Politecnico di Milano

Facoltà di Ingegneria

Dipartimento di Elettronica e Informazione

Corso di Laurea Specialistica in

Ingegneria per l'Ambiente ed il Territorio

Direct Use of Hydroclimatic Information in Water Reservoir Operation

Supervisor:

Prof. Andrea Castelletti

Co-Supervisor:

Prof. Francesca Pianosi

Master Graduation Thesis by:

Alessia Cavalli

Matr. 739834

Academic Year 2010-2011

Contents

List of Figures

List of Tables

Abstract

The optimal management of the limited amount of water available in the reservoirs (artificial or natural) is a critical issue all around the world. Indeed, while water scarcity creates economic and social problems in an everyday bigger fraction of the Earth, floods always result in property damages and loss of life. In order to more efficiently manage the different amounts of water it is necessary to improve the knowledge of hydrologic cycles and, between them, to deeply analyze the formation mechanisms of the inflow.
In this thesis the attention is focused on the water optimal management in artificial reservoirs. Usually, to solve the problems of optimal management of these reservoirs, we rely on tools able to optimize the control of complex, stochastic and non-linear systems: one of the more commonly adopted is the Stochastic Dynamic Programming (SDP). However the SDP suffers from 2 big limitations, the ”curse of dimensionality” and the ”curse of modelling”. Due to the latter the SDP can not use the information contained in the exogenous variables unless they are dynamically modeled. But the formation and trends of the inflow are influenced by several exogenous variables, including hydrological, climatic and meteorological variables. Not consider those variables in the management of a reservoir results in an inevitable loss of information.
To overcome this lack of knowledge we propose a new methodology able to exploit the exogenous information more relevant for the formation of the inflow.
In particular the objective is to identify a procedure through which: assessing the space for improving reservoir operation with respect to the level of performances reached by the traditional management systems; screening among different information sources to identify the most relevant variables and lead time for operational purposes; designing operating policies able to exploiting such information. With this new procedure we want to switch from the traditional approach of exploiting exogenous information, which includes the use of an inflow forecasting model (model-based), for a more innovative one where

the exogenous information is directly used into the management policy (model-free). Afterwards this new procedure is tested in the creation of an optimal operating policy for the management of the reservoir Hoabinh, located in the basin of the Red River in Vietnam. The proposed approach gives promising results. Thanks to the direct inclusion into the operating policy of some hydroclimatic variables, selected through the use of an Input Variable Selection (IVS) algorithm, better management performances have been achieved compared with those obtained through traditional approaches. In addition, in order to further improve the analysis previously made, it is analyzed the possibility of using variables related to phenomena of global atmospheric circulation, such as indices associated with the El-Niño phenomenon.

Finally, to verify whether the use of the significant variables selected by the IVS lead to increased performance, we create an inflow forecasting model for the inflow of the Da River using those variables. The results obtained in this particular case are not encouraging: in fact it is inferred that the exogenous information does not provide an appreciable contribution in predicting the future inflow.

Riassunto

La gestione ottimale delle quantità d'acqua disponibili all'interno dei serbatoi (artificiali o naturali) è una questione di grande importanza in tutto il mondo. Infatti, mentre la scarsità d'acqua crea problemi economici e sociali in una frazione ogni giorno più grande della Terra, le piene incontrollate sono da sempre causa di danni alle proprietà e di perdite di vite umane. Per rendere più efficiente la gestione di questi serbatoi è necessario migliorare la conoscenza dei cicli idrologici e approfondire la comprensione della formazione dell'afflusso.

In questa tesi l'attenzione è focalizzata sulla gestione ottimale dell'acqua all'interno dei serbatoi artificiali. Solitamente, per risolvere i problemi di gestione ottimale di questi serbatoi, ci si affida a strumenti in grado di ottimizzare il controllo di sistemi complessi, stocastici e non-lineari: il più usato è la Stochastic Dynamic Programming (SDP). Tuttavia la SDP è affetta da svariate limitazioni tra cui quella di non poter usare l'informazione contenuta in variabili esogene a meno che esse non siano dinamicamente modellizzate (maledizione della modellazione). La formazione e l'andamento dell'afflusso sono influenzati da svariate variabili esogene, tra cui variabili idrologiche, climatiche e meteorologiche. Non considerare queste informazioni nella gestione di un serbatoio si traduce in un'inevitabile perdita di informazione.

Per superare questa perdita di informazione si propone una nuova metodologia in grado di sfruttare l'informazione esogena, più rilevante rispetto alla formazione dell'afflusso confluente nel serbatoio in questione.

In particolare l'obiettivo è quello di identificare una procedura tramite cui: valutare se l'acquisizione di variabili esogene possa effettivamente migliorare le prestazioni delle politiche di gestione già esistenti; selezionare le variabili più rilevanti per spiegare l'afflusso al serbatoio; definire politiche di gestione ottima sfruttando le variabili selezionate, incorporandole direttamente nella politica. Con tale procedura si vuole passare dal trad-

izionale approccio di sfruttamento dell' informazione esogena, che prevede l'utilizzo di un modello di previsione per l'afflusso (model-based), ad uno più innovativo in cui l'informazione esogena viene direttamente sfruttata all'interno della politica di gestione (model-free).

In seguito questa nuova pocedura è stata testata nella creazione di una politica ottima per la gestione del serbatoio Hoabinh, posizionato nel bacino del Fiume Rosso in Vietnam. L'approccio proposto ha dato risultati promettenti: grazie all'inclusione diretta di alcune variabili idroclimatiche, selezionate tramite l'uso di un algoritmo di Input Variable Selection (IVS), nella politica di gestione, si sono ottenute prestazioni gestionali migliori rispetto a quelle ottenute tramite i metodi tradizionali. Inoltre, per poter migliorare ulteriormente l'analisi fatta in precedenza, si è analizzata la possibilità di utilizzare variabili legate a fenomeni di circolazione atmosferica globale, come gli indici associati al fenomeno El-Niño.

Infine, per verificare se l'utilizzo delle variabili significative selezionate dall'IVS conduca ad un incremento delle prestazioni, si è voluto creare un modello di previsione per l'afflusso del Fiume Da utilizzando tali variabili. I risultati ottenuti in questo caso particolare non sono incoraggianti: si è infatti dedotto che l'informazione esogena non fornisce un contributo apprezzabile nella previsione dell'afflusso futuro.

