Received: 27 August 2018 | Revised: 14 October 2019 | Accepted: 16 October 2019 | Published on: 27 November 2019

DOI: 10.1002/joc.6367

RESEARCH ARTICLE

International Journal of Climatology RMetS

Development of a near-real-time global in situ daily precipitation dataset for 0000–0000 UTC

Su Yang[1] | Phil D. Jones[2] | Hui Jiang[1] | Zijiang Zhou[1]

[1]National Meteorological Information Centre, China Meteorological Administration, Beijing, China

[2]Climatic Research Unit University of East Anglia Norwich, Norfolk, UK

Correspondence
Su Yang, National Meteorological Information Centre, China Meteorological Administration, Beijing, China.
Email: yangsu@cma.gov.cn

Phil D. Jones, Climatic Research Unit University of East Anglia Norwich, Norfolk UK.
Email: p.jones@uea.ac.uk

Zijiang Zhou, National Meteorological Information Centre, China Meteorological Administration, Beijing, China.
Email: zzj@cma.gov.cn

Funding information
National Innovation Project for Meteorological Science and Technology : Quality Control, Fusion, and Reanalysis of Meteorological Observations, Grant/Award Number: CMAGGTD003-5; National Key R&D Program of China, Grant/Award Number: 2017YFC1501801; National Natural Science Foundation of China, Grant/Award Number: 41805128

Abstract

In this study, we have developed a global in situ daily precipitation dataset based on quasi-real-time sub-daily observations of precipitation totals for the 0000–0000 UTC (Co-ordinated Universal Time) day everywhere in the world. The sub-daily precipitation data from meteorological stations are obtained via the World Meteorological Organization's (WMO) Global Telecommunication System (GTS) and China Meteorological Administration Net (CMANet) archived by the National Meteorological Information Centre (NMIC) in China and the Integrated Surface Database (ISD) released by the National Centers for Environmental Information (NCEI) in the USA. We have combined these three sources into a global dataset, referred to as NMIC. Accumulated precipitation totals (depending on the country and the WMO region) are transmitted at a variety of times on the GTS. Of these, about 4,500 stations report daily for the 0000–0000 UTC day. Here, we significantly add to this, by developing two-way accumulation algorithms to decompose other reported sub-daily totals to shorter intervals, and then re-cumulate them where possible to the 0000–0000 UTC day. Using these algorithms, we increase by 51.1% of the number of stations during 2009–2016 to around 6,800 day^{-1}. Additionally, date boundary adjustment (sliding between 1 and 6 hours either side of 0000 UTC) raises the data volume to between 7,800 and 8,700 day^{-1}. We compare our NMIC product with the First Guess Daily (FGD) product from the Global Precipitation Climatology Centre (GPCC) and GHCN-Daily from NCEI (National Centers for Environmental Information). Root mean square differences between our NMIC and GPCC FGD products over the 2009–2016 period are around 3.4–3.7 mm·day^{-1} and the average consistency percentage is about 75.1–76.8%. Greater differences between NMIC and GHCN-daily are found which are probably due to the non-uniform date boundary in GHCN-Daily.

KEYWORDS

date boundary, global dataset, near-real time, precipitation

1 | INTRODUCTION

Global in situ daily precipitation measurements provide basic data for investigations of the global water cycle, energy balance and other research areas (IPCC, 2014). These data are also important sources of information for extreme weather monitoring, climate change analysis and the validation of both new remotely-sensed observational data and Reanalyses (Adler *et al*., 2003; Hulme, 1991, 1992; Janowiak *et al*., 1998; Rudolf, 1993; Schneider, 1993; Xie *et al*., 2007; Lee and Biasutti, 2014; Beck et al., 2017a, b). Unlike temperature, precipitation is discontinuous, so data products need more stringent requirements for data values, integrity and spatial distribution.

Several research organizations and National Hydrometeorological Services (NMHSs) have developed a number of in situ global precipitation datasets (Hulme, 1991, 1992; Rudolf, 1993; Schneider, 1993; Janowiak *et al*., 1998; Daly *et al*., 2002; Klein Tank *et al*., 2002; Adler *et al*., 2003; Xie *et al*., 2007; Kamiguchi *et al*., 2010; Smith *et al*., 2011; Menne *et al*., 2012; Yatagai *et al*., 2012; Harris *et al*., 2014; Schamm *et al*., 2014). Recent efforts have moved from monthly totals to daily and sub-daily gridded datasets (Schamm *et al*., 2014; Westra *et al*., 2014; NCEI, 2017; Blenkinsop *et al*., 2018). The First Guess Daily product (Schamm *et al*., 2014) developed by the Global Precipitation Climatology Centre (GPCC) and the Global Summary of the Day (GSOD) released by the National Centers for Environmental Information (NCEI, 2017) are two of the widely used global quasi-real-time daily precipitation products. The former uses surface synoptic observational (SYNOP) reports transmitted by the World Meteorological Organization's (WMO) Global Telecommunication System (GTS) as its main data source and cumulates (where necessary) sub-daily totals (1, 3, 6, 12, 18 or 24-hr) to a daily sum for the climatological day for each station. GSOD uses the Integrated Surface Database (ISD) published by NCEI (Smith *et al*., 2011) as its main data source and utilizes the 6, 12, and 24-hr cumulative precipitation to determine daily totals. In China, the National Climate Center (Nie *et al*., 2011) established a global daily precipitation dataset based on the 24-hr cumulative precipitation from the SYNOP messages transmitted by the WMO GTS, and conducted quality control, inspection, and evaluation of these data.

The developers of GPCC's First Guess Daily (FGD) product and GSOD both realize that the definition of the day varies across the world. GPCC's aim with their FGD product is to define this as 0000–0000 UTC (Coordinated Universal Time) everywhere. GSOD tends to use accumulations for the day based on Greenwich Mean Time, however, there are a lot of records that do not end with the midnight observation. Our aim is also to define the day similarly across the world, so 0000–0000 UTC. To achieve this requires much work with the sub-daily precipitation totals available through the GTS, and to understand how this differs within and between WMO's six regions (see WMO, 2015, updated annually online for details of the different regional code practices).

NMHSs share numerous weather messages: hourly, sub-daily and daily precipitation totals are a part of the overall weather reports transmitted over the GTS. Taking 2016 as an example, the number of global sub-daily precipitation data totalled at least 50 million observations. Of these, 1.6 million were daily precipitation totals for 24-hr cumulative values for 0000 UTC. These account for only 3.2% of the total (see Section 2 for details). To increase the numbers of cumulative daily totals at 0000 UTC, other sub-daily precipitation data need to be summed from the cumulated sub-daily totals available (such as those from 3, 6, 9 and 12-hr). Unlike temperature, daily precipitation represents the cumulated amounts in a 24-hr period and it is necessary that measurements should cover the whole period. As evident from WMO (2015) observation schedules over the world are not uniform, and many stations prefer to use local times or regionally-specific schedules. However, there are several types of cumulative rainfall totals (of multi-hourly durations), that appear at various times on the GTS, some of which are overlapping and complementary (see the example shown later in Section 3.1). This means there is likely to be a large amount of sub-daily precipitation information hidden within the GTS. The only way to exploit them efficiently is to use a versatile daily sum algorithm.

The main objective of this study is to develop a dataset of near-real-time global daily precipitation values for the 0000–0000 UTC period that is as complete as possible. We refer to this as the National Meteorological Information Centre (NMIC)'s FGD product (http://data.cma.cn/data/). The data sources used in this study are described in Section 2. Section 3 introduces all the dataset procedures including the daily sum algorithm, the adjustment of the date boundary, data quality tests and comparisons of our NMIC dataset with GPCC's FGD product (the only comparable in situ daily dataset for the 0000–0000 UTC day). Section 4 discusses these evaluations and the performance of these two datasets. Finally, conclusions from this study are presented in Section 5.

2 | DATA SOURCES

2.1 | Real-time meteorological data

The principal source of all real-time meteorological data is that transmitted over the GTS of WMO. However, not all GTS data transmitted from NMHS nodes reach all other nodes (including the operational meteorological centres). Also, some may not be transmitted, sent too late or restricted to only one of the six WMO regions. In order to improve the volume of data, analyses augment the GTS using other additional sources. Some of these sources may also be real time, but some may be delayed mode data requested from NMHSs by operational centres. The ISD (https://www.ncdc.noaa.gov/isd) released by the NCEI (for precipitation) collects and integrates approximately 100 daily and sub-daily sources of precipitation time series data from around the globe, including the GTS. The additional sources used by ISD are given by Smith *et al.* (2011). The number of active station locations has reached 13,000, making ISD one of the world's most extensive global dataset of sub-daily data observations. Data received by the NMIC (http://data.cma.cn/data/) includes GTS data received in Beijing and additionally includes accumulated 1, 3, 6, 12, and 24 hr rainfall totals at 8 times (0000, 0300, 0600, 0900, 1200, 1500, 1800, 2100 UTC) from about 2,400 national meteorological sites over China every day via China Meteorological Administration Net (CMANet). We refer to the data received at NMIC (via GTS and CMANet) as the NMIC data. We expect that a more complete sub-daily data source can be produced by supplementing NMIC with the ISD data.

Figure 1 illustrates a comparison of data volumes of global sub-daily precipitation at meteorological stations in 2016. Clearly, much of the data from NMIC and ISD is in common, but some are unique to ISD, more so at some observing hours than at others. Merging NMIC and ISD sub-daily datasets improves the integrity of the hourly data source effectively with the total number of sub-daily rainfall records reaching 55.9 million adding 4.2 million sub-daily cumulative precipitation amounts to the NMIC received data (CMA and GTS).

