# A Distributed Approach to Flood Prediction Using a WSN and ML: A Comparative Study of ML Techniques in a WSN Deployed in Brazil

**Conference Paper** *in* Lecture Notes in Computer Science · October 2015

DOI: 10.1007/978-3-319-24834-9_56

Gustavo Furquim
University of São Paulo
**6** PUBLICATIONS **155** CITATIONS

Gustavo Pessin
Instituto Tecnológico Vale
**199** PUBLICATIONS **3,588** CITATIONS

Pedro Henrique Gomes
Ericsson
**48** PUBLICATIONS **1,654** CITATIONS

Eduardo Mario Mendiondo
University of São Paulo
**230** PUBLICATIONS **4,593** CITATIONS

# A Distributed Approach to Flood Prediction Using a WSN and ML: A Comparative Study of ML Techniques in a WSN Deployed in Brazil

Gustavo Furquim[1,5], Gustavo Pessin[2,5](✉), Pedro H. Gomes[3,5],
Eduardo M. Mendiondo[4,5], and Jó Ueyama[1,5]

[1] Institute of Mathematics and Computer Science (ICMC),
University of São Paulo (USP), São Carlos, São Paulo, Brazil
gafurquim@usp.br, joueyama@icmc.usp.br

[2] Vale Institute of Technology, Belém, Pará, Brazil
gustavo.pessin@itv.org

[3] Autonomous Network Research Group (ANRG),
University of Southern California, Los Angeles, CA, USA
pdasilva@usc.edu

[4] São Carlos School of Engineering (EESC), University of São Paulo (USP),
São Carlos, São Paulo, Brazil

[5] Center of Monitoring and Early Warning of Disasters, Ministry of Science,
Technology & Innovation, São Carlos, Brazil
emm@sc.usp.br, emm@cemaden.gov.br

**Abstract.** Natural disasters (e.g. floods, landslides and tsunamis) are phenomena that occur in several countries and cause a great deal of damage, as well as a serious loss of life and materials. Although very often these events cannot be avoided, their environments can be monitored and thus predictions can be made about their likely occurrence so that their effects can be mitigated. One feasible way of carrying out this monitoring is through the use of wireless sensor networks (WSNs) since these disasters usually occur in hostile environments where there is a lack of adequate infrastructure. This article examines the most recent results obtained from the use of machine learning techniques (ML) and adopts a distributed approach to predict floods using a WSN deployed in Brazil to monitor urban rivers. It also conducts a comparative analysis of ML techniques (e.g. Artificial Neural Networks and Support Vector Machines) for the task of flood prediction and discusses the results obtained from each type of technique explored so far. Finally, in the discussion of the results, a suggestion is made about how to improve accuracy in forecasting floods by adopting a distributed approach, which is based on allying computing intelligence with WSNs.

**Keywords:** Wireless sensor networks · Machine learning · Distributed systems · Flash flood forecast

## 1 Introduction

In recent years, there has been an increase in the frequency of natural disasters throughout the world. This rise is mainly caused by the increase in the number

K. Jackowski et al. (Eds.): IDEAL 2015, LNCS 9375, pp. 485–492, 2015.
DOI: 10.1007/978-3-319-24834-9_56

of occurrences of climate calamities, such as severe storms and floods. These disasters lead to a large number of victims, as well as incurring financial losses that, directly or indirectly, affect the lives of millions of people [3]. During 2013, for example, 330 disasters were recorded that were triggered by natural events throughout the world. These disasters affected 96.5 million people and caused the deaths of 21.610 victims, together with damage estimated at US$ 118.6 billion [9]. This situation has been aggravated by changes in the climatic conditions of the planet and is largely found in urban districts such as those that prevail in the State of São Paulo, Brazil. In these regions the destruction of the ecosystem caused by pollution and a lack of planning is more acute and affects the environment by altering the local climate. In this context, WSNs are an attractive way of carrying out the monitoring of urban rivers and other natural environments because of the following reasons: low costs particularly with regard to infrastructure; accessibility to inhospitable environments and the fact that they offer the prospect of using high-precision sensors; and their adaptability to environmental changes [8]. However, the collected data must be studied and their behavior understood before suitable flood predictive models can be created with a high degree of accuracy. ML techniques are very useful for this since they can produce predictive models without the need for a precise knowledge of the hydrological processes involved or the physical variables of the hydrographic basin of the region [6]. In São Carlos, Brazil, a WSN has been installed which is called REDE (*REde de sensores sem fio para Detectar Enchentes* – WSN for Flood Detection), developed by the Institute of Mathematics and Computer Sciences (ICMC) - University of São Paulo (USP) [8], which seeks to undertake the monitoring of urban rivers. One of the key features of this article is to explore this area by using and analysing data collected by the REDE system. A comparative study will also be carried out between the ML techniques and a distributive prediction which makes it possible to forecast the occurrence of floods and create an alert system for taking preventive measures.

