Received: 11 March 2019 | Revised: 9 November 2019 | Accepted: 5 January 2020

DOI: 10.1002/gdj3.88

DATA PAPER

Geoscience Data Journal RMetS WILEY

A cross-checked global monthly weather station database for precipitation covering the period 1901–2010

Dante Castellanos-Acuna | Andreas Hamann

Department of Renewable Resources, Faculty of Agricultural, Life, and Environmental Sciences, University of Alberta, Edmonton, Alberta, Canada

Correspondence
Dante Castellanos-Acuna, Department of Renewable Resources, Faculty of Agricultural, Life, and Environmental Sciences, University of Alberta, 751 General Services Building, Edmonton, AB T6G 2H1, Canada.
Email: dcastell@ualberta.ca

Funding information
Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: RGPIN 330527; Mexican Council of Science and Technology (CONACyT), Grant/Award Number: 253855

Abstract

Comprehensive monthly weather station databases are the foundation for many gridded climate data products, and they are widely used to characterize regional climate conditions, track climate change and research the impact of climate on natural and managed ecosystems. However, weather station databases are often regional in coverage, and they can have extensive gaps in station coverage over time. They may also contain errors in climate records, station coordinates or elevation. Here, we assemble a comprehensive monthly weather station database for precipitation from multiple reputable data sources. We use digital elevation models and nearby stations to search for inconsistencies in reported station locations and recorded precipitation values. We also estimated missing values in weather station time series using a linear model approach based on interpolated anomaly surfaces. The resulting station records were ranked into ten classes, according to the completeness of records, the reliability of missing value estimations and other criteria. We corrected incomplete or erroneous location and elevation information for 12% of all available station records. A total of 23% of monthly records that had missing values could be estimated with high or moderate confidence. We subsampled our global database of more than 80,000 stations with various spatial filters, so that only the highest quality station for a given area was retained. Our contribution significantly enhances global data coverage compared to individual databases currently available. Even when accepting only the stations within the top two quality ranks in our combined database, and applying the coarsest spatial filter of one station per approximately 1,600 km^2, the remaining station count of more than 20,000 stations exceeds the largest alternative database (without a spatial filter applied) by more than 50%.

Dataset

Identifier: http://doi.org/10.5281/zenodo.3520885

Creator: Dante Castellanos-Acuna and Andreas Hamann

Title: A cross-checked global monthly weather station database for precipitation covering the period 1901–2010

Publisher: Zenodo.org

Publication year: 2019

Resource type: metadata.

Version: 1.0

KEYWORDS
climate, land surface, monthly time series, precipitation, weather stations

1 | INTRODUCTION

Comprehensive monthly weather station databases are the foundation for characterizing regional climate conditions, and for tracking climate change over time. For the purpose of climatic characterizations, monthly summaries represent a good compromise between capturing seasonal climate variation without having to manage large amounts of daily weather data. Once these monthly weather summaries have been recorded for 30 years, also referred to as a climate normal period, calculating an average allows inference of long-term expectations of climate conditions that is not usually biased by cyclical or random anomalies (Guttman, 1989; Arguez and Vose, 2011). With a sufficient density of weather station data for a region, interpolation methods can be used to derive grids of baseline climate data for complex landscapes, modelling various climate phenomena, such as changes in temperature along elevation gradients, orographic precipitation and rain shadows (Hutchinson, 1995; Daly *et al.*, 2002). Once the baseline climatology of a region has been established, additional questions can be addressed with monthly time series records, such as how the climate has changed in the past, or how the climatology of a region may change in the future (Sáenz-Romero *et al.*, 2010; Ramirez-Villegas *et al.*, 2013).

Many gridded climate data products that are widely used to research the impact of climate variability and climate change on natural and managed ecosystems rely on monthly weather station databases. For the United States, the Parameter-elevation Relationships on Independent Slopes Model (PRISM) is a well-regarded database of gridded climate that benefits from the extensive network of weather stations available for this region. Gridded climate products with global coverage also include the Climate Research Unit Time Series (CRU-TS) database from the University of East Anglia (Harris *et al.*, 2014), a gridded database from the University of Delaware (Willmott and Matsuura, 1995), the Global Precipitation Climatology Centre (GPCC) product (Becker *et al.*, 2013) or the Precipitation REConstruction Land (PRECL) database from NOAA (Chen *et al.*, 2002). These databases with monthly historical resolution are limited to low spatial resolutions (0.5 degree or coarser). Alternative products, with high spatial resolutions (30 arc-seconds or approximately 1 km), are usually restricted to 30-year normal summaries and provide no interannual historical data, for example WorldClim (Hijmans *et al.*, 2005).

For researchers that develop climate grids, there are a number of important challenges in assembling the required regional or global weather station databases. First, the placement of the weather stations is usually biased towards population centres or agricultural lands, whereas climate conditions of mountainous or desert areas are usually not well documented (New *et al.*, 1999; Menne *et al.*, 2012). Another important limitation of weather station data is temporal coverage. Before the 1950s, the density of weather stations tends to be low, reaching its highest global density around the 1970s before declining again (Menne *et al.*, 2012). Additionally, many of these stations were operational only for a few years, with extensive gaps in the records or only operated seasonally, especially in mountainous regions. Finally, it is not uncommon to encounter errors in recorded climate values, errors of unit conversions in countries using the Imperial system and mistakes associated with locations, such as inaccuracies in the reported coordinates or elevation. Before the widespread use of global positioning systems, coordinates were typically recorded to the nearest minute, implying a location error of hundreds of metres, which can be problematic on mountainous terrain where the elevation and topographic gradients are an important determinant of the weather patterns.