Introduction

Water is becoming the new *gold* of our days, the *blue-gold*, and the availability of adequate amounts of water is a fundamental requirement for the sustainability of human and terrestrial life. Nowadays sustainable water resource management is a critically important priority across the globe. Thus, the importance of implementing the best possible management for the limited amounts of water is growing everyday, in every place of the Earth. Nevertheless developing efficient operating policies for the management of every water resource is always challenging due to the significant impact on performances the uncertain inflow and the variability of demands have. Inflow is related to fresh water availability for humans, animals, and plants, and to incidences of natural hazards, such as flood and drought, that occur abruptly and may result in loss of human and animal life. The forecast of its future trends has been deeply analyzed by researchers, since it provides crucial information for adaptive water resources management.

Exogenous information, especially hydroclimatic information, are usually exploited to build accurate inflow models and, then, model predictions are subsequently used to design more informed decisions. In this thesis we explore an alternative methodological approach for the incorporation of hydroclimatic data into the operating policy, without the intermediation of physical models (model-free). We will try to answer to this question: Can, the direct inclusion of exogenous hydroclimatic information in a reservoir operating policy, improves the performance of the reservoir management?

The purpose of this thesis is to establish a procedure for assessing and exploiting hydroclimatic information in water resource management, specifically in reservoir operation. We propose methodological guidelines and numerical tools for: assessing the space for improving reservoir operation respect to the level of performances reached by the traditional management systems; screening among different information sources to identify the most relevant variables and lead time for operational purposes; designing operating

policies able to exploiting such information.
The proposed approach is tested on a case study, in particular: the Hoabinh reservoir in the Red River Basin located in the Vietnam North-West territories (see Castelletti, 2011). Finally, the thesis is concluded with a study on inflow forecasting models able to exploiting the most relevant hydroclimatic variables selected by the proposed procedure. The aim of this last part is to test if the hydroclimatic variables selected for the case study, can be used also to create good inflow forecasts.

The structure of the thesis is organized as follows:

- Chapter 1: synthetic theoretical introduction about the formulation of a reservoir management problem and its possible solution methods, followed by a preliminary description of our new approach of exploiting the value of hydroclimatic information.

- Chapter 2: the new approach is fixed in a methodological and numerical procedure of 4 steps.

- Chapter 3: description and discussion about the application of the procedure, described in Chapter 3, to the case study in Vietnam, analysis of the results.

- Chapter 4: creation of an inflow forecasting model, by using an Artificial Neural Network, exploiting some hydroclimatic variables.

Chapter 1

Water Resources Management and the Role of Exogenous Information

The reservoirs, intended as regulated storage facilities, have always been a very powerful tool for the reallocation of water resources in time and space. With a proper management of them it is possible to meet several goals, even simultaneously, such as avoiding dangerous flood, accumulating water for dry seasons or exploiting it constantly over time to produce hydroelectric power. Indeed reservoirs optimization problem is by far one of the most studied subjects in the water resources research area. The basic mathematic formulation of this kind of problems is discussed following; later on several ways to solve it are proposed.

1.1 Reservoir Management Problem Formulation

The term reservoir indicates a storage and regulation structure that can be either an artificial or natural lake which is regulated by an artificial barrier. It is a system always related to natural catchments feeding it, barriers (dams or diversions), water users(e.g. hydropower plants, irrigation districts), and artificial and natural canals that connect all the above components. The task of the Reservoir Management Problem formulation consists in determining the best optimal sequence of release decisions over a defined horizon.
The physical processes that are involved in the system are time-continuous, however the model structure is always time-discrete as release decisions are taken at discrete instants

of time. The decision time-step is usually one week or one day and, in any case, not smaller than few hours, because of the physical constraints in the implementation of the decision (e.g. dam's gate operation). The system dynamics is thus given by the state transition equation:

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t, \boldsymbol{\varepsilon}_{t+1}, t) \tag{1.1}$$

where $\mathbf{x}_t \in \mathbb{R}^{n_x}$ and $\mathbf{u}_t \in U_t \subseteq \mathbb{R}^{n_u}$ are the state and control vectors at time t; and $\boldsymbol{\varepsilon}_{t+1} \in \mathbb{R}^{n_\varepsilon}$ is the disturbance[1] acting in the time interval $[t, t+1)$. The state vector $\mathbf{x}_t$ is composed of the reservoir storages and the state variables of catchments, canals, and water users. The control vector includes the release decisions at the reservoir outlet and the distribution decisions at the regulated dams. The disturbance vector $\boldsymbol{\varepsilon}$ collects the random disturbances acting in the system, e.g. climate or hydrological inputs, and error terms in the model of the system. In this thesis we consider the disturbance as a stochastic variable modeled by a pdf $\phi_t(\cdot)$, as shown in the following equation:

$$\boldsymbol{\varepsilon}_{t+1} \sim \phi_t(\cdot \,|\mathbf{x}_t, \mathbf{u}_t) \tag{1.2}$$

The system must be operated considering several m issues such as agricultural and hydropower production, flood control, ecological services. For accounting them in the structure of the model it is necessary defining them with an appropriate mathematical function called "the objective function", that should be defined to express the cost payed over the time horizon $[0, h]$

$$J^i = \underset{\varepsilon_1,\ldots,\varepsilon_h}{E} \sum_{t=0}^{h-1} [g_t^i(\mathbf{x}_t, \mathbf{u}_t, \boldsymbol{\varepsilon}_{t+1}), g_h^i(\mathbf{x}_h)] \tag{1.3}$$

where $g_t^i(\cdot)$, for $t = 1, \ldots, h-1$ are the step-cost functions created for every m issue, associated to the transitions from t to $t+1$ and $g_h^i(\cdot)$ is a penalty function over the final state. In this work the Laplace formulation has been chosen to filter the disturbance, i.e the expected value E is the statistic used to filter the disturbance and as aggregation over time the sum $\sum$ is used (for more information see Soncini-Sessa, 2007). The control vector is

