Combining different precipitation databases for flash-flood analysis in an urban watershed in Campinas, Brazil.

V. Araujo, MSc.[1*], A.E.S. Abreu, PhD.[1], P.D.P.Costa, PhD.[2] , F.A.M. Falcetta, PhD.[3], O.Yasbek, PhD.[3] , A.C.Corsi, PhD.[3], C.H.M. Porta, BSc. [4]

[1]*Institute of Geosciences, University of Campinas, Campinas, SP, Brazil*

[2]*School of Computing and Electrical Engineering, University of Campinas, Campinas, SP, Brazil*

[3] *Institute for Technological Research, São Paulo, SP, Brazil*

[4] *Federal University of Lavras, Lavras, MG, Brazil*

**Corresponding author email: vinicius.arj@hotmail.com*

Highlights

- Main challenges observed in the unification of the Brazilian rainfall databases are discussed;
- Most datasets had to be discarded because they have a high percentage of data gaps.

Introduction

In the context of flood management, it is necessary to have data on precipitation for hydrological modeling of the watershed and organization of warning and response systems. In Brazil, the main sources of rainfall data currently available are: the HidroWeb portal, managed by the National Water and Sanitation Agency (ANA); the Interactive Map portal, managed by the National Center for Monitoring and Alerting of Natural Disasters (CEMADEN); and state databases, such as the Hydrological Database of the Department of Water and Electric Energy (DAEE) of the state of São Paulo. These data sources present faults, short time periods of recordings and possible errors in measurements, and each one of them has its own characteristics of presentation and availability of data, thus complicating the process of creating a unified database. In this context, this abstract aims to present the main challenges found in the unification of rainfall data from several datasets in Brazil, based on the case study of the Ribeirão Proença Hydrographic Basin, located in the city of Campinas, São Paulo state, Brazil.

Methodology

Study area

The Ribeirão Proença Hydrographic Basin is located in the central region of the city of Campinas, Brazil, with an area of approximately 11.5km² and delimited by coordinates 7,460,000 to 7,463,000 mS and 301,000 to 298.00 mE (zone 23k), having a dense urban occupation and intense suppression of green riparian areas. Several places of the aforementioned stream are critical points of flooding, thus causing frequent damage to local commercial activities and risks to the population. In addition, the study area has two pluviometric sensors installed in its interior, which, in itself, makes it different from most Brazilian urban basins, where sensors are rarely observed in urban hydrographic basins.

Available data

Twenty-four rain gauge stations managed by different government agencies and located in the city of Campinas were evaluated, in addition to the 2 stations that exist inside the studied hydrographic watershed, as illustrated in Figure 1. These data were obtained from 4 different sources: portal Interactive

Maps CEMADEN, the Center for Teaching and Research in Agriculture (CEPAGRI), the Agronomic Institute of Campinas (IAC) and DAEE Database. Python codes were created to join the different datasets. At this stage, the selection, preprocessing and transformation processes were carried out, as proposed by Fayaad *et al.* (1996). As the final product of this stage, there is currently a unified database for carrying out the hydrological modeling of the basin in forthcoming research activities.

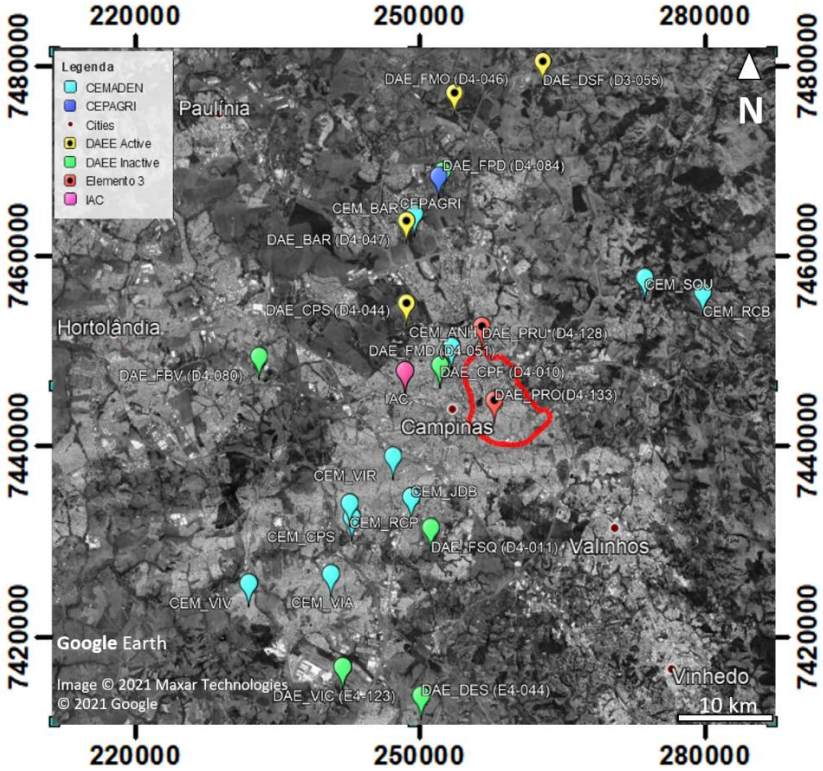

Figure 1. Location of the rain gauge stations in relation to the study area.

