Principal component analysis of pluviometric data

a) Application to outlier detection

Carlos López [1], Elizabeth González[2]

and Jorge Goyret[1]

Abstract

The techniques used for the treatment of a pluviometric data bank during the development and calibration phase of a flow-rain, flow hydrological model are presented.

The calibration phase of this type of models is considerably affected by errors (*outliers*) in the calibration set. Thus it is mandatory to either correct or eliminate those records. We applied a variety of methods for this dataset. Among them, the Principal Component Analysis (PCA) gave the best results.

The developed methodology allows real time quality control of newly acquired data with minimum computer resources requirements, which makes feasible its application in standard equipment. For the present paper, we have defined as errors only those records which differ from the value written down on paper by the observer.

However, it is believed that the PCA is able to detect also other random errors from the observer and even some type of systematic ones, which are still in the investigation phase.

1. Introduction

1.1 Sketch of the problem

In all datasets it exists at least two sources for errors: those intrinsic to the measurement operation and those generated either while keying in or during later process of the information. Both types of errors might have an important effect depending on the particular problem. According to Husain, 1989, "... the failure of many projects of considerable budget can be attributed at least in part, to the imprecision of the hydrologic information available...". In the hydrological model case, the errors propagate themselves in time, and depending on the particular characteristics of the catchment area, its effect might be considerable after significant time lags.

In the daily operation of those models, it is fairly simple for the user to notice significant outliers, because a direct evaluation can be done the day after.

In turn, during the calibration stage of the model, many empirical parameters must be fixed by analyzing thousands of values of measured vs. calculated flow; this comparison can only be made by analyzing global statistics like the standard deviation, etc.

Such fact mix those events obviously erroneous as well as other more subtle ones, which might lead to significant (and uncontrolled) bias in the parameters. For depuration purposes, it has been assumed that values written on paper by the observer are error free, so we try to detect only typing errors. However it will be clear that the method can be

[1] Centro de Cálculo, Facultad de Ingeniería, CC 30, Montevideo, Uruguay

[2] Instituto de Mecánica de los Fluidos e Ingeniería Ambiental, Facultad de Ingeniería, CC 30, Montevideo, Uruguay

easily extended for handling both random and some systematic errors, due to inappropriate sheltering of the instrument the latter and careless operation the former.

The present work should be considered as a natural extension of the task performed during the calibration phase of an hydrologic model of flow-rain, flow type for the Río Negro catchment area. For further details please refer to Silveira *et al*. (1992a y 1992b).

1.2 Methodological background

Regarding outlier detection procedures, the single national registered reference is due to the guidelines produced by the Climatology Department of the Uruguayan National Meteorological Bureau (DNM, 1988). Those specifically related with rain data are very wide and they are mostly connected with the specification of an admissible range.

At an international level, some comprehensive meteorological work has been published (Sevruk, 1982) in order to correct typical systematic errors in each station. In order to do this, they also require values of the surface wind velocity, rain rate, temperature and humidity of the air, etc.

Regarding random errors, the trend is to compare the direct measurements with a model of the phenomena (see for example, Francis, 1986; Hollingsworth *et al.*, 1986). The latter pointed out that for the case of the surface wind, the differences between observations and predictions follow approximately a gaussian distribution. In that case it is relatively simple to detect outlier values in order to analyze them carefully. An important disadvantage of this approach is the considerable amount of information required, as well as the important computer resources involved.

If we disregard (or simply it is unknown) the physical relationship between the variables, the strictly statistical procedures have to be considered. Barnett *et al*., 1984 reviewed and summarizes many techniques which might be of use in this problem. For the multivariate analysis of data he distinguishes two main methodological trends, depending on the fact wether the probability density function is assumed or not.

The first group techniques are named Discordancy Tests; they require an estimation of the parameter of the distribution. There is also some work which assumes that the theoretical distribution has one shape, and the sample another, as proposed for example by O'Hagan, 1990. He applied the idea for an example involving both a Gaussian and a Student's t distribution. Some rules might help in those cases to highlight outlying values. Our case of daily rain rate do not fit readily under such hypothesis, as follows from a simple analysis of its distribution.

The second group identified by Barnett is named as Informal Methods. They neglect the formal aspects of the probability density function, and attempt in turn to exploit certain properties of the distribution. This group includes graphic methods which look for points far from the data cloud; correlation methods, like those described by Gnanadesikan *et al.*, 1972; use of representative generalized distances, techniques usually connected with cluster analysis (see Fernau *et al.*, 1990) and Principal Component Analysis (PCA) among others, etc.