In order to support researchers that rely on monthly weather station databases to develop interpolated grids or other climate data products, we assemble and cross-check a monthly weather station database for precipitation that combines several regional and global databases that are publicly available, including the Global Historic Climate Network (GHCN) v2 database (Lawrimore *et al.*, 2011; Menne *et al.*, 2012), the station database corresponding to the Climate Research Unit (CRU) Time Series v3.21 interpolated data set (Harris *et al.*, 2014), the WMO-CLINO database of the World Meteorological Organization (WMO, 1996), the FAOCLIM 2.0 database of the Food and Agriculture Organization of the United Nations (Bogaert et al., 1995), the R- HydroNET database of the Regional Hydrometeorological Data Network for Latin America (Vorosmarty *et al.*, 1998), the European Climate Assessment (ECA) database (Tank *et al.*, 2002; Besselaar *et al.*, 2015) and the United States Forest Service (USFS) database (Rehfeldt, 2006). We use duplicate entries, digital elevation models and nearby stations to search for inconsistencies in recorded climate values in weather station records that may be due to unit conversion errors, location and elevation inaccuracies.

This study provides a consolidated database of more than 80,000 weather stations (without duplicates) ranked into ten quality classes, according to the completeness of records, the reliability of missing value estimates and other criteria. Specifically, we contribute the following corrections and enhancements for users of monthly precipitation databases: (1) when errors could be corrected without ambiguity, corrections were made and indicated by flags in the database. Alternatively, the records were flagged with the lowest quality score for removal; (2) we estimate missing values in monthly station time series records with a linear model approach based on correlations with global interpolated anomaly surfaces, where deviations of monthly weather station values from 1961 to 1990 normal climate were interpolated. The reliability of estimated values was documented with model fit statistics and quality scores. Missing value estimates are primarily provided to estimate adjusted long-term averages (e.g. 30-year climate normals); (3) lastly, we provide subsamples of this database that select the best station records for pre-determined three-dimensional filters (with different intervals for latitude, longitude, and elevation). The sub-sampled data sets retain only the most reliable and complete station records for a given area, with global coverage and with as little spatial sample bias as possible, which are meant for generating interpolated data products.

2 | METHODS

2.1 | Databases used

The public weather station databases that were used in this study (Table 1) have already been subjected to rigorous quality control methodology. The CRU and GHCN databases have been screened for duplicate records, outliers, tests for violation of logical or physical relations between variables (Tmax < Tmin), unrealistic peaks or dips in time series, spatial consistency tests by comparing with surrounding stations, etc. (New *et al.*, 1999; Durre *et al.*, 2010). We use both the daily and monthly versions of the GHCN database v2, with the daily database containing additional records of stations with shorter time series and more gaps in the record. The FAOCLIM 2 database from FAO was established in the 1980s to evaluate the global agricultural production potential in developing countries and provides additional regional coverage in Central America, agricultural areas of South America and the Sahel zone of Africa. The ECA monthly database provides good additional coverage for mountainous regions in Europe, such as the Alps, the Carpathians, the Balkans, the Caucasus and the Scandinavian mountains. The R-HydroNET database for South America provides useful additions for Amazonian precipitation data, and the USFS database has excellent additional coverage for mountainous regions in North America, including the United States, Canada and Mexico.

After removal of duplicates from the combined weather station database (as described below), temporal coverage used in this study extends from the beginning of the last century to 2010, reaching their highest spatial density from the 1960s to the 1990s for most regions of the world (Figure 1). The drop of station coverage in recent years is partially due to several databases not including recent records (Figure 1). Excellent temporal and spatial coverage for the 1961–1991 period is one reason why baseline grids are often developed for this 1961–1990 normal period (New *et al.*, 1999; Menne *et al.*, 2012). Another reason why 1961–1990 normal period is a useful reference period is that it largely precedes anthropogenic climate change (Tett *et al.*, 1999; Lawrimore *et al.*, 2011) and can therefore be used as a reference period when future climate projections are expressed as an anomaly (e.g. +2°C warming relative to a reference period). In the database that we develop in this study, we rank weather stations higher that have complete records for this 1961–1990 normal period.

2.2 | Elevation match with DEM

As a first check of station records, we compared the reported elevation for each weather station against a digital elevation model (DEM) of 1 km resolution (Gesch *et al.*, 1999). Missing elevation values in weather station records, usually indicated by flags in the elevation field, such as −9999, −999, −99 or 9,999 were replaced with the DEM value. Further, we recorded the difference between the reported elevation value and the DEM, and performed a more detailed inspection of any station that had a difference exceeding ±250 m. We checked those stations for potential errors of unit conversions, potential errors of location that may have led to an elevation discrepancy or implausible elevation values given the topography in the vicinity of the recorded station. In case of discrepancies between the DEM and reported elevation values, we normally accept the reported elevation of the weather station in mountainous terrain. Here, uncertainties in the location of weather station, for example if reported to the nearest minute, can usually explain a discrepancy with the DEM. In flat terrain, elevation differences exceeding ±250 m could usually be explained by unit conversions or other errors, and the reported elevation was replaced with the DEM value. Determining the correct value for weather stations is important, because the elevation value is normally used as a covariate in any climate modelling or interpolation effort.

2.3 | Outlier detection and missing value estimation

A useful check of weather station records is to determine consistency of records with other nearby stations to detect recording errors or unit conversion issues. In this study, we used two approaches. First, all stations were checked against one or several nearby stations within a 10 km distance (increasing the radius in 10 km increments if no station was found). Differences in monthly records were

TABLE 1 Databases included in this study with statistics describing their spatial and temporal coverage. The databases are ordered by preference, based on documented quality control efforts, accuracy of location information, temporal coverage and overlap with other databases. The latest data used were 2010 as most databases were incomplete beyond this date (Figure 1)

Database[a]	Spatial extent	Temporal extent	Temporal resolution	Number of stations
1. Climate Research Unit Time Series v3.21 observations (CRU)	Global	1901–2010	Monthly time series	11,702
2. Global Historic Climate Network Dataset v2 (GHCN)	Global	1850–2010	Monthly time series	20,541
3. FAOCLIM 2.0 global climate database (FAO)	Global	1901–1999	Monthly time series	13,529
4. World Meteorological Organization normals (WMO)	Global	1961–1990	1961–1990 normals	4,259
5. European Climate Assessment Dataset (ECA)	Europe, Russia, North Africa	1901–2010	Monthly time series	10,085
6. R-HydroNET (R-HN)	South America	1920–1990	Monthly or 1961–1990 Normal	3,256
7. Daily Global Historical Climatology Network (dGHCN)	Global	1901–2010	Daily data	45,603
8. United States Forest Service (USFS)	North America	1961–1990	1961–1990 normals	14,635

[a]References: (1) Harris *et al.* (2014), (2) Lawrimore *et al.* (2011), (3) Bogaert et al. (1995), (4) WMO (1996), (5) Van Den Besselaar *et al.* (2015), Tank *et al.* (2002), (6) Vorosmarty *et al.* (1998), (7) Menne *et al.* (2012), (8) Rehfeldt (2006).