[1] According to the notation adopted, the time subscript of a variable indicates the instant when the its value is deterministically known.

specified by a time-varying control law (also called operating rule):

$$\mathbf{u}_t = m_t(\mathbf{x}_t) \tag{1.4}$$

and the aim of the control problem is to define the sequence of control laws $m_t(\cdot)$ over the horizon $[0, h-1]$, i.e. the control policy (also called operating policy):

$$p = [m_0(\cdot), \ldots, m_{h-1}(\cdot)] \tag{1.5}$$

The optimal control problem is formulated as

$$\min_p \left[J^1, J^2, \ldots, J^m\right] \tag{1.6}$$

subject to constraints (1.1), (1.2), (1.4), (1.5), and with $\mathbf{x}_0$ given. The control problem (1.5) is a multi-objective (MO) optimization problem, whose solution is the set $\mathscr{P}$ of Pareto optimal (efficient) policies (see e.g. Miettinen, 1999). Each policy in $\mathscr{P}$ can be computed by solving the following single (aggregate) objective (SO) optimal control problem:

$$\min_p J \tag{1.7}$$

subject to constraints (1.1), (1.2), (1.4), (1.5), and with $\mathbf{x}_0$ given, and

$$J = \underset{\varepsilon_1,\ldots,\varepsilon_h}{E} \sum_{t=0}^{h-1} [g_t(\mathbf{x}_t, \mathbf{u}_t, \varepsilon_{t+1}), g_h(\mathbf{x}_h)] \tag{1.8}$$

where $g_t(\cdot)$ and $g_h(\cdot)$ are the aggregate step-cost and penalty functions obtained from $g_t^i(\cdot)$ and $g_h^i(\cdot)$ (with $i = 1, \ldots, m$) according to the aggregation method (see, e.g., Soncini-Sessa, 2007) used to re-conduct the MO problem to a SO problem.

In the water resources context, the choice of the time horizon and the penalty function $g_h(\mathbf{x}_h)$ might be critical since the life time of the system is infinite. Generally, the adoption of an infinite horizon, which vanquishes the influence of the the penalty, is recommended. When the model of the system and all the step-cost functions are cyclostationary with period T, the problem on the infinite horizon is well-posed and the solution is a periodic control policy. The SO problem over an infinite horizon is formulated as

$$\min_p \lim_{h\to\infty} J \tag{1.9}$$

subject to (1.1), (1.2), (1.4), given $\mathbf{x}_0$, and the control policy over the period T

$$p = [m_0(\cdot), \ldots, m_{T-1}(\cdot)] \tag{1.10}$$

instead of (1.5).

In this work we consider a simplified management problem formulation with only one reservoir to manage and one objective to be optimized (flood control); our task will be implementing a point-valued control policy that gives a single release decision for every decision step. For more details about the case of study see Chapter 3.

1.1.1 Traditional Problem Solution: The dynamic programming approach

The final task of the problem previously described is to define an optimal control policy (or operating policy). To accomplish this goal there are several approaches that can be used:

- *Functional Approach* that determines the optimal policy as a succession of control laws upon which no conditions are imposed. This approach is used both to determine off-line policy through *Stochastic Dynamic Programming* (SDP), with algorithms based on the numerical resolutions of the Bellman equation (see Yeh, (1985) for a review of the first applications of SDP to water resources management and Soncini-Sessa (2007) for recent improvements), and to determine on-line policy. The SDP is able to provide the optimal policy under very general assumptions. However it suffers from some critical limitations.

- *Parametric Approach* that fixes a priori the class of functions to which the control law must belong, so that a particular function, and also a particular policy, is defined by a finite number of parameters and the policy design will consist in identifying the values of the parameters that minimize the objectives (Soncini-Sessa, 2007). This approach is used to determine off-line policy when the algorithms based on SDP cannot be used due to too high computational requirements (that grow exponentially with the system dimension).

- *Learning Approach* that leaves the system to evolve under a suitable algorithm, which experiments with alternative controls until, by trial-and-error, it identifies the

optimal policy (Castelletti, 2010b). This approach, based on the ideas developed in a branch of the Artificial Intelligence which is named *Reinforcement Learning*, allows to determine the optimal operating policy considering, in the system model, also the exogenous inputs.

Stochastic Dynamic Programming (SDP) appears to be the most suitable, and one of the more commonly adopted, method for solving problem (1.7). One pillar of SDP success is its wide applicability. Indeed, the only requirements for its application are: (1) the inputs in the model can only be controls or random disturbances, which means that it is not possible to consider (and condition the policy upon) uncontrolled, exogenous, deterministic variables whose value is known in real time (e.g. rainfall measures), unless these are described by a dynamic model and so are not exogenous inputs anymore; (2) the membership-set or the pdf of the disturbance vector must be in the form as in (1.2), i.e. either the disturbance process is independent in time or, at time t, any dependency on the past could be completely accounted for by the value of the state at the same time; and (3) the step-cost functions $g_t(\cdot)$ only depends upon variables defined for the same time interval. The first condition leads to the so-called *"curse of modelling"*; this means that SDP cannot consider exogenous information.
Stochastic Dynamic Programming is an algorithm based on the calculation of the 'optimal cost-to-go' for every time instant t. Basically it is based on the resolution of the Bellman Equation, (Bellman, 1957). The Bellman equation for the SO finite horizon optimal control problem (1.7) is

$$H_t(\mathbf{x}_t) = \min_{u_t} \Psi_{\varepsilon_{t+1}} \left[\Phi \left[g_t(\mathbf{x}_t, \mathbf{u}_t, \varepsilon_{t+1}), H_{t+1}(\mathbf{x}_{t+1}) \right] \right] \tag{1.11}$$

where $H_t(\cdot)$ is the optimal cost-to-go function for the aggregate objective and only the following combinations of Φ and Ψ are considered

$$\begin{aligned} \Phi[v,w] &= v+w && \text{and } \Psi = E \\ \Phi[v,w] &= \max\{v,w\} && \text{and } \Psi = \max \end{aligned}$$

