Toward an Optimal Model based on Inequality Measures for Treatment of Historical & Real Time Flood's Dataset

El Mabrouk Marouane
LaSIT Laboratory,
Science Faculty,
UAE,
Tetuan, Morocco
elmabroukmarouane@gmail.com

Ezziyyani Mostafa
LaSIT Laboratory,
Technology and Science Faculty,
UAE,
Tangier, Morocco
ezziyani@gmail.com

Essaaidi Mohammad
LaSIT Laboratory,
Science Faculty,
UAE,
Tetuan, Morocco
essaidi@ieee.com

***Abstract*—Flood is always a problem that Morocco tries to overcome it, because of the climate. The climate in Morocco can be divided into five sub-areas, determined by the different influences that the country suffers: oceanic, Mediterranean, montagnard, continental and saharan that's why Flood forecasting becomes a challenge for Morocco. Flood forecasting and control the water flow and water level on the surface is very critical to reduce the impacts while the flood disaster events. The flood forecasting model requires the management of huge spatial datasets, which implies data acquisition, storage and processing, as well as manipulation, reporting and display results. Thus, to reach an excellent prediction in terms of accuracy, it's important to implement a model which be interested by manipulation of the historical datasets from the database in order to minimize the response time of the decision.**

In this paper, we present a new model for treatment and for comparison by using the GINI Coefficient and the Variance Coefficient in this model which has two access modes to handle historical inundations informations according to the rainfall, the runoff and the water level. The main idea is to use the Inequality Measures to compare the observed distribution with the reference distribution, in other words compare the several data received from the sensors with data already stored in the database to have an appropriate decision about flooding without going through the decision support system for Real Time Flood Forecasting and Warning.

Keywords—flood, forecasting, variation coefficient, GINI coefficient, data mining*, *database, decision.

I. INTRODUCTION

Floods are hydrological events caused by abnormally high amount of water input and insufficient discharge capacity [1]. It's also a natural seasonal phenomenon which corresponds to a rapid and temporary flow of a watercourse that doesn't cause major disruptions when its magnitude is moderate. It's described using three parameters: flow, the water level and flow velocity [2].

Classification of floods can be made as River flood, Costal flood, flash flood and urban flood. River flood occurs usually after a prolonged precipitation over large areas of the basin. This flood can last for even a week, affecting large area. The coastal floods are mainly due to Tropical cyclones, high tides and Tsunamis [2]. Preventive measures can be taken by increasing the embankment height, construction of cyclone structure as shelter for effected population [3], etc. Flash floods are a result of extreme local precipitation in limited area with high flow rate. This type of floods can occur anywhere but prove to be dangerous on steep slopes [4]. Urban floods results from extreme local rainfall combined with blocked drainage. This type of flooding depends on topographical conditions, soil conditions as well as existence of adequate and well maintained drainage facilities [5].

Globally floods are mainly caused due to long lasting rainfall (65%),brief torrential rain(15%), Tropical cyclone (10%) and monsoon rain(5%), dam break or release(1%), rain and snow melt(3%) and other(1%).[4]

Floods are the most expensive type of natural disaster regarding the material damages [6]. According to the international database on disasters EM-DAT, 2470 floods occurred internationally during the last ten years (1999 and 2009). 147,457 people lost their lives and the damage was estimated at 372.5 billion U.S Dollars [2]. In Australia, 377 million Australian dollars per year estimated over the period 1967-2005 [7]. In Morocco too and because of the water stress it's exposed to huge damages [8] [9].

During the last ten years, Morocco has experienced five severe flooding, which were 1068 deaths and affected more than 146,400 people. The damage was considerable and damaged several infrastructures [2].

That's why flood forecasting is very important in national economic construction and people's life, because flood causes huge losses every year. The flood forecasting becomes a major and practical significant research topic. In recent years, flood forecasting technology combines with computer technology, and the flood forecasting level has been greatly improved with computer database technology, artificial intelligence technology and related soft ware's development, particularly using Web technology. After 30 years development, flood

forecasting target goes through from initial simple automation of the on-line real-time, to the combination of graphics processing technology, the flood forecasting model for the mathematical precision processing technology, and to the computer consultation, human intervention Interactive forecasting system, the idea and design of the flood forecasting has great innovation [10].