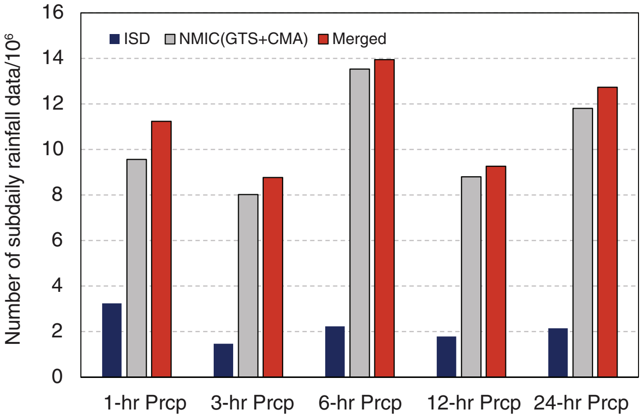


FIGURE 1 Comparison of the data volumes of global sub-daily precipitation totals from meteorological stations in 2016 between different data sources

Figure 2 shows the composition of the combined global sub-daily precipitation totals for 2016. The colour boxes represent the data volumes. It is clear that 0000, 0300, 0600, 0900, 1200, 1500, 1800 and 2100 UTC are the most common precipitation observation times across the world with data volumes of 1, 3, 6, 12, and 24-hr cumulative precipitation measurements amounting to between 8.7 and 12.7 million. Short duration records (≤3 hr) account for 35.7% of the total measurements. A total of 1.6 million daily precipitation records are 24-hr cumulative precipitation recorded at 0000:0000 UTC, making up 3.2% of the total precipitation measurements for the year, but there are another 11.2 million of 24-hr cumulative precipitation measurements recorded for different definitions of the 24-hr day (not 0000:0000 UTC).

2.2 | GPCC's FGD precipitation data

GPCC's FGD precipitation data product are used for comparison to evaluate our product. Our aim here is a comparison. With just two datasets, we are unable to say whether one is better than the other, and we are not trying to do that anyway. The GPCC's FGD precipitation data has been developed by Schamm *et al.* (2014) and has had a variety of users (Brocca *et al.*, 2014; Grams *et al.*, 2014; Schneider *et al.*, 2016; Dietzsch *et al.*, 2017; Ennenbach *et al.*, 2018; Kock *et al.*, 2018; Mellado-Cano *et al.*, 2018). The GPCC FGD product is only based on GTS data received by DWD, the German Weather Service, which is the parent organization of GPCC. GPCC FGD does not make available the station series used in their gridding, nor the specific station locations used. Instead, their gridded products provide grid-box values and the numbers of stations used for each grid-box value. GPCC uses many of the same data quality tests as used here.

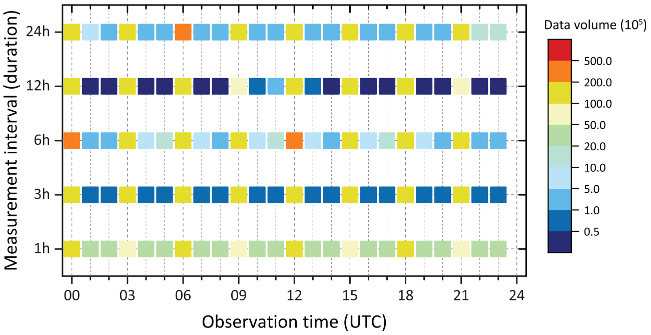


FIGURE 2 The composition of the combined NMIC and ISD global sub-daily precipitation data in 2016. The horizontal axis is observation time, and the vertical axis is the cumulative measurement duration. The colour boxes represent the data volumes

2.3 | GHCN-daily data

GHCN (Global Historical Climatology Network)-daily is an integrated database of daily climate summaries from land surface stations across the globe. GHCN-daily is comprised of daily climate records from numerous sources that have been integrated and subjected to a common suite of quality assurance tests. In general, these delayed data sources are deemed to be more reliable than real-time data, because the data institutes have had the chance to detect and eliminate the errors in real-time data. However, the date boundaries of delayed daily precipitation are non-uniform. There are some daily data that are recorded under local date boundaries and are transmitted on the GTS (and hence taken in by GHCN) directly, for example, the daily precipitation over China before 2013. They probably cause spatial inhomogeneity, particularly for regional analysis when the data from various sources/institutes are used together.

3 | METHODS

As stated, the purpose of this paper is to develop a gridded database of daily precipitation totals for 0000–0000 UTC. This section introduces the methods used to make better use of the existing sub-daily precipitation amounts to produce many more daily totals for 0000–0000 UTC (Sections 3.1 and 3.2). We also discuss quality control of the basic data (Section 3.3, with details of the tests removed to an Appendix) and the interpolation of the station data to a grid (Section 3.4).

3.1 | Methods for estimating 0000–0000 UTC daily values

In general, records of long duration at fixed observation times, for example, 24-hr cumulative precipitation at 0000 UTC and 12-hr cumulative precipitation at 0000 and 1200 UTC, are preferred for our daily precipitation calculations. Just selecting those records that meet our requirements, however, means that a large number of short duration (e.g., 1, 3 and 6-hr accumulations) or cross-day records (where accumulation periods extend across the 0000 UTC boundary) are not used because of the difficulties summing them up. In this section, we show some examples of what can be undertaken to improve the number of 0000–0000 UTC totals. In the next section, this is extended to varying the daily date boundary (0000 UTC) by up to 6 hours.

Figure 3 illustrates an example of daily precipitation calculations at WMO station 11,406 (50.068°N, 12.391°E) on second January 2016. In this case, the 12 and 24-hr cumulative precipitation appeared 6 hr later than the desirable time (0000 and 1200 UTC) so they cannot be directly used for the daily 0000–0000 UTC total. This reporting schedule is typical for WMO Region 6. Our specific method for dealing with this situation is as follows:

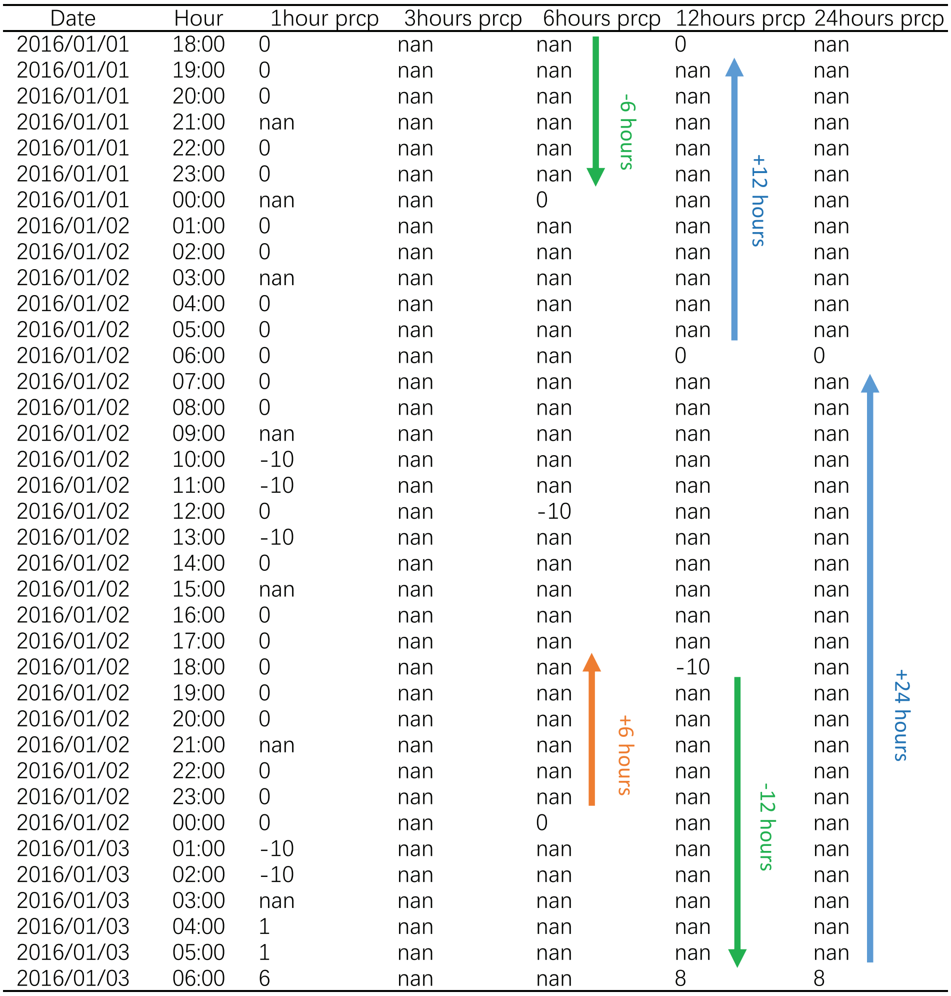
1. Add the 24-hr cumulative precipitation recorded at 0600 UTC first January and the past 12-hr cumulative precipitation recorded at 0600 UTC second January to obtain a 36-hr cumulative precipitation between 1800 UTC first January and 0600 UTC third January (backwards accumulation, upward arrows for 12 and 24-hr, respectively, in Figure 3). It is 0.8 mm.
2. Subtract the 6-hr cumulative precipitation recorded at 0000 UTC second January and the 12-hr cumulative precipitation recorded at 0600 UTC third January to obtain the 18 hr cumulative precipitation between 0000 UTC second January and 1800 UTC second January (forward accumulation, downward arrows for 6 and 12-hr, respectively, in Figure 3. This gives 0.0 mm.
3. Add the 6-hr cumulative precipitation recorded at 0000 UTC third January to obtain the 24-hr cumulative precipitation between 0000 UTC second January and 0000 UTC third January (backwards accumulation, upward arrow for 6-hr in Figure 3). The total rainfall amounts to 0.0 mm. This is consistent with the daily precipitation recorded by the GHCN-Daily (Menne *et al*., 2012) (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/).