The solution of this equation leads to an optimal control policy based on the criteria of minimizing the total expected cost of all the stages, from the time the choice is made onwards. Each control, defined by the control policy for each decision step, incurs in an immediate cost $g_t(\mathbf{x}_t, \mathbf{u}_t, \varepsilon_{t+1})$, but also impacts, through the state $\mathbf{x}_{t+1}$ that it contributes

to produce, the context in which the next control choice will be made and thus the effect that this latter will produce on all the future stages. The solution is obtained by initializing $H_h(\mathbf{x}_h)$ with $g_h(\mathbf{x}_h)$ and recursively computing $H_t(\mathbf{x}_t)$ with Eq. (1.11). Once the optimal costs-to-go have been computed for all the time instants $t = h-1,\ldots,0$, the optimal operating rule at any time t is derived as:

$$m_t(\mathbf{x}_t) = \arg\min_{u_t} \underset{\varepsilon_{t+1}}{\Psi} \left[\Phi\left[g_t(\mathbf{x}_t,\mathbf{u}_t,\varepsilon_{t+1}), H_{t+1}(\mathbf{x}_{t+1})\right]\right] \tag{1.12}$$

Thus, the result of the SDP consists in an operating policy from which it is possible to derive the single release decisions for a specific reservoir or a set of reservoirs. These decisions are made to maximize current benefits plus the expected benefits from future operation, which are represented by the recursively calculated cost-to-go function.
As anticipated, SDP is a powerful tool for solving problems such as (1.7), however it suffers from 2 big limitations. The main limit of SDP is the so called "*curse of dimensionality*", i.e. the associated computational complexity grows exponentially with the state, control and disturbance dimensions. This limits the use of SDP to small water systems where the number of reservoir is smaller than a few units (2 or 3). Then, we already anticipated that there is another critical limitation the SDP have to deal with, "*curse of modelling*": the inability to directly incorporate exogenous information unless these are properly modelled, thus enlarging the state of the system. This means that it is not possible to use exogenous information such as hydrologic, climatic and meteorological variables, directly within the structure of a SDP problem. But this restriction imposes to not consider every kind of exogenous variables that could be potentially very important in taking every release decision, such as rainfall data, air moisture or temperature, snow-pack values, soil moisture, evapotranspiration or even mid to long term climatologic phenomena.
Can the exploitation of these kind of hydroclimatic variables (as the exogenous variables will be called from now on) actually improve the reservoir management? In this work we try to answer to this question by overcoming the restriction of the SDP by creating a new procedure able to exploit the value of exogenous information. We provide methodological guidelines for: assessing the actual space for improving system operation; screening among different information sources to identify the most relevant variables and lead time for operational purposes; designing operating policies that exploits such information.

1.2 Assessing the space for improvement: The Ideal Solution

The control problem (1.5) is a multi-objective (MO) optimization problem and by solving it with an algorithm like SDP the set of Pareto optimal (efficient) policies is found.

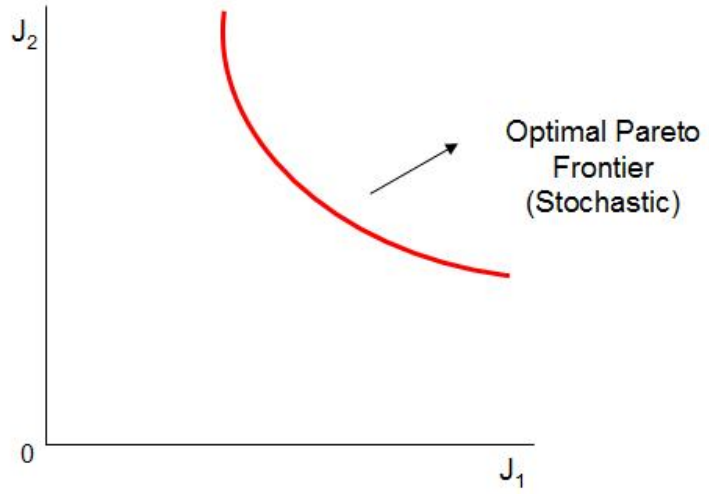


Figure 1.1: Schematic Pareto Frontier obtained by solving a SDP problem.

In the previous figure a schematic Pareto Frontier is shown; it is ideally obtained by the resolution with SDP of a management control problem related to a reservoir operation in which 2 generic objectives (J_1 and J_2) must be optimized (minimized).
Is it possible to improve the performance of this management problem? Is it possible to create a procedure able to move the frontier of Figure (1.1) towards the origin of the axes? The first step in our methodology consists in answering to these questions, i.e. in assessing the potential space for improvement stemming from the incorporation of hydroclimatic information in the decision model. This can be obtained as the difference between the operation performance computed assuming that the maximum possible information is available to the decision-maker (i.e. perfect knowledge of all present and future hydrological conditions) and the dual situation, in which decisions are taken only on the basis of the storage at a given time, with no consideration of additional hydroclimatic data. The former operation can be designed by deterministic optimization, the latter by stochastic optimization. The difference between the objective values produced by the

two optimizations is the Expected Value of Exogenous Information (EVEI) and gives an indication about the maximum space of improvement that can exists thanks to the use of exogenous information.

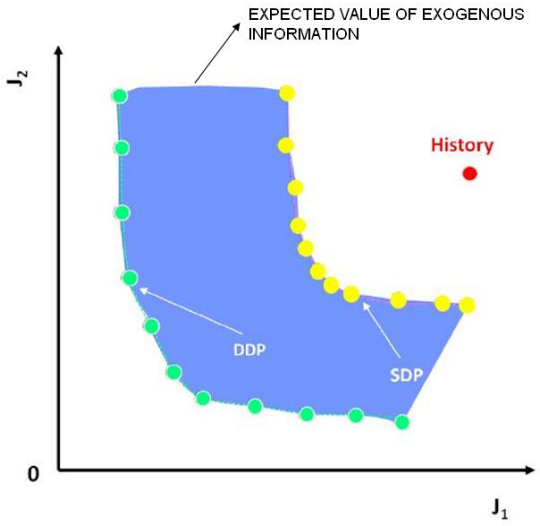


Figure 1.2: Pareto frontiers obtained by resolution of a DDP problem and of a SDP problem, and in blue is highlighted the area of 'expected value of exogenous information'.