A specific reference related to PCA is the one due to Hawkins, 1974. The author compares four statistics designed to highlight outliers. Hawkins assumed that each observation belongs to a gaussian distribution, an hypothesis which do not hold for the rain; however the concepts that can be derived are similar to the one considered here as well as the results obtained working with coal samples.

2. PCA in brief

PCA is a widely applied multivariate technique (see Richman, 1986 as a general review; Pio *et al*., 1989 for air pollution; White, 1991 for rain, etc.). It might transform one set of correlated measurements into new series of uncorrelated readings, which in turn let consider each one as an independent variable.

Moreover, the new variables minimize the remaining RMS. which might be helpful to distinguish the physics from the noise. In this work we did not attempt to rotate the obtained components, as suggested by Richman, 1986; White, 1991 among others, a process which is supposed to improve the interpretability of components more related with the physics.

2.1 Theoretical aspects

Hereinafter we will denote as $p_i(\tau_k)$ the precipitation value for time τ_k (k=1..r) at station i (i=1..n). The temporal mean at station i will be denoted with an overbar, $\overline{p}_i$.

Given a set of rain readings for a given time $p_i(\tau_k)$ they can be represented together by a vector $\mathrm{P}_{(n,1)}(\tau_k)$ which belongs to the R^n space (fig. 1). Each k-th point of the cloud corresponds to a date τ_k. The origin of coordinates is taken at the baricenter of the cloud, with components $\overline{p}_i$ which will be denoted as $\mathrm{P_M}$.

It is possible to show that it exists a direction $\vec{e}_1$ (unique in the general case) which minimizes the sum of squares S_1

$$S_1 = \sum_{k=1}^{r} \overline{M_k H_k}^2 \tag{1}$$

as sketched in fig. 1. The direction $\vec{e}_1$ does not depend on time τ_k. It will be denoted as $a_1(\tau_k)$ the projection OH_k, which is also named score in the literature. Each term in S_1 can be interpreted as the L^2 norm of the vector

$$\mathrm{P}(\tau_k) - \mathrm{P_M} - a_1(\tau_k).\vec{e}_1 \tag{2}$$

This expression shows that for any time τ_k the data vector is explained as the sum of a constant vector plus a multiple of a constant vector. The statistic S_1/r can be interpreted as the unexplained variance by an approximation by a single term.

Similarly a vector $\vec{e}_2$ can be found in order to minimize the remaining variance, so

$$S_2 = \sum_{k=1}^{r} \left| P(\tau_k) - P_M - a_1(\tau_k).\vec{e}_1 - a_2(\tau_k).\vec{e}_2 \right|^2 \tag{3}$$

being $a_2(\tau_k)$ the projection over the direction $\vec{e}_2$ of the vector $\mathrm{OM_k}$. Even from geometric arguments it can be shown that $\vec{e}_1.\vec{e}_2 = 0$.

We can apply the procedure up to $\mathrm{S_n}$. Lebart *et al.*(1977) demonstrates that $\vec{e}_\mathrm{i}$ are eigenvectors of the covariance matrix, defined as

$$C = \left\{ c_{ij} : c_{ij} = \sum_k \left(p_i(\tau_k) - \overline{p}_i \right).\left(p_j(\tau_k) - \overline{p}_j \right) \right\} \tag{4}$$

and that the eigenvalues λ_i are directly related with the sum S_i. It can be shown that the scores time series $a_i(\tau)$ and $a_j(\tau), i \neq j$, have null crosscorrelation. If we denote as D the diagonal matrix with the eigenvalues λ_i in the diagonal, and E the matrix holding the eigenvectors $\vec{e}_i$ as columns, we can prove:

$$C = E.D.E^T \tag{5}$$

In what follows we will use the term *principal components* to refer to the eigenvectors $\vec{e}_i$, and as scores the time series of the associated projections $a_i(\tau)$. It should be noticed that the index i is not related with a pluviometric station.

