USE OF MULTIVARIATE MACHINE LEARNING ANALYSIS TECHNIQUES FOR FLOOD RISK PREVENTION

D. Vito [1],

[1] Politecmico di Milano, Department of Electronic Information and Bioengineering, Milan – domenico.vito@polimi.it

Commission VI, WG VI/4

KEY WORDS: flood, risk, data, Machine Learning, Prediction

ABSTRACT:

Natural disasters such as flood are regarded to be caused by extreme weather conditions as well as changes in global and regional climate

The prediction of flood incoming is a key factor to ensure civil protection in case of emergency and to provide effective early warning system. The risk of flood is affected by several factors such as land use, meteorological events, hydrology and the topology of the land.

Predict such a risk implies the use of data coming from different sources such satellite images, water basin levels, meteorological and GIS data, that nowadays are easily produced by the availability new satellite portals as SENTINEL and distributed sensor networks on the field.

In order to have a comprehensive and accurate prediction of flood risk is essential to perform a selective and multivariate analyses among the different types of inputs.

Multivariate Analysis refers to all statistical techniques that simultaneously analyse multiple variables.

Among multivariate analyses, Machine learning to provide increasing levels of accuracy precision and efficiency by discovering patterns in large and heterogeneous input datasets.

Basically, machine learning algorithms automatically acquire experience information from data.

This is done by the process of learning, by which the algorithm can generalize beyond the examples given by training data in input. Machine learning is interesting for predictions because it adapts the resolution strategies to the features of the data. This peculiarity can be used to predict extreme from high variable data, as in the case of floods.

This work propose strategies and case studies on the application on machine learning algorithms on floods events prediction.

Particullarly the study will focus on the application of Support Vector Machines and Artificial Neural Networks on a multivariate set of data related to river Seveso, in order to propose a more general framework from the case study.

1. INTRODUCTION

1.1 Big Data Analysis

Climate change is a great determinant in the modification of the dinamic of water surface bodies.

The effect of the increase of temperature risings, and extreme metereological events reflect in the increase in the number and intensity of flood events.

A great effort of scientist and experts that deals on climate change and related effects is to predict the upcoming of this phenomena in order to furnish adeguate responses for the civil protection and mitigation of the effects.

Prediction and protection in function to mitigate the effects of extreme events relaces the concepts and ideas of “Early Warning Systems” (EWS).

A classical definition of Early Warning System is the technology and associated policies and procedures designed to predict and mitigate the harm of natural and human-initiated disasters and other undesirable events.

Early Warning Systems (EWS) are well recognized as a critical life-saving tool for floods, droughts, storms, bushfires, and other hazards.

The current classical concept of early warning systems underlines the presence of physical models of the extreme event that are keen to predict the probability of the extreme events by reproducing and calculating a set of mathematical laws related to the phenomena.

This models tent to be complex as the number of variables increses and as the number of components related to the phenomena rise up.

The WMO defined Multi-Hazard Early Warning Systems (MHWEWS)[1] the EWS that are capable to detect and predict the occurrence of different types of extreme events.

The adoption of the of the MHEWS is also a key component and counciled in the adoption of the adoption of the Hyogo Framework for Action (HFA) 2005–2015 by 168 countries.

Flood prediction is highly prone to the use of MHWEWS as the risk of flood is affected by factors such as land use, meteorological events, hydrology and the topology of the land.

A strong problem on this kind of systems is to manage the high number and the different nature of the data used for the prediction in order to have a valuable integration and an effective and reliable prediction.

Even if the use of physical and statistical models is typical the implementation of MHEWS, the current advances in data science and modelling techniques has provided the possibility to explore different approaches for the EWS (Lyong 2002,, Bell 2012), as the case of Machine learning.

This work wants to propose the application of on machine learning algorithms on floods events predicion, using in

This contribution has been peer-reviewed.

particular two strategies, the Support Vector Machines and the Artificial Neural Network.

A case study on the river Seveso in Lombardy will be also presented.

2. COMPLEX DATASET

2.1 Definition of Complex Datasets

As exposed in the introduction the prediction of the flood incomes by EWS could imply the use of data coming from different sources such satellite images, water basin levels, meteorological and GIS data, that nowadays can easily produced by the availability of new satellite portals as SENTINEL and distributed sensor networks on the field.

