A TEN TIME SMALLER VERSION OF CPC GLOBAL DAILY PRECIPITATION DATASET FOR PARALLEL DISTRIBUTED PROCESSING IN MATLAB AND R

N. POPESCU-BODORIN¹, A. BĂRBULESCU²*¹

¹Technical University of Civil Engineering, 59, Calea Plevnei, 010223, Bucharest, Romania
E-mail: bodorin@live.com
²Transilvania University of Brasov, 5, Turnului Str., 900152, Brasov, Romania
E-mail: alina.barbulescu@unitbv.ro
* Corresponding author, E-mail: alina.barbulescu@unitbv.ro

Received November 27, 2023

Abstract. A ten times smaller version of CPC Global Unified Gauge-Based Daily Precipitation Dataset (1979–2021) is released and described in this paper. Its usability is tested and proved, firstly, by illustrating that the transition to the derived smaller dataset is a case of Pearson correlation transitivity starting with the scale of global yearly data and, secondly, by using the original correlation performance criterion that the original dataset satisfies relative to the set of global actual measurements on the record. Subsequently, the daily, weekly, and monthly data cases are considered and discussed. The dataset is (re)structured for parallel processing in Matlab and R from the level of global daily data. Considering the above arguments and its reduced size, the derived dataset is appropriate to be used with Matlab and R as a replacement for the original dataset, especially for the case when much faster exploratory master-slave parallel and distributed Matlab and/or R tasks will run over locally distributed data on the slaves.

Key words: precipitation, dataset, NOAA, PSL, CPC.

DOI: https://doi.org/10.59277/RomRepPhys.2024.76.703

1. INTRODUCTION

When working with precipitation data recorded for extended periods, globally, on a refined regular spatial longitude-latitude grid, the main issue is the (un)availability of the actual data measured in the grid’s nodes. By chance, some nodes of the spatial grid are close enough to the positions of some meteorological stations to provide the actual precipitation data records for them. There is no actual precipitation data available for many other nodes on the grid, except the data recorded in a smaller or larger neighborhood around these nodes. The same is true for temperature series. The series availability is important for different studies assessing the climate change [1–4].
Different algorithms were proposed to cover the missing data in these gaps and infer plausible spatial grid values using the actual measurement data available in nearby stations. Three of these methods are most largely accepted: two of them, proposed by Cressman [5] and Shepard [6], are based on inverse-distance weighting, whereas the third one is the Optimal Interpolation (OI) method, proposed by Gandin [7]. Three different implementations, proposed by Higgins et al. [8], Legates and Willmott [9], and Xie et al. [10], one for each method mentioned above (Cressman, Shepard, Gandin), respectively, are replicated, tested, and compared in [11]. The study concludes that the OI algorithm [10] performs best overall, better than, and in strong correlation (and agreement) with Shepard’s method [9]. The performances of Cressman’s method [8] proved to be highly affected over the territories with sparse gauge/measurement data. Being an incremental development on top of the method proposed in [12], Cressman’s method is a 3-waves/stages correction algorithm, inevitably highly dependent on the distance thresholds/parameters used to establish (for each correction wave) how large is the neighborhood that propagates the corrections onto those grid points where gauge measurements are not initially available at all. Hence, the hierarchy established in [11] might result from the better choice of the calibration parameters for two of the three methods considered and compared there. Still, this is a topic of one of our future papers – a tribute to the method introduced in [12] in an era when modern AI support for such tasks was entirely missing.

The global spatial precipitation data obtained by applying the OI algorithm (onto the available daily global gauge measurements) is collected as the “CPC Global Unified Gauge-Based Daily Precipitation Dataset” [13]. As a performance metric, a Pearson correlation coefficient of 0.735 between the values calculated directly by the OI algorithm and those inferred by the OI algorithm for withdrawn gauge measurements is calculated for the entire global land area.

In this context, we extract an approximately ten times slimmer version of the original dataset, structured as a collection of daily files, highly usable for Matlab and R parallel distributed processing tasks. It still obeys the correlation performance criterion (mentioned above) and preserves the original dataset's global and local spatial and temporal statistics. The advantages of each method are also emphasized. This approach is new, given that no article offers such a comparison.

