as the inverse of its CDF using concurrent CDF probability information from neighboring stations. QM can avoid the problem of overestimating wet days in precipitation series and preserve the frequency distribution of time series, which is useful for estimating extreme events (Cannon et al., 2015).

4. Machine learning techniques have been successfully applied to infill station record gaps (Dastorani et al., 2010; Wambua et al., 2016). For example, Coulibaly and Evora (2007) estimated missing daily precipitation and temperature in northeastern Canada using six types of artificial neural networks (ANNs). Ustaoglu et al. (2008) estimated daily temperature using three ANN methods in the Geyve and Sakarya basin, Turkey. Gene expression programming was applied in the estimation of missing monthly rainfall data in Malaysia (Che Ghani et al., 2014). Sattari et al. (2017) recommended that a decision-tree algorithm can be used to estimate monthly precipitation due to its simplicity and high accuracy. Serrano-Notivoli et al. (2019) applied the *k*-nearest neighbours regression to reconstruct minimum temperature (T_{min}) and maximum temperature (T_{max}) observations in Spain to form a gridded dataset.

Previous SCDs have been developed using multiple infilling and reconstruction methods. For instance, Eischeid et al. (2000) produced a daily SCD from 1951 to 1991 for the western United States (U.S.), including 2962 precipitation stations and 2034 temperature stations; Vicente-Serrano et al. (2003) produced a daily SCD from 1901 to 2002 for northeast Spain using 3106 precipitation stations; Di Piazza et al. (2011) built a monthly SCD from 1921 to 2004 for Sicily, Italy using 247 precipitation stations; and Woldesenbet et al. (2017) produced a daily SCD of precipitation and temperature from 1980 to 2013 for the Upper Blue Nile Basin using six stations. There is currently no SCD for North America; this means that researchers often must collect station data from different databases, which is time-consuming and may cause inconsistencies between studies based on different methods.

Responding to this need, we develop a retrospective 40-year daily SCD for North America (SCDNA) of precipitation, T_{min} and T_{max} from 1979 to 2018. Central America and Caribbean are also covered by SCDNA. The three variables are selected because (1) most stations measure precipitation and temperature, while other variables, such as humidity and wind speed are measured at fewer stations, and (2) precipitation and temperature data are fundamental inputs for hydrological modeling. Station observations are collected from four global and regional databases and undergo strict quality control to eliminate dubious records. Since the performance of infilling and reconstruction methods differs in space and time, the results from 16 strategies are merged to produce a single deterministic estimate. Finally, the SCDNA is compared to four gridded products to demonstrate its performance and areas for improvement. The SCDNA is expected to have a wide variety of applications in North America, and the methodology can be used to produce SCDs in other regions of the world.

2 Datasets

2.1 Meteorological station data

This study uses precipitation, $T_{\min}$, and $T_{\max}$ station data from four databases, the Global Historical Climate Network Daily (GHCN-D; https://www.ncdc.noaa.gov/ghcnd-data-access; Menne et al., 2012), the Global Surface Summary of the Day (GSOD; https://catalog.data.gov/dataset/global-surface-summary-of-the-day-gsod), Environment and Climate Change Canada (ECCC; https://climate.weather.gc.ca/historical_data/search_historic_data_e.html), and the Mexico database from Servicio Meteorológico Nacional, under the Comisión Nacional del Agua (Livneh et al., 2015). This study uses daily precipitation totals from each dataset. Only stations with at least 8-year precipitation or $T_{\min}$ and $T_{\max}$ records between 1979 to 2018 are utilized. The requirement for minimum recording length is different among studies (e.g., Eischeid et al., 2000; Newman et al., 2015). We adopted a relatively short time limitation because (1) 8-year records are sufficient to provide basic support for missing value estimation (Fig. S1), and (2) the open-access dataset and codes enable users to design customized data selection criteria according to their research requirements.

The numbers of stations with at least 8-year records are 33026, 4619, 3634, and 4049 for GHCN-D, GSOD, ECCC, and the Mexico database, respectively (Table 1). Their spatial distributions are shown in Fig. S2. GHCN-D has complied a large amount of data from many sources including the Mexico database and ECCC. For identical stations from different sources, we keep the one with longer observation history, resulting in the exclusion of ~95% of stations from the Mexico database and adoption of ~91% of stations from ECCC. Stations with more than 30% missing values in the observation period are excluded because they could be seasonal stations or suffer serious instrumentation problems. Stations overlapping in space (same latitude and longitude) and without sufficient metadata for discrimination are merged (see Sect. 3.2). The above screening reduces the available stations from 45328 to 31772 (Table 1), yet more stations are discarded due to quality control procedures (Sect. 3.1). The final SCDNA includes 24615 ~~24721~~ precipitation, ~~19677~~ 19604 $T_{\min}$, and ~~19684~~ 19611 $T_{\max}$ stations; note that the numbers of $T_{\min}$ and $T_{\max}$ stations differ as quality controls can result in excluding the one and reserving the other in some stations.

Most stations are located in the Contiguous United States (CONUS), southern Canada, and Mexico, while few stations are located in high-latitude regions such as the Arctic Archipelago (Fig. 1b and c). The spatial distributions of precipitation and temperature stations are similar, except in eastern CONUS where precipitation stations have a higher density.

Table 1. Numbers of stations with at least 8-year records from 1979 to 2018

Station numbers	**GHCN-D**	**GSOD**	**ECCC**	**Mexico**	**Merge**	**Total**
Original numbers	33026	4619	3634	4049	0	45328
SCDNA input	24765	4331	3100	187	207	31772
SCDNA output: precipitation	19255	~~2656~~2551	2440	170	~~200~~199	~~24721~~24615
SCDNA output: $T_{\min}$	~~13445~~13394	~~3650~~3631	2219	~~167~~166	194~~6~~	19604~~77~~

SCDNA output: T_{max}	134002~~53~~	36332~~51~~	2217	166~~7~~	194~~6~~	19611~~84~~

Notification: “Merge” is derived from stations with overlapped locations from all the other data sources (Sect. 3.1.1).

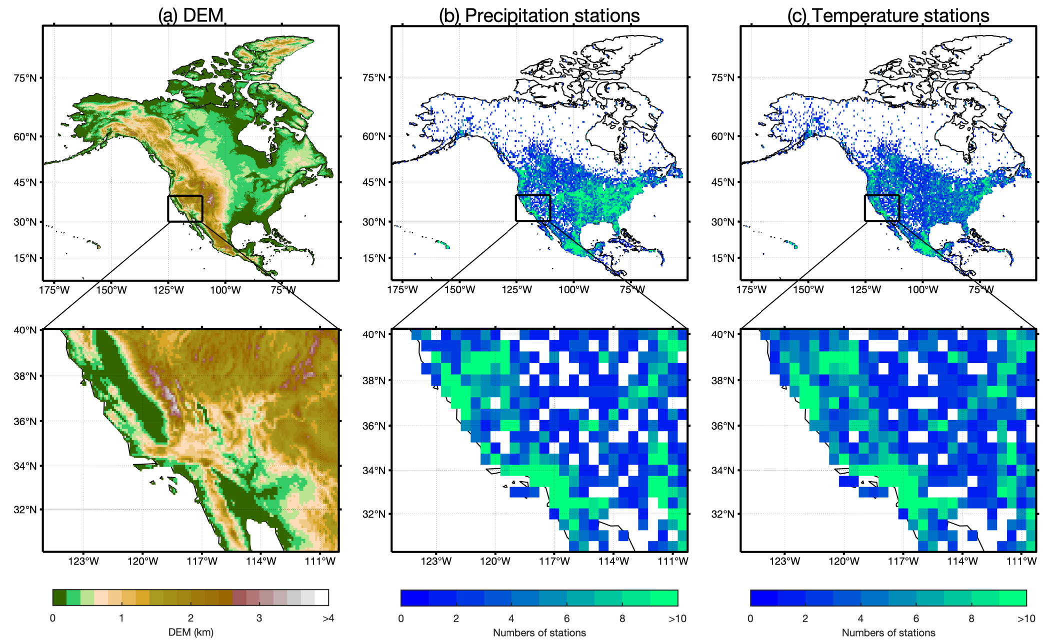


Figure 1. (a) Digital elevation model (DEM; Sect. 2.3) of North America. (b) and (c) are the densities of stations at the 0.5°×0.5° resolution for precipitation and temperature, respectively. T_{min} and T_{max} stations are highly consistent, and thus T_{min} is used to represent temperature in (c). The nested black boxes show examples of DEM and station densities.

In North America, more station observations occur in U.S. than in Canada and Mexico (Fig. 2). The number of samples in U.S. increases from 1979 to 2018, and there are more precipitation samples than temperature samples. For Canada, the numbers of precipitation and temperature samples are similar and show a decrease from 1988 to 2018; the sample number in 2018 is only 61.76% of that in 1988. Mexico has more meteorological samples than Canada, yet this number decreases after 1983. The decreasing trend is especially sharp after 2012 which may be due to the delay in data collection or termination of some stations.

Figure 3 shows the fractions of missing values for all stations during the observation period (referred as ratio-1) and during the entire period from 1979 to 2018 (referred as ratio-2). For temperature, ~20% of the stations have more than 20% missing values in the observation period (ratio-1), and ~20% of the stations have more than 70% missing values in the entire period (ratio-2). For precipitation, the fraction of missing values is larger. The fractions show strong spatial variations (Fig. S3). Ratio-2 is smaller for precipitation stations in western U.S. and temperature stations in central U.S., but larger in Canada and Alaska. Most stations in Mexico have higher ratio-1 than other regions in North America, indicating that those stations have notable fractions of missing values during the observation period.

In summary, the curves of ratio-1 indicate that a small number of missing values need infilling during the observation period, while the curves of ratio-2 indicate that extensive reconstruction is needed over the entire period.

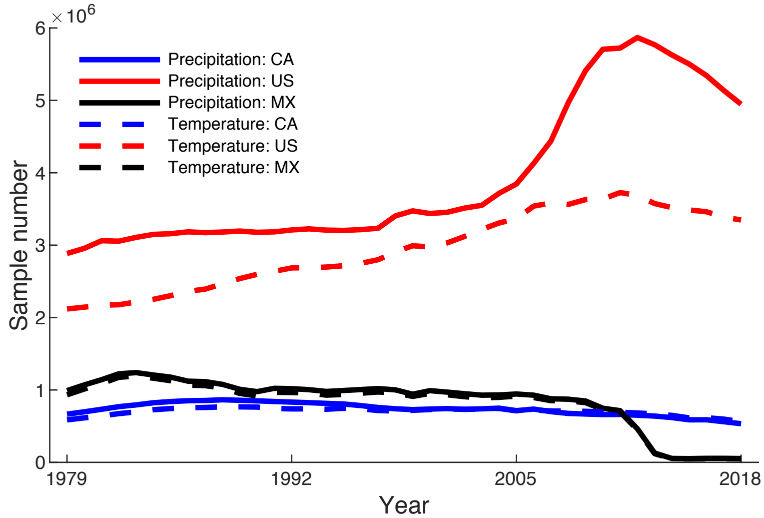


Figure 2. Sample numbers of stations for each year from 1979 to 2018. CA represents Canada, US represents United States, and MX represents Mexico. T_{max} stations are highly consistent with T_{min} stations, and thus T_{min} is used to represent temperature. The numbers of samples could be a better indicator than the numbers of stations because many stations have notable missing values.

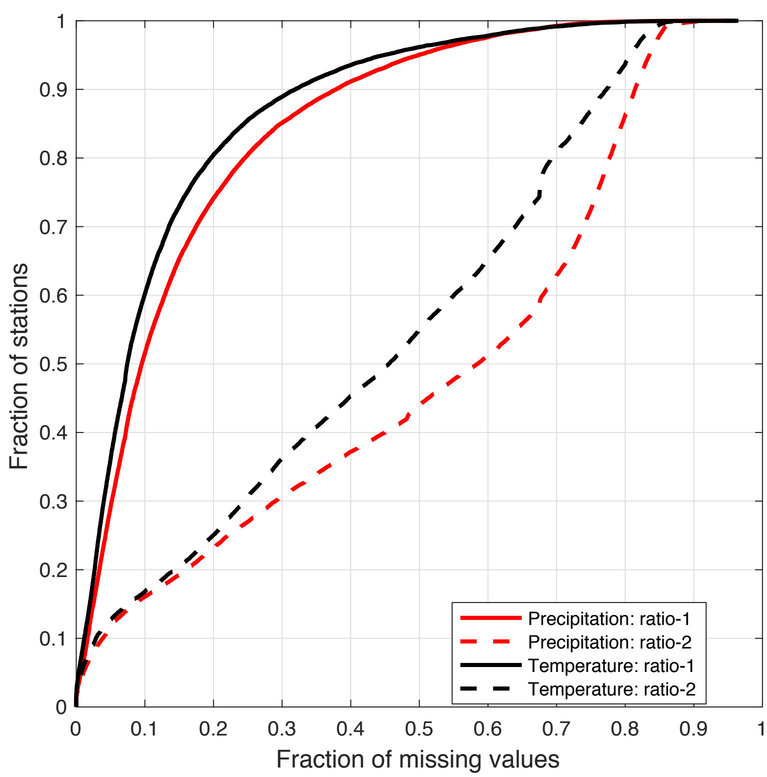


Figure 3. The fraction of missing values for stations with at least 8-year records. Ratio-1 is the degree of missingness during the observation period, and ratio-2 is the degree of missingness during the entire period of interest (1979 to 2018). T_{min} is used to represent temperature because T_{max} show almost overlapped curves with T_{min}.

Many types of precipitation and temperature measurement instruments are used at stations from different sources. For example, the Type-B rain gauge is used by Environment Canada since 1970s for most weather stations (Devine and Mekis, 2008; Wang et al., 2017), while tipping bucket and weighing rain gauges are also used in some stations (Metcalfe et al., 1997). Nipher-shielded snow gauges have been used by some synoptic stations, while ruler measurements are still used by more stations (Mekis and Brown, 2010). Station data in U.S. are from many organizations or programs with different instrument configurations. For instance, the standard rain gauge is used by the Cooperative Observer Program while Snow Telemetry uses storage-type gauges or tipping buckets. A better understanding of instrument specifications and historical changes is important for climate studies (Pielke Sr et al., 2007; Whitfield, 2014; Ma et al., 2019). A detailed summary of station instruments is provided in the documentation of the dataset (https://doi.org/10.5281/zenodo.3735533).

2.2 Reanalysis products

We use reanalysis precipitation, T_{min} and T_{max} from the fifth generation of European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric reanalyses of the global climate (ERA5; Copernicus Climate Change Service (C3S), 2017), the Japanese 55-year Reanalysis (JRA-55; Kobayashi et al., 2015), and the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2; Gelaro et al., 2017) (see Table 2). The three products are chosen because they are representative products from different international organizations and they or their predecessor (ERA-Interim, JRA-25, and MERRA) have are been widely used by researchers. The ERA5 and JRA-55 do not provide daily outputs, thus, daily precipitation is accumulated from sub-daily estimates while daily T_{min} and T_{max} are estimated by the sub-daily minimum and maximum temperature values. Gridded reanalysis precipitation is linearly interpolated to match point-scale station data, and T_{min} and T_{max} are downscaled using temperature lapse rate (TLR; see Sect. 3.1).

Table 2. Information on the three reanalysis products.