On data analysis point of view this means that EWS usually deal with complex and heterogeneus dataset.

Complex datasets are characterized to be indeed multivariate and to collect an high amount of information.

Multivariate data are sets of informations that deals with variable of different kind and nature to describe a certaint phenomenon.

In order to extract useful information that is feasible for the prediction of extreme events from complex dataset, is usually necessary to perform multivariate analysis (Anderson, 1958).

Multivariate Analysis refers to all statistical techniques that simultaneously analysis of multiple variables that could describe the phenomen.

2.2 Machine Learning for multivariate data

Machine learning, is a branch of artificial intelligence (AI), focuses on finding algorithms capable of learning and/or adapting their structure based on a set of observed data, with adaptation done by optimizing over an objective or cost function (Jin yu, 1998). In the past couple of decades it has become a common tool in almost any task that requires information extraction from large datasets (To, 2017).

Machine learning (ML) has having growing application as methodology and approach to analyse multivariate data-sets

The numbere of application of the use of machine learning in flooding prediction are going to rise in the fields of early warning systems (Jangyodsuk,Clark 2016) as they provides increasing levels of accuracy precision and efficiency by discovering patterns in large and heterogeneous input datasets.

Machine learning is usually applied to observational data, where the predictive variables are not under the control of the learner, as opposed to experimental data, where they are (Cohen, 1995).

In several sense, for the prediction of extreme events ML are going to substitute the "mechanicistic" vision related to the use of models, with a more pronounced "black-box" approach, by which the importance stands in the capability of the prediction to follow the data variabily, rather then the physical understanding of the phenomena.

In this sense ML adapts the resolution strategies to the features of the data. This peculiarity can be used to predict extreme from high variable data, as in the case of floods.

3. OVERVIEW ON MACHINE LEARNING

Machine learning has having growing application as methodology and approach to analyse multivariate data-sets as it provides increasing levels of accuracy precision and efficiency by discovering patterns in large and heterogeneous input datasets.

The range of different problems that can be faced through machine learning is clearly large, and grows as a growing number of templates are discovered to address a large set of situations (To, 2017)..

In particular Machine learning processes could be implied in three important operations:

- Classification: in classification, you will need to categorize data into predefined classes)
- Regression:regression techniques are used when the output is real-valued based on continuous variables.
- Clustering: clustering is to determine the intrinsic grouping in a set of unlabeled data.

These mechanism can be highly useful in the prediction of events.

3.1 The learning problem

Standing to the definition of Mitchell (1997) a machine learning algorithm: "A computer program is said to learn from experience E with respect to some class of tasks Tand performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." .

In general a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks.

This is the basic assumption of a machine learning problem, and the fundamental goal of machine learning is to generalized beyond the examples in the training set (Figure 1).

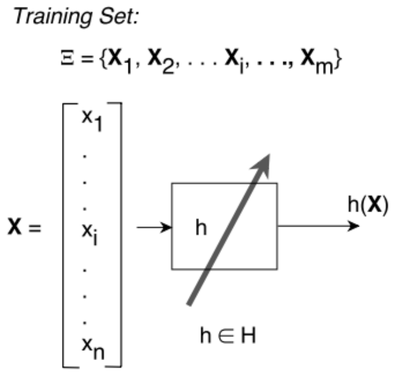

Training Set:
$\Xi = \{\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_i, \ldots, \mathbf{X}_m\}$
$\mathbf{X} =$
x_1
x_i
x_n
h
$h(\mathbf{X})$
$h \in H$

Figure 1. Machine Learning problem(To, 2017)

As machine learning can face several problems, the task could be various, but in general it can be resumed in guessing a function *f* dependent by the input.

The hypothesis about the function to be learned is denoted by *h*.

The input vector is usually called as input vector, pattern vector, feature vector, sample, example, and instance.

The function his though to be implemented by a device that has *X* as input and *h(X)* as output.

The output may be a real number, in which case the process embodying the function, h, is called a function estimator, and the output is called an output valueor estimate or target.

Both fand hare functions of a vector-valued input *X = (x1,x2, . . . ,xi, . . . ,xn)* which has n components.

Sometimes we know that f also belongs to this class or to a subset of this class. The function h is based on a training set, Ξ, of m input vector examples.