Three days' precipitation accumulations are used in this process. The daily amount is obtained by two-way accumulation of the sub-daily precipitation data. The purpose of backward accumulation is to decompose the longer sub-daily duration precipitation totals to shorter durations and create more opportunities to combine (forward accumulation) them to a 0000–0000 UTC daily total. In this study, all observational records have been decomposed to 1, 3, 6, 9, 12, 15, 18, and 21-hr cumulative totals, and then are recombined to develop totals for the 0000–0000 UTC day. Additional daily precipitation totals for 0000–0000 UTC for about 2,300 stations are revealed by this algorithm from 2009 to 2016 (these values are discussed more in Section 4.1).

3.2 | Date boundary setting

In the previous section, we showed how combining sub-daily totals produces more values for the 0000–0000 UTC day. For some stations where we cannot obtain daily

FIGURE 3 Cumulative hourly precipitation records from WMO meteorology station 11,406 (50.068°N, 12.391°E) from January 1 to January 3, 2016 (unit: 0.1 mm). "Nan" = lack of measurement and "−10" = trace precipitation. The first and second columns are date and time, respectively, and the third through seventh columns represent the cumulative precipitation from the past 1, 3, 6, 12, and 24 hr recorded at each time


-6 hours
+12 hours
+6 hours
-12 hours
+24 hours

Date	Hour	1hour prcp	3hours prcp	6hours prcp	12hours prcp	24hours prcp
2016/01/01	18:00	0	nan	nan	0	nan
2016/01/01	19:00	0	nan	nan	nan	nan
2016/01/01	20:00	0	nan	nan	nan	nan
2016/01/01	21:00	nan	nan	nan	nan	nan
2016/01/01	22:00	0	nan	nan	nan	nan
2016/01/01	23:00	0	nan	nan	nan	nan
2016/01/01	00:00	nan	nan	0	nan	nan
2016/01/02	01:00	0	nan	nan	nan	nan
2016/01/02	02:00	0	nan	nan	nan	nan
2016/01/02	03:00	nan	nan	nan	nan	nan
2016/01/02	04:00	0	nan	nan	nan	nan
2016/01/02	05:00	0	nan	nan	nan	nan
2016/01/02	06:00	0	nan	nan	0	0
2016/01/02	07:00	0	nan	nan	nan	nan
2016/01/02	08:00	0	nan	nan	nan	nan
2016/01/02	09:00	nan	nan	nan	nan	nan
2016/01/02	10:00	-10	nan	nan	nan	nan
2016/01/02	11:00	-10	nan	nan	nan	nan
2016/01/02	12:00	0	nan	-10	nan	nan
2016/01/02	13:00	-10	nan	nan	nan	nan
2016/01/02	14:00	0	nan	nan	nan	nan
2016/01/02	15:00	nan	nan	nan	nan	nan
2016/01/02	16:00	0	nan	nan	nan	nan
2016/01/02	17:00	0	nan	nan	nan	nan
2016/01/02	18:00	0	nan	nan	-10	nan
2016/01/02	19:00	0	nan	nan	nan	nan
2016/01/02	20:00	0	nan	nan	nan	nan
2016/01/02	21:00	nan	nan	nan	nan	nan
2016/01/02	22:00	0	nan	nan	nan	nan
2016/01/02	23:00	0	nan	nan	nan	nan
2016/01/02	00:00	0	nan	0	nan	nan
2016/01/03	01:00	-10	nan	nan	nan	nan
2016/01/03	02:00	-10	nan	nan	nan	nan
2016/01/03	03:00	nan	nan	nan	nan	nan
2016/01/03	04:00	1	nan	nan	nan	nan
2016/01/03	05:00	1	nan	nan	nan	nan
2016/01/03	06:00	6	nan	nan	8	8

values with the 0000 UTC date boundary constraint, an adjustment to the date boundary has been undertaken to obtain more global precipitation information. The date boundary is adjusted gradually in steps of 1 hr, up to a maximum adjustment range of ±6 hours. Duplicate daily values found by this procedure will be identified through a stuck value test (see Appendix) in our quality control procedures. A large amount of new daily data appears as the date boundary is adjusted by ±1, ±3 and ± 6 hours. The date boundary for each daily value is labelled. The results for this, discussed in Section 4.1, are expressed as NMIC (0000 ± X UTC), where X is the maximum change in date boundary up to 6 hours.

3.3 | Data quality tests

Quality assessment of all the daily totals produced was performed for all the daily precipitation totals. Due to the positive skew of daily precipitation totals, tests were performed using the cube root of the daily precipitation totals. Transformation using cube roots has long been used for making precipitation totals have a more normal distribution for statistical analysis (Stidd, 1953). Global daily precipitation data quality tests include a spike value test, a stuck value test (a test to identify long runs of the same value in the data series) and temporal and spatial consistency tests (see Appendix). Data quality results are flagged at each step and divided into three categories: credible, suspicious and erroneous. The quality test results at each step have been assembled in the final data quality level assessment. The final quality level is flagged as credible if there is no more than one suspicious test result and no erroneous test results. A value is flagged as erroneous if more than one erroneous test result has been found; otherwise, the final quality level is suspicious.

3.4 | Data gridding

In order for the daily 0000–0000 UTC dataset to be widely used, it is necessary for the stations to be combined into a gridded product. To compare with GPCC's FGD product, 1° data gridding has been undertaken. It would have been preferable to directly compare the daily 0000–0000 UTC values from GPCC for each station, but GPCC only releases data from their gridded products and not their raw station series. GPCC does release the number of stations used for each grid box. If this is zero, then extrapolation in GPCC's FGD from beyond the grid box has occurred. For our gridded product (NMIC), we only use grid boxes that have gauges in our products, so gridded values are only influenced by the measurements in the grid box. This means that our NMIC First Guess Daily product will have missing values for boxes that have no stations. So, the GPCC dataset should be more complete than our NMIC one. For NMIC, within-box interpolation uses the simple inverse distance weighting (IDW) scheme developed by Shepard (1968). Instead of the interpolation of precipitation values in mm, we interpolate percentages using a separate climatology for each station. IDW is widely used in the analysis of precipitation fields (e.g., Willimott et al., 1996 Robeson, 1997) and using percentages from a climatology is better in regions of complex topography (Jones and Hulme, 1996).

The specific steps of data gridding are as follows:

1. Use the GPCC Climatology from 2015 at 0.25° (Schneider *et al*., 2014; Meyer-Christoffer *et al*., 2015a) as the single station climatology value to calculate station sequences of daily precipitation percentage anomalies, by dividing the daily precipitation totals by the climatology.
2. Use IDW to calculate the daily sequence of daily precipitation percentage anomalies for each 1°.
3. Use the GPCC Climatology from 2015 at the 1.0° resolution as the climatology value of the 1° grid (Meyer-Christoffer *et al*., 2015b) to back transform the daily precipitation anomaly value to mm for each box.

As it is well known, precipitation is significantly inhomogeneous in space, particularly so at the daily timescale. The number of stations and their location within the box have a large effect on the gridded value. A simple test to show this influence of observational spatial density on the gridded data has been undertaken by separating the daily rainfall data over China into two different groups and calculating 1° grid precipitation using the same IDW approach. Group (1) includes the entire national meteorological observation network of 2,419 stations, while group (2) is a subset of group (1), and contains only the 189 sites that have been shared with other WMO members through GTS every day.

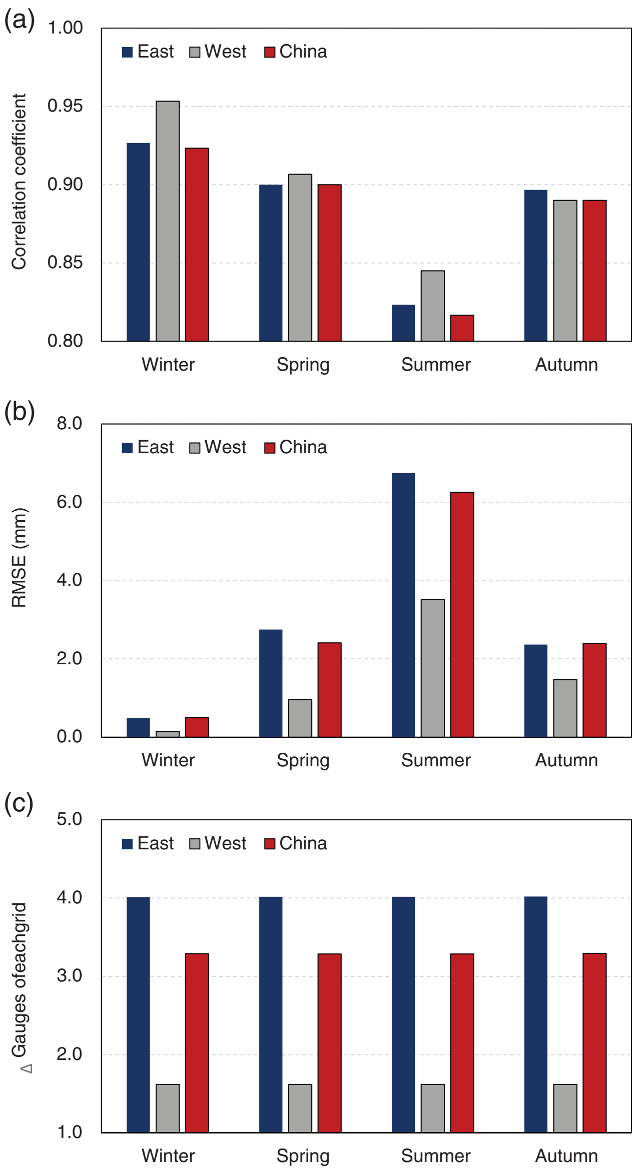


FIGURE 4 The comparison of gridded daily precipitation totals between different sampling groups for China, western and eastern China. Group (1) covers the entire national meteorological observation network (2,419 sites). Group (2) is a subset of group (1), containing only 189 sites which have been shared with WMO members through the GTS every day. Panels (a), (b), and (c) illustrate the RMSE, spatial correlation coefficient of 1°daily gridding precipitation and the difference in gauges falling in each grid between the groups, respectively