Products	Spatial resolution	Temporal resolution	Period	Agency
ERA5	0.25°×0.25°	1 h	1979-present	European Centre for Medium-Range Weather Forecasts
JRA-55	∼55~~60~~ km	3 h	1958-present	Japan Meteorological Agency
MERRA-2*	0.5°×0.625°	daily	1980-present	NASA’s Global Modeling and Assimilation Office

* MERRA-2 provides outputs in temporal resolutions from 1 h to 1 month; here we use daily values.

2.3 Auxiliary data

The Multi-Error-Removed Improved-Terrain digital elevation model (MERIT DEM) at a 3 sec (~90 m at the equator) resolution (Yamazaki et al., 2017) is used in this study. To enable temperature downscaling, the high-resolution DEM is spatially averaged to the original resolutions of ERA5, MERRA-2, and JRA-55 (Table 2). The MERIT DEM may be slightly different than the DEM data used in the three reanalysis products, and this will have a limited impact on missing data estimation (Sect. 3.3.2).

The Multi-Source Weighted-Ensemble Precipitation (MSWEP) V2.2 dataset (Beck et al., 2017, 2019) is utilized for the comparison with the SCDNA developed by this study. MSWEP merges data from ground observations, satellite products, and reanalysis models, and performs better than all products used for merging (Beck et al., 2019). The comparison can show whether the SCDNA is a better choice than MSWEP to fill gaps in station precipitation observations.

3 Methodology

The methodology to produce the SCDNA includes three primary steps (Fig. 4): (1) preparing a unified precipitation and temperature database from multiple sources (Sect. 2.1 and 3.1); (2) downscaling reanalysis estimates (Sect. 2.2 and 3.2) that are used in QM- and machine learning-based data estimation (Sect. 3.3) and comparison with the SCDNA (Sect. 4.5); and (3) producing the SCDNA from 1979 to 2018 based on 16 strategies (Sect. 3.3). The following sub-sections summarize the work in each step of the methodology (Sect. 3.1, 3.2, and 3.3) as well as the approach used to evaluate the performance of the method (Sect. 3.4).

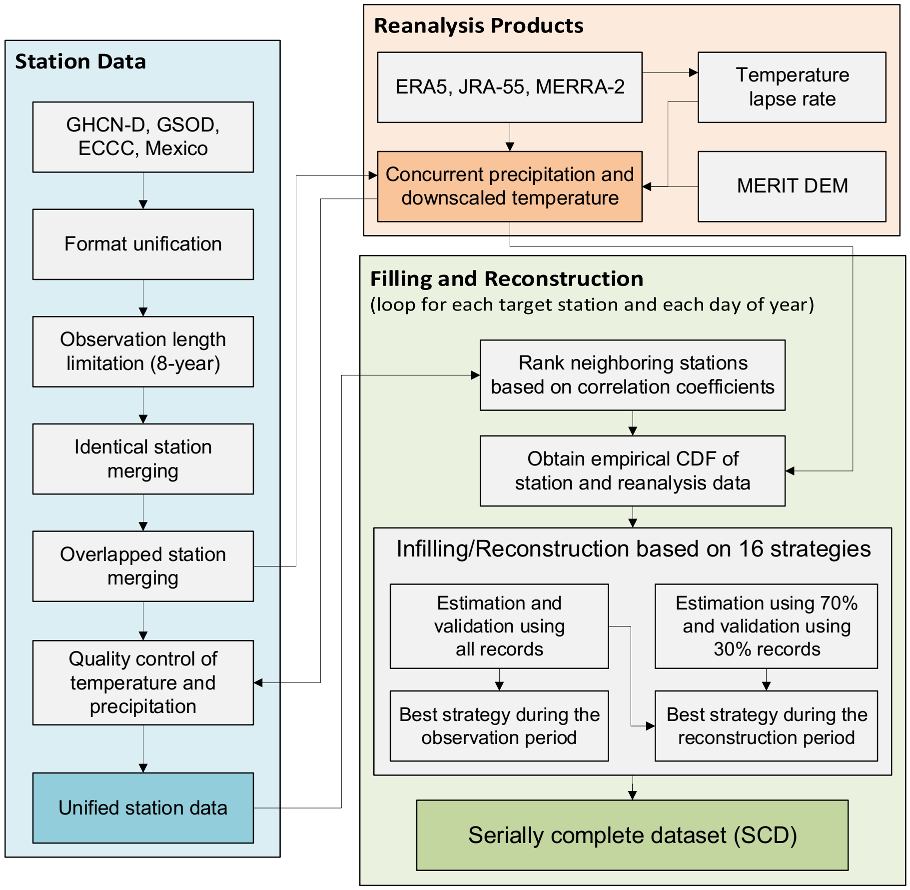


Figure 4. Flowchart of the production of the SCDNA, including station data preparation, reanalysis product processing, and missing data infilling and reconstruction.

In this study, infilling refers to the estimation of missing values during the observation period, while reconstruction refers to estimating values outside of the observation period when no station record is available (Fig. 5). Station records that fail quality control are treated as missing values.

3.1 Prepare a unified precipitation and temperature database

3.1.1 Merging of stations based on location

Stations are merged if their latitude and longitude match other stations. The problem of overlapped locations is caused by identification alteration of one station for different periods, ~~or~~ recording/rounding bias of station location information, inconsistent naming rules of different sources, and other factors. Although it is possible that multiple stations are deployed in the same location for experimental aims, location merging is done to preserve internal consistencies as inconsistent records at the same location are self-contradictory.

The method for location merging includes several steps. First, overlapping stations are extracted and grouped. Stations within the same group that have non-overlapping recording periods are simply merged into one time series. Otherwise, the Spearman's rank CC (SCC) between precipitation series from all station pairs in the group is calculated. For SCC < 0.7, the station group is discarded due to large discrepancies; for 0.7 < SCC < 0.9 the discrepancy is considered as tolerable and the station with the longest record is kept; for SCC > 0.9 stations are considered as highly correlated and their data are merged into one time series, while for overlapping periods the station with longest record is used.

Overall, 1240 stations are involved in location merging, stratified in 586 station groups. Around 10% of the groups contain more than two stations and the largest group contains five stations. After location merging, only 207 groups are kept and merged into unified times series (Table 1). Despite the steps taken above, the merged series could contain inhomogeneities due to the combination of records from multiple stations.

3.1.2 Quality control

To ensure station observations undergo strict and comprehensive quality control, we adopted the methods used to produce previous station-based datasets. For $T_{\min}$ and $T_{\max}$, we followed the method designed by Durre et al. (2010) which is adopted by GHCN-D (Menne et al., 2012). The procedures include five types of checks: integrity checks, outlier checks, internal and temporal consistency checks, spatial consistency checks, and extreme megaconsistency checks. A few of the procedures in Durre et al. (2010) require other variables such as snowfall, and thus are not adopted in this study. In addition, the quality flags in this study are partly different with those of GHCN-D because of the different sources, numbers and temporal periods of stations.

For precipitation, quality control procedures consist of three parts. The first part is similar with that for temperature. The second part (four types of checks) follows procedures designed by Hamada et al. (2011) which are adopted by the Asian Precipitation-Highly-Resolved Observational Data Integration Towards Evaluation (APHRODITE; Yatagai et al., 2012). The third part (two types of checks) adopts strategies by Beck et al. (2019) used in the production of MSWEP. Note that although Durre et al. (2010) and Hamada et al. (2011) share some common traits for precipitation, both of them are adopted to ensure quality control reliability.

~~Details~~ The details of quality checks are in Appendix B.

3.2 Downscale reanalysis data

The reanalysis temperature estimates are downscaled to match point-scale station observations using temperature lapse rate (TLR) according to

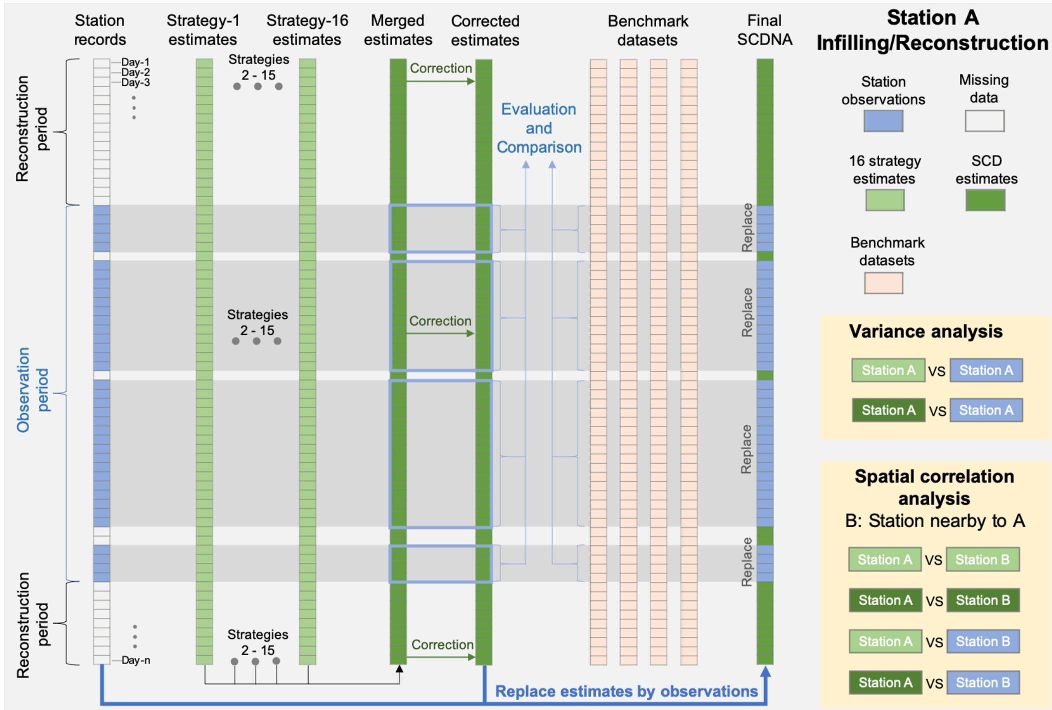
$$T_s = T_R + \mathrm{TLR} \times \Delta h \tag{1}$$

where T_R is 2-m reanalysis air temperature, T_s is downscaled temperature, Δh is the height difference between station elevation and reanalysis grid elevation. TLR shows notable spatiotemporal variations (Minder et al., 2010) and estimating TLR based on ground observations over a large domain is difficult due to the sparsity of stations. Yet recent studies show that reanalysis outputs offer an alternative in estimating gridded TLR (e.g., Gao et al., 2012). The gradient of air temperature at different pressure levels above the ground can be used to approximate near-surface TLR (Gao et al., 2012, 2018; Gruber, 2012). Tang et al. (2018) compared eight temperature downscaling methods in CONUS and found that methods based on reanalysis-derived TLR can achieve higher accuracy compared to fixed TLR (e.g., -6.5°C/km) or statistical interpolation downscaling methods. Hence, this study uses the linear regression slope between MERRA-2 air temperature and geopotential heights from 300 hPa to 1000 hPa pressure levels to represent TLR for each month at the resolution of 0.5°×0.625° (Table 2). MERRA-2 is used because it directly provides monthly data and masks temperature data if the pressure level is below land surface. The choice of pressure levels needs further investigation because relationships between vertical and near-surface temperature vary with regions. Complicated TLR phenomena such as inverse lapse rate are not considered for simplicity. The climatological mean of TLR (Fig. S4) decreases from -4.8°C/km in the northeast continent (i.e., Canadian Arctic Archipelago) to -7.2°C/km in the southwest continent (i.e., Rocky Mountains in CONUS). The smaller TLR magnitude in high latitudes is consistent with previous studies (e.g., Gardner et al., 2009; Marshall et al., 2007).

3.3 Produce the serially complete dataset

To produce the high-quality SCDNA for North America, we use 16 strategies: four based on quantile mapping with neighboring stations (QMN; e.g., Longman et al., 2019; Newman et al., 2015, 2019), four on quantile mapping with concurrent reanalysis estimates (QMR), four using spatial interpolation methods (INT; e.g., Eischeid et al., 2000; Kanda et al., 2018; Woldesenbet et al., 2017), two using machine learning methods (MAL; e.g., Dastorani et al., 2010; Wambua et al., 2016), and two multi-strategy merging methods (MRG). Merging multiple infilling/reconstruction methods can provide better estimation than individual methods, as shown by previous data merging and gap infilling studies (e.g., Eischeid et al., 2000; Beck et al., 2017, 2019; Ma et al., 2018).

We generate estimates for every station and every day from 1979 to 2018 (Fig. 5). The estimates from these 16 strategies and the SCDNA are evaluated using station observations, and the performance of the SCDNA is compared to four benchmark gridded products. Then, the estimates of the SCDNA are corrected for further accuracy improvement. Finally, estimates are replaced by station observations when observations exist and pass quality control checks. The variance and spatial correlation analyses are performed to compare the statistical properties of station observations and estimates (see Sect. 4).



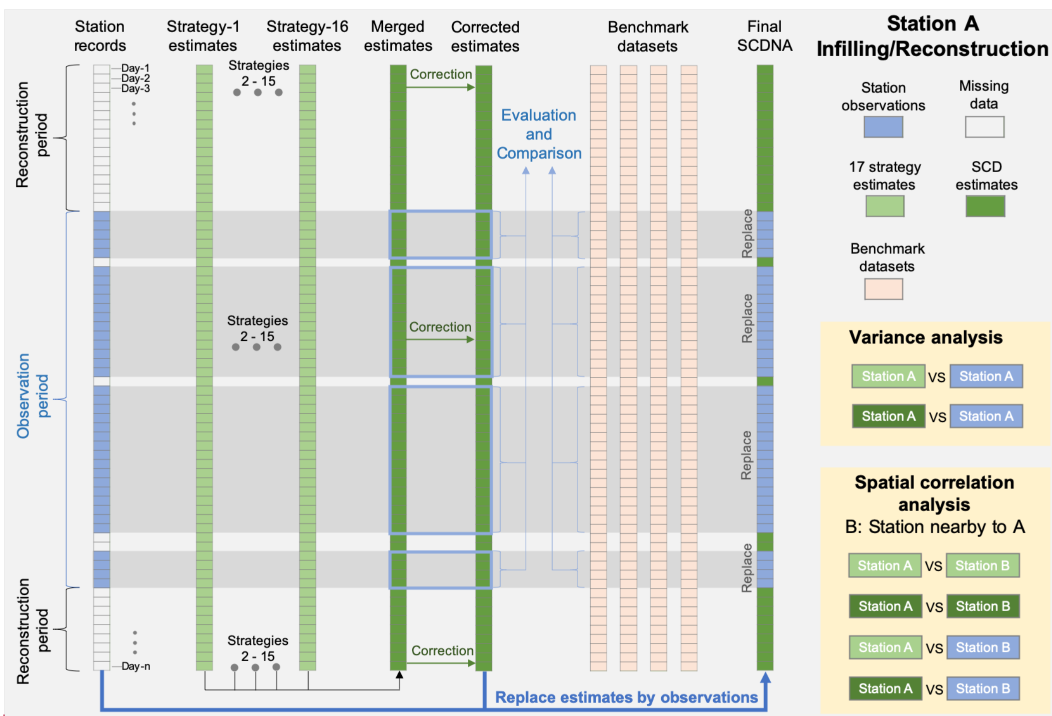


Figure 5. Diagram of the infilling and reconstruction for a specific station (referred to as A). The entire period from 1979 to 2018 is divided into the observation period and the reconstruction period. The data flows of variance and spatial correlation analyses are shown in the nested yellow boxes. Station B is a nearby station of A.