3.2 Supervised and unsupervised learning

There are two major types of learning processes.
The first one is called supervised learning: the values of f for the m samples in the training set, Ξ are know.
In the second one, called unsupervised learning, we simply have a training set of vectors without function values for them. The problem in this case is usually to partition the training set into subsets, *Ξ1, . . . , ΞR*, in a appropriate way.

3.3 Supervised and unsupervised learning

As it posed the problem of learning through machine learning has three main phases:

- *the training phase*, during which the model of learning is built using labelled data;
- *the testing phase*, during which the model is tested by measuring its classification accuracy on with held labelled data;
- *the deployment phase*, during which the model is used to predict the class of unlabeled data.

The three phases are carried out in sequence and iteratively (Figure 2) in a ML implementation.

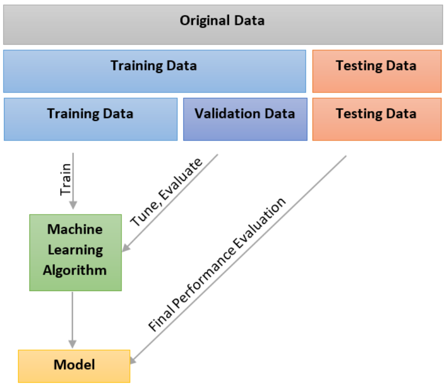


Figure 2. Machine Learning Algorithm phases

Obviously different strategies and algorithms can be used to solve the learning problem. There are literally thousands available, and hundreds more are published each year (Domingos, 2017). But basically each one of them consists of a combination of three components.

- representation: a classifier must be represented in a formal language that the computer can handle and belongs to a set of algorithms. This set coincides with the hypothesis space of the learner.
- evaluation: an evaluation function (also called objective function or scoring function) is needed to distinguish good classifiers from bad ones. The evaluation function used internally by the algorithm could be different from the external one used for optimization.
- optimization: a method to search among the classifiers in the language for the highest-scoring one. The choice of optimization technique is the key to improve the efficiency of the learner, and also helps determine the classifier produced if the evaluation function has more than one optimum.

The different machine learning techniques or learners could vary on the representation component. In the following paragraphs two special learners that has been taken as case studies for the flood predictions will be analyzed.

3.4 Support Vector Machines

Support vector machine (SVM) is a supervised machine learning algorithm that can be used for both classificationand regression challenges.
In SVM, the data points are plot in an N-dimensional space where N is the number of features and find a hyper-plane to differentiate the datapoints.
SVMs understanding can start from this simple concept: a simple way to classify a set of points in a plan is to draw a line and call points lying on one side positive and on the other side negative. If the two sets are well separated, one would intuitively draw the separating line such that it is as far as possible away from the points in both sets (Figure 3.3).

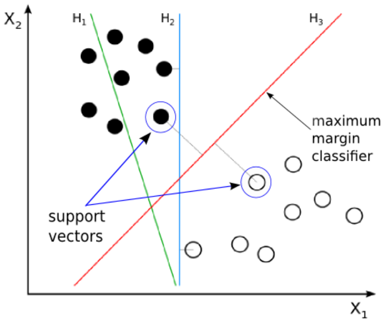


Figure 3. Support Vector Machines:concept

3.2 Artificial neural network

An Artificial Neural Network (ANN) is a learning algorithm that is inspired by the way of biological nervous systems, such as the brain, process information.
In simplified models of the brain, it consists of an input layer of neurons (or nodes, units),or one or two(or even three) hidden layers of neurons and a final layer of neurons and a final layer of output neuron (I):

$$hi=\sigma\left(\sum_{j=1}^{N} V_{ij} x_i + T_i^{hid}\right)$$

where σ() is called activation (or transfer) function, N the number of input neurons Vij the weights, xj inputs to the input neuron and the threshold terms of the hidden neurons.
ANN can be:

- Feed-forward networks: in which graphs have no loop
- Recurrent (or feedback) networks:in which loops occur because of feedback connections

Articialneural networks (ANNs) can be used as tools for prediction, classification, and decision support.

4. MACHINE LEARNING FOR FLOOD: THE CASE OF SEVESO

4.1 Case Study: The Seveso River

The Seveso is a 55-kilometre Italian river that flows through the provinces of Como, Monza e Brianza and Milan. It rises on Monte Sasso of Cavallasca, near San Fermo della Battaglia (Figure 4.1).