Deterministic optimization problem and stochastic optimization can be solved with several methods, in this thesis we always consider the Deterministic Dynamic Programming (DDP) and the Stochastic Dynamic Programming (SDP) as solving algorithms.
In Figure (1.2) it is assumed that the deterministic optimization is actually able to get significantly better operating results; this is a realistic assumption because the DDP is based on the idea that all the information about the system are knew in advance, so the implicit stochastic nature of the reservoir management problem is completely ignored.
What is expected to do in this thesis is to investigate the EVEI space shown in Figure (1.2); i.e. the possibility to improve the optimal operating policy obtained by the stochastic approach of SDP, by directly including some valuable hydroclimatic variables. Hence the idea implemented is to switch from the formulation of the operating rule (1.4) to the formulation:

$$\mathbf{u}_t = m(t, \mathbf{s}_t, \mathbf{I}_t) \tag{1.13}$$

where $\mathbf{I}_t$ is the input vector containing the most valuable hydroclimatic variables and $\mathbf{s}_t$ is the vector containing the storage of the reservoir at the different time instants t; in the

case of reservoir management the storage of the reservoir itself is generally considered as the state of the system, thus, from now on, the state variable x_t is always replaced by the variable s_t.
Then the result will be compared with the optimal release sequence $\mathbf{u}^*$ generated by the DDP that is conceptually equivalent to a feedback operating rule of the form:

$$\mathbf{u}_t \cong \mathbf{m}(\mathbf{s}_t, t, \bar{q}_{t+1}, \bar{q}_{t+2}, ..., \bar{q}_h) \tag{1.14}$$

where $\bar{q}_t$ is the measured inflow to the reservoir and h is the lead horizon.

The purpose is to design a nearly equally performing but implementable control and to accomplish this goal the future sequence of inflow must be replaced by a vector of hydroclimatic variables selected in such a way to approach as much accurately as possible the optimal sequence $\mathbf{u}^*$ generated by DDP.

1.3 A new approach

Our goal is to create an optimal operating rule able to exploit the information contained in the hydroclimatic variables and able to performs better than the law defined by the resolution of a SDP problem. To reach this goal the input vector $\mathbf{I}_t$ must be defined, and it must contains the most valuable hydroclimatic information available at time t that better works as a surrogate of the perfect knowledge of future inflow. This surrogate of the future knowledge can be defined in several ways. One of these is the traditional model based surrogate, i.e. the inflow prediction. So the control policy assumes this formulation:

$$\mathbf{u}_t = \mathbf{m}(\mathbf{s}_t, t, \hat{q}_{t+1}, \hat{q}_{t+2}, ..., \hat{q}_h) \tag{1.15}$$

where $\hat{q}_t$ is the predicted inflow to the reservoir. In this case the input vector $\mathbf{I}_t$ must contains the most valuable information in order to make the best possible inflow forecast, see Figure(1.3).

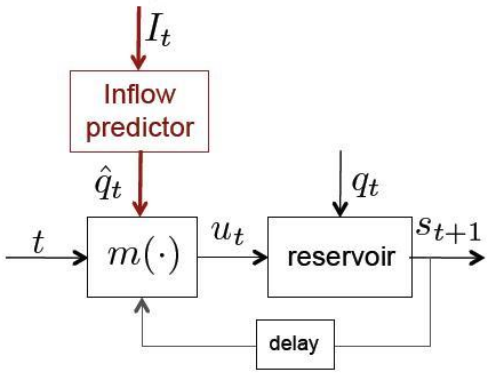


Figure 1.3: Scheme of an optimization model with model-based use of exogenous information.

The use of hydroclimatic information in inflow forecasting models has already took a real advantages in creating efficient operating policies (Makkearson, 2008). However the use of a forecasting model inevitably creates some new computational errors, so we will try to investigate the idea of directly use the same raw exogenous information employed by predictors (see Guariso, 1986). This thesis proposes a new methodology to improve reservoir operation by smartly enlarging the set of information up on which the operating policy is conditioned by basing on a new model-free approach and no more on the traditional model-based one. Indeed to reach our goal we do not follow the conventional model-based approach described above, where observations are used to identify models and model predictions are used to inform decisions. Rather we adopt a ”model-free” approach, see Figure (1.4), where data directly feed the decision model, i.e. an operating rule that provides the decision as a function of the current system conditions (the state) and of any other useful hydroclimatic information as suggested by a variable selection process. The operating rule is then calibrated using simulated system states and historical time series of hydroclimatic variables.

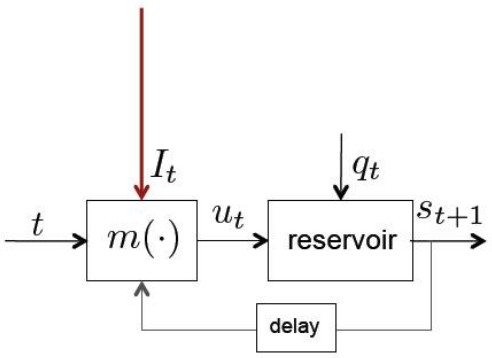


Figure 1.4: Scheme of an optimization model with model-free use of exogenous information.

Nevertheless, before implementing this process, there are a few challenges that must be analyzed. It must be decided what kind of input variables collect and analyze regarding to the system main dynamics and regarding to all the objectives that must be optimized by the operating policy. Indeed there are a lot of aspects that must be considered. First of all the objectives of the management problem: e.g. if the major objective is the flood control it may be useful a few-days-ahead information, and so, only variables with a fast dynamics should be used, instead if the major objective is irrigation, only low frequency variables should be considered because the needed information becomes seasonal. Second the main physics, geographical and technical aspects of the whole system must be observed. Third, the data availability need to be considered. After this first analysis the number of candidate variables can be very high due to the presence of multiple, possibly redundant information and spatial variability, so an empirical selection is not always effective. Hence, an instrument to select the most valuable information is necessary. In this thesis to deal with this task is used an 'Input Variable Selection' algorithm (IVS), as described in the next chapter.