Only stations with at least 3000 valid values are included in the infilling and reconstruction effort. The ~~eight~~ nine steps (termed Step-1 to Step-9~~8~~) of SCDNA production are described as below. Unless otherwise stated, the steps are implemented for each target station (s), each variable (precipitation, T_{min}, and T_{max}), and each day of the year (DOY, i.e., 1-366).

3.3.1 Data extraction

Step-1: Spatiotemporally concurrent reanalysis estimates (ERA5, JRA-55, and MERRA-2) are extracted, including precipitation, $T_{min,}$ $T_{max,}$ and TLR. Precipitation is linearly interpolated from gridded reanalysis estimates, and temperature is downscaled (i.e., corrected for the elevation difference between the reanalysis grid cell and the station elevation) based on TLR (Sect. 3.1).

Step-2: Neighboring stations (at least one and at most 30) with at least 8-year overlapped period with station s are found within the searching radius of 200 km. These stations are ranked from closest to farthest according to their CC with the target station. SCC is used for precipitation, and Pearson CC (PCC) is used for T_{min} and T_{max}. CC is calculated using data within a 31-day window centered around the current DOY from all years.

Step-3: The empirical CDFs of s, neighboring stations, and reanalysis estimates are obtained using data within the same 31-day window.

3.3.2 Infilling and reconstruction

Step-4: For each day (d) corresponding to the DOY, the estimated data are acquired based on 16 strategies which are divided into five groups.

***Group 1*: Quantile Mapping with Neighboring stations (QMN)**

- **QMN-1**: For all neighboring stations with valid records, the station with the highest CC in Step-2 is selected. The estimated data for s and d is obtained using Eq. (2).

$$X_s = F_s^{-1}(F_i(X_i)) \tag{2}$$

where X_i is precipitation or temperature for d from the selected neighboring station i, F_i is the empirical CDF of i corresponding to the DOY, F_s^{-1} is the inverse CDF of s corresponding to the DOY, and X_s is the estimated data.

- **QMN-2**: For all neighboring stations with observations, estimated values are obtained using Eq. (2) which are merged based on Eq. (3).

$$X_s = \frac{\sum_i^n W_i F_s^{-1}(F_i(X_i))}{\sum_i^n W_i} \tag{3}$$

$$W_i = CC_i^2 \tag{4}$$

where n is the number of neighboring stations, $F_s^{-1}(F_i(X_i))$ is the QM-based estimate from i, and W_i is the weight calculated using Eq. (4). CC_i is CC (SCC or PCC) between data from s and i corresponding to the DOY. W_i is assigned zero if CC_i is negative.

- **QMN-3**: Similar to QMN-2, but the weight is calculated according to the distance (D_i) between s and i based on Eq. (5). Although the exponent of distance (k) varies in different studies, -2 is the most common choice (Teegavarapu and Chandramouli, 2005).

$$W_i = D_i^k \tag{5}$$

- **QMN-4**: The median of QMN-1 to QMN-3 is used as the estimated data. The strategy of using median values is the same with Eischeid et al (2000), which could be closer to actual observations than QMN-1 to 3.

***Group 2*: Quantile Mapping with Reanalysis products (QMR)**

Reanalysis products provide useful information for SCDNA production as (1) remote regions may not have enough neighboring stations, and (2) neighboring stations also have missing values which could result in gaps of estimates at the target station.

- **QMR-1 to QMR-3**: Similar to QMN-1, but the neighboring station is replaced by concurrent ERA5, JRA-55, and MERRA-2 estimates, respectively.

- **QMR-4**: The median of QMR-1 to 3 is used as the estimated data.

***Group 3*: Interpolation (INT)**

The three interpolation methods used in this study are MLAD (referred as INT-1), NR (referred as INT-2), and inverse distance weighting (IDW, referred as INT-3). They are described below. Following Eischeid et al. (2000), neighboring stations with CC lower than 0.35 are excluded. The remaining stations are ranked from high CC to low CC. A maximum of four neighboring stations are used in the interpolation. For $T_{\min}$ and $T_{\max}$, direct interpolation from neighboring stations to s could be biased due to the elevation differences between stations. Temperature data from neighboring stations are downscaled to the elevation of s based on Eq. (1).

- **INT-1**: MLAD minimizes the sum of absolute errors. It is more robust than regression based on least squares because while least square estimation is effective when the errors are normally distributed and independent, environmental variables, especially precipitation, often violate the assumption of normality (Eischeid et al., 2000). MLAD has been well documented with better performance in gap infilling than other interpolation methods (Eischeid et al., 1995, 2000; Kanda et al., 2018; Young, 1992). The formula is shown in Eq. (6).

$$X_s = c_0 + \sum_i^n c_i X_i \tag{6}$$

where c_i (i =0, 1, …, n) is regression coefficients estimated using data within a 31-day window for each DOY. Different d corresponding to the same DOY could have different combinations of neighboring stations due to the limitation of observation availability. MLAD is performed for each combination to ensure that effective estimates are available for all days.

- **INT-2:** NR is an interpolation method proposed by Paulhus and Kohler (1952) and modified by Young (1992). The modified version is adopted in this study, which combines information from neighboring stations by replacing $F_s^{-1}(F_i(X_i))$ with X_i in Eq. (3). The weight is calculated using Eq. (7).

$$W_i = CC_i^2 \frac{N_i - 2}{1 - CC_i^2} \tag{7}$$

where N_i is the number of samples used to calculate CC_i between s and i. SCC is used for precipitation and PCC is used for temperature.

- **INT-3**: IDW is one of the most common interpolation methods. It is implemented similar to NR, where the inverse squared distance, as shown in Eq. (5), is used as the weight.

- **INT-4**: The median of INT1, INT2 and INT3 is used as the estimated data.

***Group 4*: Machine Learning (MAL)**

The two MAL methods used in this study are ANN (referred as MAL-1) and random forest (RF, referred as MAL-2; Breiman, 2001). Unlike QMN, QMR and INT that are carried out for each DOY, MAL uses complete observation records of s to ensure that ANN and RF are trained with enough values. MAL models are trained using the first 70% observations and tested using the remaining 30% observations. The MAL models' validation based on the 30% observations can indicate their performance in the reconstruction period.

The input data are from neighboring stations and concurrent reanalysis estimates. For each s, neighboring stations are determined in a way similar with Step-2, but CC is calculated using data in the entire observation period. Neighboring stations with CC lower than all reanalysis products (ERA5, JRA-55, and MERRA-2) are excluded. The remaining neighboring stations and three reanalysis products form a complete repository of input features. Then, for each day that s has no observation, the input features are extracted from the repository in three steps: (1) neighboring stations without observations for the day are excluded, (2) the remaining neighboring stations and reanalysis products are ranked according to their CC with s, and (3) at most five stations/reanalysis products with the highest CC are selected. In this way, s will have multiple combinations of input features to ensure that all days with missing values have estimates. All combinations are used to train and test the ANN and RF models, resulting in multiple estimated series for s. The final estimates of s are generated in three steps: (1) the Kling-Gupta Efficiency (KGE'; Kling et al., 2012) of all estimated series is calculated using all observations of s, and ranked from high to low KGE' (see Sect. 3.4 for definition of KGE'); (2) the series with higher KGE' is used to constitute the estimates of s in sequence; and (3) the second step is repeated until there are no missing values for s. This approach ensures that "best" and complete estimates are provided for s.

- **MAL-1**: A four-layer ANN is used. The input layer has a maximum of five nodes (depending on the number of input features), the two hidden layers both have 20 nodes, and the output layer has one node for generating precipitation or temperature estimates. The transfer functions are hyperbolic tangent sigmoid for hidden layers and linear for the output layer. The training function is resilient backpropagation. The model is trained using the first 50% data, validated using the subsequent 20% data, and tested using the final 30% data.

- **MAL-2**: A RF model with 50 trees is built with 70% training data and 30% testing data. The minimum number of samples per tree leaf is 5. The input nodes depend on the number of input features like MAL-1.

***Group 5*: Multi-Strategy Merging (MRG)**

- **MRG-1**: KGE' is used to rank the performance of the 11 strategies (QMN-1 to 3, QMR-1 to 3, INT-1 to 3, and MAL-1 to 2) as CC cannot reflect the magnitude difference (e.g., bias) between target and reference series. The first three cases of the 11 strategies are merged using squared KGE' as the weight. The individual weight is assigned zero if KGE' is negative.

- **MRG-2**: The median of the three selected strategies in MRG-1 is used as the estimated data.

3.3.3 Generating serially complete records

Step-5: In this step, Step-3 and -4 are repeated based on 70% data of s in the observation period. Then, the KGE' of estimates from all strategies are calculated using the remaining 30% observations. MAL-1 and 2 are not repeated because they are trained on the 70% observations. Although the evaluation samples are different among stations, the results are reliable and stable as shown in the results section. This step is implemented because QMN-1 to 4, QMR-1 to 4, and INT-1 in Step-4 use all data of s in the observation period to select stations, estimate empirical CDFs and carry out regression. This potential overfitting problem could lead to better performance of these strategies in the observation period but worse performance in the reconstruction period. KGE' calculated in Step-4 can represent the accuracy of estimates in the observation period, while KGE' calculated in Step-5 can represent the accuracy of estimates in the reconstruction period.

Step-6: In the observation period, the strategy with the highest KGE' in Step-4 is selected to contribute the extension/reconstruction to the SCDNA. In the reconstruction period, first, the strategy with the highest KGE' in Step-5 is determined; then, the estimates from the corresponding strategy in Step-4 are used to constitute the SCDNA because the empirical CDF and regression based on all observations in Step-4 could be more representative than the 70% observations in Step-5.

Step-7: Estimates in Step-6 are corrected for certain climatological biases using station data in the observation period. Precipitation estimates are often subjected to wet-day bias. Two methods are implemented to address this problem. First, QM is performed based on the CDF of s in Step-3. However, QM may reduce the accuracy of estimated precipitation in some cases, for which the method used in Beck et al. (2019) is adopted. This method subtracts a tiny value (0.01 mm) from the original precipitation series and rescales the series to restore the original mean value. This operation is repeated until the estimated series show equal number of wet days (>0.5 mm d^{-1}) with observations of s. In addition to wet-day bias correction, mean-value correction is implemented. The ratio between the mean values of precipitation estimates and observations is calculated in the observation period, which is used to rescale estimated series in both observation and reconstruction periods. For T_{min} and T_{max}, QM correction and mean-value correction are also implemented.

Step-8: The accuracy of the SCDNA is evaluated and compared to benchmark datasets based on actual observations (Fig. 5). Then, the estimates are replaced by observations whenever possible to generate the final SCDNA. Very occasionally, estimated T_{min} could be larger than estimated T_{max}, for which T_{max} is replaced by the maximum T_{max}, and T_{min} is replaced by the minimum T_{min} of the estimates from the 16 strategies.

Step-9: The serially complete data of SCDNA is quality controlled again using methods introduced Sect. 3.1.2 to exclude stations with unreliable estimates.

3.4 Evaluate the precipitation and temperature estimates

KGE', which is proposed by Gupta et al. (2009) and modified by Kling et al. (2012), is used to support the merging of different strategies (Sect. 3.3) and the evaluation of the estimated precipitation and temperature~~. It is a useful metric in evaluating various variables (e.g., Tang et al., 2020) and incorporates information about correlation, bias, and variance~~.

$$\begin{cases} \mathrm{KGE}' = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \\ \beta = \dfrac{\mu_s}{\mu_0} \\ \gamma = \dfrac{CV_s}{CV_o} = \dfrac{\sigma_s/\mu_s}{\sigma_o/\mu_o} \end{cases} \tag{8}$$

where r is the PCC, β is the bias ratio, and γ is the variability ratio; μ is the mean value, and σ is the standard deviation. The subscripts s and o represent estimated and reference time series, respectively. KGE' ranges from negative infinity to one. If two series exactly match, the KGE' is one. A β or γ value smaller/larger than one indicates that the mean value or variability of observations is underestimated/overestimated.

In Sect. 4, the evaluation during the observation period is based on the complete station observations (i.e., Step-4 in Sect. 3.3.2), while the evaluation during the reconstruction period is realized using 30% independent station observations (i.e., Step-5 in Sect. 3.3.3). Unless otherwise stated, SCDNA estimates in Sect. 4 are after correction (Step-7 in Sect. 3.3.3). In Sect. 4.5, SCDNA estimates are compared with gridded products (ERA5, JRA-55, MERRA-2, and MSWEP). In addition to the three SCDNA variables (precipitation, T_{min}, and T_{max}), mean temperature ($\mathrm{T_{mean}}$, the mean of T_{min} and T_{max}) and daily temperature range (T_{range}, the difference between T_{max} and T_{min}) are also included. The involvement of T_{range} can contribute to more objective comparison between SCDNA and reanalysis products because the TLR-based downscaling of reanalysis temperature contains uncertainties, which could affect the evaluation of T_{min}, T_{max}, and T_{mean}. Although there exist differences between TLR of T_{min} and T_{max}, T_{range} can reduce the effect of scale-mismatch between gridded reanalysis temperature and point station temperature on evaluation results.

4 Results

4.1 Comparison of infilling and reconstruction strategies

The value of a given infilling/reconstruction strategy can be quantified by the extent that a strategy is selected for use in the final SCDNA dataset. In this sense the contribution ratios define the proportion of estimates that come from a specific strategy. Fig. 6 shows that the contribution ratios of QMN, QMR, and INT to missing value estimation are generally smaller than 20% in North America. Please note that QMN refers to all strategies within this group unless the strategy number is specified right after QMN. This also applies to other groups. QMR shows the smallest contribution ratios for almost all stations among the five groups. Compared with other regions in North America, contribution ratios of QMR are higher for precipitation stations in western U.S. and temperature stations in Mexico. INT shows lower contribution ratios in Rocky Mountains compared with western U.S., indicating statistical interpolation without considering topographic effect is subjected to substantial uncertainties in complex terrain. MAL shows notably higher contribution ratios than QMN, QMR, and INT, particularly for T_{min} and T_{max}. The ratios of MAL are higher than 20% for ~30% precipitation stations, ~65% T_{min} stations, and ~68% T_{max} stations. MRG has the highest contribution ratios throughout North America. The average contribution ratios of MRG are 59.88%, 41.59%, and 40.56% for precipitation, T_{min}, and T_{max}, respectively. For precipitation, MRG is particularly effective in high-latitude regions (northern Canada and Alaska), western U.S. and Mexico.