The Seveso has an area of approximately 226 km2, of which about 155 km2 includes the mountain basin, quite steep, that develops almost entirely in the territory of the province of Como; while the remaining 75 km2 constitutes the valley basin, which is part of the province of Milano (Becciu, 2018).

The Seveso river belongs to the Lambro-Seveso-Olona system that is not a natural watershed since the natural hydrology altered by human regulation of the territory around Milan city area. Olona and Seveso rivers did not happen to be natural tributaries of the Lambro river although they happen to be at the present.

Olona river in fact flows in the so called southern Lambro river which flows into the Lambro river about 20 km upstream the Lambro confluence into the Po river.

Seveso river is now connected to the Lambro-Olona system since its waters flow through the channel system beneath the Milan urban area and as Redefossi channel flows into the northern Lambro river (see Figure 5).

At present, the Seveso River environment appears to be extensively compromized. The processes of urbanization and industrial development during the last 50 years have strongly reduced agricultural and natural landscapes features particullarly in the northern part near Milan (Schetke and others 2010).

Heavy rain forced levels of the Seveso River in Milan to overflow.

In almost 140 years, 342 floods occurred (i.e. 2.4 per year) and 108 since 1976.

They are frequent, but often limited (although it is difficult to accept such

a number of floods per year in a town of a self-proclaimed civilized country). However, large floods have also happened: one of the worst floods we registered was on 18 September 2010, when major damages, in addition to destroyed cars and flooded garages and shops, were recorded on subway line (Becciu, 2016)

Obviously this frequent floods of Seveso became a strong problem for the city and it is necessary to design systems that could timely predict Seveso overflows also standing the meterological conditions.

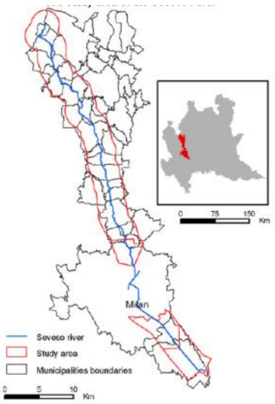


Figure 4. The Seveso river

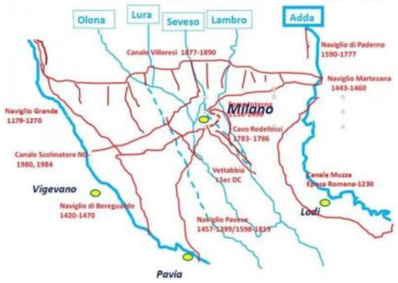


Figure 5. The Lambro-Olona-Seveso System

4.2 ML to predict Seveso floods

The risk of flood is affected by factors such as land use, meteorological events, hydrology and the topology of the land.

To have a reliable flood prediction indeed is important to consider the multivariate nature of flood prediction.

Particullarly, it can be argued that flooding is influenced by meteorological factors, but also by hydrological factors such as terrain slope, land use, vegetation, soil types, soil moisture, as well as hydrological processes related to run-off channels subject to flooding.

Figure 5.1 proposes a model for the flood prediction of Seveso rivers, that matches this two categories of input into a Machine Learning prediction algorithm based on ANN and Support Vector Machines.

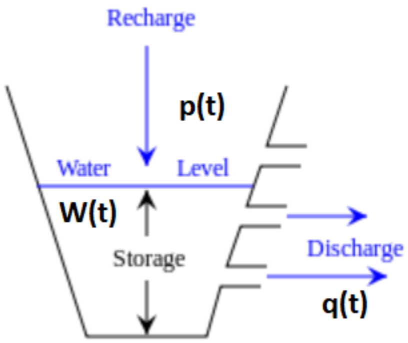
The model can be considered a ML translater of Rainfall-runoff models (Knapp, 1991) describe a portion of the water cycle and therefore the movement of a fluid - water - and therefore they are explicitly or implicitly based on the laws of physics, and in particular on the principles of conservation of mass, conservation of energy and conservation of momentum. The basic equation of the model is:

$$\frac{dW(t)}{dt} = p(t) - q(t)$$

$$W(t) = k\,q(t) \qquad \text{(II)}$$

where W(t) is the volume of water stored in the catchment at time t, p(t) is rainfall, q(t) is the river flow at time t and k is a constant parameter with the dimension of time (if the parameter was not constant the model would not be linear).