Chapter 2

A procedure for using Exogenous Information in improving Water Resources Systems Management

In the traditional approach, the use of selected hydroclimatic variables such as rainfall, snow cover, temperature, evaporation, soil moisture, etc, is confined to the creation of the future inflow sequence by the creation of a model-based inflow predictor. In this work, we explore an alternative model-free approach based on the direct use of hydroclimatic information to conditioning the release decision derived from the control law. Accordingly to this model-free philosophy of the approach, the set of candidate variables to serve for this purpose includes a wide range of information at different locations and for different time lags. Selection among such variables can be based on expert judgements or rely on statistical tools like cross-correlation or non-linear input variable selection methods, assuming the optimal deterministic release schedule as output and the candidate hydroclimatic variables as regressors. The aim of this chapter is to show the implementation of a procedure able to integrate the most valuable hydroclimatic information directly into the control law. This procedure, as shown in Figure (2.1), have been schematized in 4 step:

1. the deterministic optimization,
2. the candidate variables selection,
3. the use of the 'Input Variable Selection' algorithm,

4. the stochastic optimization.

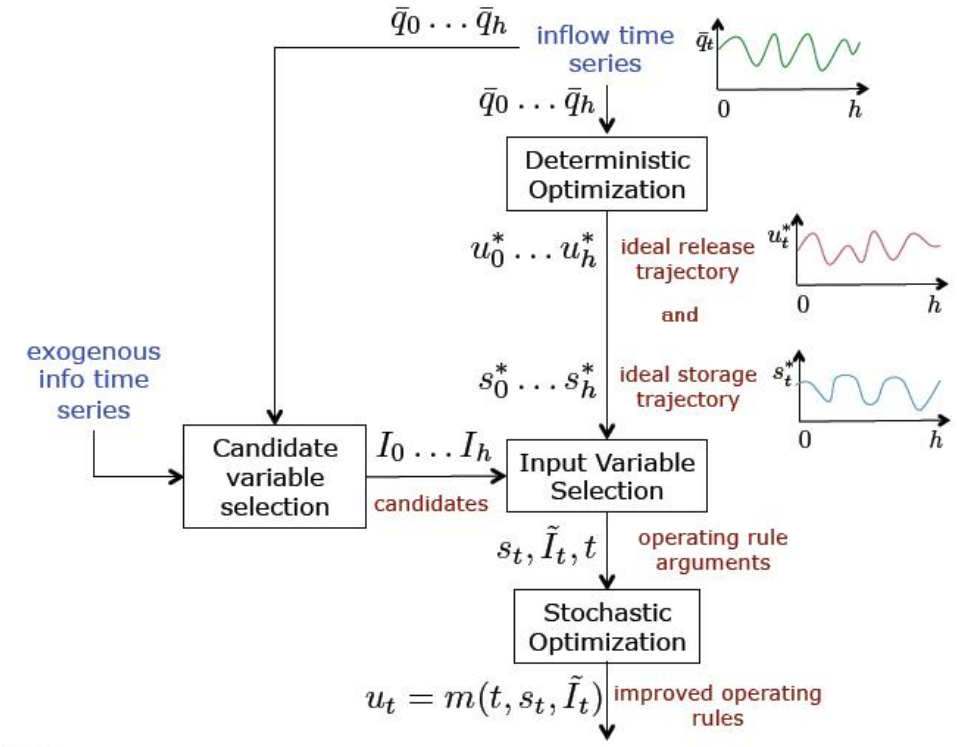


Figure 2.1: Summary of the procedure.

2.1 The deterministic optimization

The deterministic optimization problem is

$$\min_{\mathbf{u}} J(\mathbf{s}_0, \mathbf{u}, \mathbf{q}) \qquad \text{s.t. } (\mathbf{x}_0, \mathbf{q}) \text{ given} \tag{2.1}$$

where $\mathbf{u} = |u_0, ..., u_{h-1}|$ is the sequence of decisions to be taken over the optimization horizon $[0, h-1]$, and J is the objective function (cost) whose value also depends on the initial state $\mathbf{s}_0$ (i.e. the reservoir storage) and the trajectory $\mathbf{q} = |q_1, ..., q_h|$ of the uncontrolled inputs (in this thesis they are always represented by the reservoir inflow data) over the optimization horizon.

The solution of problem (2.1) consists in the optimal trajectory of release decisions, as shown in Eq. (2.2):

$$\mathbf{u}^* = |u_0^*, ..., u_{h-1}^*| \tag{2.2}$$

However this solution is optimal only under the given trajectory of future inflow. Such almost exact knowledge of the future inflow is not available, so problem (2.1) can be formulated only over past time horizon where time series of observed inputs are available. Consequently, its solution has no operational value but it only provides an upper bound of the system performances that the decision-maker could have obtained on that horizon. The problem can be solved either by non-linear programming NLP methods (e.g. gradient-based or direct search methods) or by deterministic dynamic programming DDP. The applicability and effectiveness of each method depend on the properties of the system and of the objective function. In this work, DDP was used. When applicable (i.e. when the objective function J is separable and the number of state and decision variables is limited) DDP is computationally efficient and provides very accurate solution, the only source of error being the discretization of the state and decision variables.

First step in our procedure is to solve problem (2.1) to identify the ideal solution, i.e. Eq. (2.2), in a way to assess what performances the hypothetical single decision maker could attain if (s)he had perfect knowledge of time pattern of future inflows. Starting from this ideal solution the results of all the other operational algorithms can be compared with this ideal reference. This work to identify the ideal solution using the DDP was widely carried out in the PhD thesis of Quach (2011).