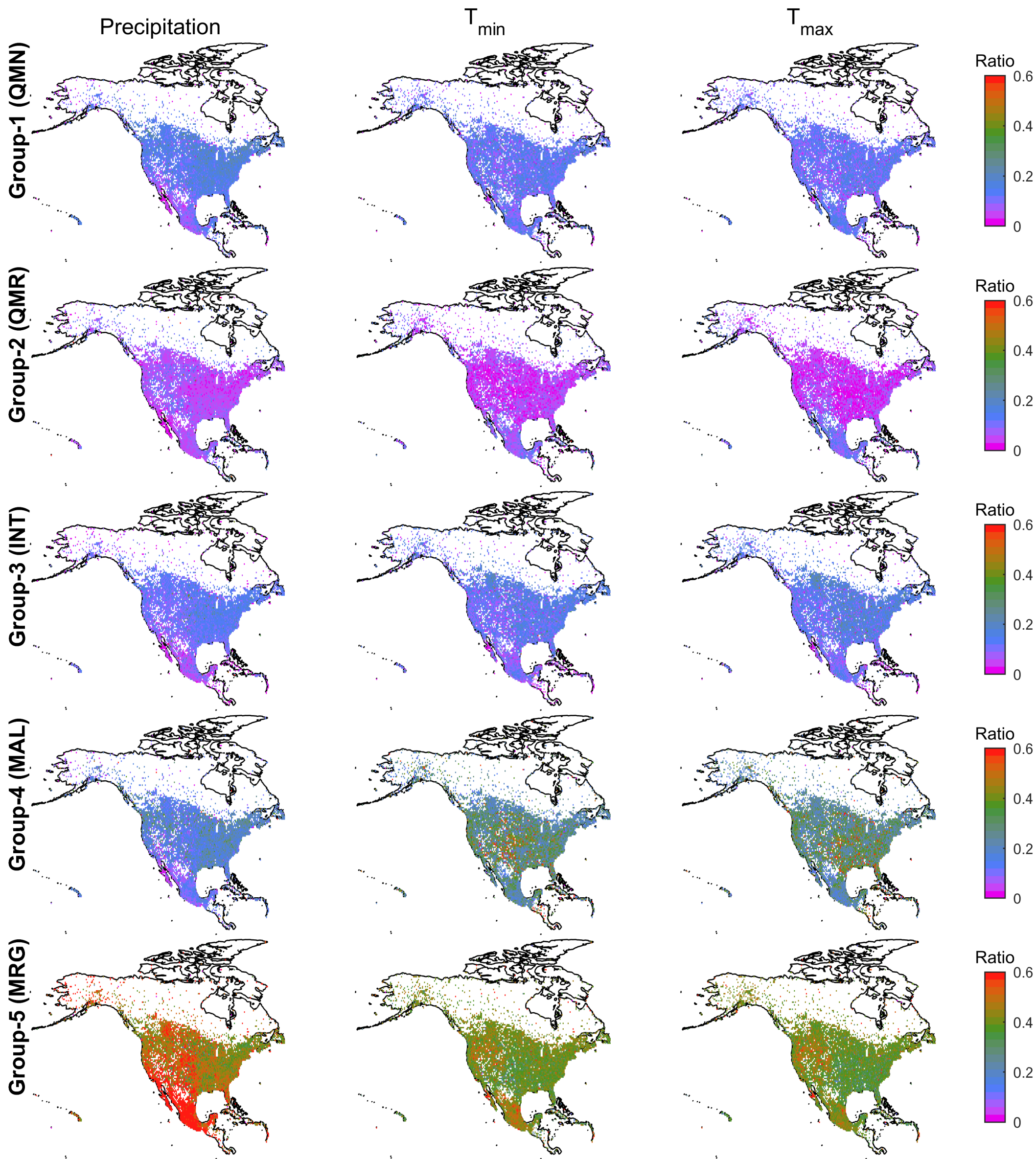


Figure 6. The contribution ratios of estimates from five infilling/reconstruction groups to the missing values of all stations from 1979 to 2018. The three columns from left to right represent precipitation, T_{min}, and T_{max}, respectively. The five rows from top to bottom represent Group-1 (QMN), Group-2 (QMR), Group-3 (INT), Group-4 (MAL), and Group-5 (MRG), respectively. The maps are at the resolution of 0.5°. The ratio for each grid cell is the mean value of all stations within this grid cell.

Figure 7 shows the KGE′ and contribution ratios of 16 strategies. The KGE′ of estimated precipitation is lower than that of estimated T_{min} and T_{max} due to the stronger spatial and temporal homogeneity of temperature (Fig. 7). The median KGE′ values of T_{min} and T_{max} are generally above 0.9, and the accuracy of estimated T_{max} is higher than that of T_{min}. The KGE′ during the reconstruction period is smaller than that during the observation period, which is particularly obvious for QMN, QMR, and INT-1 compared with other strategies, because QMN and QMR transfer CDF during the observation period to other periods, and INT-1 transfers regression relationship during the observation period to other periods. MAL suffers a slight degradation in the reconstruction period, and the better performance of MAL-2 than MAL-1 shows that RF could be a better choice than ANN in estimating missing data. For MRG, the differences of KGE' between the two periods are relatively small. For example, the median KGE' values of MRG-1 for T_{max} are 0.99 and 0.98 for observation and reconstruction periods, respectively. MRG also shows higher KGE' and a narrower quantile ranges than other strategies, particularly for precipitation, benefiting from merging estimates from multiple strategies

Regarding contribution ratios (Fig. 7), strategies with higher KGE' often have larger contributions to the estimated series. However, this is not always true because the selection of strategies is performed for each DOY. Note that the contribution ratios of MAL-2 are even higher than MRG-1 during the observation period for T_{min} and T_{max}, although MRG-1 achieves higher KGE' than MAL-2 for most stations. This is because MAL-2 could be the best choice for more DOY than MRG-1 even though MRG-1 may achieve the best overall performance. An example using T_{min} data from one station is shown in Fig. S5.

In the reconstruction period when observations are absent, the contribution ratios of MAL-2 decrease drastically compared with the observation period, contributing to the increased ratios of other strategies (particularly MRG-1). Although QMR shows the lowest contribution ratios, reanalysis products have implicit contributions to other strategies (e.g., MAL and MRG). Overall, MRG-1 shows much higher contribution ratios than all the other strategies (including MRG-2) during the reconstruction periods, indicating that it is the most important strategy in missing value estimation. Hence, combining information from multiple strategies is more reliable, and KGE'-based merging is more effective than the median-value-based estimation.

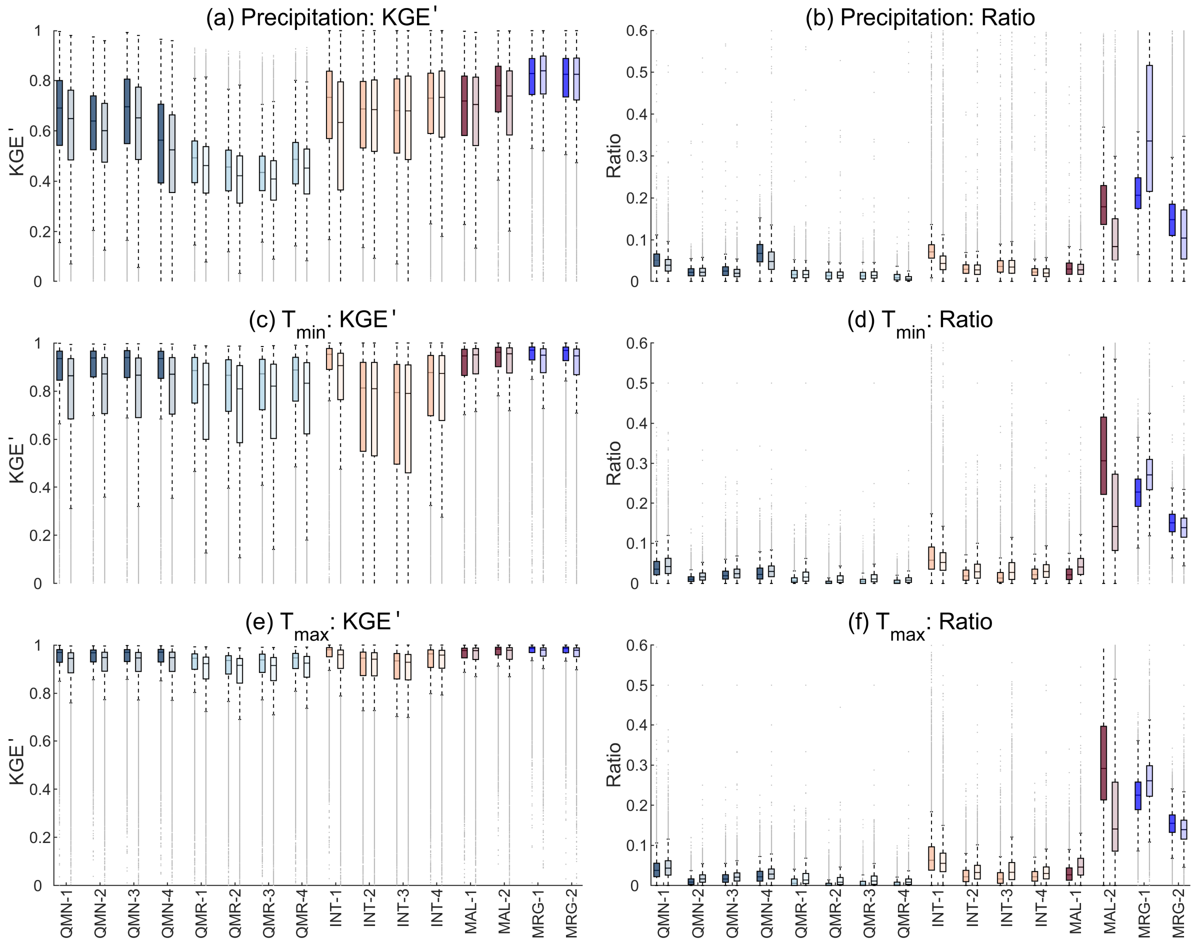


Figure 7. Boxplots of (a, c, and e) the KGE' and (b, d, and f) the contribution ratio of 16 strategies for all stations. Each strategy corresponds to two boxes in each sub-figure; the left one with darker color represents the observation period, and the right one with lighter color represents the reconstruction period. The line within the box is the median. The upper and lower edges of the box represent the 25th and 75th percentiles, respectively. Values more than 1.5 times the interquartile range away from the upper or lower edges are outliers (dots).

4.2 Impact of reconstruction on spatial correlation and series variance

All infilling/reconstruction strategies except QMR rely on information from neighboring stations; this could affect the spatial correlation structure and the variance of SCDNA series. Space-time correlations and other properties (e.g., intermittency of precipitation) are important considerations because they can influence the performance of follow-on applications that use the SCDNA as input. Theoretically, QMN strategies could significantly inflate spatial correlation but retain variance of station observations. The spatial correlation inflation in INT strategies could be lower but the variance would be underestimated due to smoothing. QMR-1 is used as an example to demonstrate the effect of QM on spatial correlation and series variance (Fig. S6), because QMN uses different station combinations for every DOY which would mask the effect of QM on final estimates. If the ERA5 used by QMR-1 is replaced by station observations, the results should be generally consistent. According to Fig. S6, the spatial correlation is substantially inflated by

QMR-1, particularly for T_{min} and T_{max}, while the standard deviation of QMR-1 estimates is very close to that of observations. This supports the design of estimating missing data using neighboring stations for each DOY as otherwise the inflation of CC could be very substantial for the entire period.

The spatial correlation based on station observations (Fig. 8a, d, and g) shows obvious seasonal variations, with CC lower in the warm season and higher in the cold season. The seasonality of CC for T_{max} is weaker compared with that for precipitation and T_{min}. The SCDNA estimates capture the seasonal patterns but underestimates the variation (Fig. 8b, e, and h) because the inflation of spatial CC is larger in the warm season than cold season (Fig. 8c, f, and i). Moreover, the inflation is larger for neighboring stations with lower correlation with the target station. We tested selecting neighboring stations according to their distance from the target station, and similar results were acquired. For precipitation, the median CC differences of all stations are close to 0.1 in the cold season and ranges between 0.1 and 0.15 in the warm season. For T_{min}, the median CC differences are generally between 0.05 and 0.15. The CC differences of T_{max} are relatively homogeneous for different seasons and generally fluctuate between 0.05 and 0.1. The inflation of CC is because (1) the estimates from the 10 neighboring stations and the target station are generally derived using highly overlapped information (Sect. 3.3.1), and (2) estimation is realized for each DOY for all strategies except MAL, meaning that calculating CC for each DOY show the inflation to the largest extent.

The final SCDNA replaces estimates by observations, which can largely relieve the inflation of spatial correlation (Fig. S7), depending on the degree to which observations are present in the record. For T_{min} and T_{max}, CC is very close to that based on observations; for precipitation, correlation in wintertime is even lower than that based on observations.

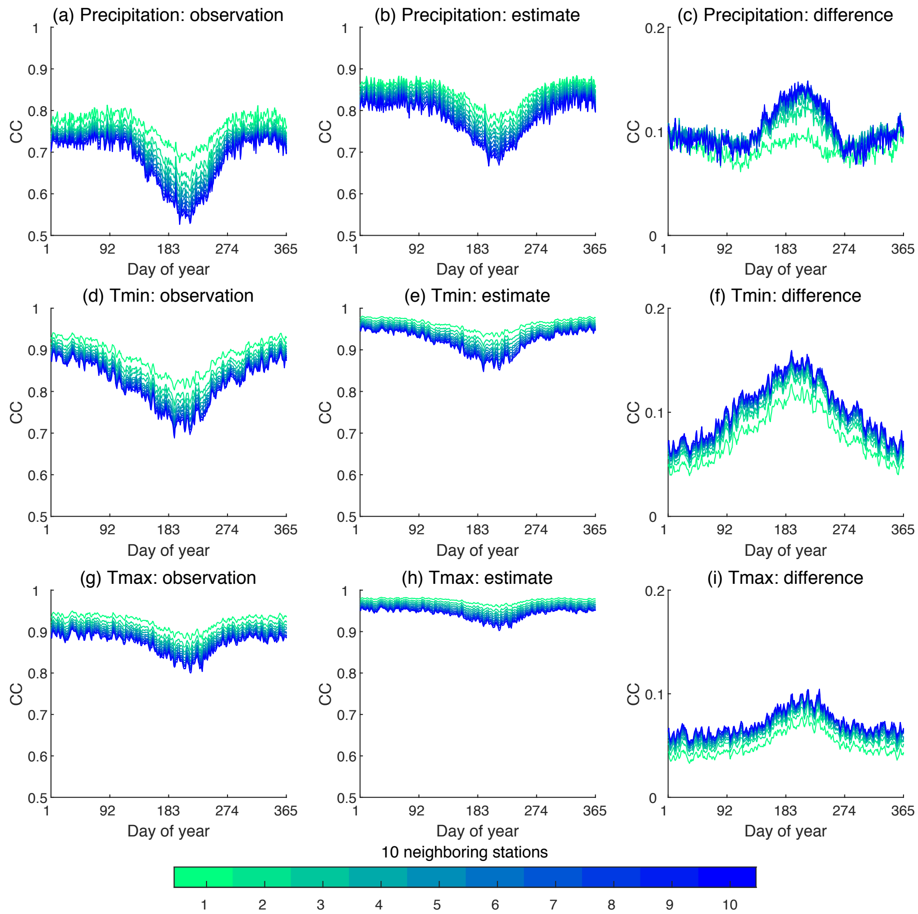


Figure 8. CC between target and neighboring stations for all DOY using station observations (the first column), SCDNA estimates (second column), and differences between SCDNA- and observation-based CC (the third column). CC is calculated in the observation period. For each target station, 10 neighboring stations are selected according to the correlation between time series from target and neighboring stations. Smaller numbers represent higher correlation. For example, station 1 represents the neighbor with the highest CC with the target station. Each curve represents the median CC of all stations.

