An update of Xavier, King and Scanlon (2016) daily precipitation gridded data set for the Brazil

Alexandre C. Xavier[1]
Carey W. King[2]
Bridget R. Scanlon[3]

[1] Federal University of Espírito Santo, Espírito Santo
Department of Rural Engineering, Alto Universitário s/n - 29500-000 - Alegre - ES, Brazil
alexandre.xavier@ufes.br

[2] The University of Texas at Austin, Energy Institute
2304 Whitis Ave Stop C2400, Austin, Texas, USA
careyking@mail.utexas.edu

[3] The University of Texas at Austin, Bureau of Economic Geology,
The Jackson School of Geosciences,
P O Box X, Austin, Texas, USA
bridget.scanlon@beg.utexas.edu

Abstract. The objective of this work is to present an update to the daily precipitation gridded set developed by Xavier, King and Scanlon (2016), where the previous dataset is namely v2 and the new one is v2.1. The v2.1 gridded data uses 9,259 rain gauges relative to 3,630 in v2. We also extend the period of the gridded data by two years (v2.1 is from Jan/01/1980 to Dec/12/2015 while v2 is from Jan/01/1980 to Dec/12/2013). In the generation of v2.1 gridded data set, we tested two interpolation methods: the angular distance weighting (ADW), and the inverse distance weighting (IDW). We selected the ADW interpolations because it presents better skill score in the cross-validation analysis. The v2.1 gridded data are derived using 155% more rain gauges than v2. Almost all skill scores from cross-validation of v2.1 are better than those from v2, and we make this update of the gridded set available to the community.

Keywords: precipitation, interpolation, Brazil .

1. Introduction

Precipitation is the main input variable in modeling of studies about hydrology, meteorology and crops yields. Usually, the computational tools used for hydrologic and crop modeling (e.g., SWAT (NEITSCH; ARNOLD; WILLIANS, 2011) and CROPWAT (SMITH, 1992)) require precipitation data to be both organized and continuous over time (e.g., no missing data in the time series). Generally, to have precipitation data with these characteristics the following tasks are required: i) acquire data from responsible agencies, ii) fill in data gaps (e.g., days with no data at a rain gauge) using data from neighboring stations, and iii) finally, format the data according to the needs of computational analysis tools.

Prior to 2015, a high-quality and available precipitation data set did not exist for Brazil. Recently, in order to provide meteorological data accessible to the scientific community, Xavier, King and Scanlon (2016)[1] published a daily gridded data set for the following variables: precipitation, evapotranspiration, maximum and minimum temperature, solar radiation, relative humidity, and wind speed. These gridded data extend over the period of Jan/1/1980 to Dec/12/2013, and they are continuous in space and time across Brazil. The initial gridded

[1] published online on Oct/2015

data set, version one "v1", was not in the proper format for many common software packages. A second version of the data, "v2," was formatted by the gridded shape and the date format such that it could be opened readily in software such as Panoply[2], Ncview[3], GrADS[4] and Basemap Matplotlib Toolkit[5]. Some recent studies have already used the data: Davi *et al.* (2015) performed a comparison of the v2 precipitation data with the precipitation of the Tropical Rainfall Measuring Mission (TRMM) (MELO et al., 2015); Scarpare *et al.* (2016) used the data to assess water requirements and yield of the sugar cane expansion area in Brazil (SCARPARE et al., 2016); and Davi *et al.* (2016) studied droughts and water resources in the Paraná river basin (MELO et al., 2016).

In an effort to maintain and improve the gridded data set of Xavier, King and Scanlon (2016), this work generates an update of the precipitation variable. This is done by using more data via both additional rain gauges and extending the time period of the data to Dec/31/2015.

2. Data and methods

2.1. The rain gauges data

The rain gauge data were collected during June/2016, from the *Sistemas de Informações Hidrológicas* (Hidroweb: `http://hidroweb.ana.gov.br/default.asp`) and from the *Instituto Nacional de Meterologia* (INMET). The data are originally in millimeters of rain per day or hour (mm/day or mm/hour). The number of rain gauges collected from Hidroweb and INMET were, 8515 and 744, respectively, totaling 9,259. The agencies responsible for the larger number of rain gauges are presented in Figure 1a, where we can cite: ANA (Agéncia Nacional de Águas), DAEE-SP (*Departamento de Águas e Energia Elétrica*) and SUDENE (*Seperintendência do desenvolvimento do Nordeste*). Figure 1b presents the spatial distribution of the rain gauges within the context of river basin boundaries. From visual inspection it is clear that the rain gauges are not uniformly distributed across Brazil or within river basins. The Amazon river basin presents the lowest gauge density while the Paraná river basin the highest density.