Summing up, it exists a lineal transformation which relates the observed time series $p_i(\tau), i = 1..n$, with the scores $a_i(\tau)$ which can be written in matrix form as

$$P(\tau) = P_M + E.A(\tau) \tag{6}$$

being P_M the vector holding the mean precipitation of the period, and $A(\tau)$ a vector holding the scores.

$$P(\tau) = \begin{bmatrix} p_1(\tau) \\ \vdots \\ \vdots \\ \vdots \\ p_n(\tau) \end{bmatrix}; P_M = \begin{bmatrix} \overline{p}_1 \\ \vdots \\ \vdots \\ \vdots \\ \overline{p}_n \end{bmatrix}; A(\tau) = \begin{bmatrix} a_1(\tau) \\ \vdots \\ \vdots \\ \vdots \\ a_n(\tau) \end{bmatrix}; E = \begin{bmatrix} \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ \vec{e}_1 & \vec{e}_n & \cdots & \vec{e}_{n-1} & \vec{e}_n \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \end{bmatrix} \tag{7}$$

Except in pathological cases, matrix E is not singular, thus once the rain measurements $p_i(\tau), i = 1..n$ are given, it is possible to obtain the scores $A(\tau)$ by applying the expression:

$$A(\tau) = E^{-1}.(P(\tau) - P_M) \tag{8}$$

It will be useful later to show that Eq. 6 can be rewritten as

$$P(\tau) = P_M + \sum_{i=1}^{i=n} a_i(\tau).\vec{e}_i \tag{9}$$

2.2 The need for a progressive depuration

The eigenvectors $\vec{e}_i$ (denoted also as *patterns*) are calculated using an available cloud of data points. It might exist a small number of unlikely values (outliers) which might affect to some extent the patterns themselves. In Silveira *et al.*, 1991 it has been shown that even in a population of r=4000 points, only two outliers might significantly affect the patterns. This fact makes mandatory that a recursive depuration effort strategy is to be adopted. On early stages we will look only for those more evident values. As it will be shown, the process can go later to look for more subtle cases.

3. Application of the technique to a particular case: the Río Tacuarembó catchment area

3.1 General characteristics of the study area

Despite the work considered a substantially greater area, we restrict ourselves for this analysis to the Río Tacuarembó catchment area, of 20.000 km^2, located at 32° S, 55°W at 400 km of the Atlantic Ocean. The area can be characterized by a smooth orography, with heights lower than 500 m, few valleys and lakes. The monthly mean for the rain is within 74 and 120 mm/month. The study period is nearly of 15 years, from Jan 1$^{st.}$ 1975 to Dec 2$^{nd.}$ 1989, value clearly over the threshold suggested by Hawkins, 1974.

3.2 A brief description of the compared methods

a) Outlying values of the univariate series

This method is fairly simple, and requires the calculation of a "feasible" range for the values recorded in each station: whenever any record is outside it, it is pointed out as a candidate to be in error. In the given dataset it is usual to mistype records taken in mm/day as taken in tenths of mm/day. This values could be found only if the mistyped record is over 100 mm/day, but the procedure is impractical for other cases.
For the daily rain example this method can detect only events clearly outlying by excess, but on the other hand it is impossible to suggest a zero value reading as an error, because over 80% of the population is exactly zero.

b) Discrepancy of the Thiessen's spatial mean series

The first stage requires that the mean average of the rain is calculated by the Thiessen method (Jácome Sarmento *et al.*, 1990) using different subsets of stations taken from the n available for each day. Thus different time series arise, and when compared if they differ "too much" the particular day is checked.
The results obtained (not presented here) let say that this method gives a much powerful test that the one before; true errors exist in nearly 30% of the selected dates (Silveira *et al*., 1991). Moreover, the errors themselves need not to be outlying.

c) Outlying values of the multivariate series

For the Río Tacuarembó dataset, typically two out three days have some missing value. Then, we must distinguish two situations for each time τ:

c.1) All n stations have readings

c.2) Some values are missing

In the first case, it is possible to calculate the n scores $a_i(\tau)$. If for some i, $a_i(\tau)$ is not within the i-th specified range, all n records used for calculate the scores should be checked. The specified ranges were determined by analyzing the probability density function of the scores for the whole period.

In the second case, an imputation procedure is required. It might be nearest neighbor or any other. Using the same dataset López *et al.*, 1994 analyzed the performance of four methods for missing value imputation are compared, being the most efficient the Penalty of Principal Components, so we apply it here.

Once imputated the missing values, we are in the position to apply the criteria of c.1) by checking each of the scores. However, both here and at Silveira *et al.*, 1991 we relaxed the criteria, and the date was checked if any of the imputated records is negative or bigger than 100 mm/day. For further details, please see López *et al.*, 1994.

In figures 3 and 4 the typical probability density function (pdf) for both the weakest and strongest scores are shown. For the range determination, we restrict ourselves to symmetric ones with a single parameter α_i. For each i, α_i is selected in order to make valid over 96% of the events. If the pdf is nearly symmetric (as for example patterns 2, 3, ... 17, see fig. 3 and 4) this rule implies to reject approximately 2% of each tail of the distribution. For heavily skewed distributions (pattern 1, fig. 3) we reject only from one side of the pdf.