The Early Warning Systems (EWS) enables to make decision oriented selections and to check weak areas within Logistics. Various types of flood warning systems are worldwide in operation and their complexity depends on the local flood history, available operators and financial constraints. The essential primary elements of all flood warning systems are weather forecast, meteorological and hydrological observations. The EWS can be categorized as Manual flood warning systems, Simple automated flood warning systems and sophisticated flood warning systems.

Sophisticated flood warning systems use real time meteorological and hydrological data in theoretical and empirical as well as hybrid hydraulic/hydrological models to predict the possibility of flooding. The European Flood Alert System (EFAS) [7] employs a similar method to the US IFLOWS that differs only by infrastructure and computer software. Catchment information such as soil moisture and substrate are also included in a sophisticated model such as Sacramento, MIKE [11].

So, it is very important to establish the expert knowledge base and flood forecasting. Considering the above facts, it is therefore necessary to develop an effective DSS to facilitate decision making for flood warning using all recent advances in DSS theory and computer science, and combining all necessary elements of a DSS into one system, and since Morocco has several watersheds that are similar in terms of morphology, topography and climate, so we must incorporate the reusability of basin's information to be used in other watersheds that are similar.

For the Real Time Flood Forecasting, it's well known that we need a decision support system for forecasting and for warning, however for a given region there may be an event of flooding occurs in the same way as a previous case, if the DSS forecast, it may slow the forecasting and warning process because it may fall in the case of redundant treatment and thus overloading the base station.

Therefore we must consider reducing the load on the base station before reaching the real time forecasting and warning stage by treating the flood's historical dataset in an offline mode.

In this paper, we propose a model which can compare between the historical data and the real-time data received from the sensors based on the rainfall, the runoff and the water level by using the Standard Deviation, the Variance and the GINI Index for the comparison in order to optimize our Decision Support System regarding the decision quality and response time.

II. INEQUALITY MEASURES

According to Lazarsfeld a problem occurs during matching multidimensional objects since we want to treat them as an all, thus it is necessary to combine the partial measures into a single global measure, which summarizes them.

A. Inequality measures

Inequality is a statement to compare the size, or the order of two objects.

- The notation $a = b$ means that a is **equal to** b.
- The notation $a \neq b$ means that a is **not equal to** b.

The inequality can be measured by the difference (a - b), the ratio a / b or a transformation of these. [12]

Desirable properties of an inequality measure: [12]

1. Non-negative
2. Zero if and only if observed distribution identical to the reference distribution.
3. All observations processed in the same manner.
4. Independent of the average value of the variable.
5. Observation aggregation showing the same degree of specificity should not change the value of the measure.
6. Principle of Pigou-Dalton transfer: an inequality measure should decrease if the distribution is modified in a way that the inequality clearly reduces.

In statistics the dispersion of the observed values is often interpreted as a reflection of the concentration or inequality of the measured property. [13]

In the following section, we present the dispersion measures in descriptive statistics

The **standard deviation (σ)** [13] (represented by the Greek letter sigma, σ) shows how much variation or dispersion from the average exists. A low standard deviation indicates that the data points tend to be very close to the mean (also called expected value), a high standard deviation indicates that the data points are spread out over a large range of values.

It is calculated by the following equation:

$$\sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (1)$$

The **Variance (V)** [13] measures how far a set of numbers is spread out. (A variance of zero indicates that all the values are identical.) Variance is always non-negative: A small variance indicates that the data points tend to be very close to the mean (expected value) and hence to each other, while a

high variance indicates that the data points are very spread out from the mean and from each other.

$$V = (\sigma)^2 \tag{2}$$

The **Variation Coefficient (C)** [13] also called **Relative Standard Deviation (RSD)** is a measure of relative dispersion.