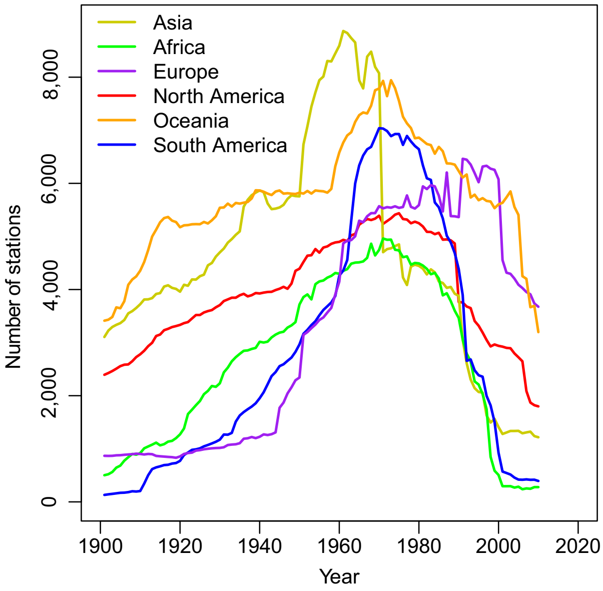

FIGURE 1 Temporal coverage of weather station records from all databases listed in Table 1 combined, after removal of duplicate station records

calculated, and then, the stations were sorted based on the maximum difference per station pair. A histogram of the differences revealed a bimodal distribution, with far outliers exceeding a difference value of 1,000 mm. This represented a problem primarily confined to a small percentage of stations from the CRU database and was recorded as a station quality flag.

Second, we compare station values to interpolated grids of monthly anomalies from 1901 to 2010 of the CRU-TS data product (Mitchell and Jones, 2005; Harris *et al.*, 2014). Our criterion for potentially problematic station records was low correlations of station data with the CRU-TS monthly anomalies for the location of the weather station. For this purpose, we calculated correlation coefficients for each month of the year between weather station records and the corresponding CRU grid cell. Stations with low correlation coefficients were flagged as potentially problematic.

If the correlation between CRU-TS data and weather station records were high, we used a simple linear model with a fixed intercept at zero to predict missing values in the temporal record of station data. The weather station record needed to have at least 20 years for the 1961–1990 period and the linear model an $R^2 > 0.7$ for to be considered for estimating missing values. A less reliable missing value estimation, resulting in a lower quality rank, was based on linear models with R^2 values between 0.5 and 0.7 and at least 27 years of data for the 1901–2010 period. A special case for missing value estimation was stations located in desert areas, where a linear model could not be established due to the majority of monthly precipitation values being zero. For stations with at least 10 years of monthly data for the 1901–2010 period, but located in desert areas, we filled any missing values in the observed weather station data with the corresponding CRU-TS values directly (i.e. not using a linear model). These estimates were flagged as filled and assigned a lower quality score (see next section), allowing users of our database to select various quality criteria to filter the database according to their needs.

2.4 | Quality criteria

For each station, we assigned a quality score based on the completeness of the station record, the quality of the linear model to estimate missing values and a number of other criteria (Table 2). The best station quality score (1) was assigned to stations that had at least 90% complete data for the 1961–1990 period, either as monthly time series, or reported as average for the 1961–1990 normal period (i.e. from WMO or R-HydroNET databases). The next best score (2) was assigned to the stations that had at least 66% of the data for the 1961–1990 period complete and where missing values could be estimated with a linear model that had an R^2 of at least 0.7. The following score (3) was assigned to stations with a similar criteria, but between 33% and 66% of the data of the 1961–1990 was complete (i.e. 10 years), plus a total of 25% of the data complete for the 1901–2010 period (i.e. 28 years of data in total), and an R^2 of at least 0.7 for estimation of missing values. The fourth score (4) was given to records that did not report monthly time series, but only 1961–1990 normal period averages (i.e. from WMO and R-HydroNET) and with completeness of annual records between 66% and 90%. The next score (5) was given to station records that did not cover the 1961–1990 period well, but that still contained a substantial time series with at least 25% of the data complete for 1901–2010 time series (i.e. 27 years), and with a total of 90% of data either observed or estimated for 61–90 time series with $R^2 > 0.7$. Score (6) was assigned to stations with at least 25% of the data complete for 1901–2010 time series (i.e. 27 years and missing values estimable with $R^2 > 0.5$. Score (7) includes all seasonal stations that covered three to ten months of the year and that otherwise covered at least quality score 6. The score of (8) was applied to entries that did not have monthly time series but only a 1961–1990 normal period average with between 33% and 66% completeness of the data (i.e. applicable to some entries of the WMO database) or data completeness was not reported (applicable to the USFS database). A score of (9) was given to stations that exhibited far outliers in monthly data that were inconsistent with nearby stations. Finally, the score of (10) was given to all remaining stations that did not fulfil any of the above-stated requirements, usually stations with very short time series.