Figure 4 illustrates the difference caused by sampling (comparing the gridded daily precipitation totals based on the data from the two groups). We show results for all of China (190 1° grid boxes) and for East China (East of 105°E, 57 1° grids), and West China (West of 105°E, 133 1° grids). Separating China into these two parts (east and west of 105°E) is common practice in China (see, for example, Shi and Xu, 2006; Zhao et al., 2014). For China, monsoon influences are very important, particularly the seasonal characteristics of rainfall in China, so results have also been separated into the four standard seasons. Panels (a), (b), and (c) represent the root mean square error (RMSE, mm·day^{-1}), spatial correlation coefficient and the average difference in gauge numbers between daily grid rainfall datasets based on the groups (1) and (2) data, respectively. Values plotted are the average values of all the grid boxes in each of the three parts of China. The formula for RMSE, together with that for the spatial correlation coefficient and the differences in gauge numbers are defined in the Appendix.

The sample size and distribution barely affect the gridded results in the winter owing to the small precipitation amounts in this season. The RMSE is lower than 0.3 mm·day^{-1} and the spatial correlation coefficient reaches 0.92. In summer, in contrast, the gridded precipitation series is much more influenced by the number of gauges. RMSE is between 4.0 and 6.0 mm·day^{-1} and the spatial correlations are lower at around 0.82. The significantly regional character appears in panels (a) and (b). Therefore, despite the much lower station density in the western region, differences are smaller than those in the much more highly sampled eastern region. This is due to the much drier climate in the western as opposed to the eastern region. As expected also, gridded precipitation amounts, seasonally and regionally, are much more influenced during the wet season. Also, average results for China are strongly influenced by what happens in the wetter eastern part of China. This is important to bear in mind when looking at global-scale comparisons later in Sections 4.2 and 4.3.

4 | RESULTS

In this section, we compare the results from our NMIC FGD dataset, with those from GPCC's FGD precipitation data product. Hereafter, the daily products from NMIC and GPCC will be identified by their organization, with the date boundary adjustment of NMIC product given in the parenthesis.

4.1 | Data volume and coverage

Figure 5 shows the number of daily precipitation stations during 2009–2016. Three different versions of the NMIC dataset are compared to the GPCC dataset and to the lower total of combining the GTS, CMA and ISD gauges where just the 24-hr cumulative precipitation records at 0000 UTC are provided. First, it is clear that the lowest curve is for gauges that provide the 0000–0000 UTC daily precipitation totals directly, at around 4,500 day^{-1}. Our two-way accumulation algorithm without shifting the 0000 UTC boundary enhances the number of stations by 51.1% to around 6,800 day^{-1} that is similar to GPCC. Adjusting the date boundary by 3- and 6-hr adds further to the daily data volume by 14.4 and 26.9%, respectively, giving daily precipitation data volumes reaching about 7,800 day^{-1} (0000 ± 0300 UTC) and 8,700 day^{-1} (0000 ± 0600 UTC). Apart from the lowest curve showing the count of direct 0000–0000 UTC totals, all curves show significant upward trends from 2009 to 2016. This indicates that the increased sub-daily precipitation accumulations in recent years are not caused by more 24-hr cumulative records at 0000 UTC but by the fragmented sub-daily precipitation data which are used efficiently by NMIC's and GPCC's daily statistical methods (the two-way sum up algorithms).

Figure 6 shows the spatial distribution of the NMIC data product on a 1° grid over the globe averaged over the period 2009–2016. Panel a represents the results for 0000 UTC, with panels b and c showing the additional stations provided by the 0000 ± 0300, 0000 ± 0600 UTC date boundary adjustments. Asia, America and parts of Europe are the regions with the highest density for the 0000 UTC date boundary (panel a), but most of Australia is blank, and there are few data in central and eastern Europe and Africa (≤0.25 stations day^{-1} on average). Much daily data emerge for Australia and more data can be found in Russia, India, central and eastern Europe as the date boundary is adjusted to 0000 ± 0300 UTC. As a result of further date boundary adjustments (0000 ± 0600UTC), a considerable amount of new data appears in central and eastern Europe and Africa. To some extent, the adjustment of the date boundary resolves the issue caused by inconsistent observation schedules and improves the data volumes and coverage efficiently.

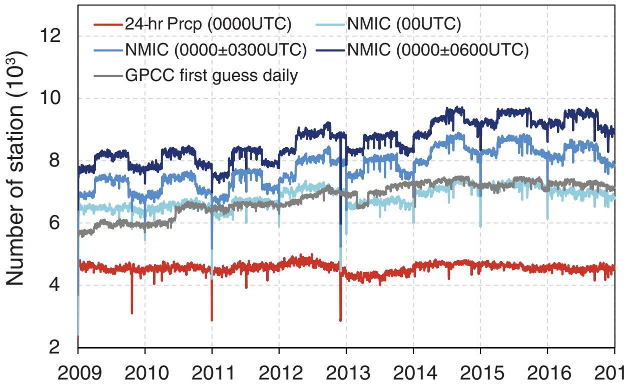


FIGURE 5 The number of daily precipitation data during 2009–2016

Figure 7 shows the differences between NMIC and GPCC (NMIC-GPCC). The darkest shades stand for the grid boxes where the daily precipitation data can only be found in either the NMIC or the GPCC dataset. Without the date boundary adjustment, NMIC appears to have a significant data shortage in Australia and the whole of Europe, but has significant advantages in China and similar performances compared with GPCC over North and South America. Due to the date boundary adjustment, the data scarcity issue in Australia and Europe has been improved. In addition, a considerable number of stations emerge in India and central Africa. Study of Figures 6 and 7 indicate how observing schedules differ between countries.

4.2 | Comparing with real time product (GPCC Frist guess)

In this section, only the grid boxes which have gauges in both NMIC and GPCC are evaluated. Both absolute (RMSE) and relative (Cr, consistency ratio, formula in the Appendix) methods are used in this assessment. All comparisons are based on the daily values in order to reflect all the biases in the records. For example, the bias

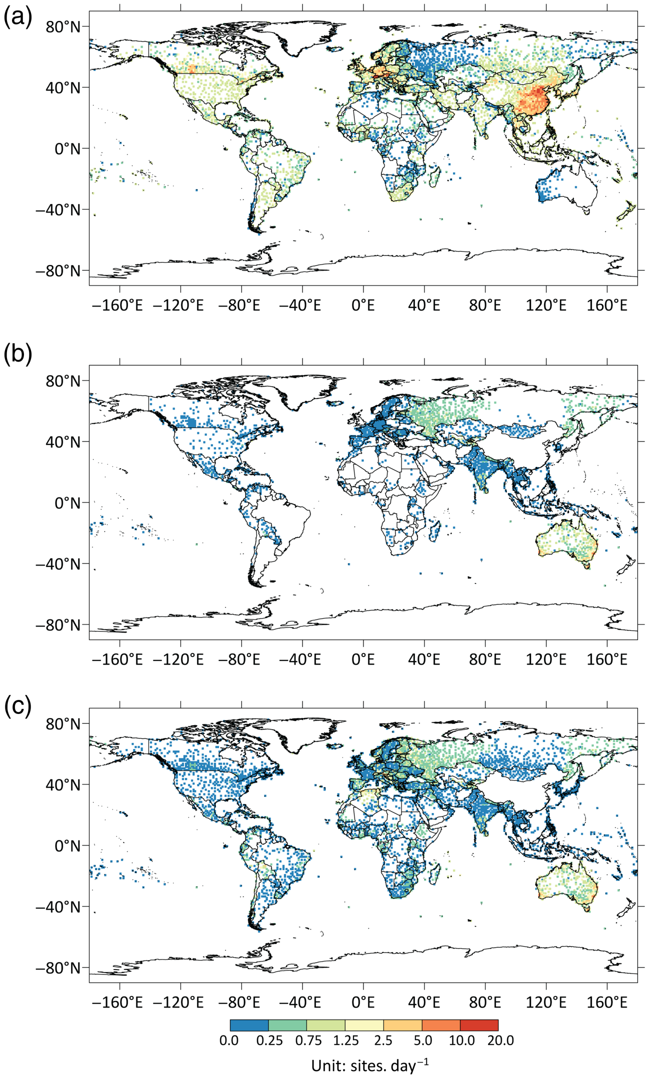


FIGURE 6 Spatial distribution of the NMIC product on a 1° grid over the globe during the period 2009–2016. Panel a represents the number of the sites in each 1° grid using the 0000 UTC date boundary. Panels b and c show the additional site counts under 0000 ± 0300 UCT and 00 ± 0600 UCT date boundaries relative to panel a (0000 UTC). Units are sites day^{-1}