2. DATA SERIES AND METHODOLOGY

2.1. DATA BASE

The “CPC Global Unified Gauge-Based Daily Precipitation Dataset” is a dataset of yearly files, the list of fully recorded years being indexed from 1979 to 2021. Each file contains a 3-D array $720 \times 360 \times 365$ indexed by longitude, latitude, and day, respectively, filled with daily precipitation data inferred by applying the OI
algorithm [10] onto the actual gauge measurements recorded in meteorological stations worldwide. It has 2.47 GB and contains netCDF (network Common Data Form) files in netcdf4_classic standard format [14].

Despite being a short and more usable name, the internal title of the dataset, “CPC Global Precipitation,” can create confusion regarding other datasets containing only actual measurements taken in some meteorological stations and not data inferred from them. Therefore, we will further refer to the “CPC Global Unified Gauge-Based Daily Precipitation Dataset” as the “original database” or “original dataset” and, eventually, the “original float dataset.” In contrast, the smaller one we derived from it will be called the “new integer dataset” or the “integer version.”

2.2. MOTIVATIONS FOR REORGANIZING THE DATA

The impediments we found relevant concerning the original dataset are (1) the size, (2) the structure of the collection grouped on years, and (3) the impossibility of organizing efficient Matlab and R parallel processing tasks starting with the level of daily data.

The size of the original dataset is firstly conditioned by the massive redundancy of spatial points with no data inferred there. There are 166555 useless data points in each daily layer (one such daily layer is displayed in Fig. 1), all encoded as −9.969209968386869e+36 value. These points should be, in fact, treated, encoded, and stored only once for all daily data layers as a binary mask of unavailable data (the white areas in Fig. 1) and outside the files encoding/storing the layers, excluded from the storage of all daily layers.

Fig. 1 – Original CPC Global Precipitation Dataset: The data is not available over the large bodies of water (no data on 64.26% of any daily data layer).

To prepare the dataset for Matlab and R parallel processing tasks from the level of the daily data layer, each such layer should be stored as an individual file and not
as a layer within a 3D array of annual data as the original dataset. Moreover, each daily data file in the new smaller dataset version should contain only the data found outside the unavailability index/mask. Additionally, occasionally missing values within the daily data files should be uniquely encoded by zeros to allow future users of the new dataset to handle them properly.

Secondly, the size of the original dataset is conditioned by the data type chosen to encode the inferred precipitation values on the grid. The original datasets store these values using the single-precision floating-point format [15] on 32 bits. Without insisting on the fact that, for some purposes, even storing the data on only four bites is enough (by working with up to 16 fuzzy labels spanning from severely dry to wet), it can be tested that 85.33% of the data within the original dataset is integer data. Only 14.67% of the data has a proper fractional part and requires floating-point storage. The idea that the accuracy of the original dataset resides in 14.67% of its values that happen to have a proper fractional part is doubtable. It suggests the possibility of switching the entire database storage to an integer data type without essential losses in the accuracy and meaning of the dataset.

The truncation from float to integer for the precipitation values over the driest regions (those under 1 mm of rain) is a tiny but meaningful loss of accuracy. This points out the necessity of encoding these values differently to analyze anything concerning these regions. Still, this data’s social and economic value is very limited; consequently, it is not our concern here. We prepare these data mainly for tracing and analyzing the dynamic and quality of surface and groundwater resources over time in different regions of significant social and economic interest in the first place, and not necessarily for keeping the maximum accuracy of the precipitation data over the deserts.

Rounding up and down the float data to integer, on the other hand, will balance the total information loss and gain for the entire globe and period 1979–2021 to a total average relative loss of 0.0066 mm per data point, respectively a yearly relative average loss value of 1.5416E-04 mm of rain per data point, and a daily relative average loss value of 4.2236E-07 mm of rain per data point.

When choosing the integer format for the daily data files in the new dataset, we noted that from 15,706 daily data layers within the original database, 15,676 round to at most 255 integer values and, therefore, qualify for uint8 (unsigned integer on 8 bits) storage. The data in the remaining 30 files can be stored as uint16 (unsigned integer on 16 bits) values. How weak the integer truncation of the original data affects the yearly total precipitation is visually illustrated in Fig. 2. It is not only that the float and the integer versions of total yearly precipitation are unitary correlated, but it happens that for all the years in the dataset, the float and the integer versions of the yearly data are also unitary correlated.