The monthly consolidated database contains all entries from our source databases (Table 1), without duplicates removed at this stage. In addition to the original entries, we report the DEM value for the station location, a linear model estimate for any missing value for the 1901–2010 period, the R^2 value for the linear model estimate and a flag that indicates whether the precipitation value was recorded, estimated, filled with CRU estimates for desert stations or not estimable. Additional columns specify per cent of complete records for the 1901–2010 period, the per cent of complete records for the 1961–1990 period, a database quality score according to Table 1, a station quality score according to Table 2 and a combined quality score that ranks database scores within station quality scores (e.g. 23 would indicate a station quality score of 2 for an FAO database entry). Smaller numbers indicate overall higher quality records.

2.5 | Duplicate removal and database subsets

Duplicate stations among the databases were common and were removed based on reported weather station IDs and or based on identical latitude and longitude values. Most databases used station IDs derived from those of the World Meteorological Organization. Where possible, we parsed the ID field to generate station IDs that conformed to the WMO format. This occasionally resulted in duplicate station IDs among different databases that were located in different states, countries or continents that used similar but independent ID schemes. To avoid these false-positive duplicate detections, we assigned to each station a global code related to the country, state or province where they were located. The final ID-based duplicate removal retained the station with the highest overall quality score for a given station ID in the same jurisdiction. This step did not remove all duplicates, as some databases did not use the WMO station IDs for some or all of their records.

The second duplicate removal step was location-based. We generated grids of 2.5 arcminutes (~5 km), 5 arcminutes (~10 km), 10 arcminutes (~20 km) and 20 arcminutes (~40 km). In addition, we want to retain stations in the same general area that are located at different elevations. For this purpose, we created elevation intervals of 100 m for each of the above grids. To sub-sample the original database at different spatial densities and to remove any additional duplicates that were missed in the previous step, we retained a single station with the highest overall quality score in each of the three-dimensional grid cells.

3 | RESULTS AND DISCUSSION

3.1 | Recorded station elevation versus DEM

From the total consolidated record of 123,000 stations (no duplicates removed), 9% had missing values for elevation indicated by a flag of −9999 or similar. In addition, several databases contained a sizable number of records that had zero recorded for elevation. (0.11% of CRU, 3.5% of GHCN, 0.4% of FAO, 0.05% of ECA, and 1.5 of dGHCN). Not all of these zero values were plausible measurements, for example indicated in blue in Figure 2, typically located across India, Australia and Brazil. Here, zero values were presumably used

TABLE 2 Data quality scores based on the completeness of the station record for the 1961–1990 period, the completeness of station records for the 1901–2010 period, the quality of the linear model to estimate missing values and a number of other criteria

Score	Data requirements for station quality score
1	At least 90% complete for 61–90 time series or normal averages (i.e. 3 missing values allowed)
2	At least 66% complete for 61–90 time series & missing values estimated with $R^2 > 0.7$
3	At least 33% complete for 61–90 time series (i.e. 10 years), 25% complete for 1901–2010 time series (i.e. 27 years), and missing values estimated with $R^2 > 0.7$
4	At least 66% complete values for reported 61–90 normal average (i.e. uncorrected)
5	At least 25% complete for 1901–2010 time series (i.e. 27 years), and 90% observed or estimated for 61–90 time series with $R^2 > 0.7$
6	At least 25% complete for 1901–2010 time series (i.e. 27 years), and missing values estimable with $R^2 > 0.5$
7	Seasonal stations (three to ten months) that otherwise ranked at least quality score 6
8	At least 33% complete values for reported 61–90 normal average (i.e. uncorrected), or completeness of record unreported.
9	Stations with a quality score of at least 6, but that contained individual monthly observations that were identified as a far outlier and as inconsistent with nearby stations.
10	All remaining stations that did not meet at least quality criteria 8, usually containing short time series or high numbers of missing values.

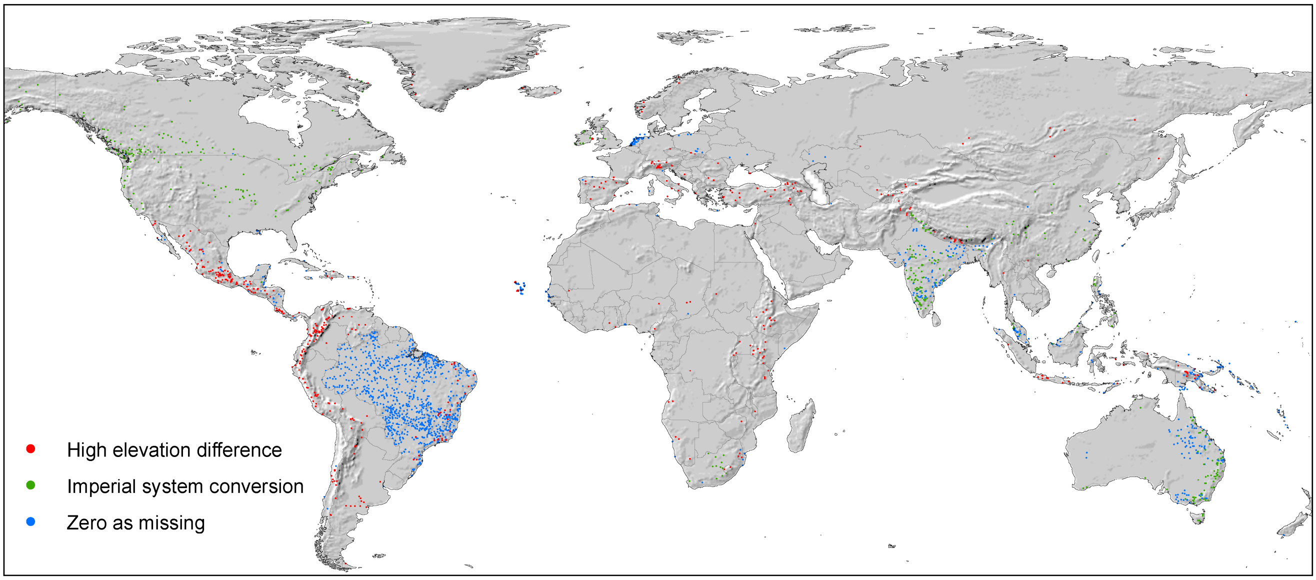

FIGURE 2 Stations with recorded elevations of zero (blue), with double conversions or omission of conversions from feet to metres (green), and other station with elevation differences >250 m (red) between a digital elevation model and the value recorded for the station location

when the record should indicate missing values. For simplicity, we replaced all zero values with records from the digital elevation model, even when zero values were plausible measurements, that is for the Netherlands and other coastal locations. The rationale for this replacement was that many databases had at least some records, where an elevation value of zero actually indicated a missing value as well. Recorded elevation values of zero that represented a correct measurement were usually located near the coast, where the DEM replacement resulted in a very similar elevation estimate.