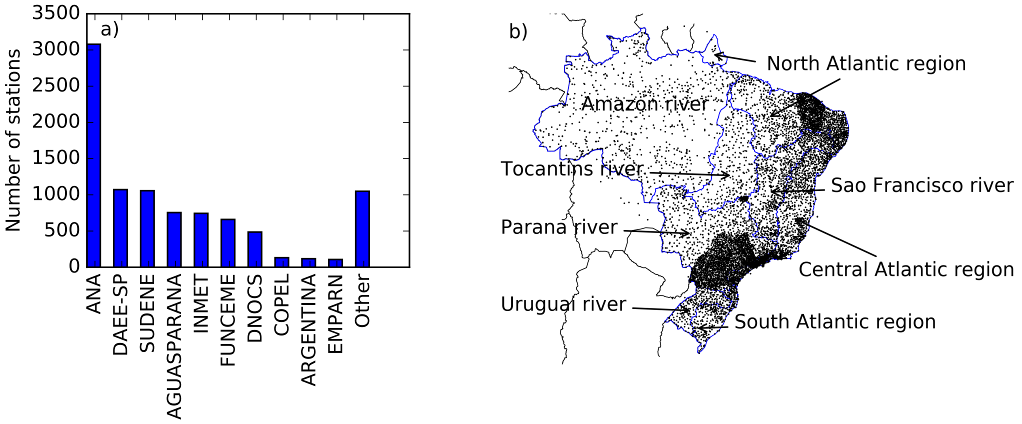

Figure 1: Number of rain gauges per responsible agency (a), and the spatial distribution of rain gauges within river basins in Brazil (b).

After collecting the data, we checked the presence of rain gauge data that are duplicated

[2] see: `http://www.giss.nasa.gov/tools/panoply/`
[3] see: `http://meteora.ucsd.edu/~pierce/ncview_home_page.html`
[4] see: `http://cola.gmu.edu/grads/`
[5] see: `http://matplotlib.org/basemap/)`

(e.g., have the same value and position). As in the previous work of Xavier, King and Scanlon (2016), we did not perform a homogeneity analysis for precipitation data. We only eliminated precipitation data exceeding 450 mm/day (LIEBMANN; ALLURED, 2005) and less than 0 mm/day.

2.2. Interpolation methods and cross-validation

For the v2 gridded data set, Xavier, King and Scanlon (2016) tested six different methods to interpolate precipitation: angular distance weighting (ADW); inverse distance weighting (IDW); average inside the area of each grid of 0.25° x 0.25°; thin plate spline; natural neighbor; and ordinary point kriging. Among them, they verified, using cross-validation analysis that the ADW and the IDW were superior to the others. Thus, in this paper, we only tested those two methodologies to estimate precipitation data. We used the most accurate interpolation method to generate new gridded data (section 3).

The IDW method is a common interpolation technique where each data point is weighted inversely proportional to the distance between the interpolation point and the location of the data informing the interpolated estimate. The ADW method uses two weights: one based on the correlation decay distance (CDD) and the other based in the position of the rain gauges in relation of the query point where we want to do the estimation. For more details on the IDW and ADW interpolation methods see Ly, Charles and Degré (2011), New, Hulme and JonesJones (2000) and Hofstra and New (2009).

We use a cross-validation analysis to determine the best interpolation method to estimate precipitation for each data point in our data set. The cross-validation procedure has two steps. First, the data point is deleted from the rain gauges data set. Second, an interpolation is made for this removed data point (e.g., for its position and day) using both the IDW and ADW procedures.

2.3. Statistics

With the observed and estimated daily data, we use a set of statistics to test the performance of the two interpolation methods. We used the statistics and procedures described in Hofstra et al. (2008) and Xavier, King and Scanlon (2016) (Table 1).

Table 1: Statistics used in cross-validation analysis.