Results and discussion

It was necessary to create specific Python codes for each of the different data sources, since each one had particular characteristics and in Brazil there is no standardization in the processes of data acquisition, registration and storage that facilitates the unification of the different databases.

The main challenges observed in the database unification process were: transforming the different types of encodings into a single format (Figure 2A); corrections of different names for the same rain gauge station (Figure 2B); standardization of the date presentation format (Figures 2C and 2D); and standardization in the way of presenting the data (Figure 2D).

A

```
CAMPINAS_PLUVIO_JULHO_2014.csv is iso-8859-1
CAMPINAS_PLUVIO_AGOST_2014.csv is utf-8
CAMPINAS_PLUVIO_SET_2014.csv is iso-8859-1
CAMPINAS_PLUVIO_OUT_2014.csv is iso-8859-1
```

B

```
['vila rica' 'sousas' 'vila aeroporto' 'vila vitoria'
 'barÃ£o geraldo' 'campinas' 'jd das bandeiras'
 'ribeirão das cabras' 'campinas11' 'corrégo serafim'
 'rio capivari']
```

C

DAEE Database

Month/Year (Mês/Ano) \ DAY	1	2	3	4	5	6	7	8	9	10	11
03/2002	---	---	---	---	0,0	0,0	0,0	0,6	7,3	0,0	0,0
04/2002	0,0	0,0	0,0	12,7	0,0	0,0	11,2	3,3	0,0	0,0	0,0
05/2002	3,9	0,0	5,9	27,5	0,0	0,0	12,2	0,0	0,0	0,0	0,0

D

CEMADEN Database

Year/Month/Day	municipio	codEstacao	uf	nomeEstacao	datahora	valorMedida
	CAMPINAS	350950208A	SP	Jd das Bandeiras	2020-07-11 13:10:00	0.00
	CAMPINAS	350950205A	SP	Sousas	2018-03-07 14:10:00	0.00
	CAMPINAS	350950230H	SP	Córrego Serafim	2019-06-13 20:10:00	0.00

Figure 2. Main issues observed in the database unification process. A: Different types of encodings; B: Different names for the same rain gauge station (highlighted in red); C and D: Different standards in data presentation format and different date formats.

It should be noted that conflicting characteristics were observed even between data from the same database, such as the different types of encoding present in the CEMADEN database files (Figure 2A) or the same station having different names in the same database (Figure 2B).

Furthermore, each data source has its own characteristics regarding the temporal frequency in the rainfall records, varying from daily to sub-hourly records. Therefore, all sub-hourly and hourly data were grouped and converted into daily measurements, in order to maintain a correlated measurement pattern in the final database.

Since the sensors located inside the basin operated from 2016 to 2020, a period containing 1826 days, this interval was defined as the study period and, from this, it was possible to evaluate the other sensors according to the availability of data in the aforementioned period and the respective number of failures contained in each dataset.

Thus, from the initial total of 24 rain gauge stations, only 9 were deemed adequate to be used during the hydrologic modelling of the study area, since the others either did not present measurements in the study period or had high values of daily failures (data gaps). Table 1 illustrates the sensors chosen to compose the final research database.

Table 1. Characteristics of the datasets that were chosen to form the pluviometric database of the area.

Database	Name/Code	Dataset	Acquisition Rate	Data Gaps (%)
DAEE	D4-128	2016 – 2020	Hourly	31.9
DAEE	D4-133	2016 – 2020	Hourly	9.94
DAEE	D4-044	1941 – 2020	Daily	18.3
DAEE	D4-046	1958 – 2020	Daily	18.4
DAEE	D4-047	1958 – 2020	Daily	26.4
DAEE	D3-055	2002 – 2020	Daily	11.9
CEMADEN	Vila Vitória	2014 – 2020	Sub-Hourly	21.8
CEMADEN	Sousas	2014 – 2019	Sub-Hourly	15.7
CEMADEN	Vila Aeroporto	2014 – 2020	Sub-Hourly	33.1

Regarding the selected data, the historical series of DAEE presented an average of 18.7% of failures, while those of CEMADEN presented 23.5%. It should be noted that all selected historical series have failures, including the sensors located inside the basin.

Conclusions and future work

There are growing efforts in Brazil to acquire data that enables facing flash-floods in a safer way and precipitation data is becoming more and more available. However, the lack of standardization regarding data acquisition protocols renders the combination of different databases a difficult task. So far, in the study area a considerable part of the acquired data had to be disregarded, due to the high percentage of data gaps. Even though some of the temporal series of the stations chosen for the continuity of the research are short, the tendency is that more data will be available in the next years for stakeholders to reach decisions on flash-flood preparedness and alert.

Acknowlegments

The authors acknowledge the support from the São Paulo Research Foundation (FAPESP, grants number 2017/50343-2 and 2020/00058-2).

References

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37. https://doi.org/10.1609/aimag.v17i3.1230.