2.2 Candidate Variable Selection

Before implementing an Input Variable Selection algorithm it is necessary to define the set of candidate hydroclimatic variables. As anticipated in the previous chapter, we are searching for variables able to give valuable information about the inflow formation and trend and able to play as surrogate of the future inflow for a specific reservoir. Theoretically the inflow is physically influenced by a variety of different variables with which it can have linear or non-linear relationships; it is possible to summarize these relationships

with this generic identity

$$q_t = f(\mathbf{I}_t) \tag{2.3}$$

where f is a generic set of functions, each one of them describing the relation between every components of vector $\mathbf{I}_t$ with the inflow formation and trend. Our final task is to fulfill the vector $\mathbf{I}_t$. To select appropriately these components of $\mathbf{I}_t$ it is fundamental a deep study of the main characteristics and dynamics of the reservoir under exam.
To empirically choose what kind of variables can be candidates, different evaluations must be done about:

- The objectives of the management problem and the related dynamics
- The data availability

First the objectives of the management problem and the related evaluation horizon must be analyzed; indeed if we need to solve a management problem to optimize a flood control objective, we will need to consider variables with a rapid dynamics. In this case variables such as the measure of precipitation at any or multiple steps should be analyzed. While if the optimization is related to the hydropower production or to the delivery of water for irrigation, variables able to give information with a seasonal dynamic should be considered such as snow pack/cover information, climatic variability indices like El-Niño Southern Oscillation (ENSO), soil moisture or solar activity data should be considered. Last but not least we have to face with the data availability.
Until now the most studied approach to include hydroclimatic information in reservoir management was the one of creating an inflow predictor, which, then, will be included in the optimization model. Recently, for making improved inflow forecasts, a lot of researchers have concentrated their attention in the study of large-scale atmospheric circulation phenomena and towards the different indices related to them. Such indexes are for example the ENSO indices. Thanks to these study it was stated that there is an effective improvement in making inflow forecasting by exploiting this kind of exogenous information, see e.g. Maity (2008). This proves that these variables are valuable in explaining inflow formation and trends, and then, that there is a potential to expand the scope of study by incorporating these more comprehensive sources of hydro-meteorological information directly in optimization of reservoir operation.

2.3 Input Variable Selection (IVS)

The optimal release sequence of equation (2.2) generated by deterministic dynamic programming is conceptually equivalent to a feedback operating rule as shown in Eq. (1.14).

According to our model-free approach, to design a nearly equally performing but implementable operation, the future sequence of inflow must be replaced by a vector $\mathbf{I}_t$ of hydroclimatic variables selected in such a way to characterize as much accurately as possible the optimal sequence $\mathbf{u}^*$. Candidate variables to serve as surrogate of future inflows include past and/or cumulated values of the inflow, precipitation, snow cover, and any other hydroclimatic information observed in the basin and relevant to the operation objectives. When the number of candidate variables is higher than few units, the selection process can considerably benefit from the use of input variable selection algorithms, which generally outperform expert judgment and cross-correlation analysis in presence of many redundant inputs and strongly non-linear underlying causal relationships. In this work, we used the tree-based Iterative Input Selection (IIS) algorithm introduced in the PhD thesis of Galelli (2010), which is an input selection algorithm that holds three features particularly useful in the problem we are dealing with: flexibility (i.e. the ability of modelling strongly non-linear functions), computational efficiency (i.e. the ability of processing large data-sets), and scalability with respect to the input dimensionality (i.e. the ability of handling several input variables with a different range of variability). This algorithm is based on a *reiterative input selection*, i.e. it selects the most valuable input variables basing on a cross validation procedure. Then the performance index of whole model is calculated as an average value of n models; where n is a parameter of the algorithm (called n-fold) and represents the number of section in which we decide to split the data-set. We prefer a *reiterative input selection* instead of a *direct input selection* approach because in the direct method the training data-set has to be split in a calibration and a validation set basing only on the expert's experiment. So the accuracy resulting from the experiments depends on the way of division made into the original data-set. To overcome such limit it is possible to use the reiterative procedure based on a random division of the data-set made by the algorithm itself and not by a human operator. The considered key performance parameter is R^2, the explained variance (also called coefficient of determination), that is a statistic providing a measure of how well future outcomes are likely to be predicted by the model

used, i.e.

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} \tag{2.4}$$

with SS_{err} as the sum of squares of residuals

$$SS_{err} = \sum_i (\hat{Y}_i - \bar{Y}_i)^2 \tag{2.5}$$

and SS_{tot} as the total sum of squares

$$SS_{tot} = \sum_i [\bar{Y}_i - (\frac{1}{N} \sum_i^N (\bar{Y}_i)]^2 \tag{2.6}$$

where N is the total number of observation. The better the model, the closer the value of R^2 is to one.
In the next section a specific IVS algorithm, used for the experiments taken in this thesis, will be explained.

2.3.1 Iterative Input Selection Algorithm

To fulfill all the requirements mentioned in the previous section, a model-free, forward-selection algorithm was chosen, its name is Iterative Input Selection (IIS). This algorithm was developed within the PhD thesis of Galelli (2010).
Given the output r_t and the vector z_t of candidate features, the IIS algorithm first globally ranks the elements of z_t accordingly to a statistical measure of significance that accounts for non-linear dependencies, and then refines the ranking by evaluating the individual contribution of the features ranked in the first p positions. This parameter p must be chosen 'a priori' and it represents the number of variables with the highest rank related to the output. Then the most significant feature is selected and employed as regressor for a predefined model. To account for features redundancy, the algorithm proceeds by repeating these operations on those data that still have to be explained, namely on the residuals of the model built at the previous iteration. The algorithm iterates these operations until the selection of new features does not further improve the performance of the model being built. Thus the IIS algorithm requires to select an effective statistical measure of significance, which, on its turn, influences the choice of the model class. The only parameters to

be specified are thus p and a tolerance ε used to terminate the algorithm. The steps of the IIS algorithm are shown here:

Step 1 Set $k = 0$ and $\tilde{z}_t$ as empty vector.

- Rank, in decreasing order, the features in the vector z_t according to their statistical measure of significance in explaining the output r_t.
- Select the features $z_t^1, ..., z_t^p$ ranked in the first p positions. For $i = 1, ..., p$ identify a model of the form $\hat{r}_t^{k,i} = c(z_t^i)$ and evaluate its performance R_i in explaining r_t.
- Denote as z_t^k and $\hat{r}_t^k$ the feature and the estimate of r_t corresponding to the model with the highest performance R^k. Store z_t^k in $\tilde{z}_t$.
- Compute the residual $e_t^k = r_t - \hat{r}_t^k$.

Step 2 Set $k = k + 1$.