$\mathrm{R} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}\sqrt{(X_i - \bar{X})^2}\sqrt{(Y_i - \bar{Y})^2}}$	$\mathrm{bias} = \bar{Y} - \bar{X}$
$\mathrm{RMSE} = \sqrt{\dfrac{\sum_{i=1}^{n}(X_i - Y_i)^2}{n}}$	$\mathrm{MAE} = \dfrac{1}{n}\sum_{i=1}^{n}\lvert X_i - Y_i \rvert$
$\mathrm{CRE} = \dfrac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$	$\mathrm{CSI} = \dfrac{a}{a+b+c}$
$\mathrm{PC} = \dfrac{a+d}{a+b+c+d}$	

$\bar{X}$ and $\bar{Y}$ are the mean of X and Y, respectively, of the observed and estimated data; n is the number of observed data available; R is the correlation coefficient; RMSE is the root mean square error, MAE is the mean absolute error; CRE is the compound relative error; CSI is the critical success index; and PC is percent correct. PC is used to evaluate the state of precipitation as "wet" or "dry" where a wet day (at a rain gauge) is defined by precipitation greater than 0.5 mm/day; a is number of hits (correct forecast), b is number of false alarms (event was forecast but not observed), c is number of missed forecasts (event occurred but was not forecast), and d is number of correct rejections. CSI is used to evaluate if the interpolated data can to replicate precipitation greater than 0.5 mm/day and the extreme high precipitation days which are those that fall above the 95th percentile (CSI high, CSIH) in the observed and estimated data (see Hofstra et al. (2008)).

We use the aforementioned statistics in two main ways. First, we used the statistics to determine whether ADW or IDW is the better interpolation method when considering all data within the cross-validation process. Second, with the selected interpolation method, we present the statistics, per basin, to indicate the accuracy of the interpolation scheme.

To determine which interpolation method is better, we calculate a skill score based upon the ranking of the statistics. For example, if ADW's R is greater than that from IDW, then ADW would be rank number 1 and IDW ranked number 2. We repeat this procedure for the other statistics, and select the one that has the lowest overall skill score. We either present the results of cross-validation of the previous version, v2 to estimate the improvement (or lack thereof) of this new "v2.1" data set.

3. The grid data set generation

After we selected the best interpolation method as described in the Section 2.3, we used it to generate the new gridded data set of rainfall. The Brazil gridded data has the resolution of $0.25°$ per $0.25°$ such that Brazil has a total of 11,299 cells. For each cell/day of the grid, we calculate a single precipitation value from the unweighted average of 25 individual interpolations within that cell, taken at $0.05°$ spacing.

The codes to evaluate the cross-validation and to generate the new precipitation gridded data set were written in Python[6] lanquage, with the aid of the packages: Numpy (WALT; COLBERT; VAROQUAUX, 2011), Joblib[7], netCDF4[8] and Matplotlib (HUNTER, 2007).

4. Results and discussion

4.1. Rain gauges data set summary

When checking the data, we found 29 pairs of stations with the same coordinates and data, and we deleted one rain gauge for each pair. Thus, we have a total of 9,249 rain gauges in the rain gauge dataset. The amount of deleted precipitation data values exceeding 450 mm/day and less than 0 mm/day were 541 and 486, respectively. The total number of data points with observed data for this updated gridded data set, v2.1, is $\approx$63.2 million, while in the previous data set, v2, had $\approx$32 million (XAVIER; KING; SCANLON, 2016). The total increase of the number of rain gauges used for v2.1 in relationship to v2 is 155% (v2 was done with 3625).

Figure 2 presents the temporal behavior of the number of rain gauges with data to generate each gridded data set v2 and v2.1. The major difference in the number of rain gauges occurs in the beginning of the series, when the number of rain gauges for v2.1 is approximately 4,000 more than the number used in v2. The difference in the number of rain gauges is less in the years 2007-2012, but even in these years, the number for v2.1 is at least 1,000 greater. The steep decrease in the available rain gauges at the end of each time series may be due ANA, which is responsible for maintaining Hidroweb, taking time to update the rain gauge data from the other agencies and make them available via Hidroweb.

[6] see: www.python.org
[7] see: https://pythonhosted.org/joblib/index.html
[8] see: http://unidata.github.io/netcdf4-python/

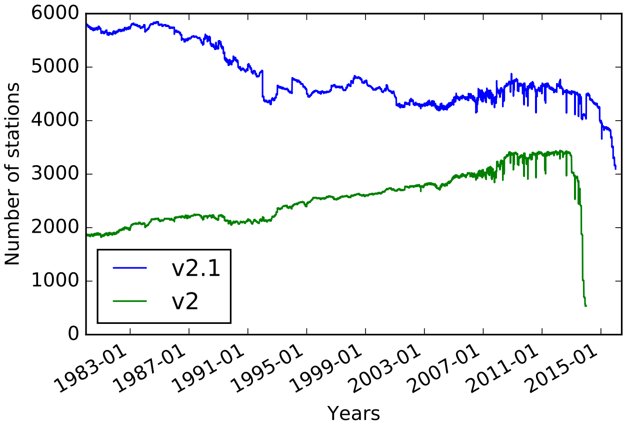

Figure 2: Temporal behavior of the number of rain gauges with observed data for the v2 and v2.1.