For all other records, we screened for substantial discrepancies between the digital elevation model and the recorded station elevation. This yielded a number of stations where conversions from the imperial system to the metric system were either omitted, or applied twice (Figure 3. rows of green points). We found that most databases were affected by this type of error, but to varying degrees (0.7% of CRU, 4.4% of GHCN, 0.6% in FAO, 0.1% of ECA, 0.4% of dGHCN and 0.3% of USFS). Even within local regions, only a subset of stations had these conversion errors. The errors were corrected by either multiplying by 3.281

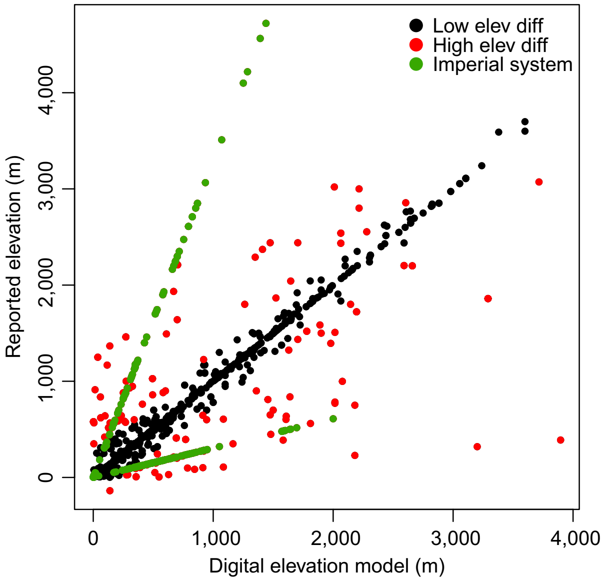

FIGURE 3 Scatter plot of recorded station elevation over the elevation value from a digital elevation model for the station. Double conversions or omission of conversions from feet to metres are visible as off-diagonal rows of green dots, other stations with an elevation difference >250 m are indicated in red. The location of these stations are mapped with the same colours in Figure 2

or dividing by 0.305. We only carried out the corrections for countries that in some point in their history used the Imperial system, and where these errors were almost exclusively located (Figure 2). For stations with elevation conversion issues, we also checked if precipitation conversions may have been incorrect, but we could not find instances where precipitation may have plausibly been incorrectly converted from inches to millimetres.

Lastly, other stations with large elevation discrepancies were flagged, but retained unchanged. These stations are usually located in mountainous regions (Figure 2. red circles), and the recorded elevation is likely a more reliable indicator of the true elevation of the weather station than the DEM value for these locations. A large elevation difference was therefore not incorporated into the quality score, but is reported as a separate elevation difference statistic for each station.

3.2 | Missing value estimation

For the purpose of calculating long-term climate averages, we provide missing value estimations that may be used in lieu of accepting a certain number of missing values in estimation of climate normals. The missing value estimation relies on a linear model with interpolated CRU-TS anomaly grids that have a coarse resolution (30 arcminutes), but nevertheless often yield strong correlations with recorded weather station data. The interpolated grids allow missing value estimation because of spatial interpolation from nearby stations that have records for the missing target value. While the correlations of station values with the monthly interpolated grids are often quite high and suitable for prediction (Figure 4a), the relationship is also often biased with a slope considerably deviating from the diagonal (e.g. Figure 4c). A moderate proportion of missing values could be estimated based on a linear model with R^2 values between 0.5 and 0.7 (Figure 4b, and Table 3).

3.3 | Final spatially sub-sampled databases

Spatial and elevational sub-sampling by elevation intervals and various grid sizes (2.5, 5, 10, and 20 arcminutes) is meant to provide users with databases where local duplications (nearby stations) are removed, retaining only a single station with the

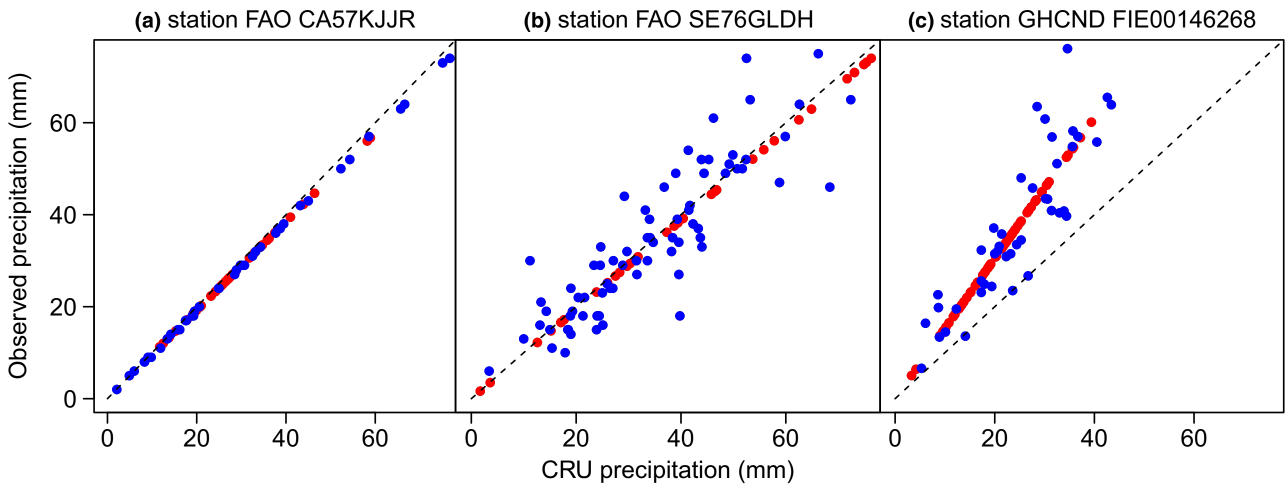

FIGURE 4 Examples for estimation of missing values in January precipitation (red circles) using a linear model with high confidence indicated by an $R^2 > 0.7$ (a), moderate confidence with an R^2 between 0.5 and 0.7 (b), and moderate confidence as well as a biased relationship (c), leading to different quality scores as outlined in Table 2