The original data can be reorganized for Matlab and R parallel processing tasks by rounding the entire original dataset to integer (uint8 – most of the time, or uint16 – if required by the situation) and splitting the data into daily precipitation data files.
3. METHODOLOGY

Given the above motivation, the structure of each daily precipitation data file will consist of three internal parts:

1. The integer value 8 or 16, signaling what kind of integers is encoded in the file (uint8 or uint16). If a daily data file counts less than 255 unique integer values, this value will be set to 8; otherwise, to 16.

2. The dictionary of the sorted unique integer values appearing in the daily data. Its functionality is associating the values greater than the upper limit of the integer domain used for encoding a value under this limit. For example, suppose the values to be encoded are (7, 43, 7, 512, 43, 7, 512). In that case, the dictionary will be {(7, 1), (43, 2), (7, 1), (512, 3), (43, 2), (7, 1), (512, 3)}, and equivalently (7, 43, 512), meaning that any value of 7, 43 and 512 occurring in the data will be replaced with 1, 2, and 3, respectively.

3. Within the above dictionary, a zero value will be used to encode occasionally missing data (we recall that systematically missing data is encoded in a mask). All the other integers in the dictionary will appear as one unit larger than the initial values in the original dataset. It is the way we choose to differentiate between occasionally missing data encoded as zeros and zero-measurements that will be encoded as 1.

4. The short integer data (uint8 or uint16) to which the original data has been translated. For the above example, this will be (1, 2, 1, 3, 2, 1, 3).

The fifth component of all global daily precipitation data files is common to all. It is stored externally, outside them, as one single file containing the data unavailability mask illustrated in Fig. 1.
Conveniently, in our integer dataset release, each daily data file will be named “data_YEAR_DDD.mat,” where “data” announces the name of the internal variable stored within the file as being “data,” YYYY is the year encoded in four digits. DDD is the day’s count within the current year – usually ranging between 001 and 365.

The next step is to show that this dataset is equally “good enough” to replace the original dataset for analyzing global yearly data and statistically or fuzzy “good enough” for analyzing global daily, weekly, monthly, and trimestral data.

To prove this assertion, consider an arbitrary data (Pearson) correlated with the original dataset with a correlation value of $c_1$. Denote by $c_2$ the correlation between the original yearly data and its integer version, and by $c_3$ the correlation between the arbitrary data and the integer version derived from the original dataset. Since, in general, the Pearson correlation is not transitive, the condition that should be checked now is [16]:

\[ c_1^2 + c_2^2 > 1 \]  

and then, moreover, the following relation [16] holds true:

\[ c_1 c_2 - \sqrt{(1 - c_1^2)(1 - c_2^2)} \leq c_3 \leq c_1 c_2 + \sqrt{(1 - c_1^2)(1 - c_2^2)}. \]  

4. RESULTS AND DISCUSSION

The first result of encoding the original dataset is that the output of this encoding is an integer dataset of only 188 MB – more than ten times slimmer than the original one (2.47 GB). It contains only a small “*.mat” global daily data files that are easy to use in Matlab and R parallel distributed processing tasks.

The second result is that the encoding procedure produces this much smaller integer dataset.

To prove that this dataset is “good enough” at the yearly global level, one should verify (1) and (2). Since $c_2 = 1$ for all yearly data, the condition (1) is satisfied, and (2) rewrites to:

\[ c_1 = c_1 c_2 \leq c_3 \leq c_1 c_2 = c_1. \]  

Consequently, at the yearly global level, the integer truncation of the original dataset preserves any performance criterion formulated in terms of Pearson correlation between the original dataset and any other similarly-shaped arbitrary data that could be considered. The similarly-shaped arbitrary data include the data obtained by removing some data-point values (considered initially and directly as simple gauge-measurements points) from the global map and their replacement by values computed from neighbor available data using the OI method – as it is the
case of the cross-validation procedure and correlation-based criterion used in [11] to validate the reliability of both the OI method and the original dataset obtained by applying the OI method. Hence, the new integer version of the dataset is equally good enough for purposes concerning the global yearly data as the original data is.