Rainfall-Runoff modeling is a classical approach in hydrology: such kind of models may include other input variables, like temperature, information on the catchment or others.



This contribution has been peer-reviewed.

https://doi.org/10.5194/isprs-archives-XLII-3-W4-549-2018 |

Figure 6. The Rainfall- Runoff Model

Several variants of the linear reservoir modeling scheme can be introduced, for instance by adopting a non linear relationship between discharge and storage. This non linearity can be due for example to unpredictability of extreme events or non linear correlation between rainfalls and flow, on which can be called as the Quantitative Precipitation Forecasting model (QPF) ().

The use of QPFin flood forecasting plays an important role, allowing for extension of the lead-time for the river flow forecast, which enables timelier implementation of flood control measures (Wang, 3000).

A reliable QPF is not an easy task to obtain due to rainfall being one of the most difficult elements of the hydrological cycle to forecast. Much uncertainty still affects the performance of rainfall prediction models (Reynolds, 2003). However, numerical weather prediction models such as the timely use of remote sensing observations (for example radar data and satellite images) allows the issue of short-term forecasts (Xue et al., 2000).

In the recent past, empirical, nonlinear, data driven models, like Artificial Neural Network (ANN) and Support Vector Machine (SVM) are being widely used to address the shortcomings of the parametric approach.

As Machine learning algorithrms their performance depends on the available data to be "learned", without any a priori hypothesis about the kind of relationship, which is allowed to be complex and nonlinear.

A the machine learning algorithm learns about its environment or a dynamic system through an iterative process of adjustments applied to its weights and biases.

The environment is characterised by a set of exemplars, which is typically a group of patterns of „environmental‟ variables. The algorithm becomes more "knowledgeable" about its environment after each iteration of the learning process. Like learning in human beings learning is an inferred process which cannot be perceived directly, but can be assumed to have happened by observing changes in performance (Zurada, 1992).

4.3 Use of Satellite Data

As mentioned in the previous paragraph, satellite multisensor-data can can be used to investigate the evolution in time and space of water bodies and water balance

Particullarly the internal calibrationof a distributed hydrological water balance model using satellite land surface temperature images (Mancini, 2013).

Although the output from satellite and radar images provides useful information on precipitation patterns they do not usually provide a satisfactory assessment of rain intensities.

Figure 5.2 depicts the role of remote sensing in a rainfall- runoff model (Mancini, 2013).

The strength of remote sensing techniques lies in the possibility to provide both spatial and temporal views of surface water quality parameters that is typically not possible from in situ measurements.

Besides water quality parameters, remote sensing enables us to investigate land cover dynamics and evolution; depending on the extent of target areas, a wide variety of satellite instruments, mostly coming from spatial medium resolution sensors are available for describing land cover at catchment scales (Giardino, 2010).

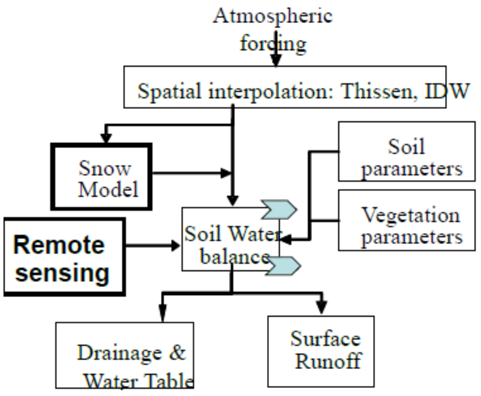


Figure 7. Use of Remote Sensing Images for flood prediction

For the EU countries indeed there is the possibility to access also to satellite data of Copernicus.

Copernicus, previously known as GMES (Global Monitoring for Environment and Security), is the European Programme for the establishment of a European capacity for Earth Observation.

Copernicus can provide a complex set of systems which collect data from multiple sources.

It consists of a complex set of systems which collect data from multiple sources: earth observation satellites and in situ sensors such as ground stations, airborne sensors, and sea-borne sensors.

Particullarly SENTINEL-3, can provide optical and altimeter mission monitoring sea and land parameters.

5. METHODS

5.1 Feature Set and Problem Setup

For the Seveso river case study the feature vector has been composed by the set of metereological variables

particullarly discharge d(t), temperature T(t) and precipitation P(T) (especially for snow events), and the hydrodynamic variables W(t) volume of water stored and q(t) river flow. Furthemore LIDAR satellite images of the river shapes has been taken from the Copernicus portal in order to get informations on evotraspiration, and solar radiation reflected and transmitted.