highest overall quality for a given grid cell. As the grid size for subsamples increases, we therefore retain slightly higher proportions of high-quality stations while the overall database size decreases (Table 4). The WMO and CRU databases have the highest proportion of high-quality station records, but their database size is relatively small compared to our combined and sub-sampled databases. Even when sampling one station at the coarsest 20 arcminute grid resolution, we retain more than 20,000 stations with a high-quality score of 1 or 2 (Table 4).

The spatial distribution of stations using the coarsest subsample at 20 arcminute resolution is shown in Figure 5, where the quality is indicated by a colour legend. Low-quality stations are typically restricted to mountainous areas and specific countries. For example, most of India's weather station coverage has gaps after 1970 in all databases, leading to intermediate quality scores. For North America, there are equal number of high and low-quality stations, but we should note that the low-quality stations primarily represent Quality 8 stations from the USFS database. The USFS database contains 1961–1990 normal averages, without reporting the number of observations that went into these estimates. The low-quality score in this case was given for lack of information, that is where duplicate records are available, other databases with more information would be preferred.

Virtually, all databases included in this study contributed a significant amount of stations to the final combined database (Table 5). The individual contributions to some degree reflect their size of each database (*cf.* Table 1), with the daily GHCN database contributing the largest number of stations. However, with spatial filters applied that subsample the best stations for a three-dimensional grid cell (area and elevation band), the contributions of individual databases become more equalized. Also, the best quality stations (right section of Table 5) are sourced from all databases except USFS, which lacked documentation to achieve a high-quality score in this compilation.

3.4 | Applications and limitations

In this data management and data cleaning effort, we made a number of subjective choices that are guided by particular

TABLE 3 Percentage of observed records and estimable missing values by station database

Database	Observed records	Predicted ($R^2 > 0.7$)	Predicted (R^2 0.5–0.7)	Zero & filled	Not estimated
CRU	60.3	28.9	3.5	4.1	1.5
GHCN	44.6	29.7	9.7	3.7	6.7
FAO	43.6	28.8	6.2	5.5	12.1
ECA	36.2	46.4	3.0	0.0	13.6
R-HN	14.4	29.2	10.0	1.1	38.9
dGHCN	29.2	33.5	7.9	1.0	25.1

TABLE 4 Size and the percentage of records with different quality scores for the individual databases used in this study, for all databases combined prior to the removal of duplicates and for subsets that select the highest quality station for various grid sizes

		Station quality (see Table 2)									
Database	Number of stations	1	2	3	4	5	6	7	8	9	10
CRU	11,702	71	20	<1	0	0	5	0	0	5	<1
GHCN	20,541	36	14	3	0	4	34	2	0	<1	7
FAO	13,493	45	13	5	0	1	23	1	0	<1	12
WMO	4,149	93	0	0	5	0	0	0	0	0	1
ECA	9,998	47	12	4	0	<1	20	3	0	<1	14
R- HydroNET	3,256	59	9	0	4	0	8	0	0	0	20
dGHCN	44,763	19	11	6	0	5	31	2	0	<1	26
USFS	14,629	0	0	0	0	0	0	0	100	<1	0
Combined	122,531	33	11	4	<1	3	22	1	12	<1	14
No dups	98,510	32	9	4	<1	3	23	2	13	<1	14
2.5' Grid	79,045	31	9	4	<1	3	24	2	13	<1	14
5' Grid	71,716	32	9	4	<1	3	24	2	14	<1	13
10' Grid	60,484	33	9	4	<1	3	23	1	15	<1	12
20' Grid	46,040	37	9	3	<1	2	21	1	16	<1	11