Figures 9 and 10 show CC between estimates at the target station and observations at the neighboring station. For precipitation, most strategies exhibit similar spatial correlation structure with observations for most stations. QMR largely underestimates CC compared with observations, which should be attributed to the differences between precipitation of reanalysis products and stations. There are notable differences for different strategies within one group. For example, QMN-1 shows larger inflation when observation-based CC is higher, which is not seen in QMN-2 to 4. This is probably because QMN-1 only uses information from the one neighboring station with the highest correlation with the target station for each DOY. Higher observation-based CC in Fig. 9 means this neighboring station could be more frequently used by QMN-1 to estimate data for the target station, resulting in the larger inflation of CC. Another example is that INT-1 underestimates the CC for 68.75% stations, whereas INT-2 to 4 overestimates the CC for almost all stations. For SCD-1, inflation of CC is observed for 76.60% stations, whereas the magnitude of overestimation is smaller than that in Fig. 8. The mean values of observation-based and estimate-based CC are 0.71 and 0.77, respectively. SCD-2 replaces estimates by observations and is the final dataset of this study. It reduces the mean estimate-based CC to 0.70. The overall spatial correlation structure of observations is generally preserved by SCD-2. However, SCD-2 calculates CC for the entire period which is different from the period of observation-based CC, resulting in uncertainties such as the underestimation for some stations when observation-based CC is larger than 0.7.

The spatial correlation of T_{min} is much stronger than that of precipitation (Fig. 10). Most strategies overestimate the CC for most stations, whereas the magnitude is quite small. For example, SCD-1 inflates the CC for 96.96% stations, while the mean CC values for observations (0.95) and SCD-1 (0.96) are very close to each other. QMR still underestimates CC similar to Fig. 9 for precipitation. CC based on SCD-2 is generally consistent with that based on observations, while slight underestimation exists for some stations when observation-based CC is higher than 0.9. T_{max} shows similar spatial correlation patterns with T_{min} (Fig. S8).

In summary, inflation of CC is inevitable particularly when estimates are obtained using information from sole data source such as one neighboring station or one reanalysis product. The inflation is larger if each DOY is treated separately (Fig. 8 and S7), but smaller if CC is calculated for all years (Fig. 9, 10 and S8). Combining information from multiple sources (stations and reanalysis) and combining multiple strategies for each DOY are beneficial in estimating the overall spatial correlation structure. The spatial correlation structures vary for different strategies, and further studies are needed to clearly demonstrate how and why the estimate-based CC differs from observation-based CC.

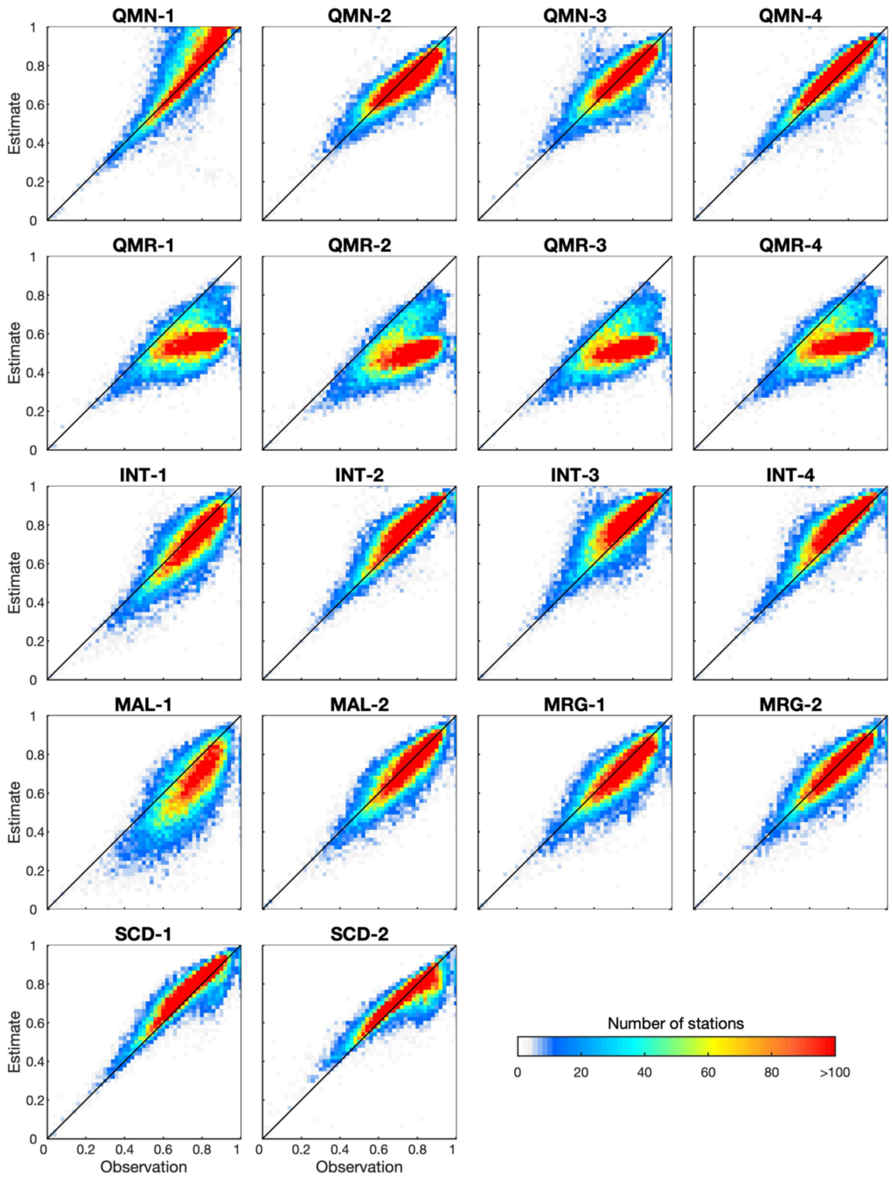


Figure 9. Scatter density plots of CC between precipitation from the target station and neighboring stations. For each target station, the neighboring station with the highest correlation with the target station is selected. X-axis represents the CC between observed precipitation from target and neighboring stations. Y-axis represents the CC between estimated precipitation from the target station and the observed precipitation from the neighboring station. Each sub-figure corresponds to one strategy in Sect. 3.3.2. SCD-1 represents SCD estimates after correction, while SCD-2 replaces estimates by observations. CC is calculated during the overlapped observation period between target and neighboring stations, and the only exception is SCD-2 which calculates CC using precipitation from target and neighboring stations during the entire period.

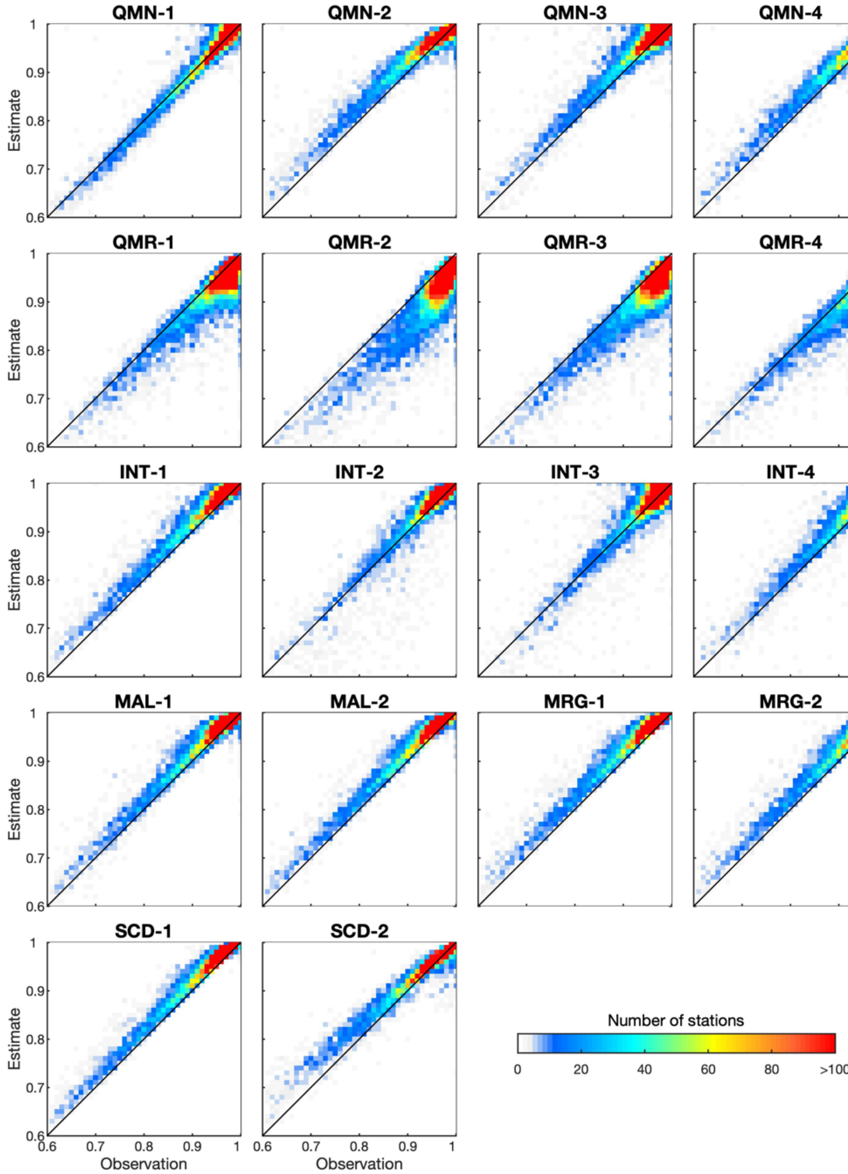


Figure 10. Similar with Fig. 9, but for T_{min}.

The variability of observations and of the corrected and uncorrected SCDNA estimates (Step-7 in Sect. 3.3.3) are compared using the standard deviation of the observation period (Fig. 11). The standard deviation of uncorrected SCDNA precipitation is lower than that of observations, while after correction, the standard deviation agrees very well with observations. The mean values of standard deviation are 7.36, 6.30, and 7.36 for observations, uncorrected SCDNA, and corrected SCDNA, respectively. For T_{min} and T_{max}, corrected and uncorrected SCDNA estimates both show consistent variability with observations.

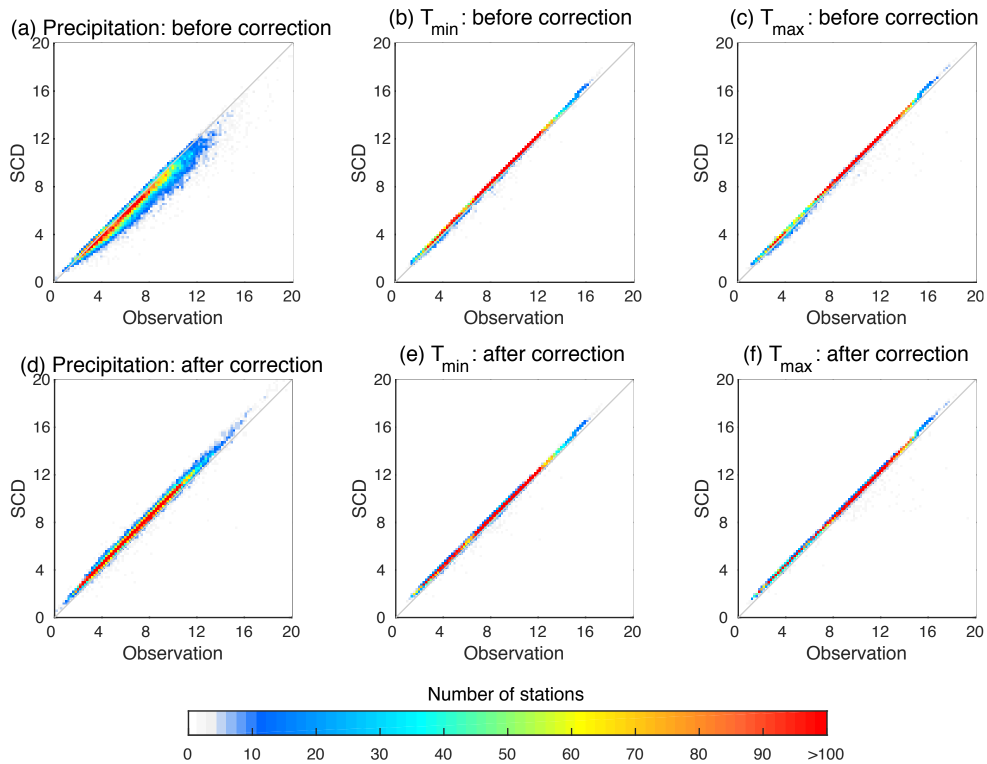


Figure 11. The standard deviation of observations and SCDNA estimates before and after correction. Data in the observation period are used.

4.3 The performance of the serially complete dataset

Uncorrected SCDNA estimates show high accuracy in North America (Fig. 12). For precipitation, the median KGE' of all stations is 0.87, and the median values of r, β, and γ are 0.91, 0.92, and 0.96, respectively. The KGE' for Mexico stations generally ranges between 0.6 and 0.8, which is smaller than that in U.S. and southern Canada. Some stations in Rocky Mountains, Caribbean, Alaska and northern Canada (regions with complex topography or climate), also show lower KGE' for precipitation estimates. The spatial distribution of r is similar with that of KGE', while the magnitude is higher. According to γ, most stations underestimate precipitation variability which is consistent with Fig. 11; β is generally lower than one in most regions of North America, particularly in Rocky Mountains and Mexico where SCDNA underestimates precipitation.