4. Results

4.1 With simulated errors

In order to test the ability of the method for this problem, we select a subset of n = 13 carefully revised stations which have less than 5% of missing records for the period of r = 5450 days (nearly 15 years).

We selected at random a set of 2832 ternas of station-date-value which is around 4% of the population. The wrong values for rain were generated by a mechanism which attempts to replicate the pdf of the real data. In fig. 5 we show the distribution for positive values.

We applied the suggested method in order to detect the artificial errors. In tables I and II we presented the total number of error detected discriminated by step. Between the first and second column, the difference is in the recalculation of the limits α_i. The detected errors in the first columns were ignored in order to calculate the new α_i, but they are expected to be detected in the second sweep. Neither the covariance matrix nor the eigenvectors were recalculated.

Another possibility is detect-correct-recalculate. The results are presented in the first column. For the second depuration, we eliminate the outlying values detected and both the covariance and its eigenvectors are recalculated.

We show in bold the results for days with missing values. In table I, we express the results in relation to the *number of revised values checked against paper*. In table II, we present the results in relation with the total *number of errors yet in the population.*

The results show that it is more convenient to change the limits α_i rather than recalculate the eigenvectors. Thus for two sweeps it can be found 81% of the wrong values, which affects 49% of the revised days.

If we want to recalculate the pattern as soon as we detect the first 571 errors, in the second depuration we found only 186 errors, which account only for 21% of the days to check.

Such behavior was not observed while working with the raw data: even very few errors affected significantly the patterns, requiring in turn a couple of iterations in order to stabilize them.

	First sweep $\frac{\text{total detected}}{\text{total revised}}.10^3$	*Second sweep* $\frac{\text{total detected}}{\text{total revised}}.10^3$
First depuration	$\frac{360+\mathbf{211}}{7644+\mathbf{6318}}.10^3 = 41$	$\frac{2065+\mathbf{215}}{54067+\mathbf{6435}}.10^3 = 38$
Second depuration	$\frac{151+\mathbf{35}}{5798+\mathbf{5863}}.10^3 = 16$	$\frac{1784+\mathbf{40}}{50924+\mathbf{5837}}.10^3 = 32$
Third depuration	$\frac{68+\mathbf{36}}{4966+\mathbf{7514}}.10^3 = 8$	$\frac{276+\mathbf{39}}{9555+\mathbf{7397}}.10^3 = 19$

Table I: Evolution of the depuration process in relation with the data to be checked. Terms in the table follow the schema $(A+B)/(C+D).10^3$, being A: wrong values detected in full days; B: wrong values detected in incomplete days (**in bold**); C: number of records revised in complete days and D: number of records revised in incomplete days (**in bold**).

	First sweep $\frac{\text{total detected}}{\text{total not yet found}}$	*Second sweep* $\frac{\text{total detected}}{\text{total not yet found}}$
First depuration	$\frac{360+\mathbf{211}}{2832} = \frac{571}{2832} = 0.20$	$\frac{2065+\mathbf{215}}{2832} = \frac{2280}{2832} = 0.81$
Second depuration	$\frac{151+\mathbf{35}}{2832-571} = \frac{186}{2261} = 0.08$	$\frac{1784+\mathbf{40}}{2261} = \frac{1824}{2261} = 0.81$
Third depuration	$\frac{68+\mathbf{36}}{2261-186} = \frac{104}{2075} = 0.05$	$\frac{276+\mathbf{39}}{2075} = \frac{315}{2075} = 0.15$

TableII: Evolution of the depuration process in relation with the remaining errors. Terms in the table follow the schema $(A+B)/C$, being A: wrong values detected in full days; B: wrong values detected in incomplete days (**in bold**) and C: wrong values in the database yet to be found.

4.2 Over real errors

In a real situation a table like Table II cannot be created. It is required also a criteria to stop the procedure: we decided to stop as soon as no new errors (*true* errors) were found. We define as *true* error all those cases which the number in the files do not coincide with the one written on paper.

In early stages we worked for full days (with no missing values) over a set of 21 stations. Two phases could also be distinguished.

In the first one, after performing the PCA calculations, we removed the worse errors. They were identified because even not significantly affecting the mean vector, the first and second patterns were completely distorted (see Silveira *et al*., 1991 for details). This stage corresponds with rows 1, 2 and 3 of Table III.