The outputs are the flood prediction risk R considered as a boolean variable in a binary classification problem and the flood return period Tm.

Particullarly the binary classification problem has been addressed-

In its simplest form, the binary classification reduces to: given a pattern x drawn from a domain X, estimate which value an associated binary random variable $y \in \{\pm1\}$ will assume..

The output set has been evaluated applying a ANN and SVM algorithms.

5.4 Evaluation

All the proposed algorithms have been evaluated in term of power of prediction and accuracy. The basis for the evaluation has been the confusion matrix as illustrated in Figure 5.9.

On the columns is represented the Predicted Class while in the rows the Actual Class.
Particullarly precision, recall and f-measure defined as:

6. RESULTS

Table 1 resumes the results of the perfomance evaluation of the two techniques.

	PRC	RCL	f-measure
ANN	77,23%	83,80%	80,41%+3,13%
SVM	79,23%	82,90%	87,52%

Table 1. Machine Learning algorithms prediction result

6. CONCLUSIONS

Machine Learning techniques provide a feasible example of alghoritms that could easily predict extreme events by using multivariate datasets.

The work has explored particullarly the use of two algorithms indeed Artificial Neural Networks and Support Vector Machines.

In particular the work has demonstrated the possibility to address the non-linearities of the predicition of extreme events by data driven prediction techniques.

An ANN is able to handle non-linearity and automatically adjusts to new information, while generally requiring little computational effort (Rietjes and de Vos, 2008).

ANNs are widely accepted as powerful ways of modelling complex non-linear and dynamical systems for which there are large amounts of sometimes noisy data (Chen et al., 2002).

They are pedagogic as opposed to decompositional; whereas decompositional methods knowledge of the domain or the physical characteristics of the problem are needed, a pedagogical method requests only data and does not depend on knowledge of the relationships between factors that affect the problem.

ANN can thus be trained without having intimate knowledge of the hydrological or other aspects of flood forecasting

On the other way SVM can represent a potential alternative which relieves the user from a time-consuming trial and-error procedure of synthesizing the suitable network architecture as in ANN without compromisingon the prediction accuracy.

REFERENCES

WMO, 2017, Multi-Hazard Early Warning Systems (MHEWS), Available at: http://www.wmo.int/pages/prog/drr/projects/Thematic/MHEWS/MHEWS_en.html

Liong, S. Y., Sivapragasam, C. 2002. Flood stage forecasting with support vector machines. JAWRA Journal of the American Water Resources Association, 38(1), pp. 173-186.

Bell B., Wallace, B., Zhang, D. 2012. Forecasting river runoff through support vector machines. In Cognitive Informatics & Cognitive Computing (ICCI* CC), Proceeding 2012 IEEE 11th International Conference on Machine Learning, pp. 58-64.

Anderson T. W. 1958. An introduction to multivariate statistical analysis (Vol. 2,). New York: Wiley, pp. 5-3 .

William W Cohen. 1995. Fast effective rule induction. In Proceedings of the twelfth international conference on machine learning, pp.153-123.

Nasrabadi, N. M. 2007.. Pattern recognition and machine learning. Journal of electronic imaging, 16(4), pp. 049901 .”

Domingos, P. 2012. A few useful things to know about machine learning. Communications of the ACM, 55(10), pp. 78-87.

Jangyodsuk, P., Seo, D. J., Elmasri, R., & Gao, J. 2016. Comparative Presentation of Machine Learning Algorithms in Flood Prediction Using Spatio-Temporal Data. In Proceedings of the 2015 International Conference on Communications, Signal Processing, and Systems, pp. 1015-1023.

Clark , R. A., Flamig, Z., Gourley, J. J., Hong, Y. 2016. Machine Learning Predictions of Flash Floods. In AGU Fall Meeting Abstracts.

Ceppi, A., Ravazzani, G., Salandin, A., Rabuffetti, D., Montani, A., Borgonovo, E., Mancini M. 2013. Effects of temperature on flood forecasting: analysis of an operative case study in Alpine basins, Natural Hazards and Earth System Sciences, 13(4), pp. 1051-1062.