The initial analysis of the data was conducted by employing the concepts of Time Series or more precisely, the Immersion Theorem of Takens [7]. This theorem creates a time series with overlapping data in a simplified and multidimensional representation and following this procedure, employs modelling techniques that can allow observations to be related over a period of time and for predictions to be made with a greater degree of accuracy. Furthermore, three ML techniques were employed for this modelling: Support Vector Machine (SVM), the Gaussian Process (GP) and Multilayer Perceptron (MLP). In this way, this article examines these ML techniques when used for flood prediction and show the comparative results of this study. The local forecasts carried out by models present in some of the sensor nodes, can help in the forecast carried out by the other nodes. In this way, an analysis was conducted that sought to determine when the predictions carried out by some nodes, could be used to increase the accuracy of the prediction of the other nodes, leading to a distributed approach for flood prediction. The adoption of this distributed predictive control approach, together with the comparative analysis of the ML techniques, made it possible to increase the accuracy of the predictions as well as to partition the processing needed to implement the various models that had been created. The rest of this article is divided in the following manner: Sect. 2 outlines the studies related to flood prediction. Section 3 discusses the methods employed and describes the following: (a) how the data collected by the WSN was handled, (b) the Immersion Theorem of Takens and (c) the ML techniques that were employed and compared in the studies. The article concludes with an analysis of the obtained results (Sect. 4); the final considerations and suggestions for further studies are made in Sect. 5.

## 2 Related Works

Hossain et al. [5] make use of a system based on adaptive neuro-fuzzy inference (ANFIS) to make predictions of floods, as well as the likely persistence of this flooding in the River Meghna, Bangladesh, by using data collected from a WSN. Although good results were attained in this article, only a system based on a WSN was used and no wireless sensor network was in fact installed in the region. In addition, there is also a description in this article of the features that the sensors used in the experiment, which leads to their application being restricted.

Still on the question of flood prediction, Damle and Yalcin [2] describe an approach based on a combination of chaos theory and data mining that is adopted to obtain features of the time series containing the daily output of the Mississipi River in the US. This methodology follows a case flow which is initiated by using the Takens Theorem to reconstruct the time series. The reading of the data was carried out through a seasonal collection and not by means of WSN. Despite achieving good results, the method only predicts the use of data from a point on the river and thus ends up by not analyzing the existing relationship between the different collection points of the whole river. Elizabeth et al. [1] set out a system for prediction and flood alerts by using statistical models and a WSN made up of 4 types of different nodes.The predictive system was tested with historical data collected over a period of 7 years from the Blue River, in Oklahoma, although the article does not state how the collection was carried out. In addition, a reduced version of the proposed WSN, which has few nodes and a predictive model for centralized processing, was installed in the River Charles, in Massachussets, and River Agúan, located in the north of Honduras. The simplified installation of the WSN served to test the communications system and data collection.

## 3 Methods of Flood Forecasting

Takens [7] observed that a given time series $x_0, x+1, ..., x_{n-1}$, can be reconstructed in multidimensional space $x_n(m, \tau) = (x_n, x_{n+\tau}, ..., x_{n+(m-1)\tau})$, called time-delay coordinate space, where $m$ represents the embedded dimension and $\tau$ represents the dimension of separation. In other words, the time series is reconstructed in vectors with $m$ values, in which each value corresponds to an observation spaced with an equal time delay to $\tau$. These vectors represent the interdependence relationship that exists between the observations and simplifies the study of the behaviour of the time series. After this reconstruction, ML can be employed to produce prediction models of the time series with a greater degree of accuracy. It is difficult to determine the values of $m$ and $\tau$ with precision in some kinds of time series, in particular those extracted from real-world environments and subject to noise. However, they can be estimated by means of methods like Auto-Mutual Information (for $\tau$) and False Nearest Neighbors (for $m$).