Estimated temperature shows much higher KGE' compared with precipitation. The median KGE' and r of T_{min} are 0.97 and 0.99, respectively. For T_{max}, the median of KGE' and r are 0.99 and 0.99, respectively. The median γ and β are both between 0.99 and 1 for T_{min} and T_{max} with small variations, particularly for T_{max} (Fig. 12); the KGE' of T_{min} and T_{max} is lower in Caribbean and Mexico. For T_{min}, the KGE' for some stations around 45°N and Rocky Mountains is lower than surrounding regions although γ is spatially homogeneous for the same region. This is because the mean T_{min} is close to zero for some stations in this region, resulting in the large magnitude of β and γ. In contrast, T_{max}

exhibits homogeneous performance in the same region for all metrics.~~. The discrepancies between T_{min} and T_{max} need further investigation.~~

Corrected SCDNA estimates (see Step-7; Fig. S9) have higher accuracy than uncorrected estimates (Fig. 12). For example, the median KGE' for precipitation is improved from 0.87 to 0.90 after correction. The KGE' for T_{min} and T_{max} is also improved but not as significant as precipitation. β equals to one for all stations due to the mean-value correction. γ for precipitation changes from negative to positive for all stations, whereas magnitude of bias (deviation from one) is smaller after correction. As a result, t~~T~~he spatial distribution of metrics for T_{min} is also more homogeneous. ~~Therefore, the correction procedures are effective.~~

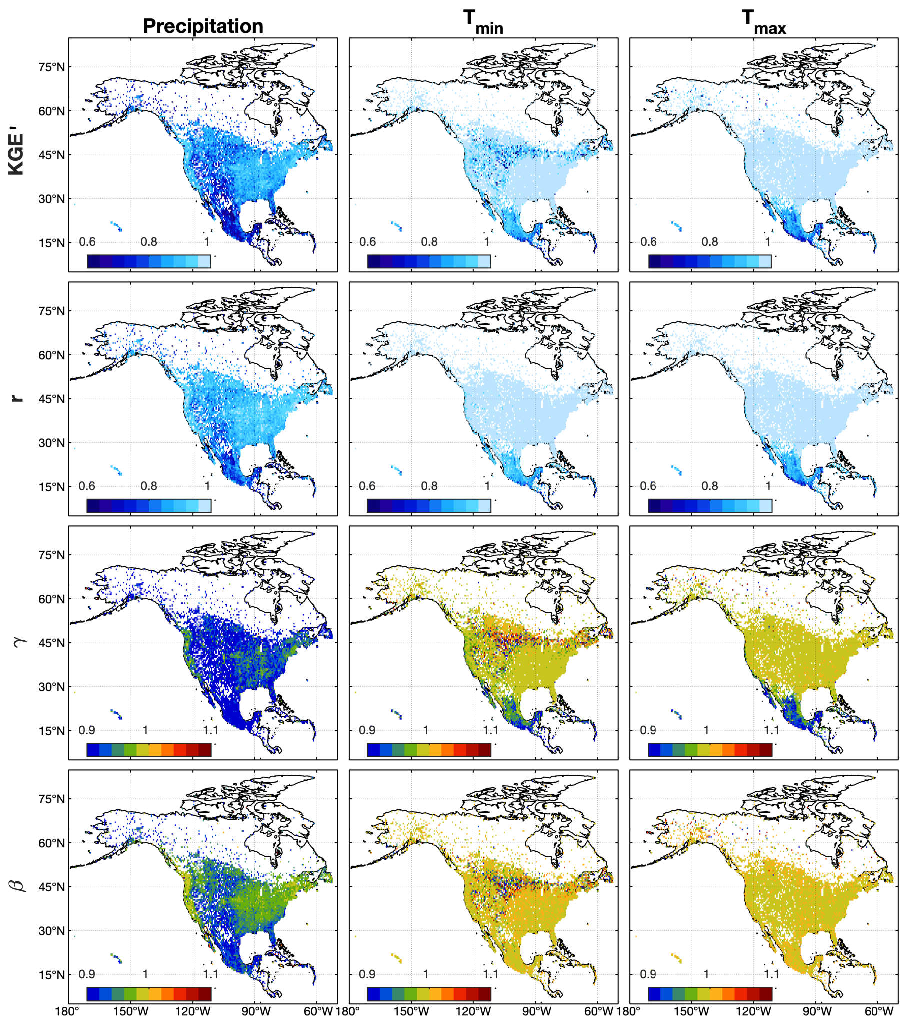


Figure 12. The spatial distributions of KGE' and its three components (r is CC, β is the bias ratio, and γ is the variability ratio) for uncorrected SCDNA estimates over North America during the observation period. The maps are at the resolution of 0.5°. The value for each grid cell is the median value of all stations within this grid cell.

The distributions of KGE' vary during the year (Fig. 13). For precipitation, more stations show lower KGE' during summer (DOY 150 to 250) than at other times of the year, which may be due to the variability of summertime convective precipitation. For T_{min}, some stations show lower KGE' from DOY 100 to 250. The seasonal variation of KGE' for T_{max} is relatively weak, although KGE' is slightly more concentred at higher level during spring and autumn than winter and summer. The overall performance of T_{max} is better than T_{min} and precipitation.

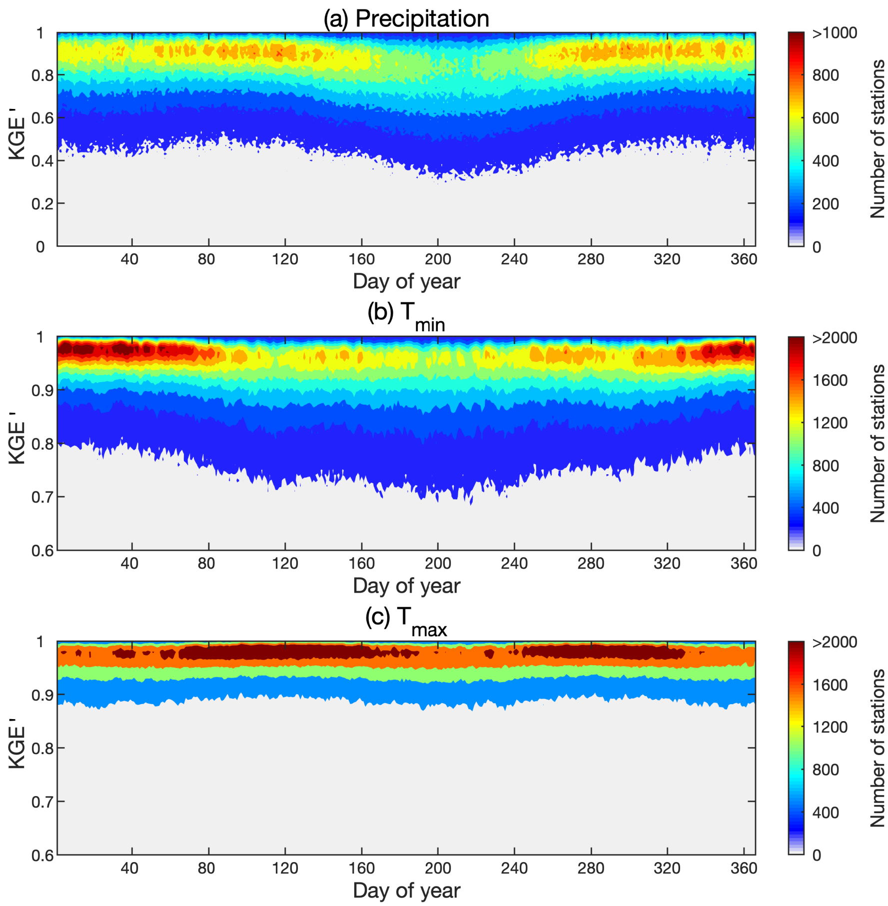


Figure 13. The distribution of KGE' for each day of year for (a) precipitation, (b) T_{min}, and (c) T_{max}. Corrected SCDNA estimates are used.

4.4 Comparison between the serially complete dataset and gridded products

SCDNA precipitation and temperature are compared with benchmark gridded products to demonstrate whether the SCDNA is a good choice when station data are unavailable. Actual station observations are used as reference. Although assessing gridded products using point-scale station data contains uncertainties (Tang et al., 2018a), the objective of this section is to illustrate their agreement with station observations in lieu of provide an exhaustive quantitative assessment of their real-world accuracy.

Overall, the SCDNA achieves much higher KGE' than reanalysis products for all variables (Fig. 14). For precipitation, the median KGE' differences between the SCDNA and ERA5, JRA-55 and MERRA-2 are 0.48, 0.57, and 0.54, respectively. The corresponding KGE' differences for T_{min} are 0.46, 0.61, and 0.36, respectively. The improvement for T_{max} is smaller, particularly in eastern U.S. where topography is relatively flatter compared with western U.S. The KGE' differences of T_{mean} are lower than T_{min} but higher than T_{max} due to the offset effect. T_{range} suffers little from the elevation differences between stations and reanalysis grids, and is suitable to demonstrate the differences between SCDNA and reanalysis products. The median KGE' differences for T_{range} between the SCDNA and ERA5, JRA-55 and MERRA-2 are 0.31, 0.48, and 0.31, respectively.

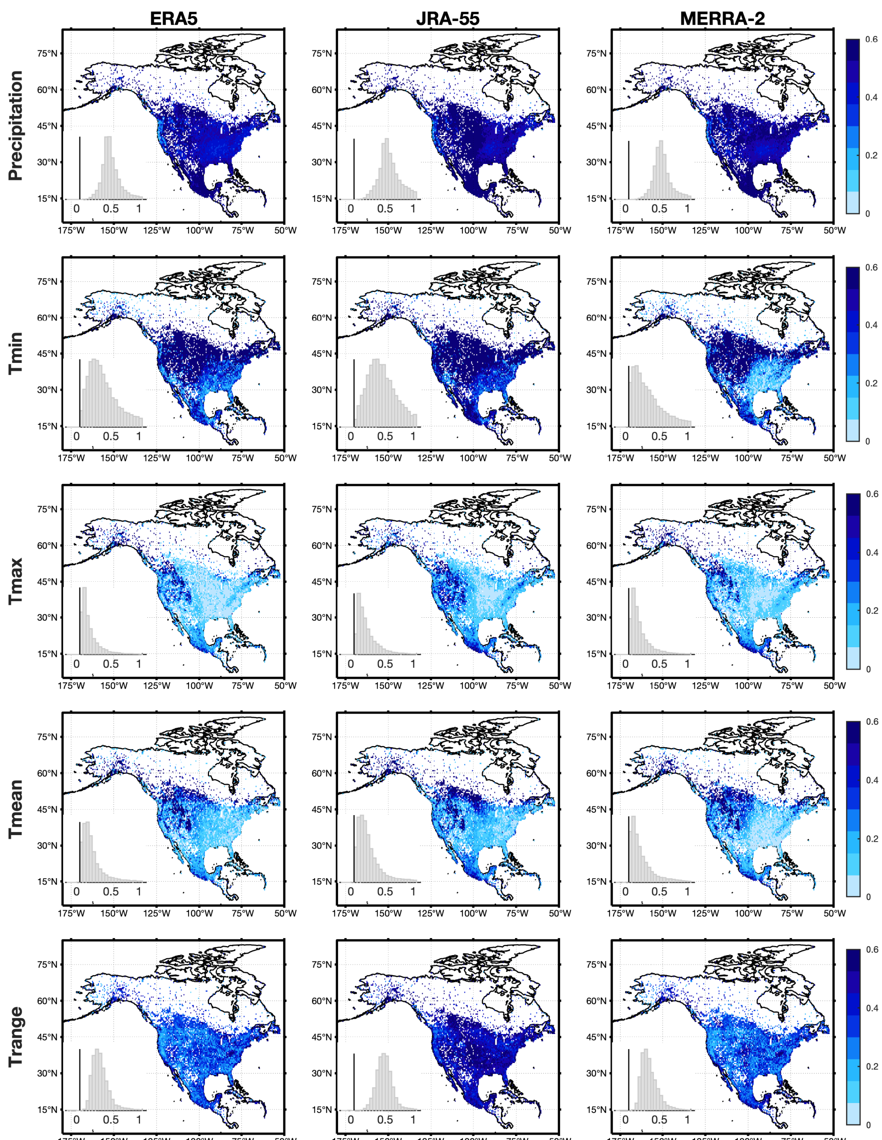


Figure 14. Spatial distributions of KGE' differences between SCDNA estimates and three reanalysis products (ERA5, JRA-55, and MERRA-2). The nested histograms show KGE' differences between the SCDNA and reanalysis products. Corrected SCDNA estimates are used.

SCDNA and MSWEP precipitation is compared (Fig. 15). Since MSWEP merges data from numerous stations, the evaluation of MSWEP based on station data is not independent, which could result in the overestimation of its KGE'. Even so, SCDNA precipitation shows higher KGE' than MSWEP for 98.97% stations with a median KGE' difference of 0.31. Fig. 15 shows notable differences between MSWEP and SCDNA at the Canada-USA border and the USA-Mexico border. This is because MSWEP infers gauge reporting time by searching for the highest correlation between gauge data and the temporally shifted reanalysis/satellite estimates (Beck et al., 2019). ~~Fig. 15 shows notable differences between Canada, U.S. and Mexico~~ The estimated temporal shift could vary with countries, which results in distinct differences of station-based evaluation results along national boundaries ~~which could be due to the differences in observation time of stations in different countries~~. The accumulation periods of station and MSWEP precipitation are inconsistent in some cases, which could affect the evaluation of MSWEP (see Sect. 5.1).

Note that the evaluation does not indicate that the SCDNA has higher accuracy than the gridded products; rather, the results show that SCDNA is a better substitute than gridded products when station observations are unavailable.

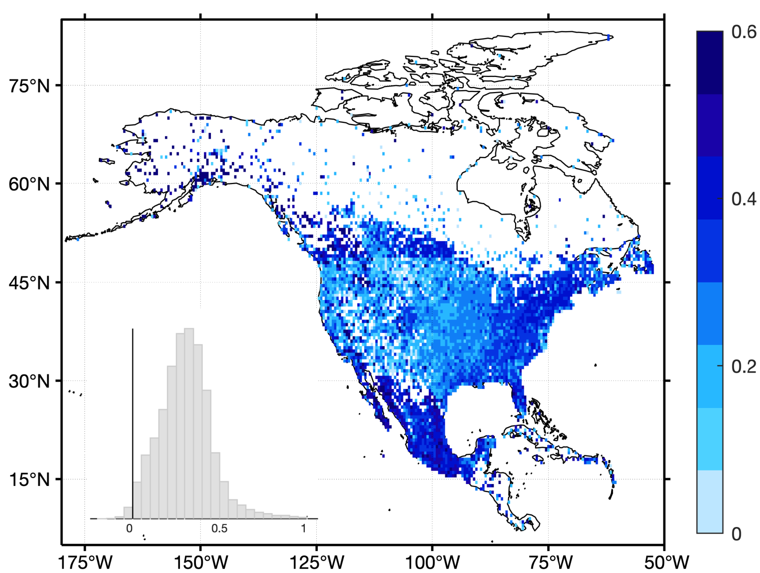


Figure 15. Spatial distributions of KGE' differences between SCDNA and MSWEP precipitation. Corrected SCDNA estimates are used.

5. Discussion

5.1 Observation time of stations

Meteorological stations in different countries usually have different local observation time, and stations in the same country may also experience change of observation time (Vincent et al., 2012). Most station databases including those used in this study do not account for reporting-time inconsistencies due to lack of hourly observations and well-documented station metadata. Vincent et al. (2009) examined several methods to adjust the time of daily precipitation observations, which, however, often altered observed precipitation intensity. Beck et al. (2019) inferred the reporting time of daily precipitation observations by calculating SCC between the series of stations and gridded products, which is useful to correct the bias of gridded products. A simple experiment is carried out using the method of Beck et al. (2019) to infer the lag day of station series. For precipitation, 6418 stations show nonnegligible time shift from the reporting date (Fig. S10). However, this method may be unsuitable for temperature because the estimated lag day is mostly zero, and the inferred reporting time cannot be directly applied to adjust station observations.

The inconsistent reporting time has different impact on precipitation, T_{min}, and T_{max}. For example, if a station records data from 8:00 a.m. on January 1st to 8:00 a.m. on January 2nd, the station will probably use January 2nd as the reporting time. However, two thirds of the 24-h time are within January 1st, indicating that the accumulated precipitation could mostly occur on January 1st. T_{max} could also occur during the daytime on January 1st, but it is hard to determine on which day T_{min} occurs, which makes it challenging to adjust precipitation, T_{min} and T_{max} at the same time. The difference between universal and local time makes this problem more complicated. Thus, the reporting time of stations is not corrected here due to aforementioned difficulties.

5.2 Homogenization

Inhomogeneities in station observations are defined as variations that are not caused by weather and climate factors. Long-term station records are often subjected to inhomogeneities due to factors like station re-location, observation time change, instrument change, and surrounding environment change (Venema et al., 2012). Many methods have been developed to identify breakpoints and homogenize station series in annual, monthly or even daily scales (e.g., Ma et al., 2008; Vincent et al., 2002, 2012). Different methods could generate different estimates of inhomogeneities as shown by many comparison studies (e.g., Beaulieu et al., 2008; Reeves et al., 2007; Venema et al., 2012). The four station databases (Sect. 2.1) used in this study provide original station records without homogenization. The SCDNA would inherit the potential inhomogeneities contained in these databases, and the infilling/reconstruction may also lead to discontinuities. The homogenization of the SCDNA is challenging considering that (1) the dataset covers a broad range of climate, topography, and countries, (2) the number of stations is large and differences between station periods (ranging from 8 to 40 years) are substantial, and (3) whether existing methods are suitable for homogenization of infilling/reconstruction estimates needs exploration. Therefore, homogenization is not carried out in this study, which, however, is an important direction of future studies.