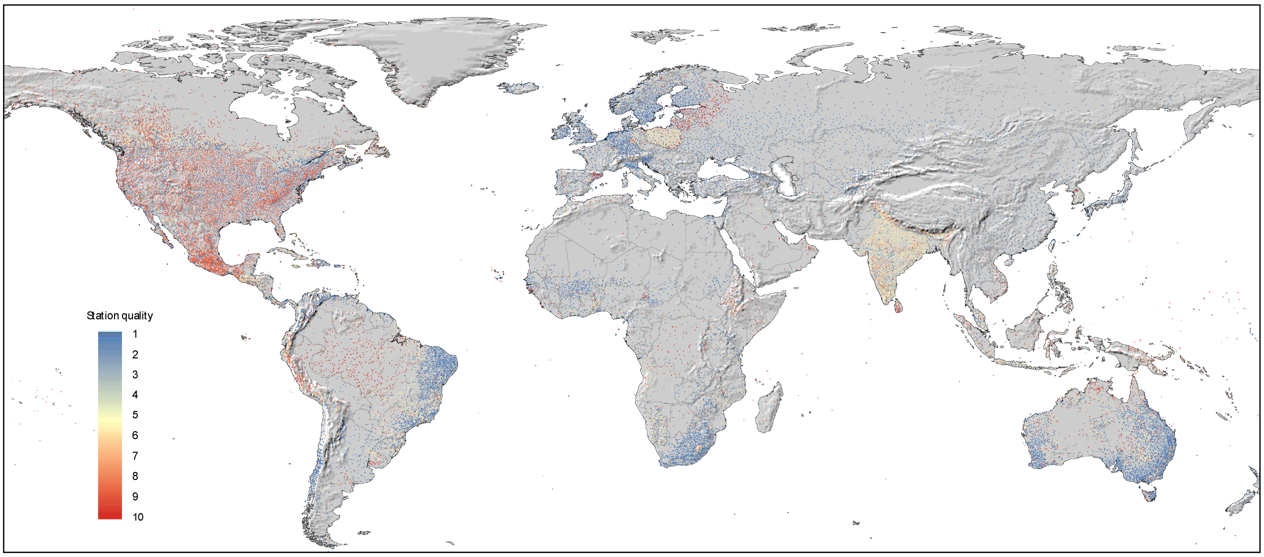

FIGURE 5 Map of stations coloured by quality score (blue = high quality, red = poor records, for details see Table 2), for the subsample where one station is selected for each 20 arcminute grid cell (approximately 1,600 km^2) and 100 m elevation interval

TABLE 5 Contributions of individual databases to the combined database developed in this study. Values represent percentages relative to all stations of the combined database with and without spatial filters applied

	All weather stations				Best stations (quality 1 & 2)			
Database	**No filter**	**5' filter**	**10' filter**	**20' filter**	**No filter**	**5' filter**	**10' filter**	**20' filter**
CRU	11.2	14.6	16.7	20.3	10.5	13.8	15.8	19.5
GHCN	13.9	16.1	16.8	17.7	4.5	4.5	4.5	4.6
FAO	10.5	6.9	5.9	5.4	5.5	2.9	2.4	2.1
WMO	2.7	2.0	2.1	2.3	2.5	1.9	1.9	2.1
ECA	9.9	11.7	11.9	11.0	5.8	6.5	6.6	6.7
HydroNET -R	2.5	2.6	2.8	3.1	1.9	2.1	2.1	2.4
dGHCN	36.0	32.2	28.6	24.3	10.4	9.0	8.7	8.2
USFS	13.4	13.9	15.3	15.9	0.0	0.0	0.0	0.0

applications that this database may be useful for, namely for the development of long-term climate normal surfaces that can serve as reference periods for ecological research on adaptation of organisms with climate, biological response of organisms to interannual climate variability and response of organisms to historical and future climate trends. As a useful normal reference period, we advocate the use of the 1961–1990 climate normal, which strikes a good balance between excellent global weather station coverage, and largely preceding a strong anthropogenic warming signal (Tett *et al.*, 1999; Lawrimore *et al.*, 2011; Estrada *et al.*, 2013). Therefore, our station quality ranks specifically take this period into account. Nevertheless, users that are interested in other periods can easily modify the ranking system. All records of the combined database were retained, and all decision criteria for quality ranks for each station are included to specify other preferences.

For spatial interpolation of weather station data, we recommend the use of station subsets, where only one station with the best quality score is selected for various grid sizes. For spatial interpolation, multiple records in close proximity are generally not needed, and the disadvantage of potentially poor quality records may outweigh the advantages of dense spatial coverage. For generating interpolated surfaces of climate normal periods other than the 1961–1990 period, or for generating surfaces with a monthly resolution, for example to study response of organisms to climate variability or climate trends at a monthly time step, we recommend using an anomaly or delta approach, described for example by Mitchell and Jones (2005) and Wang *et al.* (2006). Relying on our missing value estimation for stations up to a quality score of 8, an adjusted 1961–1990 normal average can be obtained for a large majority of the stations contained in this database (75% of stations). Deviations from this 1961–1990 normal estimate can then be calculated for all observed values in station time series. Interpolated monthly anomalies or interpolated normal anomalies can then provide robust climate estimates for years outside the 1961–1990 period, even if weather station coverage is not as dense.

The complete global database of monthly precipitation records from 1901 to 2010 is available from http://doi.org/10.5281/zenodo.3520885. This repository also contains regional files, climate normal summaries for the 1961–1990 period and subsamples based on spatial filters, where the highest quality stations are selected for a grid cell. Our general recommendations for developing global interpolated climate products are to use the 20 arcminute sub-sample and station records with a quality score of at least 8. For the 1961–1990 period, this subset would include stations with the following summary statistics (derived from data underlying Tables 3 and 4): 69% of records complete, 17% of monthly records being estimated with high confidence (linear model estimates with an $R^2 > 0.7$), 6% of monthly records being estimated with moderate confidence (linear model estimates with an R^2 between 0.5 and 0.7) and 2% of records being filled with CRU estimates for desert stations. For the development of local climatology products, a higher density station coverage may be used (i.e. one station per 2.5, 5, or 10 arcminute grid cell), especially if local station coverage is generally poor.

4 | CONCLUSION

The databases that we generated in this study should be of value to a variety of users who create gridded precipitation data or other climate data products derived from weather station data. We corrected a sizable number of errors in reported station elevations due to unit conversions, or due to missing values being reported as zero. Elevation errors can be detrimental for the quality of interpolated surfaces, as elevation is almost always used as a predictor variable or covariate in interpolation techniques for climate data. In addition, the estimation of missing values with linear models should render some stations useful that did not have appropriate coverage to calculate a specific 30-year climate normal due to missing values, but that had enough records from other years available to infer long-term climate conditions. Finally, the sub-sampling procedure guarantees that poorer records from various source databases are replaced by other, better quality records for nearby locations within the same elevation band. Our contribution significantly enhances the global data coverage compared to individual databases currently available (Table 5). This applies in particular for high-quality stations. The combined database contains almost 40,000 stations within the quality ranks 1 and 2, that is 41% of 97,112 stations (Table 4). Even when applying the strictest spatial filter, selecting one station per 20 arcminutes (or approximately 1,600 km^2), the resulting station count of more than 20,000 stations in the combined database exceeds the count of the top two quality stations in any individual database (without a spatial filter applied) by at least 58%.