The variation coefficient is defined as the ratio of the standard deviation to the mean:

$$C_v = \frac{\sigma}{\mu} \tag{3}$$

The variation coefficient is useful because the standard deviation of data must always be understood in the context of the mean of the data. In contrast, the actual value of the variation coefficient is independent of the unit in which the measurement has been taken, so it is a dimensionless number. For comparison between data sets with different units or widely different means, one should use the coefficient of variation instead of the standard deviation. Only the Variation Coefficient has the six desired properties already presented. [13]

However, inequality or concentration can be measured without referring to the mean. [14] proposed to compare each individual with each other. In the descriptive statistics field [14] introduced the concept of mean difference defined as:

$$g = \frac{1}{n(n-1)} \sum_{i,j}^{n} |x_i - x_j| \tag{4}$$

III. PROPOSED MODEL

This section will introduce the main idea of reducing the load on the base station before reaching the real time forecasting and warning stage by treating the flood's historical dataset in an offline mode.

As already said, if we want to compare all the data between them to have a decision without going through the DSS, then we must compare two multidimensional objects. So we propose to compare all factors data received from the sensors with the factors data noted in the flood historical statistics, furthermore instead of comparing each variance coefficient of the factor in the historical data with its corresponding data in real time, we conceived to propose an equation that includes all these factors into a single formula which give as a single value for each recordset as follows:

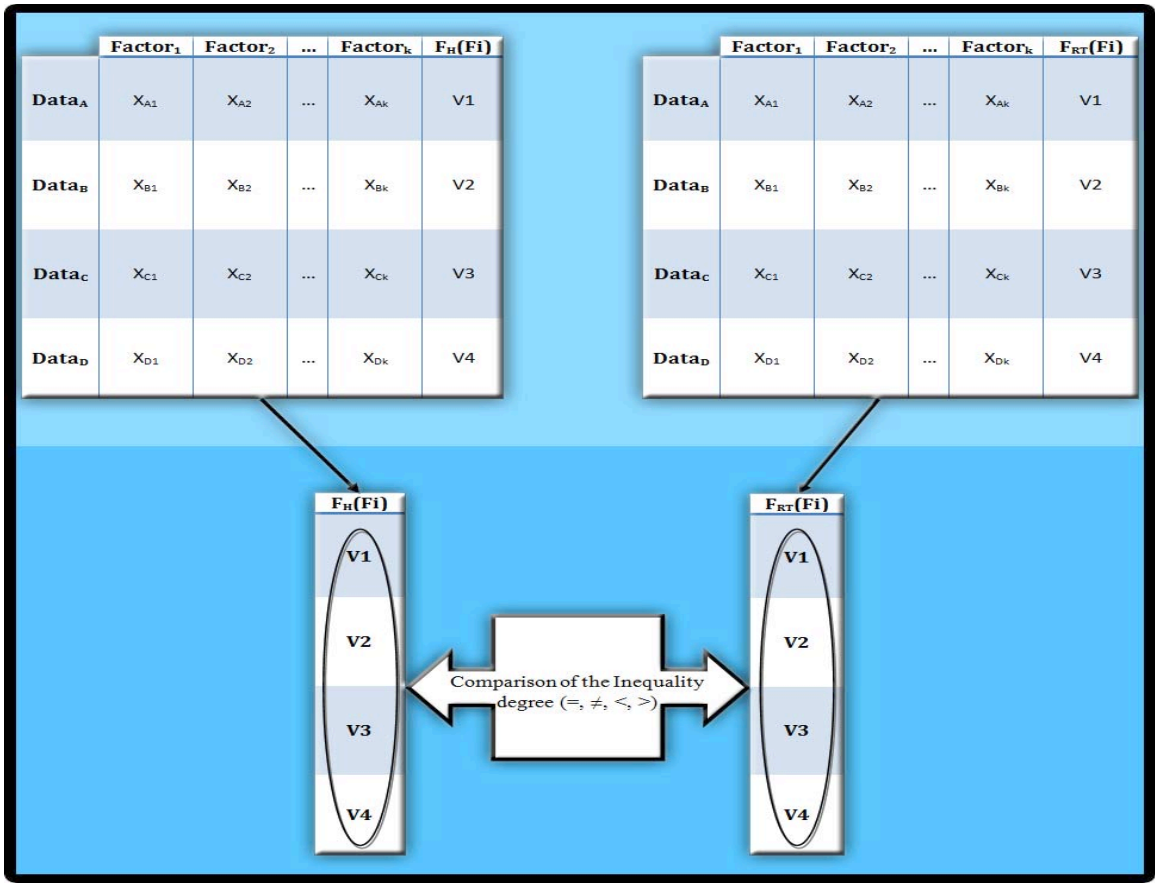

Fig. 1. Concept of the Proposed Inequality Comparison

Where

$$F(Z) = f(Fi) \quad (5)$$

With

- $i \in \mathbf{R}^*_+$ and F_i = { all factors that lead to flooding }

In our proposal, we work with the rainfall, runoff and the water level as study factors, so:

$$F(Z) = \frac{aF1 \times cF3}{bF2 \times TSFV} \quad (6)$$

With

- F1, F2, F3 and TSFV are respectively rainfall (mm), runoff (m/s), water level (m) and the time between the previous forecast value and the current forecast value in second (s). For the first record TSFV = 1, and FH, FRT are respectively Historical data and Real Time data functions for factors which give us values (V1, V2, V3, V4,…, Vn) to compare them in our Historical Processor Model which has 2 levels for comparison.
- (a, b, c) are coefficients and (a, b, c) $\in \mathbf{R}^*_+$. These coefficients are the percentages of influence for each factor leading to flooding on a flooded area.

A. *Architecture of the proposed model*

Our system of Real Time flood forecasting and warning consists of two access modes, online and offline, in our proposed model in this paper, we consecrated the offline mode that will support the clustering of historical data and the matching process for decisions based on historical data. First, we cluster the historical data, then we divide the Clustered Historical data into distributions according to this interval :] Ending of flood, Beginning of flood], we will explain this splitting process in the next section, after we store it in the database, next we load the distributions into the Historical Processor Model. Here is the diagram of the Offline Mode:

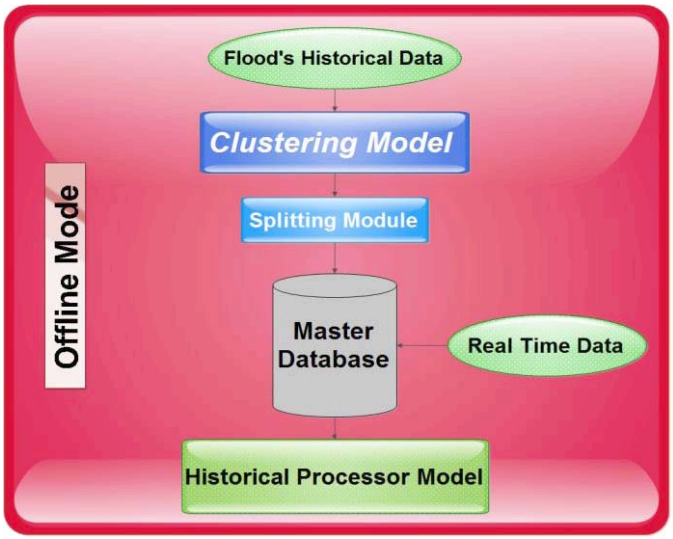

Fig. 2. Offline Mode Process

B. *Splitting Module*

Historical data of an area includes all the information regarding the impact of flooding. Each record includes the values of the factors at a given time, the date of data reception and of course the decision there was flood or not. However, in the inequality measures, we don't compare a record with another one horizontally, so it risk that the decision be unknown about flooding at a given moment, consequently we can never predict based on these historical data.