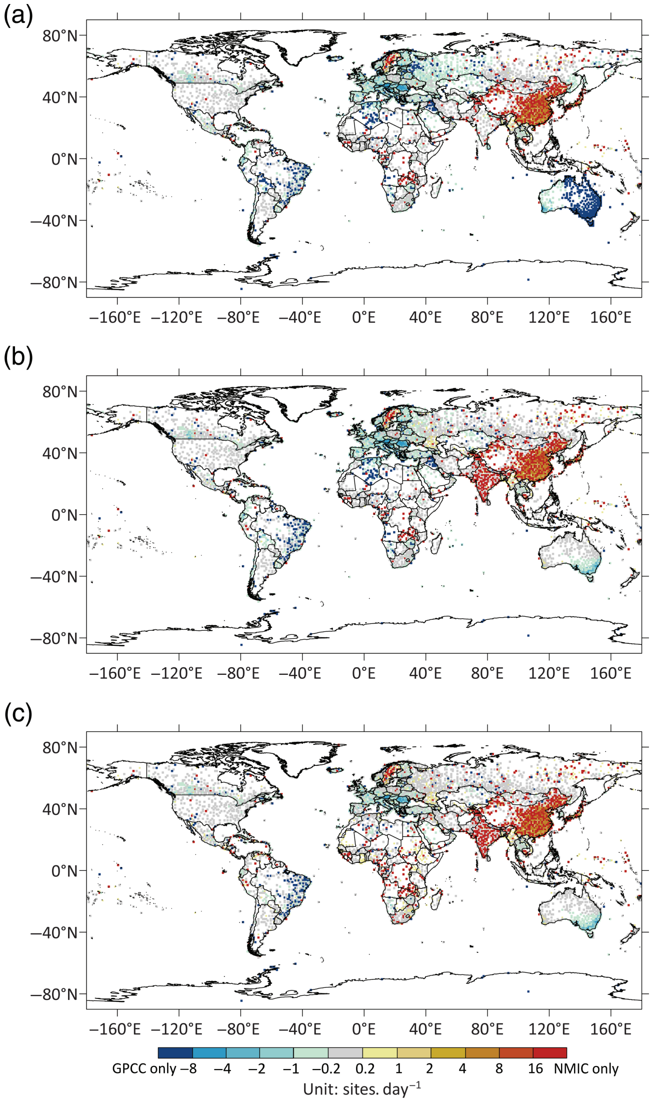

(a)
(b)
(c)
80°N
40°N
0°N
−40°N
−80°N
−160°E −120°E −80°E −40°E 0°E 40°E 80°E 120°E 160°E
GPCC only −8 −4 −2 −1 −0.2 0.2 1 2 4 8 16 NMIC only
Unit: sites. day−1

FIGURE 7 As Figure 6 but for the differences in the number of sites for each 1° grid between NMIC and GPCC (NMIC-GPCC). Panels a, b and c represent the differences between GPCC and NMIC (0000 UTC), GPCC and NMIC (0000 ± 0300 UTC), GPCC and NMIC (0000 ± 0600 UTC), respectively

induced by the date boundary is obvious in the daily total comparisons but is inconspicuous in the monthly amounts comparison because the monthly totals are barely affected by the shifted hours.

Figure 8 displays the RMSE of daily precipitation totals (in mm day^{-1}) for each grid box between NMIC and GPCC during 2009–2016. Panels a, b, c represent the results for the 0000, 0000 ± 0300, 0000 ± 0600 UTC date boundaries, respectively. Averages of RMSE are 3.4 (0000 UTC), 3.4 (0000 ± 0300 UTC) and 3.7 mm·day^{-1} (0000 ± 0600 UTC). In panel a (without date boundary change), a small RMSE (≤2.5 mm·day^{-1}) is evident across Europe, Asia, Canada and Africa, while some grid points' RMSE in South America and South Asia are higher than 10 mm·day^{-1}. As discussed in the previous section (with respect to China) RMSE values are generally low when precipitation totals are low, despite sparse coverage. High RMSE values result when precipitation totals are high, even when the spatial density of coverage is high if the sources differ. China is the best example of this, where RMSE increases from the dry regions (northern part) to the humid regions (southern part). As described in the previous section, the NMIC product has much higher spatial data density over China relative to GPCC. Thus, it can be concluded that the difference in data volumes and their spatial distribution is the main cause of the bias between NMIC and GPCC over China. As noted in

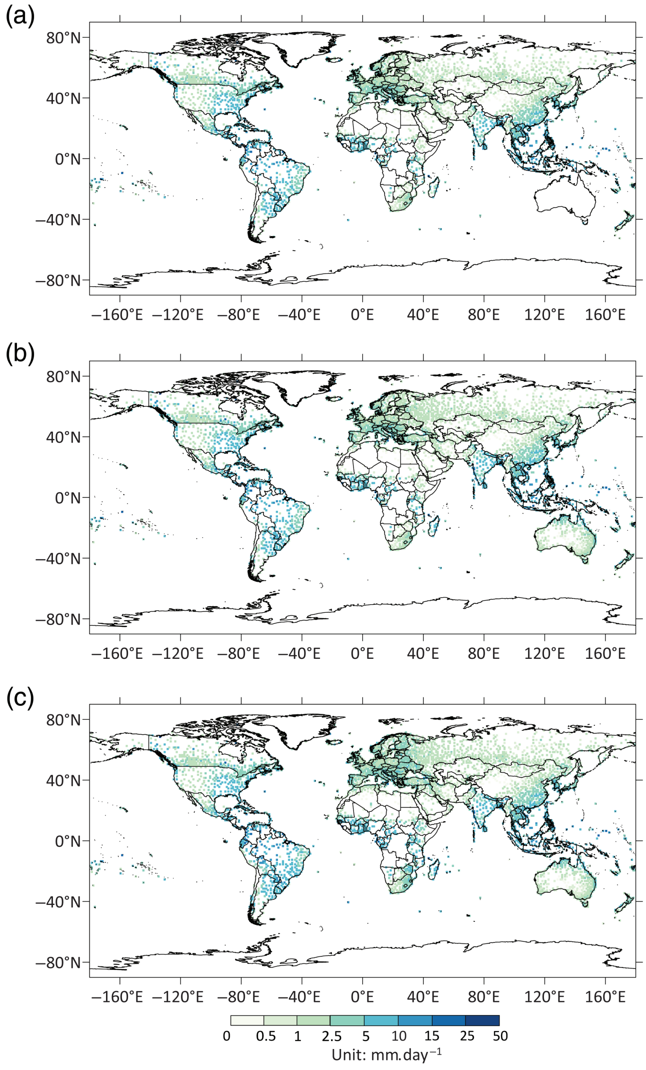


FIGURE 8 Root-mean-square error (RMSE) of daily precipitation totals for each grid box between NMIC and GPCC during 2009–2016. Panels a, b and c stand for the results for 00 UTC, 0000 ± 0300 UTC and 0000 ± 0600 UTC date boundaries. Units are mm day^{-1}

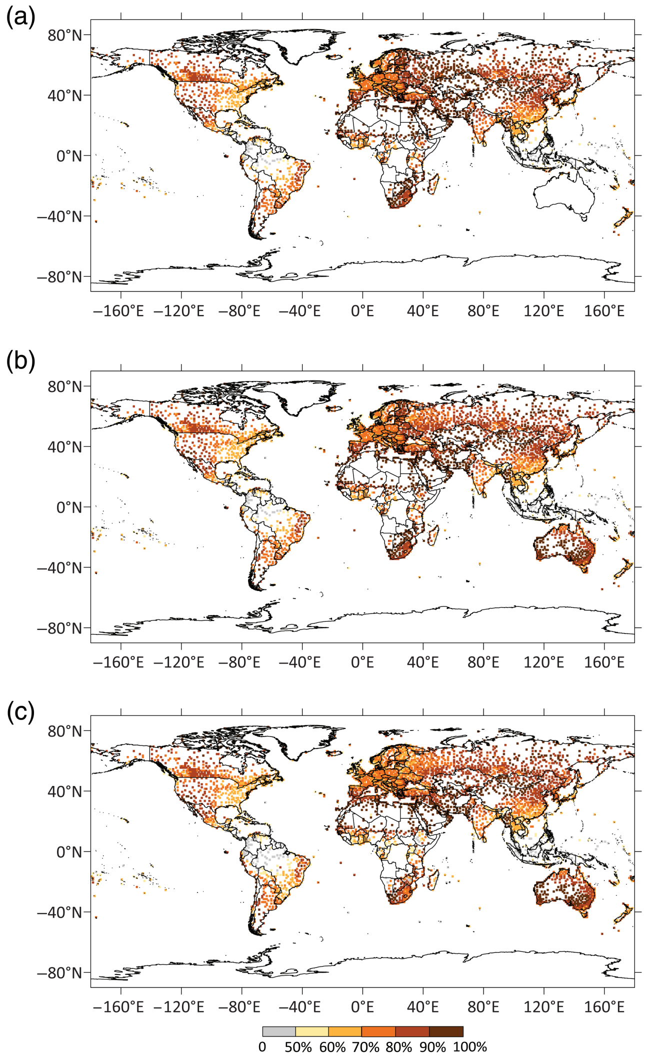


FIGURE 9 The consistency ratio (Cr in %) for each grid box between NMIC and GPCC during 2009–2016. Panels a, b and c stand for the results of 0000 UTC, 0000 ± 0300 UTC and 0000 ± 0600 UTC date boundaries

Section 3.4, the differences are more significant in the wet/humid regions compared to dry regions.