ACKNOWLEDGEMENTS

Funding for this study was provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada through a Discovery Grant RGPIN-330527 to AH, and by the Mexican Council of Science and Technology (CONACyT) though a scholarship to DCA (#253855). We thank two anonymous reviewers for their excellent suggestions that contributed to substantial improvements over an earlier version of this database.

OPEN PRACTICES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at http://doi.org/10.5281/zenodo.3520885. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

ORCID

Dante Castellanos-Acuna https://orcid.org/0000-0001-9016-4049
Andreas Hamann https://orcid.org/0000-0003-2046-4550

REFERENCES

Arguez, A. and Vose, R.S. (2011) The definition of the standard WMO climate normal: the key to deriving alternative climate normals. *Bulletin of the American Meteorological Society*, 92, 699–704.

Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Schamm, K., Schneider, U. and Ziese, M. (2013) A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present. *Earth System Science Data*, 5, 71–99.

Bogaert, P., Mahau, P. and Beckers, F. (1995) *The Spatial Interpolation of Agro-Climatic Data, Cokriging software and source data. User's Manual v1.0b*. Agrometeorology Series Working Paper 12. Rome, Italy: Environmental Information Management Service, Sustainable Development Department, Agrometeorology Group, FAO.

Chen, M., Xie, P., Janowiak, J.E. and Arkin, P.A. (2002) Global land precipitation: a 50-yr monthly analysis based on gauge observations. *Journal of Hydrometeorology*, 3, 249–266.

Daly, C., Gibson, W.P., Taylor, G.H., Johnson, G.L. and Pasteris, P. (2002) A knowledge-based approach to the statistical mapping of climate. *Climate Research*, 22, 99–113.

Durre, I., Menne, M.J., Gleason, B.E., Houston, T.G. and Vose, R.S. (2010) Comprehensive automated quality assurance of daily surface observations. *Journal of Applied Meteorology and Climatology*, 49, 1615–1633.

Estrada, F., Perron, P. and Martinez-Lopez, B. (2013) Statistically derived contributions of diverse human influences to twentieth-century temperature changes. *Nature Geoscience*, 6, 1050–1055.

Gesch, D.B., Verdin, K.L. and Greenlee, S.K. (1999) New land surface digital elevation model covers the Earth. *EOS, Transactions: American Geophysical Union*, 80, 69–70.

Guttman, N.B. (1989) Statistical descriptors of climate. *Bulletin of the American Meteorological Society*, 70, 602–607.

Harris, I., Jones, P.D., Osborn, T.J. and Lister, D.H. (2014) Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset. *International Journal of Climatology*, 34, 623–642.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. and Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25, 1965–1978.

Hutchinson, M.F. (1995) Interpolating mean rainfall using thinplate smoothing splines. *International Journal of Geographical Information Systems*, 9, 385–403.

Lawrimore, J.H., Menne, M.J., Gleason, B.E., Williams, C.N., Wuertz, D.B., Vose, R.S. *et al*. (2011) An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *Journal of Geophysical Research: Atmospheres*, 116, D19121. https://doi.org/10.1029/2011JD016187

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E. and Houston, T.G. (2012) An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29, 897–910.

Mitchell, T.D. and Jones, P.D. (2005) An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *International Journal of Climatology*, 25, 693–712.

New, M., Hulme, M. and Jones, P. (1999) Representing twentieth-century space-time climate variability. Part I: development of a 1961–90 mean monthly terrestrial climatology. *Journal of Climate*, 12, 829–856.

Ramirez-Villegas, J., Challinor, A.J., Thornton, P.K. and Jarvis, A. (2013) Implications of regional improvement in global climate models for agricultural impact research. *Environmental Research Letters*, 8, art024018. https://doi.org/10.1088/1748-9326/8/2/024018

Rehfeldt, G.E. (2006) *A spline model of climate for the Western United States. General Technical Report RMRS-GTR-165*. Fort Collins, CO: Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Sáenz-Romero, C., Rehfeldt, G.E., Crookston, N.L., Duval, P., St-Amant, R., Beaulieu, J. *et al*. (2010) Spline models of contemporary, 2030, 2060 and 2090 climates for Mexico and their use in understanding climate-change impacts on the vegetation. *Climatic Change*, 102, 595–623.

Tank, A.M.G.K., Wijngaard, J.B., Konnen, G.P., Bohm, R., Demaree, G., Gocheva, A. *et al*. (2002) Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology*, 22, 1441–1453.

Tett, S.F.B., Stott, P.A., Allen, M.R., Ingram, W.J. and Mitchell, J.F.B. (1999) Causes of twentieth-century temperature change near the Earth's surface. *Nature*, 399, 569–572.

Van Den Besselaar, E.J.M., Tank, A.M.G.K., Van Der Schrier, G., Abass, M.S., Baddour, O., Van Engelen, A.F.V. *et al*. (2015) International climate assessment & dataset: climate services across borders. *Bulletin of the American Meteorological Society*, 96, 16–21.

Vorosmarty, C.J., Jauregui, C.F. and Donoso, M.C. (1998) *A regional, electronic hydrometeorological data network for South America, Central America, and the Caribbean*. Durham, NH: University of New Hampshire.

Wang, T., Hamann, A., Spittlehouse, D.L. and Aitken, S.N. (2006) Development of scale-free climate data for Western Canada for use in resource management. *International Journal of Climatology*, 26, 383–397.

Willmott, C.J. and Matsuura, K. (1995) Smart interpolation of annually averaged air temperature in the United States. *Journal of Applied Meteorology*, 34, 2577–2586.

WMO (1996) *Climatological Normals (CLINO) for the period 1961–1990*. WMO Series No. 847, ISBN 9263008477. Geneva, Switzerland: Secretariat of the World Meteorological Organization.

How to cite this article: Castellanos-Acuna D, Hamann A. A cross-checked global monthly weather station database for precipitation covering the period 1901–2010. *Geosci Data J*. 2020;7:27–37. https://doi.org/10.1002/gdj3.88