So to remedy this problem, we thought to divide the historical data into distributions according to the interval] Ending of flood, Beginning of flood]. Each distribution will have a case of flooding, so without checking the information concerning the decision, we will know very well that the region has experienced flood in every divided distribution. Before the Slitting Process, we calculate $F_H(F_i)$ for every single record for the Comparison Process according to the function (6). Here is a diagram representing the Splitting Process:

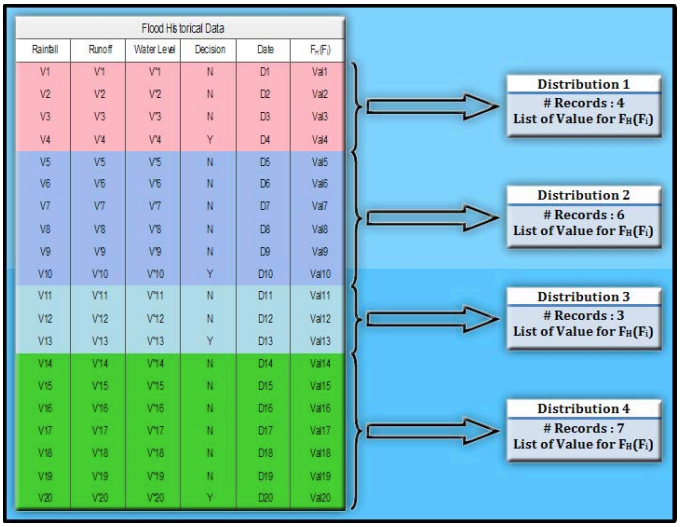

Fig. 3. Splitting process

C. *Historical processor Model*

The Historical Processor Model is for the Double Matching Validation, it's divided into two levels of validation. The first level is for the validation by the Variance Coefficient, the distributions are loaded from the Master Database, and stored in a local database dedicated for this Model. This level includes four modules, the Historical Data Module serves to arrange the Historical Distributions loaded from the Master database in a list to be sent to the local database, and the Real Time Data Module serves to calculate the Variance Coefficient for the Real Time Distribution. We have also other two modules, the Calculator Module and the Comparison Module which will be presented in the next section.

The second level is for the double validation by the GINI Coefficient, the selected distributions in the level 1 are loaded

from the local database by using the list sent from the level 1. This level has six modules, first we have the the Real Time Data Module, it serves to calculate the GINI Coefficient for the Real Time Distribution, second we have the Calculator module which has the same process as the Calculator Module in the level 1 with one difference, it calculate the GINI Coefficient, third we have the Comparison Module and we have the Loopback Module as well as Warning Message Module for the warning messages and finally the Final Warning Message Module for the final message. Here is the diagram of the Historical Processor Model:

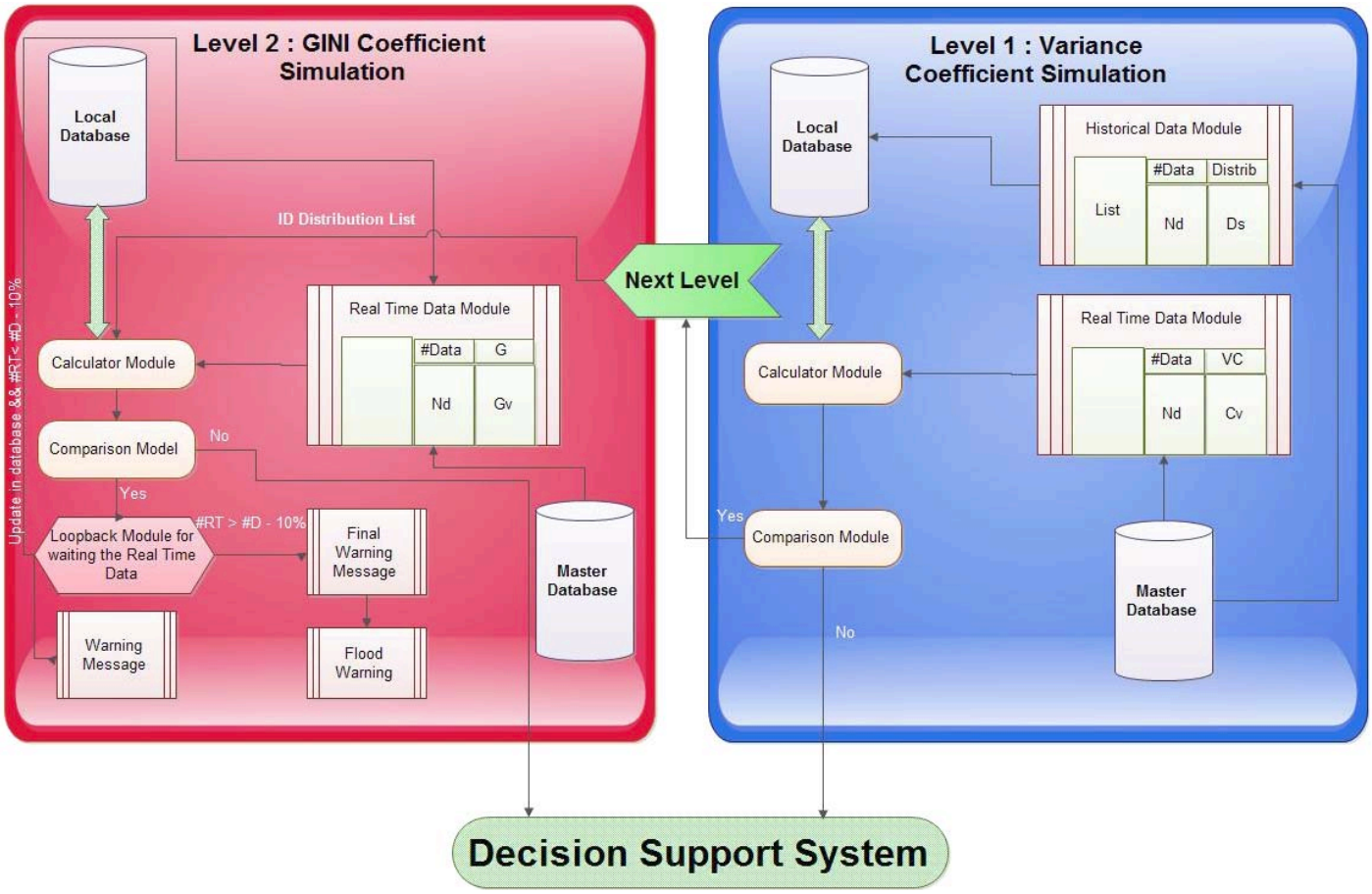

Fig. 4. Historical Processor Model

D. Double validation process

To activate the Historical Processor Model, it must be an update in the real time data table on the database. In this case the system sends the number of the real time distribution to the Calculator Module in the level 1. This module calculates the Variance Coefficient for each distribution taking just the record number which is equal to the number of record data in real time, in other word it searches for appropriate distributions for comparison according to Real Time record's number. For example, if there is a historical distribution which had 1000 records and real-time data distribution which has 400 records, so the module calculates the Variance Coefficient for the historical distribution just for the first 400 records. The module calculates the entire Variance Coefficient for all the distributions as already said, and then it transfers the Variance Coefficient list to the Comparison Model. This latter model serves to identify the distributions that have the most similar Variance Coefficient to decide about the impact of flooding ($C_H \sim C_{RT}$) if this is the case then we go to the next level to achieve the double validation and to be sure that the values are similar to an old event in the historical data and the flood will be occurs else there is no similar case, so it must go through the decision support system for forecasting and warning. This module sends a list that contains the ID distributions which are similar to the Real Time distribution proved in the Comparison Model for not to repeat the same work that was done by the calculator module in the Level 1. The second level is for the validation by the GINI Coefficient. The Calculator Module in Level 2 receives the ID distribution's list to load the distributions from the local database to calculate the GINI Coefficient for the Comparison Module. This module calculates the GINI Coefficient for the distribution list received from the level 1 taking just the record number which is equal to the number of record data in real time, then the module sends the GINI Coefficient of the distributions to the Comparison

Module to decide about the impact of flooding ($G_H \sim G_{RT}$), if this is the case then the system reaches the Loopback Module. In this module, the system waits for an update in the Real Time Table on the database and the Real Time Distribution still lower than (90% of the Historical Distribution) to repeat the calculation in the Calculator Module while displaying the warning messages. If the Real Time Distribution is higher than (90% of the Historical Distribution), so the alarm and warning messages are triggered. In this next section we present the flow chart of this process.