It is interesting to note that the daily precipitation values obtained by the 3-hr date boundary adjustment in panel b (mainly in Australia, Central and Eastern Europe, Russia and India) does not display generally higher RMSE relative to panel a (0000 UTC), while the new points (mainly in Central and Eastern Europe and Africa) in panel c (0000 ± 0600 UTC) show relatively larger differences when compared to GPCC. This means that while the date shift produces a large amount of new data, it does not induce much larger biases when the date boundary adjustment is less than 3 hr and probably resolves some issues caused by different regional recording time rules and transmission. Although more new data appear when the date boundary shifts further (greater than 3 hr), biases compared to GPCC's FGD product increase.

Figure 9 displays the distribution of the consistency ratio (Cr) during 2009–2016 (for the definition of the index, see the Appendix). Cr presents the ratio of the difference between GPCC and NMIC FGD products falling in the tolerated threshold and may not be as useful as the RMSE with respect to absolute rainfall amounts. In comparison with RMSE, Cr is a relative evaluation for the daily rainfall and therefore complements RMSE. Panels a, b and c stand for the results of different date boundaries. The overall average Cr values are 76.8% (0000 UTC), 76.8% (0000 ± 0300 UTC) and 75.1% (0000 ± 0600 UTC), respectively. A similar spatial distribution is also evident in Figures 6 and 8, where Northern China, Russia and Europe show high correspondence (Cr ≥ 80%) between NMIC and GPCC products. Regions with large data volume differences (Figure 6) have humid climates such as Southern China, the northern part of South America and the central part of Africa and display lower consistency values. Although NMIC and GPCC have similar data volumes for North America (Figure 7), low Cr (≤70%) can be found in the wetter eastern parts of North America indicating that the observations in this region in NMIC and GPCC are not same.

For all four global-scale plots (Figures 6–9), white areas indicate missing data in either the NMIC or less likely the GPCC dataset. The largest are over Greenland and the Antarctic. Complete terrestrial coverage has been achieved for the MSWEPv2 dataset (Beck et al., 2017b, 2019) by merging in Reanalysis data (from ECMWF) for missing observing grid boxes. MSWEPv2 is also considerably longer (1979–2015) than our NMIC data (2009–2016), but our aim is an in situ-based daily precipitation dataset for 0000–0000 UTC.

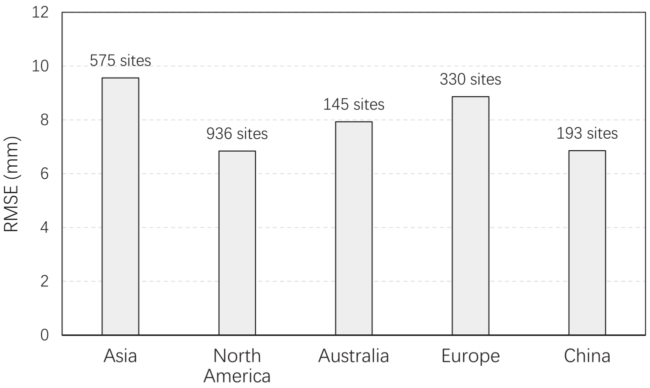


FIGURE 10 The station-based RMSE over five regions between the data from GHCN-daily V3.25 and NMIC daily precipitation dataset for 2012 (2,179 WMO sites in total). The date boundaries of the NMIC product are 00 UTC except for Australia where the date boundary has been shifted up to 3 hours because few daily rainfall amounts can be summed under 0000 UTC. The four continental regions are defined in Table 1

4.3 | Comparison with non-real time product (GHCN-daily)

Figure 10 illustrates the station-based RMSE over five regions (the spatial range of four continental regions are defined as that in Table 1) between the daily data from GHCN-Daily V3.25 and NMIC daily precipitation dataset in 2012 (2,179 WMO sites in total). The date boundaries of the NMIC product are 0000 UTC except for Australia where the date boundary has been shifted by 3 hr because few daily rainfall amounts were obtained under 0000 UTC. The station-based RMSE between NMIC and GHCN-daily changes from 5.9 to 9.6 mm·day^{-1}. Asia and North America are the regions with maximal and minimal diversity, respectively. The difference in date boundary causes a 6.8 mm·day^{-1} RMSE in the Chinese daily value.

Furthermore, Figure 11 illustrates the influence of date boundaries for daily rainfall amounts in Beijing from June to August in 2012 when heavy rainfall occurred frequently. The left panel in Figure 11 is the daily time series. The curves with and without the highest peak are totals summed for 0000 UTC (used in NMIC) and 2000 local mean time (contained in GHCN-Daily = 1200

TABLE 1 : The spatial range of four continental regions in Figure 10

Region	Spatial range
Asia	Lat: 1–81°N, Lon: 45–170°E
North America	Lat: 25–85° N, Lon: −170 to − 20°E
Australia	Lat: −44 to 10° N, Lon: 90–160°E
Europe	Lat: 36–82°N, Lon: −25–66°E

Abbreviations: Lat, latitude; Lon, longitude.

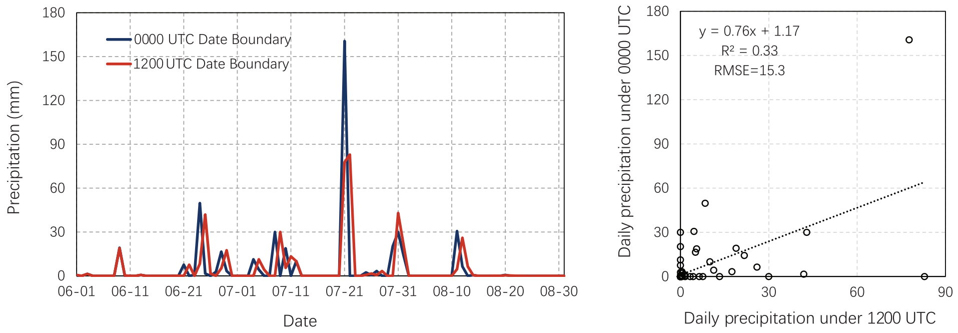


FIGURE 11 Comparison of daily precipitation sums using different date boundaries. The data are from Beijing (China) meteorological observatory from June to August in 2012. The left panel is the time series of daily precipitation under 0000 and 1200 UTC (20:00 local time) date boundary. The right panel is the scatter plot between daily precipitation totals using two date boundaries. The dashed line in the panel is the linear regression and the 1:1 line, respectively

UTC), respectively. Fluctuation of the lines are very similar, but time shifts between them can be found. The huge daily precipitation value (>150 mm) only appears for 0000 UTC. The right panel shows the scatter of daily precipitation, the vertical and horizontal axes represent the values under 0000 and 1200 UTC date boundaries, respectively. Low correspondence between the daily values summed under the two date boundaries can be found. Most of the dots on rainy days (≥0.1 mm) lie far away from the 1:1 line, the correlation coefficient is only 0.33 and the RMSE reaches 15.3 mm·day^{-1}. Slight variations of the date boundary do not change the distribution of daily precipitation totals series too much, but they significantly affect the absolute daily amounts, for example, the maximum daily rainfall amounts in Figure 11. These indicate that it is necessary to develop a uniform date boundary for a daily rainfall dataset, otherwise, there are probably serious spatial inhomogeneity issues for daily precipitation totals induced by various date boundaries used in different countries.

5 | SUMMARY AND DISCUSSION

In this study, we have established a global in situ daily precipitation NMIC dataset-based on quasi-real-time sub-daily data for the 0000–0000 UTC day. To obtain as many daily precipitation series as possible, we have combined the sub-daily data from the CMA, GTS and ISD into a single data source and developed two-way accumulation algorithms for the non-uniform observation schedules and various cumulative durations at various observation times. Furthermore, date boundary adjustment (by up to 6 hr either side of 0000 UTC) has reduced the missing data in some areas caused by different observing and transmission schedules within the six WMO regions. Every record in this dataset has a quality code based on the results of several data quality tests. The NMIC FGD dataset for the 2009–2016 has been compared with GPCC's FGD dataset (the only other near-real-time 0000–0000 UTC precipitation product) and also with GHCN-daily, which is not available in real time. The main results of this study show that:

1. Adding in ISD supplements 4.2 million sub-daily cumulative precipitation measurements per year to the NMIC received data (CMA and GTS).
2. Daily precipitation volumes during 2009–2016 obtained by the algorithm developed in this study are around 6,800 day^{-1}. This is equivalent to the level of GPCC's FGD precipitation data product (Schamm *et al.*, 2014).
3. Adjusting the 0000 UTC date boundary by between 3 and 6 hours further increases daily data volumes by 14.4 and 26.9%. This particularly resolves missing data evident in Australia, but date boundary shifts greater than 3 hr increase biases compared to GPCC's FGD product.
4. The average RMSE between NMIC and GPCC's FGD product is around 3.4–3.7 mm·day^{-1} and the average consistency percentage is about 75.1–76.8%. Regions with differences in the number of gauges used between NMIC and GPCC and areas with humid climates always have larger discrepancies. The difference between NMIC and GPCC in China denotes that the data size and distribution significantly affects the daily gridding precipitation dataset.
5. For both our NMIC and for GPCC's FGD dataset, the number of 0000–0000 UTC series increases during the 2009–2016 period, but this increase does not come from more stations reporting their 0000–0000 UTC totals directly on the GTS during this period.
6. There is a relatively greater diversity between NMIC and GHCN-daily than between NMIC and GPCC's

FGD. This is probably due by the non-uniform date boundary in this non-real time product that can introduce large biases in daily amounts (e.g., the lack of some Chinese data before 2013).