Currently, the WSN used for this study (the REDE system) consists of 7 sensors; six of them are level sensors which are used to measure the water pressure and change the height in centimeters and 1 sensor measures the pluviosity and shows the volume of rainfall in the region (Fig. 1).

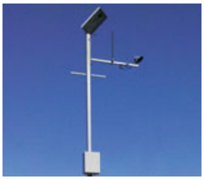


**Fig. 1.** A REDE WSN node deployed in São Carlos - SP, Brazil.

The data considered in this article were obtained by three sensors from the REDE system (called here $s1$, $s2$ and $s3$) during the whole month of May 2014. The readings of the water level were carried out at 5 min intervals and the sensors were arranged in the following way: Sensors $s1$ and $s2$ are in different rivers which converge to form the water flow that passes by Sensor $s3$. If the proposed approach is adopted, the first stage is to measure the overlapping of the time series formed by the readings of the river levels given by the three sensors. Auto-Mutual Information techniques were employed for this, and these were designed to estimate the separation dimension and the False Nearest Neighbors, to estimate the embedded dimension. In this way, the separation and embedded dimensions that were used to estimate the overlapping of the time series that was read are as follows: $\tau = 1$ and $m = 2$ for sensors s1, s2 and s3. For sensor s2 we also evaluate $\tau = 5$ with $m = 2$. After carrying out the time series with overlapping data, three ML techniques were employed and evaluated before being implemented to produce prediction models, namely: SVM (with epsilon-SVR and nu-SVR kernels), GPs and the MLP. These techniques were employed and compared to produce local forecast models which only use the data from the sensors themselves. This analysis made it possible to determine the best prediction model and the overlapping parameters for each sensor, as well as to allow comparisons to be made with the distributed approach described in the following paragraph. If the current time is considered to be $t$, the predictions were carried out for the $t + 1$ instant, when $\tau = 1$ (5 min in the future), or $t + 5$ (25 min in the future), when $\tau = 5$. In the second stage of the experiments, a distributed approach for predictions was employed, which explores the readings and the processing capacity of multiple sensors (instead of only one). In this way, the processing capacity of the WSN nodes can be explored in a more satisfactory way as well as giving the sensors a greater degree of freedom for the task of forecasting floods, particularly in situations where communication with the central database has been lost. In this approach, the $s1$ an $s2$ carry out a part of the processing and send the results to sensor $s3$. This processing involves making forecasts of the river level for instant $t + 1$ (5 min in the future), with regard to its respective locations and only using the readings themselves. After being sent to sensor $s3$, the local predictions $(t + 1)$ of sensors $s1$ and $s2$ were combined with the local readings of sensor $s3$, where the rest of the processing was executed. In this study, the ML techniques were applied again to create forecast models for sensor $s3$, this time by making use of the combined data and predictions of the 3 sensors. The distributed approach to flood prediction set out here, also seeks to analyze whether the predictions made by the groups of sensors ($s1$ and $s2$, in this case), can improve the more long-term forecasting carried out by the other sensors ($s3$, in this case). As well as creating forecast models for the $t + 1$ instant (5 min in the future), the same combination of data and data techniques was used for this in sensor $s3$, to create predictive models for instant $t + 2$ (10 min in the future). It should be noted that this article only makes use of the data collected by the real-world WSN to simulate and analyze the approach put forward, since the predictive models and distributed approach are not implemented in the nodes of the REDE system. Nonetheless, the nodes of the REDE system have the capacity to execute the forecast models and carry out the necessary communication. In addition, the results obtained from these experiments (which are shown in the next section), have encouraged us to make a real-world implementation of the proposed approach.

## 4 Results and Discussion

After the extent of the overlap of the time series was read by sensors $s1$, $s2$ and $s3$, the SVM implementations (epsilon-SVR and nu-SVR kernels), GP and MLP were used to create flood prediction models. For overlapping values equal to $\tau = 1$ and $m = 3$, two readings of the river level were used as input. The output is the forecast of a point in the future $(m = 3)$, where all these values (from the input to the prediction) are consecutive readings above the river levels $(\tau = 1)$. The Waikato Environment for Knowledge Analysis (WEKA) [4] was employed to model and evaluate both techniques. The results are obtained by employing a 10-fold cross validation procedure.