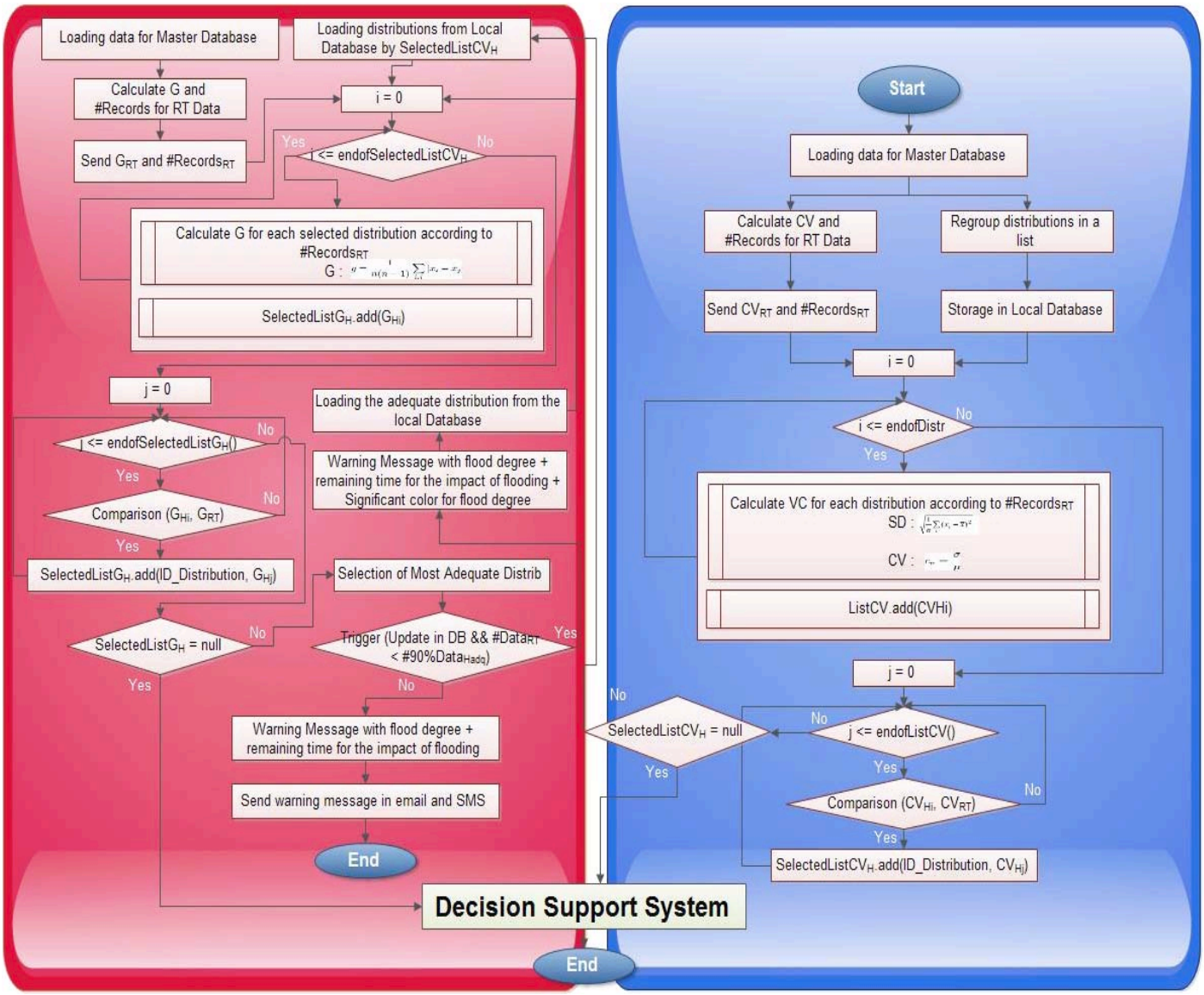

Fig. 5. Flow Chart of the Historical Processor Model

IV. CONCLUSION

In this paper, we have presented a new optimal model based on Inequality Measures for treatment of Historical and Real Time Dataset in order to reduce the overload in the base station, also to reduce the response time of warning in order to overcome the flood disaster. To make it, we consecrated the offline mode that will support the clustering of historical data and the matching process for decisions based on historical data by using a proposed function F(Z) to match. Before matching these data, we proposed a Splitting Module which divide the historical data into distributions according to the interval] Ending of flood, Beginning of flood], so each distribution will have a case of flooding. We presented also The Historical Processor Model which is for the Double Matching Validation to decide about the impact of flooding since this disaster is the most expensive type of natural disaster regarding the damages.

The further work will focus on the proposal of a model for classification of historical data and real-time data received from wireless sensors by using Clustering approach, it will also focus on the implementation of a simulator to simulate the results and to do necessary comparative studies.

REFERENCES

[1] J.K.Roy, D. Gupta, and S. Goswami, "An improved flood warning system using, wsn and artificial neural network*," in Proceeding of the India Conference (INDICON), 2012 Annual IEEE*, India, 2012, pp. 770–774.

[2] D. de la surveillance et de la prévention des risques, *Etude pour la réalisation dune cartographie et d'un système d'information géographique sur les risques majeurs au Maroc, Mission 1 Identifications des risques, Le risque dinondation*. Maroc: Ministère de l'Energie, des Mines, de l'Eau et de l'Environnement, 2008.

[3] Manavalan, S. Chattopadhyay, Mangala, and Y. S.Rao, "Emerging trends of computational grid based near real time/real time flood assessment and forecasting models," *in Proceeding of the Third International Conference on Emerging Trends in Engineering and Technology*, 2010, pp. 471–475.