4.2. Cross-validation analysis

Table 2 shows the statistical results of the cross-validation process for the IDW and ADW interpolation methods. We also show the results of the IDW interpolation method that was used to generate v2 of the gridded precipitation data set (XAVIER; KING; SCANLON, 2016). The cross-validation for v2.1 and v2, respectively, used a total of ≈63.2 and ≈32 million pairs of observed and estimated data. The ADW method for v2.1 has the best average rank compared to the IDW with v2 (from (XAVIER; KING; SCANLON, 2016)) and v2.1 calculated in this paper. The exceptions are PC and CSI, where ADW ranked in second and third position, respectively.

Because ADW has the best (lowest value) average skill score, it was selected as the interpolation method for this v2.1 data set. The change between the v2.1 and v2 cross validation analysis is also presented in Table 2, where we cite, for example, an improvement (increase) of 6% in R and an improvement (reduction) of 34% in the bias. No improvements exist for PC, while for CSI results are worse in this new cross-validation.

Table 2: Statistics and their respective skill score for precipitation v2.1 and v2. The Δ value is the change (%) in the statistics between ADW (v2.1) and IDW (v.2, Xavier, King and Scanlon (2016))

	Interpolation method						
Statistics	ADW (v2.1)	Rank (#)	IDW (v2.1)	Rank (#)	IDW (v2)	Rank (#)	Δ (v2,v2.1(ADW)) (%)
R	0.648	1	0.633	2	0.609	3	6
bias	0.0027	1	0.0029	2	0.0040	3	-34
RMSE	8.541	1	8.822	2	9.141	3	-7
CRE	0.593	1	0.632	2	0.666	3	-11
MAE	3.384	1	3.366	2	3.709	3	-9
PC	0.784	2	0.798	1	0.783	3	0
CSI	0.520	3	0.530	2	0.534	1	-3
CSIH	0.325	1	0.323	2	0.290	3	12
AV rank		1.38		1.88		2.75	

It is useful to consider the spatial performance of ADW across Brazil. Figure 3a-c shows how three of the statistics observed from the cross-validation, R, PC and CSIH, vary across Brazil. The spatial behavior is similar among statistics, where the eastern regions of Brazil that have a higher density of rain gauges have higher skill score. This behavior of increased skill score in regions with highest rain gauge density was also observed by Hofstra et al. (2008) in Europe.

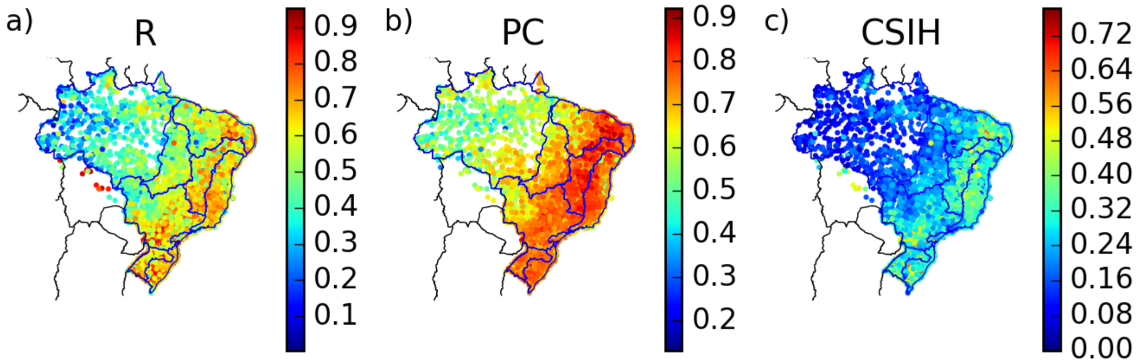

Figure 3: Statistics of cross-validation per rain gauge station: the coefficient of correlation (a), critical success index for high values (b); and percent correct (c).