As can be seen in Table 1, the best results for all the sensors were obtained with values of $\tau = 1$ and $m = 2$. Thus in the experiments that follow, where more data for distributed prediction sensors were used, these parameters ($\tau = 1$ and $m = 2$) were kept to carry out the overlapping of the time series of sensors $s1$ and $s2$ and the technique was used to ensure a greater degree of accuracy for both sensors (in the case of the GP). In the distributed predictive approach, the best forecasting models for instant $(t + 1)$ were executed in sensors $s1$ and $s2$ that consider the current instant as being $t$. The forecasts made by these sensors

**Table 1.** Performance of the ML techniques that consider the estimated values for overlapping data.

| Sensor | $\tau$ | $m$ | SVM (epsilon-SVR) | | SVM (nu-SVR) | | GP | | MLP | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE |
| $s1$ | 1 | 2 | 0.5456 | 0.2132 | 0.5439 | 0.1956 | **0.6069** | **0.1804** | 0.5935 | 0.1836 |
| $s2$ | 1 | 2 | 0.3484 | 1.4142 | 0.3481 | 1.4257 | **0.6051** | **1.0934** | 0.5679 | 1.1547 |
| | 5 | 2 | 0.2829 | 1.4699 | 0.2825 | 1.4807 | 0.3583 | 1.3815 | 0.2948 | 1.4563 |
| $s3$ | 1 | 2 | 0.6059 | 1.7499 | 0.6040 | 1.7486 | 0.8731 | 0.9796 | **0.8794** | **0.9334** |

**Table 2.** Results of the ML techniques for the distributed prediction in $t+1$.

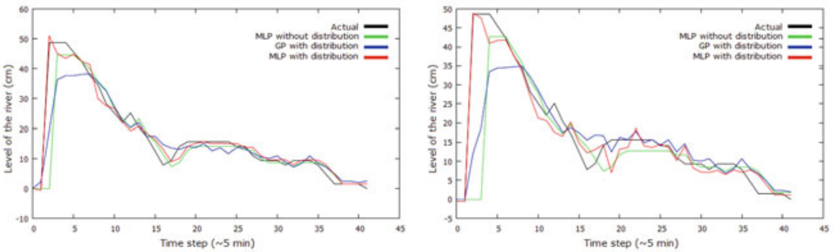
| | SVM (epsilon-SVR) | | SVM (nu-SVR) | | GP | | MLP | |
|---|---|---|---|---|---|---|---|---|
| | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE |
| $s3$ | 0.6059 | 1.7499 | 0.6040 | 1.7486 | 0.8731 | 0.9796 | 0.8794 | 0.9334 |
| Distributed | 0.5864 | 1.7979 | 0.5844 | 1.7986 | **0.8858** | 0.9534 | 0.8805 | **0.9270** |

were then sent to sensor $s3$, where new forecast models were created for instants $t+1$ and $t+2$ using the data itself and the forecasts made by sensors $s1$ and $s2$. Table 2 shows the performance of the ML techniques to create new forecast models of a distributed approach for instant $t+1$. The parameters of the ML techniques were left in the Weka standard deviation, since the MLP used the learning rate $= 0.3$ and a single layer with 6 hidden neurons. The first line of the table shows the results obtained without the use of distribution for purposes of comparison and the better performance is highlighted in bold.

As can be seen in Table 2, the performance indices diverge with regard to what is the best model created by the ML techniques. This is because the GP achieves a better performance when account is taken of $R^2$ and MLP has a better performance when the RMSE is considered. Another point that is worth highlighting is that both of the forecast models for $t+1$, produced by MLP or the GP, improved their performance when a distributed approach was adopted. However, both the SVM kernels had a worse performance with the proposed approach. Table 3 makes use of the same distributed approach by using the forecasts for instant $t+1$ carried out by sensors $s1$ and $s2$. However, in these experiments, the forecast models created by the ML techniques for sensor $s3$, make forecasts for instant $t+2$. In this way, the forecasts were made 10 min in

**Table 3.** Results of the ML techniques of distributed prediction in $t+2$.

| | SVM (epsilon-SVR) | | SVM (nu-SVR) | | GP | | MLP | |
|---|---|---|---|---|---|---|---|---|
| | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE |
| $s3$ | 0.5487 | 1.8463 | 0.5482 | 1.8453 | 0.7661 | 1.3167 | **0.7784** | **1.2613** |
| Distributed | 0.5872 | 1.7943 | 0.5856 | 1.7937 | 0.7961 | 1.2302 | **0.8065** | **1.1787** |