[4] K.Grust, *Usefulness of early flood warning systems*. European flood Alert System (EFAS), 2006.

[5] S. A. Khan, "Rainfall-runoff modelling using data driven and statistical methods," *in Proceeding of the International Conference of the Advances in Space Technologies*, September 2006, pp. 16–20.

[6] J. L. Jin, X. L. Zhang, and J. Ding, "Projection pursuit model for evaluation grade of flood disaster loss," *Systems Engineering Theory & Practice*, vol. 22, no. 2, pp. 140–144, 2002.

[7] J. Wang, "Development of a decision support system for flood forecasting and warning a case studyon the maribyrnong river," Ph.D. dissertation, Victoria University, Melbourne, Australia, January 2007.

[8] M. ELMabrouk, L. Cherrat, M. Ezziyyani, and M. Essaaidi, "Conception of real-time flood forecasting system approach based on anytime techniques," *in Proceeding Engineering & Technology (PET) of the International Conference on Control, Engineering & Information Technology (CEIT13)*, June 2013, pp. 207–211.

[9] M. ELMabrouk, O. Kassara, S. ELMamoune, M. Ezziyyani, M. Essaaidi, S. Gaou, and A. Abajja, "An adaptive intelligent decision support system for real-time flood forecasting," *International Journal of Research in Computer Applications and Robotics*, vol. 1, no. 7, pp. 14–23, October 2013.

[10] Y. Yan and Y. Cao, "A mode of storm flood forecasting dss establish ion," CCIS, vol. 237, pp. 261–266, 2011.

[11] M. D.F.Lekkas, C.Onof and E.A.Ballas, "Application of artificial neural networks for flood forecasting," *Global Nest: The International Journal*, vol. 6, no. 3, pp. 205–211, 2004.

[12] A. Valeyre, "Formes et propriétés des indices d'inégalité entre proportions," *Mathématiques et sciences humaines*, vol. 132, pp. 13–37, 1995.

[13] G. Saporta, *Probabilités, analyse des données et statistique*. Editions Technip, 2006.

[14] C. Gini, "Measurement of inequality of income," *Economic Journal*, vol. 31, pp. 22–43, 1921.