In the second phase we selected those days which scores $a_i(\tau_k)$ exceed the allowable value. The measurements for such day were checked against paper, and corrected if any discrepancy exist. Then we recalculate the Principal Components and the process start again.

For each score a_i the limits were estimated either as three times the standard deviation, or were simply ignored. Despite the criteria of the *three times* is a well known boundary valid for the gaussian distribution, the method do not requires neither imply it.

During the task it has been observed that some days were systematically pointed out as suspicious, even though they agree with the paper. We performed a subjective analysis in a case per case basis, and we classified further the values as *consistent* and *dubious*. The former were associated normally to heavy rain events concentrated in space; the latter show very different values even in very near stations. They were temporally removed from the database in order to not affect the PCA calculations (see González *et al.*, 1991).

The process ends when all dubious values coincide with paper. In table III the evolution of the depuration process is shown. In the first three stages, we only look for gross errors, checking essentially the scores associated with the leading patterns, which explains the low number of days affected.

Another important point is the measurement of efficiency. The column headed with η in Tables III and IV shows a number which even being independent of the number of stations seems to be pessimistic; in practice it is more representative the one indicated by column G (measured as errors per revised day), because in most cases the error was so obvious that by merely checking one or two our of the 21 values were enough to locate the error.

Stage	A	B	C	E	F	G	η
1	9	21	6	51			34
2	354	326	154	448		186	87
3	222	267	336	475	395	174	83
4	70	83	206	286	219	126	60
5	72	60	8	132	105	106	51
6	41	2	29	111	18	65	31
7	9	1	12	115	13	19	9
8			1	113	2	1	0
9				109		0	0

Table III: Evolution of the depuration of real errors for the full days (i.e. without missing values). We analyzed 21 stations. Keys to table: A.- Wrong values; B.- The digital value do not exist on paper; C.- Dubious value; E.- Total number of days checked; F.- Days not considered before; G.- (A+B+C)/E*100 Total number of errors for each 100 days revised; η = (A+B+C)/(21*E)*1000 Total number of errors for 1000 values checked.

Regarding the days with missing values, we applied the Penalty of Principal Components method (described by López *et al.* 1994). In those days with zero rain readings we simply assign zero to the missing values. In other case, we penalized the 10 weakest scores using as weights the reciprocal of the variance. Table IV shows the work in different stages, being all percentages to the total number of values revised in each stage.

Stage	A	B	C	D	E	F	G	η
1	344	314	220	945	457		399	210
2	56	27	65	57	495	94	41	22
3	117	118	138	37	558	179	73	39
4	52	69	118	21	536	94	49	26
5	17	36	36	10	586	53	17	9
6	21	12	34	6	560	30	13	7
7	19	20	9	1	659	15	7	4
8					659		0	0

Table IV: Evolution of the process of real errors for days with missing values. We analyzed 19 stations. Keys to table: A.- Wrong values; B.- The digital value do not exist on paper; C.-Dubious value; D.- Data exist on paper but were not digitized; E.- Total number of days checked; F.- Days not considered before; G.- (A+B+C+D)/E*100 Total number of errors for each 100 days revised; η =(A+B+C+D)/(19*E)*1000 Total number of errors for 1000 values checked

5. Conclusions

We have described and presented a methodology for multivariate quality control based upon Principal Component Analysis (PCA). The results, considering the effort involved can be regarded as satisfactory. In a controlled experiment we succeeded in identify one error every two days checked, finding that way over 80% of the known errors.

The required computer time can be considered minimal. The heaviest part is the calculation of the covariance matrix and its associated eigenvectors, an operation which is performed a limited number of times.

Considering that for each event it is only required a linear transformation, it is possible to apply the method in real time even with hand held computers.

6. Acknowledgments

Carlos López designed both the methodology and the experiment. Elizabeth González was in charge of the real errors phase, and also conducted the bibliographic search. Both authors jointly analyzed the obtained results. Jorge Goyret implemented in part the algorithms, and performed all calculations related with the described experiment.

It should be mentioned that this work is an extension of the task accomplished under the contract "Development of an hydrological model for the Río Negro catchment area" funded by UTE. The permission for using the information and publish the results is gratefully acknowledged.

7. References

Barnett, V.; Lewis, T., 1984. "Outliers in statistical data" John Wiley and Sons, 463 pp.

DNM, 1988. "Procedimientos para el control de calidad climatológico" Informe interno de la Dirección Nacional de Meteorología, Nov. 1988, 20 págs.