Table 3 summarizes the cross-validation results per major river in basin. The basins with higher rain gauge density, for example, Paraná and Uruguai basins, have better skill scores, specifically higher values of R, PC, CSI and CSIH and lower values of bias, RMSE, CRE and MAE. On the other hand, Amazon and Tocantins river basins, with lower rain gauge density, have lower skill scores. This trend was also observed with v2 of the data. Generally, the basin-scale skill scores of v2.1 are better than those observed in v2. We can cite for example, the R values for the new data set are better for v2.1 than those in v2, with exception of the São Francisco river basin and Central Atlantic region (XAVIER; KING; SCANLON, 2016). When we compare with the skill scores of Hofstra et al. (2008) for Europe, our skill scores are similar only in the basins for the Paraná, Uruguai, South Atlantic, those with higher rain gauge density.

Table 3: Cross-validation results for interpolation methods per variable and per basin.

Basin	R	bias	RMSE	CRE	MAE	PC	CSI	CSIH
Amazon river	0.375	0.009	12.804	0.923	6.754	0.627	0.483	0.138
Tocantins river	0.507	-0.008	10.438	0.770	4.587	0.743	0.496	0.208
North Atlantic region	0.590	0.005	7.845	0.670	2.895	0.790	0.445	0.285
São Francisco river	0.625	0.022	7.253	0.624	2.399	0.831	0.481	0.342
Central Atlantic region	0.653	0.004	7.454	0.585	2.879	0.777	0.518	0.339
Parana river	0.715	0.008	8.007	0.496	3.185	0.808	0.561	0.378
Uruguay river	0.733	0.001	9.064	0.470	3.596	0.798	0.542	0.412
South Atlantic region	0.730	-0.059	8.561	0.472	3.406	0.791	0.595	0.404

Figure 4a-b shows the statistics for R and bias of the daily cross-validation analysis, from period of Jan/01/1980 to Dec/12/2015. The other statistics (e.g. CRE and CSIH) can be found in the supplementary material (XAVIER; KING; SCANLON, 2016) available at: `https://utexas.app.box.com/v/xavier-etal-ijoc-data`. For each statistic we present (in the raster image) the results of the cross-validation statistics for each day of the year (DOY) and a line plot indicating the average of the statistic for each DOY over the 36 years. For all statistics there is no obvious long-term trend from year to year. For example, the values of R (in each season) are similar in the recent years (e.g., 2005-2015) when compared to the early years (e.g., 1980-1990). R and bias have almost a linear behavior over the DOY, where on average, they are almost 0.6 and 0.0 mm/day respectively (Figure 4a) and b).

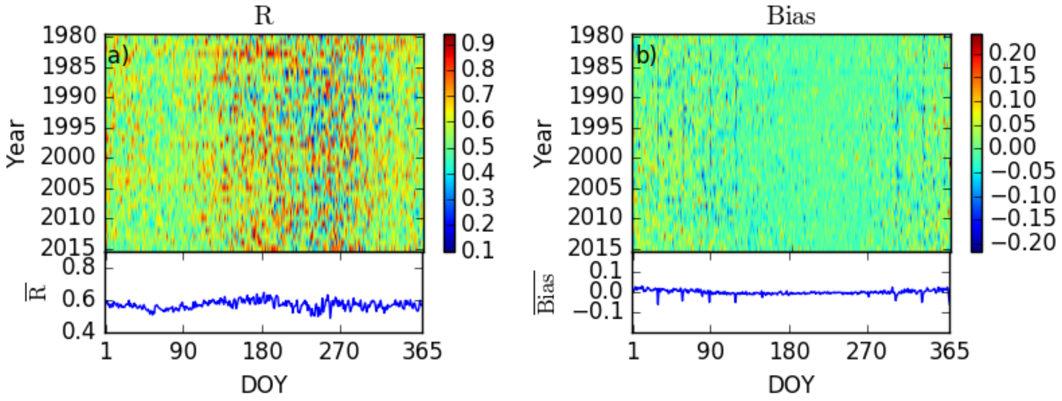

Figure 4: Daily skill scores of the relationship between observed and estimated precipitation when interpolating using ADW.

4.3. Precipitation gridded data set

The precipitation gridded data set v2.1 is available to download at: `https://utexas.app.box.com/v/xavier-etal-ijoc-data`. The files are in the Network Common Data Form (NetCDF), where we include coordinates, dates and other relevant information. The controls files, with the number of stations in the cell and distance of the nearest station with data to the center of the cell, are also available at the site.