5.3 Limitations of the KGE' statistic

We use KGE' because it incorporates information about correlation, bias, and variability, and hence provides more information on methodological performance than an individual metric. For example, the PCC between temperature estimates and observations is usually close to one and cannot reflect the bias term, while the mean square error is prone to the effect of extreme values (or outliers). However, KGE' also has limitations. For example, the values of KGE' depend on the units of measurement (e.g., Santos et al., 2018) – in our case, the β values for temperature are clearly always close to one if the units of measurement for temperature are in Kelvin. Since these statistics incorrectly indicate very small temperature biases, we used °C for all KGE' calculations in this study, ensuring that β has more leverage in the KGE' statistic. Moreover, and critical for our analysis, the normalization used in the KGE' formula (β and γ) means that the KGE' values are low when the denominators of β and γ are close to zero (e.g., Santos et al., 2018). This problem is especially acute for temperature – for instance, we found that KGE' values were very small for cases where μ_o is close to zero. Nevertheless, the number of cases where μ_o is close to zero is rather small, where ~0.5% of all cases (based on all stations and all DOY) show absolute values of mean T_{min} smaller than 0.1°C. For cases with μ_o close to zero, the ranking based on KGE' is similar to the ranking based on mean absolute error, which means that KGE' can still function as a ranking indicator when its value is low. Further work is needed to both comprehensively evaluate the alternative infilling strategies presented in this paper and evaluate more advanced multi-method merging strategies.

5.4~~3~~ Potential improvement directions

Several steps could be taken to improve the SCDNA. First, the optimal strategy could be different for each station as shown by the results in this study. Therefore, the quality of SCDNA may be further improved by using more infilling/reconstruction methods, which would yield diminishing returns at some point. For example, the long short-term memory (LSTM) could be suitable to impute missing station observations. Optimizing the configuration of various strategies will be necessary to balance computation efficiency and estimation accuracy, particularly when the number of stations is large. Second, some stations suffer from undercatch, which depends on gauge type, precipitation phase, environmental conditions, etc. The bias caused by undercatch can be substantial for stations located in high latitudes and in the mountains (Yang et al., 2005; Scaff et al., 2015). Third, the SCDNA does not distinguish between rainfall and snowfall. Considering that a large part of North America has frequent snowfall in winter, precipitation phase classification will be useful for hydrometeorological studies. Auxiliary data from reanalysis and satellite products could be used to partition precipitation into rain and snow. Finally, although the SCDNA agrees well with station observations, long-term trends are difficult to reconstruct when actual observations are unavailable, meaning the SCDNA may not be suitable for climate trend analysis in the reconstruction period. Some gridded datasets use only stations with long-term records (e.g., (Wood, 2008; Werner et al., 2019) to achieve temporally consistent estimates, whereas such stations are very few. Reasonable trend estimation is challenging but meaningful for SCD.

Furthermore, other variables such as wind and humidity observed by stations also suffer from the same problems faced by precipitation and temperature. Future studies should explore whether the current methodology is applicable to other variables. A SCD covering more variables would be useful for research in various fields.

6 Data availability

The SCDNA dataset is available at https://doi.org/10.5281/zenodo.3735533~~https://doi.org/10.5281/zenodo.3735534~~ (Tang et al., 2020) in netCDF format. The basic variables are station identification, latitude, longitude, elevation, date, and TLR derived in Sect. 3.2. Stations that undergo location merging (Sect. 3.1.1) are identified and all relevant stations are included in the data file. For precipitation, T_{min}, and T_{max}, the variables in the netCDF4 file include original station observations, quality flags provided by original station databases, quality flags provided by this study, estimates from 16 strategies, uncorrected SCDNA estimates, corrected SCDNA estimates, the final SCDNA with estimates replaced by observations, data source flags indicating the source of each record in SCDNA (observations or 16 strategies), and accuracy metrics (KGE' and its three components) for all estimates (16 strategies and SCDNA).

Scripts used to produce the SCDNA are available at https://github.com/tgq14/GapFill. The dataset will be regularly updated to cover latest periods.

7 Conclusions

This study developed a daily SCD of precipitation, T_{min}, and T_{max} for ~~27280~~27276 stations from 1979 to 2018 over North America (SCDNA). The original station data are compiled from multiple sources and undergo strict quality control. Many stations have nonnegligible fractions of missing values in observation and reconstruction periods. For each station, the infilling and reconstruction are implemented using 16 strategies (quantile mapping, statistical interpolation, and machine learning) based on information from neighboring stations and concurrent reanalysis estimates (ERA5, JRA-55, and MERRA-2). The final SCDNA combines estimates from the 16 strategies and is corrected using station observations. The spatial correlation is preserved and might be slightly inflated. The SCDNA estimates reproduce the variance of original station observations very well, particularly for temperature. The median KGE' of the final precipitation, T_{min}, and T_{max} for all stations is 0.90, 0.98, and 0.99, respectively. The comparison with four benchmark gridded products shows that the SCDNA has much better agreement with station observations. The SCDNA will be useful for a variety of hydrometeorological studies in North America.

Author contributions: GT and MC designed the study. GT performed the analyses and wrote the paper. All authors contributed to data analysis, discussions about the methods and results, and paper improvement.

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

Acknowledgements: The study is funded by the Global Water Futures (GWF) program in Canada. The authors appreciate the extensive efforts from the developers of the ground and reanalysis datasets to make their products available. The authors also thank Zenodo (https://zenodo.org/) for publishing our dataset as open access to users.

Appendix A

Table A1. Acronyms used in this paper

Acronym	Full name
ANN	Artificial neural network
APHRODITE	Asian Precipitation-Highly-Resolved Observational Data Integration Towards Evaluation
CC	Correlation coefficient
CDF	Cumulative distribution function
CONUS	Contiguous United States
DEM	Digital elevation model
DOY	Day of year
ECCC	Environment and Climate Change Canada
ERA5	the fifth generation of ECMWF atmospheric reanalyses of the global climate
fD	Fraction of days without precipitation
GHCN-D	Global Historical Climate Network Daily
GSOD	Global Surface Summary of the Day
IDW	Inverse distance weighting
INT	Interpolation
JRA-55	Japanese 55-year Reanalysis
KGE'	Kling-Gupta Efficiency
LSTM	Long short-term memory
MAL	Machine learning
MLAD	Multiple regression based on the least absolute deviation criteria
MERIT DEM	Multi-Error-Removed Improved-Terrain digital elevation model
MERRA-2	Modern-Era Retrospective analysis for Research and Applications, Version 2
MRG	Multi-strategy merging
MSWEP	Multi-Source Weighted-Ensemble Precipitation
NR	Revised normal ratio
PCC	Pearson CC
QM	Quantile mapping
QMN	QM using neighboring stations
QMR	Quantile mapping with concurrent reanalysis estimates
RF	Random forest
SCC	Spearman CC

SCDs	Serially complete datasets
TLR	Temperature lapse rate
T_{max}	Maximum temperature
T_{mean}	Mean temperature
T_{min}	Minimum temperature
T_{range}	Daily temperature range
U.S.	United States
UTC	Universal Time Coordinated

Appendix B

Five types of checks (Durre et al., 2010) are adopted for the quality control of temperature.

1. Integrity checks. The first type of integrity check is *a duplication check* to identify duplicated records for time series in different time periods. The second type of integrity check includes *the streak check* to identify consecutive identical values and *the frequent-value check* to identify close but not necessarily consecutive identical values. The *world record exceedance check* sets lower (-89.4°C) and upper (57.7°C) bounds of temperature.

2. Outlier checks, including *the gap check* that examines the frequency distributions for all calendar months, and the *climatological outlier check* that is based on the traditional z-score (e.g., Hubbard and You, 2005).

3. Internal and temporal consistency checks, including *the iterative temperature consistency check*, to ensure some inherent relationships are abided (e.g., T_{min} cannot be larger than T_{max}); *the spike/dip check*, identifies temperatures which deviate from previous and following days by at least 25°C; and *the lagged temperature range check*, which identifies abnormally large differences between T_{min} and T_{max} during a 3-day time window.

4. Spatial consistency checks, including *the regression check* and *the spatial corroboration check. The regression check* builds regression relationships between temperature at the target location and selected nearby stations to determine whether temperature at the target station should be flagged according to regression residuals and standardized residuals. *The spatial corroboration check* flags temperature at the target station if the value deviates far from the temperature at neighboring stations.

5. Extreme megaconsistency checks to ensure that certain relationships hold for the entire records of stations. For example, T_{max} cannot be higher than the lowest T_{min} for the calendar month, and vice versa.

For precipitation, quality control strategies are from three studies. The first part is similar with temperature, but does not include the third type of checks (internal and temporal consistency checks). The second part is from Hamada et al. (2011).

1. Repetition checks. The non-zero check identifies constant daily values (> 10 mm d^{-1}) that occur for more than four days. The zero check compares the annual zero-precipitation frequency with its climatological value to spot unusual frequencies of zero.

2. Duplicated monthly or sub-monthly record check. The temporal CC and the number of days with equal precipitation are used to identify whether two different months have the same records caused by human errors.

3. Z-score-based outlier check. Daily precipitation is flagged if its difference with the mean value from precipitation within a 15-day window of all years is larger than nine standard deviations. This step is repeated until no outlier is identified.

4. Spatiotemporally isolated value check. Extremely large precipitation is identified in both space and time based on the percentiles of precipitation differences between the target station and neighboring stations within a radius of 400 km.

The third part is from Beck et al. (2019).

1. Empirical criterion based on the fraction of days without precipitation (*f*D). This was designed to identify the long series of erroneous zero precipitation contained in GSOD station records. However, we found that this criterion misidentifies some acceptable records in GHCN-D. Therefore, the *f*D-based check is only implemented for GSOD.

2. Discarding stations with fewer than 15 unique values or more than 99.5% dry records (<0.5 mm d^{-1}).

References

Alexander, L. V., Zhang, X., Peterson, T. C., Caesar, J., Gleason, B., Tank, A. M. G. K., Haylock, M., Collins, D., Trewin, B., Rahimzadeh, F., Tagipour, A., Kumar, K. R., Revadekar, J., Griffiths, G., Vincent, L., Stephenson, D. B., Burn, J., Aguilar, E., Brunet, M., Taylor, M., New, M., Zhai, P., Rusticucci, M. and Vazquez-Aguirre, J. L.: Global observed changes in daily climate extremes of temperature and precipitation, J. Geophys. Res. Atmospheres, 111(D5), doi:10.1029/2005JD006290, 2006.

Anderson, B. T., Wang, J., Salvucci, G., Gopal, S. and Islam, S.: Observed Trends in Summertime Precipitation over the Southwestern United States, J. Clim., 23(7), 1937–1944, doi:10.1175/2009JCLI3317.1, 2009.

Beaulieu, C., Seidou, O., Ouarda, T. B. M. J., Zhang, X., Boulet, G. and Yagouti, A.: Intercomparison of homogenization techniques for precipitation data, Water Resour. Res., 44(2), doi:10.1029/2006WR005615, 2008.

Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B. and de Roo, A.: MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data, Hydrol. Earth Syst. Sci., 21(1), 589–615, doi:10.5194/hess-21-589-2017, 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(3), 473–500, doi:10.1175/BAMS-D-17-0138.1, 2019.

Breiman, L.: Random Forests, Mach. Learn., 45(1), 5–32, doi:10.1023/A:1010933404324, 2001.

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(17), 6938–6959, doi:10.1175/JCLI-D-14-00754.1, 2015.

Che Ghani, N. Z., Abu Hasan, Z. and Tze Liang, L.: Estimation of Missing Rainfall Data Using GEP: Case Study of Raja River, Alor Setar, Kedah, Adv. Artif. Intell., doi:10.1155/2014/716398, 2014.

Copernicus Climate Change Service (C3S): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, Copernic. Clim. Change Serv. Clim. Data Store CDS July 2019 Httpscdsclimatecopernicuseucdsapphome, 2017.

Coulibaly, P. and Evora, N. D.: Comparison of neural network methods for infilling missing daily weather records, J. Hydrol., 341(1), 27–41, doi:10.1016/j.jhydrol.2007.04.020, 2007.

Dastorani, M. T., Moghadamnia, A., Piri, J. and Rico-Ramirez, M.: Application of ANN and ANFIS models for reconstructing missing flow data, Environ. Monit. Assess., 166(1), 421–434, doi:10.1007/s10661-009-1012-8, 2010.

Devi, U., Shekhar, M. S., Singh, G. P., Rao, N. N. and Bhatt, U. S.: Methodological application of quantile mapping to generate precipitation data over Northwest Himalaya, Int. J. Climatol., 39(7), 3160–3170, doi:10.1002/joc.6008, 2019.

Devine, K. A. and Mekis, E.: Field accuracy of Canadian rain measurements, Atmosphere-Ocean, 46(2), 213–227, 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(2), 475–497, doi:10.1175/2007JAMC1356.1, 2008.

Di Piazza, A., Conti, F. L., Noto, L. V., Viola, F. and La Loggia, G.: Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy, Int. J. Appl. Earth Obs. Geoinformation, 13(3), 396–408, doi:10.1016/j.jag.2011.01.005, 2011.

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G. and Vose, R. S.: Comprehensive Automated Quality Assurance of Daily Surface Observations, J. Appl. Meteorol. Climatol., 49(8), 1615–1633, doi:10.1175/2010JAMC2375.1, 2010.

Eischeid, J. K., Bruce Baker, C., Karl, T. R. and Diaz, H. F.: The Quality Control of Long-Term Climatological Data Using Objective Data Analysis, J. Appl. Meteorol., 34(12), 2787–2795, doi:10.1175/1520-0450(1995)034<2787:TQCOLT>2.0.CO;2, 1995.

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(9), 1580–1591, doi:10.1175/1520-0450(2000)039<1580:CASCND>2.0.CO;2, 2000.

Gao, L., Bernhardt, M. and Schulz, K.: Elevation correction of ERA-Interim temperature data in complex terrain, Hydrol. Earth Syst. Sci., 16(12), 4661–4673, doi:10.5194/hess-16-4661-2012, 2012.

Gao, L., Wei, J., Wang, L., Bernhardt, M., Schulz, K. and Chen, X.: A high-resolution air temperature data set for the Chinese Tian Shan in 1979–2016, Earth Syst. Sci. Data, 10(4), 2097–2114, doi:10.5194/essd-10-2097-2018, 2018.

Gardner, A. S., Sharp, M. J., Koerner, R. M., Labine, C., Boon, S., Marshall, S. J., Burgess, D. O. and Lewis, D.: Near-Surface Temperature Lapse Rates over Arctic Glaciers and Their Implications for Temperature Downscaling, J. Clim., 22(16), 4281–4298, doi:10.1175/2009jcli2845.1, 2009.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A.

M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M. and Zhao, B.: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), J. Clim., 30(14), 5419–5454, doi:10.1175/jcli-d-16-0758.1, 2017.