Fernau, M.E.; Samson, P.J., 1990. "Use of Cluster analysis to define periods of similar meteorology and precipitation chemistry in eastern North America. Part I: Transport Patterns" Journal of Applied Meteorology, V 29, N 8, 735-750.

Francis, P.E., 1986. "The use of numerical wind and wave models to provide areal and temporal extension to instrument calibration and validation of remotely sensed data" In Proceedings of A workshop on ERS-1 wind and wave calibration, Schliersee, FRG, 2-6 June, 1986 (ESA SP-262, Sept. 1986)

Gnanadesikan, R.; Kettenring, J.R., 1972. "Robust estimates, residuals and outlier detection with multiresponse data" Biometrics, V 28, 81-124.

González, E.; Morales, C., 1991. "Depuración de la base de datos pluviométricos de la cuenca del Río Tacuarembó". Informe interno preparado para el Departamento de Hidrología del Instituto de Mecánica de los Fluidos e Ingeniería Ambiental. 11 pp.

Hawkins, D.M., 1974. "The detection of errors in multivariate data, using Principal Components" Journal of the American Statistical Association, V 69, 346, 340-344.

Hollingsworth, A.; Shaw, D.B.; Lonnberg, P.; Illari, L.; Arpe, K. and Simmons, A.J., 1986. "Monitoring of observation and analysis quality by a data assimilation system" Monthly Weather Review, V 114, N 5, 861-879.

Husain, T., 1989. "Hydrologic uncertainty measure and network design" Water Resources Bulletin, V 25, N 3, 527-534.

Jácome Sarmento, F.; Sávio, E.; Martins, P.R., 1990. "Cálculo dos coeficientes de Thiessen em microcomputador". En las Memorias del XIV Congreso Latinoamericano de Hidráulica, Montevideo, Uruguay (6-10 Nov., 1990). V 2, 715-724.

Lebart, L.; Morineau, A.; Tabard, N. 1977. "Techniques de la Description Statistique: Méthodes et logiciels pour l'analyse des grands tableaux". Ed. Dunod, París. 344 pp.

López, C.; González, J. F.; Curbelo, R., 1994. "Análisis por componentes principales de datos pluviométricos. b) Aplicación a la eliminación de ausencias" Estadística, 46, 146, 147, pp. 55-83 También Publicación Técnica del Centro de Cálculo PTCECAL2/92, Centro de Cálculo, Facultad de Ingeniería, CC 30, Montevideo, Uruguay.

O'Hagan, A., 1990. "Outliers and credence for location parameter inference" Journal of the American Statistical Association: Theory and Methods, V 85, N 409, 172-176.

Pio, C.A.; Nunes, T.V.; Borrego, C.S.; Martins, J.G., 1989. "Assesment of air pollution sources in an industrial atmosphere using principal components and multilinear regression analysis" The Science of the Total Environment, V 8, 279-292.

Richman, M.B., 1986. "Review article: Rotation of principal components" Journal of Climatology, V 6, 293-335.

Sevruk, B., 1982. "Methods of correction for systematic error in point precipitation measurement for operational use" World Meteorological Organization WMO 589, Operational Hydrology Report 21, 89 pp.

Silveira, L.; López, C.; Genta, J.L.; Curbelo, R.; Anido, C.; Goyret, J.; de los Santos, J.; González, J.; Cabral, A.; Cajelli, A., Curcio, A.,

1991. "Modelo matemático hidrológico de la cuenca del Río Negro" Informe final. Parte 2, Cap. 4. 83 pp.Silveira, L.; Genta, J.L.; Anido Labadie, C., 1992. "HIDRO URFING - Modelo hidrológico para previsión de caudales en tiempo real- Parte I: Simulación de los procesos hidrológicos en el suelo" . Publicación Interna del Dpto. Hidrología, IMFIA 1/92, Instituto de Mecánica de los Fluidos, Facultad de Ingeniería, CC 30, Montevideo, Uruguay.

Silveira, L.; Genta, J.L.; Anido Labadie, C., 1992. "HIDRO URFING - Modelo hidrológico para previsión de caudales en tiempo real- Parte II: Transformación en cuenca, ruteo y criterios de calibración y verificación" . Publicación

Interna del Dpto. Hidrología, IMFIA 2/92, Instituto de Mecánica de los Fluidos, Facultad de Ingeniería, CC 30, Montevideo, Uruguay.

White, D., 1991. "Climate regionalization and rotation of principal components" International Journal of Climatology, V 11, 1-25