In this study, we utilized sub-daily rainfall data at meteorological stations from the NMIC (GTS+CMA) and ISD. However, there are a large number of non-meteorological observatories (i.e., not run by NMHSs) around the world, for example, hydrological stations in China. These are independent of NMHS meteorological stations, but these sources have not been exploited in this study. Their use could potentially provide an effective way to further enrich data sources for 0000–0000 UTC daily precipitation amounts by combining non-meteorological and meteorological stations into a new data source.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (Grant No. 41805128), National Innovation Project for Meteorological Science and Technology:Quality Control, Fusion, and Reanalysis of Meteorological Observations (Grant No. CMAGGTD003-5) and National Key R&D Program of China (Grant No. 2017YFC1501801). We would like to thank Yu Yu, Lei Xue and Shuli Han for their insightful comments.

AUTHOR CONTRIBUTIONS

S. Y. designed and established the two-way accumulation algorithms of daily rainfall amount, the date boundary adjustment method and the daily rainfall data quality tests. S. Y. and P.D. J. analysed and evaluated the performance of the merged global hourly data and NMIC FGD and wrote the paper. H. J. merged the global hourly data from ISD and GTS. Z. J. Z. helped with the analysis and interpretation.

REFERENCES

Adler, R.F., Huffman, G.J., Chang, A., Ferraro, R., Xie, P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J. and Arkin, P. (2003) The version 2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979-present). *Journal of Hydrometeorology*, 4, 1147–1167.

As for a publisher, NCEI is published by NOAA (National Oceanographic and Atmospheric Administration) and the part of NOAA we are referring to is based at Asheville, NC, USA

Beck, H.E., Van Dijk, A.I.J.M., De Roo, A., Dutra, E., Fink, G., Orth, R. and Schellekens, J. (2017a) Global evaluation of runoff from 10 state-of-the-arthydrological models. Hydrology and Earth System Sciences. *CopernicusGmbH*, 21(6), 2881–2903 https://doi.org/10.5194/hess-21-2881-2017.

Beck, H.E., Van Dijk, A.I.J.M., Levizzani, V., Schellekens, J., Miralles, D.G., Martens, B. and De Roo, A. (2017b) MSWEP:3-hourly 0.25° global gridded precipitation (1979-2015) by merging gauge,satellite, and reanalysis data. *Hydrology and Earth System Sciences.Copernicus GmbH*, 21(1), 589–615. https://doi.org/10.5194/hess-21-589-2017.

Beck HE, Wood EF, Pan M, FisherCK, Miralles DG, Van Dijk AIJM, McVicar TR, Adler RF. 2019. MSWep v2 Global3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bulletinof the American Meteorological Society*, 100(3): 473–500. https://doi.org/10.1175/BAMS-D-17-0138.1.

Blenkinsop, S., Fowler, H.J., Barbero, R., Chan, S.C., Guerreiro, S. B., Kendon, E., Lenderink, G., Lewis, E., Li, X., Westra, S., Alexander, L., Allan, R.P., Berg, P., Dunn, R.J.H., Ekström, M., Evans, J.P. and Holland, G. (2018) The INTENSE project: using observations and models to understand the past, present and future of sub-daily rainfall extremes. *Advances in Science & Research*, 15, 117–126.

Brocca, L., Ciabatta, L., Massari, C., Moramarco, T., Hahn, S., Hasenauer, S., Kidd, R., Dorigo, W., Wagner, W. and Levizzani, V. (2014) Soil as a natural rain gauge: Estimating global rainfall from satellite soil moisture data. *Journal of Geophysical Research: Atmospheres*, 119(9), 5128–5141. https://doi.org/10.1002/2014JD021489.

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(2), 99–113. https://doi.org/10.3354/cr022099.

Dietzsch, F., Andersson, A., Ziese, M., Schröder, M., Raykova, K., Schamm, K. and Becker, A. (2017) A global ETCCDI-based precipitation climatology from satellite and rain gauge measurements. *Climate*, 5(1), 9. https://doi.org/10.3390/cli5010009.

Ennenbach, M.W., Concha Larrauri, P. and Lall, U. (2018) County-scale rainwater harvesting feasibility in the United States: climate, collection area, density, and reuse considerations. *Journal of the American Water Resources Association*, 54(1), 255–274. https://doi.org/10.1111/1752-1688.12607.

Grams, C.M., Binder, H., Pfahl, S., Piaget, N. and Wernli, H. (2014) Atmospheric processes triggering the central European floods in June 2013. *Natural Hazards and Earth System Sciences*, 14 (7), 1691–1702. https://doi.org/10.5194/nhess-14-1691-2014.

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 (3), 623–642. https://doi.org/10.1002/joc.3711.

Hulme, M. (1991) An intercomparison of model and observed global precipitation climatologies. *Geophysical Research Letters*, 18, 1715–1718.

Hulme, M. (1992) A 1951-80 global land precipitation climatology for the evaluation of general circulation models. *Climate Dynamics*, 7, 57–72.

IPCC. (2014) *Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Janowiak, J.E., Gruber, A., Kondragunta, C.R., Livezey, R.E. and Huffman, G.J. (1998) A comparison of the NCEP-NCAR reanalysis precipitation and the GPCP rain gauge-satellite combined dataset with observational error considerations. *Journal of Climate*, 11(11), 2960–2979.

Jones, P.D. and Hulme, M. (1996) Calculating regional climatic time series for temperature and precipitation:methods and illustrations. *International Journal of Climatology*, 16(4), 361–377. https://doi.org/10.1002/(SICI)1097-0088(199604)16:4<361::AID-JOC53>3.0.CO;2-F.

Kamiguchi, K., Arakawa, O., Kitoh, A., Yatagai, A., Hamada, A. and Yasutomi, N. (2010) Development of APHRO_JP, the first Japanese high-resolution daily precipitation product for more than 100 years. *Hydrological Research Letters*, 4, 60–64. https://doi.org/10.3178/hrl.4.60.

Klein Tank, A.M.G., Wijngaard, J.B., Können, G.P., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Müller-Westermeier, G., Tzanakou, M., Szalai, S., Pálsdóttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., van, E.A.F.V., Forland, E., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., López, J.A., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L.V. and Petrovic, P. (2002) Daily dataset of 20th-century surface air temperature and precipitation series for the European climate assessment. *International Journal of Climatology*, 22(12), 1441–1453.

Kock, R.A., Orynbayev, M., Robinson, S., Zuther, S., Singh, N.J., Beauvais, W., Morgan, E.R., Kerimbayev, A., Khomenko, S., Martineau, H.M., Rystaeva, R., Omarova, Z., Wolfs, S., Hawotte, F., Radoux, J. and Milner-Gulland, E.J. (2018) Saigas on the brink: multidisciplinary analysis of the factors influencing mass mortality events. *Science Advances*, 4, eaao2314. https://doi.org/10.1126/sciadv.aao2314.

Mellado-Cano, J., Barriopedro, D., García-Herrera, R., Trigo, R.M., Álvarez-Castro, M.C., Mellado-Cano, J., Barriopedro, D., García-Herrera, R., Trigo, R.M. and Álvarez-Castro, M.C. (2018) Euro-Atlantic atmospheric circulation during the late maunder minimum. *Journal of Climate*, 31(10), 3849–3863. https://doi.org/10.1175/JCLI-D-17-0261.1.

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(7), 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1.

Meyer-Christoffer A, Becker A, Finger P, Rudolf B, Schneider U, Ziese M. 2015a. GPCC Climatology Version 2015 at 0.25°: Monthly Land-Surface Precipitation Climatology for Every Month and the Total Year from Rain-Gauges built on GTS-based and Historic Data.

Meyer-Christoffer A, Becker A, Finger P, Rudolf B, Schneider U, Ziese M. 2015b. GPCC Climatology Version 2015 at 1.0°: Monthly Land-Surface Precipitation Climatology for Every Month and the Total Year from Rain-Gauges built on GTS-based and Historic Data.

NCEI. 2017. *Global Surface Summary of the Day*. https://data.nodc.noaa.gov/cgibin/iso?id=gov.noaa.ncdc:C00516.

NCEI (2017) is an online dataset called GSOD, which is referred to in the paper. There is a link below. There is nothing on the website about who or how to cite.

NCEI though is the National Centers for Environmental Information.

Nie S. P., Luo Y, Li WJ, Wu TW. 2011. A gauge-based global daily precipitation dataset from 1980 to 2009: Quality control and evaluation. *(in Chinese) Advances in Climate Change Research*., 7: 235J242.

Lee, D.E. and Biasutti, M. (2014) Climatology and Variability of Precipitation in the Twentieth-CenturyReanalysis. *Journal of Climate*, 27(15), 5964–5981. https://doi.org/10.1175/JCLI-D-13-00630.1.

Robeson, S.M. (1997) Spherical methods for spatial interpolation: review and evaluation. *Cartography and Geographic Information Systems*, 24, 3–20.

Rudolf B. 1993. Management and analysis of precipit at ion dat a on a routine basis. *International Symposium on Precipitat ion and Evaporation*. Bratislava, 20–24.

Schamm, K., Ziese, M., Becker, A., Finger, P., Meyer-Christoffer, A., Schneider, U., Schröder, M. and Stender, P. (2014) Global gridded precipitation over land: a description of the new GPCC first guess daily product. *Earth System Science Data*, 6(1), 49–60. https://doi.org/10.5194/essd-6-49-2014.