For the entire study period, at the global daily data level, the Pearson correlation between the original and integer dataset ranges from 0.9979 to 1, with an average of 0.9994. Hence, the nature of \( C_3 \) varies from a crisp number identical to \( C_1 \) (when \( C_2 = 1 \)) to a random variable centered in \( C_1 C_2 \) whose values vary from \( C_1 C_2 - \sqrt{(1 - C_1^2)(1 - C_2^2)} \) to \( C_1 C_2 + \sqrt{(1 - C_1^2)(1 - C_2^2)} \), when \( C_2 < 1 \).

For \( C_1 = 0.735 \) (the correlation criterion in [11]) and the average case of \( C_2 = 0.9994 \), the condition (3) rewrites as:

\[
0.7111 \leq C_3 \leq 0.7580 \text{ with } C_1 C_2 = 0.7346. \tag{4}
\]

For \( C_1 = 0.735 \) (the correlation criterion in [11]) and for the worst case of \( C_2 = 0.9979 \), the condition (3) rewrites as:

\[
0.6895 \leq C_3 \leq 0.7774 \text{ with } C_1 C_2 = 0.7335. \tag{5}
\]

Consequently, at the level of global daily data, the integer truncation of the original dataset is slightly weakening (instead of exactly preserving) any performance criterion formulated in terms of Pearson correlation between the original dataset and any other similarly-shaped arbitrary data. The existing data degradation consists of slightly lowering the correlation coefficient (in the worst case considered above, from 0.735 to 0.7335) and weakening the nature of this coefficient from a crisp value to a random variable taking values between limits calculated by (2) – in the interval [0.6895, 0.7774] in the worst case given in (4).

Figure 3a illustrates \( C_2 \) as a random variable for all global daily data (1979–2021). Figure 3b shows the histogram of all \( C_2 \) values and illustrates how and where the random variable \( C_2 \) takes values according to the available data. Figure 3c assumes that for each daily global data, the random variable \( C_3 \) is uniformly distributed between the limits indicated in (2) – a hypothesis that allows us to further illustrate by histogram, in Fig. 3d, the distribution of the random variable \( C_3 \) at the level of all data.

A second view on the experimental results in Fig. 3 is the following. Since the previous results were obtained reasoning under incertitude, \( C_2 \) and \( C_3 \) are, actually, nothing more than fuzzy numbers/words – instead of being random variables. Thus, the histograms (Figs. 3b and 3d) illustrate the shape of the fuzzy membership function for these two fuzzy numbers, \( C_2 \) and \( C_3 \). The same interpretation is valid for Figs. 4 and 5, which illustrate the experimental results for the weekly and monthly global data, respectively.
Fig. 3 – Correlation for global daily float and integer data, and the degradation of a correlation performance criterion from 0.735 crisp to a random variable: a) the random variable $C_2$ (min = 0.9979, mean = 0.9994, max = 1) for all global daily data; b) the histogram of $C_2$; c) the random variable $C_3$ for global daily data under the uniformity hypothesis when $C_2$ varies as shown in (a); d) the histogram of $C_3$ (min = 0.6892, mean = 0.7346, max = 0.7776) under the uniformity hypothesis.

Fig. 4 – Correlation for global weekly float and integer data, and the degradation of the correlation performance criterion from 0.735 crisp to a random variable: a) the random variable $C_2$ (min = 0.9966, mean = 0.9994, max = 1) for all global weekly data; b) the histogram of $C_2$; c) the random variable $C_3$ for all global weekly data under the uniformity hypothesis, when $C_2$ varies as illustrated in (a); d) the histogram of $C_3$ (min = 0.6770, mean = 0.7346, max = 0.7881) under the uniformity hypothesis.

Fig. 5 – Correlation for global monthly float and integer data, and the degradation of the correlation performance criterion from 0.735 crisp to a random variable: a) the random variable $C_2$ (min = 0.9921, mean = 0.9987, max = 1) for all global monthly data; b) the histogram of $C_2$; c) the random variable $C_3$ for all global monthly data under the uniformity hypothesis, when $C_2$ varies as illustrated in (a); d) the histogram of $C_3$ (min = 0.6438, mean = 0.7340, max = 0.8145) under the uniformity hypothesis.
The first incertitude affecting the reasoning here is that we don’t know for sure the total space of possible events for \( C_2 \), i.e. the total range where \( C_2 \) can vary more than we know now, regardless it is computed for daily (Fig. 3), weekly (Fig. 4), monthly (Fig. 5) or yearly (Fig. 6) data.