(a)

(b)

**Fig. 2.** Graphic comparison between the real-world readings and the forecasts made by the (a) $t + 1$ and (b) $t + 2$ instants.

advance, which allows warnings to be given with more time for decision-making and thus go further in saving lives and reducing financial losses. The parameters of the ML techniques were kept the same as in the previous experiment. For comparative purposes, the first line of the Table 3 shows the results obtained without the use of the distributed approach and the better performance was highlighted in bold. In a different way from the previous experiment (and as can be seen in Table 3), all the models formed by the ML techniques achieved a better performance when a distributed approach was adopted to make forecasts for instant $t + 2$. Since the best results were obtained from the MLP, this time it improved its performance in both the performance indexes followed by the GP. The SVM, which had both the kernels, managed to improve its performance by adopting a distributive approach despite creating models that had a performance that was inferior to that of the other techniques. The performance of the forecast models created by the ML techniques can be illustrate graphically as shown in Fig. 2(a) (for the predictions in $t + 1$) and Fig. 2(b) (for the predictions in $t + 2$). As can be seen, in both cases there is a discrepancy between the values estimated by the MLP and the real-world readings of the river level when the distributed approach is not employed. This discrepancy is corrected when the distributed approach is adopted and as well as this, the peak values are represented in a better way. However, even when the distributed approach is employed, the GP ends up by underestimating the peak values, and shows a better performance when the readings are more stable. This analysis is of great value because in examining flood prediction, the peak values are very important since they are at the points where the flooding occurs.

## 5 Conclusion and Future Works

A distributed approach has been set out and examined to increase the accuracy of flood prediction in urban rivers by using data collected by means of WSNs and ML. In addition, three ML techniques were analyzed and compared (SVM with two different kernels, GP and MLP). These techniques were employed and evaluated both to create local forecast models and to create models for a distributive approach. This way, the combination of these techniques with the distributed approach for flood prediction is the main contribution of this paper. This study has made it possible to assess the performance of each ML technique, when applied to flood prediction to select the best combination of methods for the forecasting carried out. In our distributed approach for flood prediction, it is also possible to share the processing required to execute the various forecast models selected among the WSN nodes. As a result, the proposed distributive approach is able to achieve a greater degree of accuracy using the processing power that already exists in the WSN, as well as being able to better represent the behavior of the river at times of great importance when peaks occur on the water level. It thus offers a promising means of tackling the problem of flood prediction. In future work we intend to analyze in greater depth the relationship between the readings and predictions made by the sensors at different points of the hydrographic basin. Our aim is also to embed the distributed approach proposal in the nodes of the REDE system.

## References

1. Basha, E.A., Ravela, S., Rus, D.: Model-based monitoring for early warning flood detection. In: Proceedings of 6th ACM SenSys (2008)
2. Damle, C., Yalcin, A.: Flood prediction using time series data mining. J. Hydrol. **333**(2–4), 305–316 (2007)
3. Guha-Sapir, D., Hoyois, P., Below, R.: Annual disaster statistical review 2013: the numbers and trends. Technical report, Centre for Research on the Epidemiology of Disasters (2014)
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
5. Hossain, M., Turna, T., Soheli, S., Kaiser, M.: Neuro-fuzzy(nf)-based adaptive flood warning system for bangladesh. In: ICIEV International Conference (2014)
6. Kar, A., Winn, L., Lohani, A., Goel, N.: Soft computing-based workable flood forecasting model for ayeyarwady river basin of myanmar. J. Hydrol. Eng. **17**(7), 807–822 (2012)
7. Takens, F.: Detecting strange attractors in turbulence. Dyn. Syst. Turbul. **898**, 366–381 (1981)
8. Ueyama, J., Hughes, D., Man, K.L., Guan, S., Matthys, N., Horre, W., Michiels, S., Huygens, C., Joosen, W.: Applying a multi-paradigm approach to implementing wireless sensor network based river monitoring. In: 2010 CDEE International Symposium (2010)
9. Wallemacq, P., Herden, C., House, R.: The human cost of natural disasters 2015: a global perspective. Technical report, Centre for Research on the Epidemiology of Disasters (2015)