Schneider, U. (1993) *The GPCC Quality Control System for Gauge Measured Precipitation Data. GEWEX Workshop on Analysis Methods of Precipitation on Global Scale*. WMO: Koblenz.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M. and Rudolf, B. (2014) GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. *Theoretical and Applied Climatology*, 115(1–2), 15–40. https://doi.org/10.1007/s00704-013-0860-x.

Schneider, U., Ziese, M., Meyer-Christoffer, A., Finger, P., Rustemeier, E. and Becker, A. (2016) The new portfolio of global precipitation data products of the global precipitation climatology Centre suitable to assess and quantify the global water cycle and resources. *Proceedings of the International Association of Hydrological Sciences*, 374, 29–34. https://doi.org/10.5194/piahs-374-29-2016.

Shepard, D. (1968) Two- dimensional interpolation function for irregularly- spaced data. *Proc 23rd Nat Conf*, 517–524. https://doi.org/10.1145/800186.810616.

Shi, X. and Xu, X. (2006) The spacialcharacteristics of decadally climatical turnover pattern in winter and summerover China. *Chinese Science Bulletin*, 51(17), 2075–2084.

Smith, A., Lott, N. and Vose, R. (2011) The integrated surface database: recent developments and partnerships. *Bulletin of the American Meteorological Society*, 92(6), 704–708. https://doi.org/10.1175/2011BAMS3015.1.

Stidd, C.K. (1953b) Cube-root-normalprecipitation distributions. *Eos, Transactions American Geophysical Union*, 34(1), 31–35. https://doi.org/10.1029/TR034i001p00031.

Stidd, C.K. (1953a) Cube-root-normal precipitation distributions. *Eos, Transactions American Geophysical Union*, 34(1), 31–35 https://doi.org/10.1029/TR034i001p00031.

The GSOD is a dataset released by NCEI (https://data.noaa.gov/dataset/dataset/global-surface-summary-of-the-day-gsod).

Westra, S., Fowler, H.J., Evans, J.P., Alexander, L.V., Berg, P., Johnson, F., Kendon, E.J., Lenderink, G. and Roberts, N.M. (2014) Future changes to the intensity and frequency of short-duration extreme rainfall. *Reviews of Geophysics*, 52, 522–555. https://doi.org/10.1002/2014RG000464.Received.

Willmott, C.J., Robeson, S.M. and Janis, MJ. (1996) Comparison of approaches for estimating time-averaged precipitationusing data from the USA. International Journal of Climatology. *JohnWiley and Sons Ltd*, 16(10), 1103–1115. https://doi.org/10.1002/(SICI)1097-0088(199610)16:10<1103::AID-JOC78>3.0.CO;2-P.

WMO. (2015) *Manual on Codes -International Codes*. Geneva.

Xie, P., Chen, M., Fukushima, Y., Yang, S., Liu, C., Yatagai, A. and Hayasaka, T. (2007) A gauge-based analysis of daily precipitation over East Asia. *Journal of Hydrometeorology*, 8(3), 607–626. https://doi.org/10.1175/JHM583.1.

Yatagai, A., Kamiguchi, K., Arakawa, O., Hamada, A., Yasutomi, N. and Kitoh, A. (2012) Aphrodite constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges. *Bulletin of the American Meteorological Society*, 93(9), 1401–1415. https://doi.org/10.1175/BAMS-D-11-00122.1.

Zhao, Z., Luo, Y. and Huang, J. (2014a) Are Extreme Weather and Climate Events Affected by Global Warming? *Progressus Inquisitiones De Mutatione Climatis(in Chinese)*, 10(5), 388–390.

Zhao, Z., Luo, Y. and Huang, J. (2014b) Are Extreme Weather and Climate Events Affected by Global Warming? *PROGRESSUSINQUISITIONES DE MUTATIONE CLIMATIS (in Chinese)*, 10(5), 388–390.

How to cite this article: Yang S, Jones PD, Jiang H, Zhou Z. Development of a near-real-time global in situ daily precipitation dataset for 0000–0000 UTC. *Int J Climatol*. 2020;40:2795–2810. https://doi.org/10.1002/joc.6367

APPENDIX: FORMULAE USED FOR DATA QUALITY AND GRID COMPARISONS

Data quality tests

Spike test

$$\mathrm{qc}=\left\{\begin{array}{l}\text{credible, } x_i \le x_{\text{extreme}} \\ \text{wrong, } x_i \ge x_{\text{extreme}}\end{array}\right\}$$

$$x_{\text{extreme}}=\min[1.5*\max(\bar{x}), 1{,}800\,\text{mm}]$$

where the subscript i stands for the measurement on the ith day; x_{extreme} is the threshold value for the record and is the smaller value between 1,800 mm and 1.5*max ($\bar{x}$), where $\bar{x}$ represents the subset of the historic measurements in the month (Jan, Feb,...Dec) which removes the largest 5% of values.

Temporal consistency test

$$\mathrm{qc}=\left\{\begin{array}{l}\text{credible, } \sqrt[3]{\mathrm{x_i}} \le \mu+3.5\sigma \\ \text{wrong, } \sqrt[3]{\mathrm{x_i}} > \mu+3.5\sigma\end{array}\right\}$$

$$\mu=\sqrt[3]{}(\text{median}(\boxtimes))$$

$$\sigma=\text{std}\left(\sqrt[3]{}(\boxtimes)\right)$$

where the μ and σ are the median value and *SD* of $\sqrt[3]{}(\boxtimes)$, respectively.

Spatial consistency test

$$\mathrm{qc}=\left\{\begin{array}{r}\text{credible}, \delta_i \le \min[\bar{\delta}+2.5\nabla\delta+10\,\text{mm},\ 200\,\text{mm}] \\ \text{double}, \min[\bar{\delta}+2.5\nabla\delta+10\,\text{mm}, 200\,\text{mm}] < \delta_i < \min[\bar{\delta}+3.5\nabla\delta+20\,\text{mm}, 400\,\text{mm}] \\ \text{wrong}, \delta_i \ge \min[\bar{\delta}+3.5\nabla\delta+10\,\text{mm}, 400\,\text{mm}]\end{array}\right\}$$

$$\delta_i=\sqrt{\frac{\sum_{j=1}^{m}\left(x_j-x_i\right)^2}{m}}$$

$$\bar{\delta}=\text{median}(\delta_i)$$

$$\nabla\delta=\text{std}(\delta_i)$$

where the subscript j stands for the jth neighbouring sites (within the 100 km radius around the candidate site) and m is the total number of neighbouring sites. It should be noted that the $\bar{\delta}$ of a site in an arid region is close to zero. Sudden rain is probably misjudged as wrong data due to small $\bar{\delta}$ and $\nabla\delta$. So 10 and 20 mm are added to the threshold.

Stuck values test

$$\mathrm{qc}=\left\{\begin{array}{r}\text{wrong, } \sigma=0 \text{ and } \mu>0 \\ \text{credible, else}\end{array}\right\}$$

$$\sigma=\text{std}(x_{i-1}, x_i, x_{i+1})$$

$$\mu=\text{mean}(x_{i-1}, x_i, x_{i+1})$$

where σ and μ are the *SD* and smoothing average of 3 days measurements, respectively. Continual no rain days in a dry season are not treated as stuck values.

Grid comparison tests

The root mean square error (RMSE)

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1,n}\left(P_{\text{NMIC}}^{i}-P_{\text{GPCC}}^{i}\right)^2}{n-1}}$$

where the superscript stands for the i-th measurement with the subscript NMIC and GPCC representing the sources of the datasets, respectively. n is the total number of measurements and P is the gridded value.

Spatial correlation coefficients

1. Calculate the 1° gridded values using 2,419 (all CMA sites) and 189 (GTS subset) for China, respectively, producing two gridded daily precipitation datasets.

2. Calculate the partial correlation coefficient between the two-gridded dataset every day over different regions during 2009–2016 and get the average value during the whole period.

$$r_i = \text{correlate}\left(P_i^{\text{all}}, P_i^{\text{sub}}\right)$$

where P_i^{all} represents all gridded daily values (calculated by 2,419 sites) for the *i*th day, P_i^{sub} is similar to P_i^{all} but for the gridded values calculated from 189 sites.

Gauge count differences

1. Calculate the difference in the number of gauges for each 1° grid every day between the two gridded daily precipitation datasets.

2. Calculate the average difference over the chosen region.

Consistency measure (Cr)

In order to evaluate the correspondence with GPCC, a parameter Cr is defined using formulae (1) to (4), where *a* and *b* represent precipitation data from NMIC and GPCC dataset, respectively; the subscript *j* denotes the *j*th day; δ_j is the absolute error; ε_j is the minimum relative error. When ε_j is less than 15% or δ_j is smaller than 1 mm, the two datasets are considered to be consistent, and counter_j is given the value of 1. The $C_r$is the ratio of consistency of each grid.

$$\delta_j = \left|Preci_j^a - Preci_j^b\right| \tag{A1}$$

$$\varepsilon_j = \min\left(\frac{\delta_j}{Preci_j^a}, \frac{\delta_j}{Preci_j^b}\right) \tag{A2}$$

$$\text{counter}_j = \begin{cases} 1, & \delta_j \le 1\,\text{mm or}\, \varepsilon_j \le 15\% \\ 0, & \delta_j > 1\,\text{mm and}\, \varepsilon_j > 15\% \end{cases} \tag{A3}$$

$$C_r = \frac{\sum_{j=1}^{n} \text{counter}_j}{n} \tag{A4}$$