![Fig. 6 – Correlation for global yearly float and integer data and the preservation of the correlation performance criterion as 0.735 crisp: a) the random variable \( C_2 \) (constant 1) for all global yearly data; b) the histogram of \( C_2 \); c) the random variable \( C_3 \) for all global yearly data under the uniformity hypothesis, when \( C_2 = 1 \); d) the histogram of \( C_3 \) (constant = 0.735).](image-url)

Without saying it explicitly, we assumed the input data for the OI algorithm is statistically well-enough accumulated along so many years and regions, and the OI algorithm is efficient, accurate, and stable enough, such that one can safely assume that the range of \( C_2 \) is precisely known from the accumulated experimental data (Figs. 3–6a). Thus, it is assumed unchangeable; therefore, the list of total possible events for \( C_2 \) is known. Hence the discussion about \( C_2 \) and \( C_3 \) can be done in terms of probability and random variables, regardless they are computed for daily, weekly, monthly, or yearly data. Another assumption we made, covering for a different species of incertitude, is that \( C_3 \) is uniformly distributed within the limits mentioned in (2) – a slightly pessimistic (hence better and safer) hypothesis than assuming normality (for example), but still an assumption to cover for incertitude.

Introducing such hypotheses happens when translating experimental results in terms of probability and statistics, and actually, they are not generally true. Therefore, to allow \( C_2 \) and \( C_3 \) residing within the larger weaker class possible, where they go without any restrictive assumptions, we will better say that, accordingly to the available experimental data, \( C_2 \) and \( C_3 \) are, at the best guess possible now, just fuzzy numbers. As results from the available global daily data, the fuzzy membership function of \( C_2 \) cannot be less or contrary to the histogram in Fig. 3b.

For global daily data, under the uniformity hypothesis, the fuzzy membership function of \( C_3 \) – as it is known from data available for \( C_2 \), relation (2), and uniformity hypothesis – cannot be less or contrary to the histogram in the Fig. 3d. Similar results can be formulated for weekly and monthly data.

The largest incertitude affecting the analysis of any long or short-term precipitation data is the future behavior of the global ocean and atmosphere circulation. Anything could change (going out of some expected patterns known
from the past) in the future, in which case even the distribution of the input data for the OI algorithm will change dramatically; hence, the calibration of the algorithm can go questionable, or at least the algorithm could produce data distributed differently than in the past. Both statistical and fuzzy perspectives will be hardly affected in this case, asking for other probabilistic approaches or higher-order fuzzy numbers to be considered as instruments for encoding the random variables’ evolution or fuzzy numbers, respectively, in time and space. Still, both perspectives are valuable and applicable for analyzing past data, regardless of possible future data changes.

A correlation coefficient of 0.735 is used in [11] to estimate the quality of the OI algorithm and the original dataset, defining the requirements for a dataset to be considered “good enough.” The presented results allow the reader to select the statistical or fuzzy perspective (summarized in Figs. 3–6). One of these two perspectives may be preferred for the case of global daily/weekly/monthly data and the purposes relative to them. The new, integer, ten times smaller database is still either statistically good enough or fuzzy good enough. We presented the arguments for both interpretations and tilted the balance for the second one. For the case of global yearly data, the correlation coefficient is exactly preserved by the new dataset. Hence, in this case, the new dataset is equally good enough by comparison with the original dataset.

5. CONCLUSION

This paper releases a new, ten times smaller, integer version of the „Global Unified Gauge-Based Daily Precipitation Dataset” to allow efficient parallel and distributed processing tasks in Matlab and R.

As long as a correlation performance criterion is adopted, like in [11], since the correlation is preserved exactly (Fig. 6) in the case of global yearly data, the new integer dataset is an ideal replacement for the original dataset for all the purposes concerning global yearly data. For the case of analyzing global daily, weekly, or monthly data, the new dataset is still highly usable. However, the correlation criteria weaken from crisp to fuzzy or from crisp to probable (depending on the reader’s perspective).

The database is available at https://lmrec.org/bodorin/data/noaa/001/ and can be freely downloaded and used.

REFERENCES