5. Conclusion

In this work we presented an update of the precipitation (only) gridded data set of Xavier, King and Scanlon (2016). For this new version, v2.1, we used 155% more rain gauges than used to create the previous data set (v2). In addition, we also extend the gridded set range by two more years (from 2013 to 2015), and the entire v2.1 daily precipitation data set spans Jan/01/1980 to Dec/31/2015, while the previous version, v2, ranged from Jan/01/1980 to Dec/31/2013.

The angular distance weighting (ADW) interpolation scheme provides better statistics skill score than those obtained when using the inverse distance weighting (IDW). Thus, we used the ADW interpolation method for all years and locations generated in the v2.1 data while in Xavier, King and Scanlon (2016) the IDW method was used. Overall, the cross-validation performed for each major river basin scale provides more accurate results for v2.1 (the present study) relative to those observed for v2.

6. Acknowledgements

First author thanks to CNPq for financial support.

References

HOFSTRA, N. et al. Comparison of six methods for the interpolation of daily, european climate data. **Journal of Geophysical Research: Atmospheres**, v. 113, n. D21, p. n/a–n/a, 2008. ISSN 2156-2202. Available from Internet: <http://dx.doi.org/10.1029/2008JD010100>.

HOFSTRA, N.; NEW, M. Spatial variability in correlation decay distance and influence on angular-distance weighting interpolation of daily precipitation over europe. **International Journal of Climatology**, John Wiley & Sons, Ltd., v. 29, n. 12, p. 1872–1880, 2009. ISSN 1097-0088. Available from Internet: <http://dx.doi.org/10.1002/joc.1819>.

HUNTER, J. D. Matplotlib: A 2d graphics environment. **Computing in Science Engineering**, v. 9, n. 3, p. 90–95, May 2007. ISSN 1521-9615.

LIEBMANN, B.; ALLURED, D. Daily precipitation grids for south america. **Bulletin of the American Meteorological Society**, v. 86, p. 1567–1570, 2005.

LY, S.; CHARLES, C.; DEGRÉ, A. Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the ourthe and ambleve catchments, belgium. **Hydrology and Earth System Sciences**, v. 15, n. 7, p. 2259–2274, 2011. Available from Internet: <http://www.hydrol-earth-syst-sci.net/15/2259/2011/>.

MELO, D. C. D. et al. Reservoir storage and hydrologic responses to droughts in the paraná river basin, southeast brazil. **Hydrology and Earth System Sciences Discussions**, v. 2016, p. 1–19, 2016. Available from Internet: <http://www.hydrol-earth-syst-sci-discuss.net/hess-2016-258/>.

MELO, D. C. D. et al. Performance evaluation of rainfall estimates by trmm multi-satellite precipitation analysis 3b42v6 and v7 over brazil. **Journal of Geophysical Research: Atmospheres**, v. 120, n. 18, p. 9426–9436, 2015. ISSN 2169-8996. 2015JD023797. Available from Internet: <http://dx.doi.org/10.1002/2015JD023797>.

NEITSCH, S.; ARNOLD, J.; WILLIANS, J. **Soil & Water Assessment Tool: theorical documentation**. Texas A&M University, 2011. Available from Internet: <http://swat.tamu.edu/media/99192/swat2009-theory.pdf>.

NEW, M.; HULME, M.; JONESJONES, P. Representing twentieth-century space-time climate variability. part ii: Development of 1901-96 monthly grids of terrestrial surface climate. **Journal of Climate**, v. 13, p. 2217–2238, 2000.

SCARPARE, F. V. et al. Sugarcane land use and water resources assessment in the expansion area in brazil. **Journal of Cleaner Production**, v. 133, p. 1318 – 1327, 2016. ISSN 0959-6526. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S095965261630751X>.

SMITH, M. **CROPWAT: A Computer Program for Irrigation Planning and Management**. Food and Agriculture Organization of the United Nations, 1992. (FAO irrigation and drainage paper). ISBN 9789251031063. Available from Internet: <http://books.google.com.br/books?id=p9tB2ht47NAC>.

WALT, S. van der; COLBERT, S. C.; VAROQUAUX, G. The Numpy array: A structure for efficient numerical computation. **Computing in Science Engineering**, v. 13, n. 2, p. 22–30, March 2011. ISSN 1521-9615.

XAVIER, A. C.; KING, C. W.; SCANLON, B. R. Daily gridded meteorological variables in brazil (1980-2013). **International Journal of Climatology**, John Wiley & Sons, Ltd, v. 36, n. 6, p. 2644–2659, 2016. ISSN 1097-0088. Available from Internet: <http://dx.doi.org/10.1002/joc.4518>.