Gruber, S.: Derivation and analysis of a high-resolution estimate of global permafrost zonation, The Cryosphere, 6(1), 221–233, doi:10.5194/tc-6-221-2012, 2012.

Gupta, H. V., Kling, H., Yilmaz, K. K. and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377(1–2), 80–91, 2009.

Hamada, A., Arakawa, O. and Yatagai, A.: An Automated Quality Control Method for Daily Rain-gauge Data, Glob. Environ. Res., 15(2), 183–192, 2011.

Hasanpour Kashani, M. and Dinpashoh, Y.: Evaluation of efficiency of different estimation methods for missing climatological data, Stoch. Environ. Res. Risk Assess., 26(1), 59–71, doi:10.1007/s00477-011-0536-y, 2012.

Hubbard, K. G. and You, J.: Sensitivity Analysis of Quality Assurance Using the Spatial Regression Approach—A Case Study of the Maximum/Minimum Air Temperature, J. Atmospheric Ocean. Technol., 22(10), 1520–1530, doi:10.1175/JTECH1790.1, 2005.

Kanda, N., Negi, H. S., Rishi, M. S. and Shekhar, M. S.: Performance of various techniques in estimating missing climatological data over snowbound mountainous areas of Karakoram Himalaya, Meteorol. Appl., 25(3), 337–349, doi:10.1002/met.1699, 2018.

Kemp, W. P., Burnell, D. G., Everson, D. O. and Thomson, A. J.: Estimating Missing Daily Maximum and Minimum Temperatures, J. Clim. Appl. Meteorol., 22(9), 1587–1593, doi:10.1175/1520-0450(1983)022<1587:EMDMAM>2.0.CO;2, 1983.

Kenawy, A. E., López-Moreno, J. I., Stepanek, P. and Vicente-Serrano, S. M.: An assessment of the role of homogenization protocol in the performance of daily temperature series and trends: application to northeastern Spain, Int. J. Climatol., 33(1), 87–108, doi:10.1002/joc.3410, 2013.

Kling, H., Fuchs, M. and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, J. Hydrol., 424, 264–277, 2012.

Knowles, N., Dettinger, M. D. and Cayan, D. R.: Trends in Snowfall versus Rainfall in the Western United States, J. Clim., 19(18), 4545–4559, doi:10.1175/JCLI3850.1, 2006.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K. and Takahashi, K.: The JRA-55 Reanalysis: General Specifications and Basic Characteristics, J. Meteorol. Soc. Jpn. Ser II, 93(1), 5–48, doi:10.2151/jmsj.2015-001, 2015.

Livneh, B., Bohn, T. J., Pierce, D. W., Munoz-Arriola, F., Nijssen, B., Vose, R., Cayan, D. R. and Brekke, L.: A spatially comprehensive, hydrometeorological data set for Mexico, the U.S., and Southern Canada 1950–2013, Sci. Data, 2(1), 150042, doi:10.1038/sdata.2015.42, 2015.

Longman, R. J., Frazier, A. G., Newman, A. J., Giambelluca, T. W., Schanzenbach, D., Kagawa-Viviani, A., Needham, H., Arnold, J. R. and Clark, M. P.: High-Resolution Gridded Daily Rainfall and Temperature for the Hawaiian Islands (1990–2014), J. Hydrometeorol., 20(3), 489–508, doi:10.1175/JHM-D-18-0112.1, 2019.

Ma, C., Fassnacht, S. R. and Kampf, S. K.: How Temperature Sensor Change Affects Warming Trends and Modeling: An Evaluation Across the State of Colorado, Water Resour. Res., 55(11), 9748–9764, 2019.

Ma, L., Zhang, T., Li, Q., Frauenfeld, O. W. and Qin, D.: Evaluation of ERA-40, NCEP-1, and NCEP-2 reanalysis air temperatures with ground-based measurements in China, J. Geophys. Res., 113(D15), doi:10.1029/2007jd009549, 2008.

Ma, Y., Hong, Y., Chen, Y., Yang, Y., Tang, G., Yao, Y., Long, D., Li, C., Han, Z. and Liu, R.: Performance of Optimally Merged Multisatellite Precipitation Products Using the Dynamic Bayesian Model Averaging Scheme Over the Tibetan Plateau, J. Geophys. Res. Atmospheres, 123(2), 814–834, doi:10.1002/2017jd026648, 2018.

Maraun, D.: Bias Correction, Quantile Mapping, and Downscaling: Revisiting the Inflation Issue, J. Clim., 26(6), 2137–2143, doi:10.1175/JCLI-D-12-00821.1, 2013.

Marshall, S. J., Sharp, M. J., Burgess, D. O. and Anslow, F. S.: Near-surface-temperature lapse rates on the Prince of Wales Icefield, Ellesmere Island, Canada: implications for regional downscaling of temperature, Int. J. Climatol., 27(3), 385–398, doi:10.1002/joc.1396, 2007.

Mekis, É. and Brown, R.: Derivation of an adjustment factor map for the estimation of the water equivalent of snowfall from ruler measurements in Canada, Atmosphere-Ocean, 48(4), 284–293, doi:10.3137/AO1104.2010, 2010.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E. and Houston, T. G.: An overview of the global historical climatology network-daily database, J. Atmospheric Ocean. Technol., doi:10.1175/JTECH-D-11-00103.1, 2012.

Metcalfe, J. R., Routledge, B. and Devine, K.: Rainfall measurement in Canada: changing observational methodsand archive adjustment procedures, J. Clim., 10(1), 92–101, 1997.

Minder, J. R., Mote, P. W. and Lundquist, J. D.: Surface temperature lapse rates over complex terrain: Lessons from the Cascade Mountains, J. Geophys. Res., 115(D14), doi:10.1029/2009jd013493, 2010.

Newman, A. J., Clark, M. P., Craig, J., Nijssen, B., Wood, A., Gutmann, E., Mizukami, N., Brekke, L. and Arnold, J. R.: Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous United States, J. Hydrometeorol., 16(6), 2481–2500, doi:10.1175/JHM-D-15-0026.1, 2015.

Newman, A. J., Clark, M. P., Longman, R. J., Gilleland, E., Giambelluca, T. W. and Arnold, J. R.: Use of Daily Station Observations to Produce High-Resolution Gridded Probabilistic Precipitation and Temperature Time Series for the Hawaiian Islands, J. Hydrometeorol., 20(3), 509–529, doi:10.1175/JHM-D-18-0113.1, 2019.

Papalexiou, S. M. and Montanari, A.: Global and regional increase of precipitation extremes under global warming, Water Resour. Res., 55(6), 4901–4914, 2019.

Papalexiou, S. M., AghaKouchak, A., Trenberth, K. E. and Foufoula-Georgiou, E.: Global, regional, and megacity trends in the highest temperature of the year: Diagnostics and evidence for accelerating trends, Earths Future, 6(1), 71–79, 2018.

Pappas, C., Papalexiou, S. M. and Koutsoyiannis, D.: A quick gap filling of missing hydrometeorological data, J. Geophys. Res. Atmospheres, 119(15), 9290–9300, 2014.

Paulhus, J. L. H. and Kohler, M. A.: Interpolation of missing precipitation records, Mon. Weather Rev., 80(8), 129–133, doi:10.1175/1520-0493(1952)080<0129:IOMPR>2.0.CO;2, 1952.

Pielke Sr, R., Nielsen-Gammon, J., Davey, C., Angel, J., Bliss, O., Doesken, N., Cai, M., Fall, S., Niyogi, D. and Gallo, K.: Documentation of uncertainties and biases associated with surface temperature measurement sites for climate change assessment, Bull. Am. Meteorol. Soc., 88(6), 913–928, 2007.

Ramos-Calzado, P., Gómez-Camacho, J., Pérez-Bernal, F. and Pita-López, M. F.: A novel approach to precipitation series completion in climatological datasets: application to Andalusia, Int. J. Climatol., 28(11), 1525–1534, doi:10.1002/joc.1657, 2008.

Reeves, J., Chen, J., Wang, X. L., Lund, R. and Lu, Q. Q.: A Review and Comparison of Changepoint Detection Techniques for Climate Data, J. Appl. Meteorol. Climatol., 46(6), 900–915, doi:10.1175/JAM2493.1, 2007.

Rubin, D. B.: Inference and missing data, Biometrika, 63(3), 581–592, doi:10.1093/biomet/63.3.581, 1976.

Santos, L., Thirel, G. and Perrin, C.: Technical note: Pitfalls in using log-transformed flows within the KGE criterion, Hydrol. Earth Syst. Sci., 22, 4583–4591, doi:10.5194/hess-22-4583-2018, 2018.

Sattari, M.-T., Rezazadeh-Joudi, A. and Kusiak, A.: Assessment of different methods for estimation of missing data in precipitation studies, Hydrol. Res., 48(4), 1032–1044, doi:10.2166/nh.2016.364, 2017.

Scaff, L., Yang, D., Li, Y. and Mekis, E.: Inconsistency in precipitation measurements across the Alaska–Yukon border, The Cryosphere, 9(6), 2417–2428, doi:10.5194/tc-9-2417-2015, 2015.

Serrano-Notivoli, R., Beguería, S. and Luis, M. de: STEAD: a high-resolution daily gridded temperature dataset for Spain, Earth Syst. Sci. Data, 11(3), 1171–1188, doi:https://doi.org/10.5194/essd-11-1171-2019, 2019.

Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data, in Proceedings of the 1968 23rd ACM national conference, pp. 517–524, ACM., 1968.

Simolo, C., Brunetti, M., Maugeri, M. and Nanni, T.: Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach, Int. J. Climatol., 30(10), 1564–1576, doi:10.1002/joc.1992, 2010.

Stooksbury, D. E., Idso, C. D. and Hubbard, K. G.: The Effects of Data Gaps on the Calculated Monthly Mean Maximum and Minimum Temperatures in the Continental United States: A Spatial and Temporal Study, J. Clim., 12(5), 1524–1533, doi:10.1175/1520-0442(1999)012<1524:TEODGO>2.0.CO;2, 1999.

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

Tang, G., Behrangi, A., Ma, Z., Long, D. and Hong, Y.: Downscaling of ERA-Interim Temperature in the Contiguous United States and Its Implications for Rain–Snow Partitioning, J. Hydrometeorol., 19(7), 1215–1233, doi:10.1175/jhm-d-18-0041.1, 2018b.

Tang, G., Clark, M. P., Newman, A. J., Wood, A. W., Papalexiou, S. M., Vionnet, V., Whitfield, P. H.: SCDNA: a serially complete precipitation and temperature dataset for North America from 1979 to 2018 [Dataset], Zenodo, https://doi.org/10.5281/zenodo.3735533, 2020.

Tang, Q., Wood, A. W. and Lettenmaier, D. P.: Real-Time Precipitation Estimation Based on Index Station Percentiles, J. Hydrometeorol., 10(1), 266–277, doi:10.1175/2008JHM1017.1, 2009.

Teegavarapu, R. S. V. and Chandramouli, V.: Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records, J. Hydrol., 312(1), 191–206, doi:10.1016/j.jhydrol.2005.02.015, 2005.

Ustaoglu, B., Cigizoglu, H. K. and Karaca, M.: Forecast of daily mean, maximum and minimum temperature time series by three artificial neural network methods, Meteorol. Appl., 15(4), 431–445, doi:10.1002/met.83, 2008.

Venema, V. K. C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J. A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C. N., Menne, M. J., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Prohom Duran, M., Likso, T., Esteban, P. and Brandsma, T.: Benchmarking homogenization algorithms for monthly data, Clim. Past, 8(1), 89–115, doi:10.5194/cp-8-89-2012, 2012.

Vicente-Serrano, S. M., Saz-Sanchez, M. A. and Cuadrat, J. M.: Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): application to annual precipitation and temperature, Clim. Res., 24(2), 161–180, doi:DOI 10.3354/cr024161, 2003.

Vincent, L. A., Zhang, X., Bonsal, B. R. and Hogg, W. D.: Homogenization of Daily Temperatures over Canada, J. Clim., 15(11), 1322–1334, doi:10.1175/1520-0442(2002)015<1322:HODTOC>2.0.CO;2, 2002.

Vincent, L. A., Milewska, E. J., Hopkinson, R. and Malone, L.: Bias in Minimum Temperature Introduced by a Redefinition of the Climatological Day at the Canadian Synoptic Stations, J. Appl. Meteorol. Climatol., 48(10), 2160–2168, doi:10.1175/2009JAMC2191.1, 2009.

Vincent, L. A., Wang, X. L., Milewska, E. J., Wan, H., Yang, F. and Swail, V.: A second generation of homogenized Canadian monthly surface air temperature for climate trend analysis, J. Geophys. Res. Atmospheres, 117(D18), doi:10.1029/2012JD017859, 2012.

Wambua, R. M., Mutua, B. M. and Raude, J. M.: Prediction of Missing Hydro-Meteorological Data Series Using Artificial Neural Networks (ANN) for Upper Tana River Basin, Kenya, Am. J. Water Resour., 4(2), 35–43, doi:10.12691/ajwr-4-2-2, 2016.

Wang, X. L., Xu, H., Qian, B., Feng, Y. and Mekis, E.: Adjusted daily rainfall and snowfall data for Canada, Atmosphere-Ocean, 55(3), 155–168, 2017.

Werner, A. T., Schnorbus, M. A., Shrestha, R. R., Cannon, A. J., Zwiers, F. W., Dayon, G. and Anslow, F.: A long-term, temporally consistent, gridded daily meteorological dataset for northwestern North America, Sci. Data, 6(1), 1–16, doi:10.1038/sdata.2018.299, 2019.

Whitfield, P. H.: Climate Station Analysis and Fitness for Purpose Assessment of 3053600 Kananaskis, Alberta, Atmosphere-Ocean, 52(5), 363–383, doi:10.1080/07055900.2014.946388, 2014.

Woldesenbet, T. A., Elagib, N. A., Ribbe, L. and Heinrich, J.: Gap filling and homogenization of climatological datasets in the headwater region of the Upper Blue Nile Basin, Ethiopia, Int. J. Climatol., 37(4), 2122–2140, doi:10.1002/joc.4839, 2017.

Wood, A. W.: The University of Washington Surface Water Monitor: An experimental platform for national hydrologic assessment and prediction, in 22nd Conf. on Hydrology., 2008.

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S. and Bates, P. D.: A high-accuracy map of global terrain elevations, Geophys. Res. Lett., 44(11), 5844–5853, doi:10.1002/2017GL072874, 2017.

Yang, D., Kane, D., Zhang, Z., Legates, D. and Goodison, B.: Bias corrections of long-term (1973-2004) daily precipitation data over the northern regions, Geophys. Res. Lett., 32(19), n/a-n/a, doi:10.1029/2005gl024057, 2005.

Yatagai, A., Kamiguchi, K., Arakawa, O., Hamada, A., Yasutomi, N. and Kitoh, A.: APHRODITE: Constructing a Long-Term Daily Gridded Precipitation Dataset for Asia Based on a Dense Network of Rain Gauges, Bull. Am. Meteorol. Soc., 93(9), 1401–1415, doi:10.1175/bams-d-11-00122.1, 2012.

Young, K. C.: A Three-Way Model for Interpolating for Monthly Precipitation Values, Mon. Weather Rev., 120(11), 2561–2569, doi:10.1175/1520-0493(1992)120<2561:ATWMFI>2.0.CO;2, 1992.