- Rank, in decreasing order, the features in the vector z_t according to their statistical measure of significance in explaining the output e_t^{k-1}.
- Select the features $z_t^1, ..., z_t^p$ ranked in the first p positions. For $i = 1, ..., p$, identify a model of the form $e_t^{k-1} = c(z_t^i)$ and evaluate its performance R^i in explaining e_t^{k-1}.
- Denote as z_t^k the feature corresponding to the model with the highest performance. Store z_t^k in $\tilde{z}_t$.
- Identify a model of the form $\hat{r}_t^{k,i} = c(z_t^i)$ and evaluate its performance R^k in explaining r_t.
- Compute the residual $e_t^k = r_t - \hat{r}_t^k$.

Step 3 Termination test
If $(R^k - R^{k-1}) < \varepsilon$, the algorithm stops. The selected features are stored in $\tilde{z}_t$, with dimensionality $\tilde{z}_z = k - 1$. Otherwise, return to Step 1.

As for the model class and the statistical measure of significance, it is employed a class of tree-based methods, named *Extremely Randomized Trees* (Guerts, 2006), and an

Extra-Trees based ranking procedure.

Tree-based methods stand out as a class of non-parametric methods that can provide modelling flexibility, computational efficiency, interpretability and good accuracy in both regression and classification problems. They are all based on the idea of decision tree models, which are tree-like structures representing a cascade of rules leading to numerical values. These structures, composed of decision nodes, branches and leaves, are obtained by first partitioning at the top decision node, with a proper splitting criterion, the set of the regressors into two sub-sets, thus creating the former two branches. The splitting process is then repeated in a recursive way on each derived sub-set, until the numerical values belonging to a sub-set vary just slightly or only few elements remain. When this process is over, the branches represent the hierarchical structure of the sub-sets partition, while the leaves are the nest sub-sets associated to the terminal branches.

The Extremely Randomized Trees, or Extra-Trees, are a recent method for classification and regression problems proposed by Guerts (2006). Extra-Trees methods build ensembles of unpruned regression trees according to a top down approach that starts from the top decision node and systematically explores the regressors set. Extra-Trees, with respect to the other randomization methods (e.g. Random Forest (Breiman, 2001), PERT (Cutler and Guohua, 2001)), exploit the original training data-set and split the nodes by selecting the cut-point and the regressor totally (or partially) at random. The rationale behind these two characteristic is that the use of the original training data-set is motivated to minimize bias, while the randomization of both the cut-point and the regressor selection can reduce variance more efficiently than other randomization methods (see Guerts, 2006). The Extra-Trees based procedure used here has a few parameters that must be a priori fixed on the basis of the problem specifics, and by empirical or trail-and-errors evaluations:

- n_{min} : is the minimum number of observations needed to split a node. Large values of n_{min} lead to small trees (few leaves), with high bias and small variance. Conversely, low values of n_{min} lead to fully-grown trees, which may over-fit the data. The optimal value of n_{min} depends not only on the risk aversion to over-fitting, but also on the level of noise in the output of the training data-set: the noisier is the output, the higher should be the optimal value of n_{min}.

- *ScoreTh*: is another parameter varying between 0and 1 that controls the termination

criteria. It works together with the parameter n_{min} and the possible situations that should be considered are:

1. $n_{min} = 2; ScoreTh \sim 0.98 \rightarrow$ the termination criteria is lead by $ScoreTh$
2. $n_{min} = 100; ScoreTh \sim 0.98 \rightarrow$ the termination criteria is lead by the both parameters
3. $n_{min} = anynumber; ScoreTh = 0 \rightarrow$ the termination criteria is lead by n_{min}

- M: is the number of trees in the ensemble, influences the strength of the variance reduction and the behavior of the estimation error, which is a decreasing function of M (see e.g. Breiman, 2001). High values of M increase the accuracy of estimates, because more trees randomly are built and the final forecast is certainly more robust. However, the greater of M increases considerably the calculation time, thus a compromise must be found between high accuracy and computing time.

- *n-fold*: is the number of calibration and validation sets in which the original data-set is decided been split. The higher the *n-fold* parameter, the higher the precision of the model, but, again, the higher is the computing time.

- p: is the number of variables with the highest rank related to the output modeling that the human operator want to be shown.

2.4 Optimization

Once the hydroclimatic information vector $\mathbf{I}_t$ has been selected, the next step in our procedure is the identification of an optimal operating rule conditioned upon this information, i.e.

$$\mathbf{u}_t = \mathbf{m}(\mathbf{x}_t, t, \mathbf{I}_t) \tag{2.7}$$

The stochastic nature of the system is now preserved by describing all the uncontrolled inputs (inflow) as stochastic processes. The resulting stochastic optimization problem is inherently much more difficult than the deterministic problem (2.1) since it is a search in the infinite-dimensional space of functions $\mathbf{m}(\cdot)$ rather than a search in the space of vectors $\mathbf{u}$. Stochastic Dynamic Programming is by far the most widely used method to solve such a problem. However, its application is subject to the limitations imposed by

the so called curses of modelling and of dimensionality.
There are several other approaches available to compute the operating policy exploiting the exogenous information of the vector $\mathbf{I}_t$. Among the parametric approaches there is the Fixed Class Policy method that assumes the control policy belongs to an a-priori fixed class of functions, within which a particular policy can be selected by specifying a vector of parameters. In doing so, the Optimal Control (OC) problem is traced back to a Mathematical Programming problem. The idea at the base of this method is to assume some regularity on the shape of the control law, i.e. the control law belongs to a given class of functions (e.g. piecewise linear, polynomial, non-linear, ecc.). Generally the solution obtained by a Fixed Class Policy method is suboptimal, since it is decided 'a priori' the class of function to which the policy belongs.
Another approach able to exploiting the exogenous information belongs to the Learning ones. There is a specific algorithm named, *Q-Learning*, able to exploit the information included within the vector $\mathbf{I}_t$; it was originally developed in the branch of the Artificial Intelligence (AI) which goes under the name of Reinforcement Learning (RL), but it is also a relative of SDP. In fact it is solved by the recursive resolution of the Bellman Equation modified by the introduction of the Q-factor. Hence, the Bellman Equation belonging to the SDP problems has transformed from this previous formulation:

$$\mathbf{H}_t(\mathbf{s}_t) = \min_{u_t} \underset{\varepsilon_{t+1}}{E} [\mathbf{g}_t(s_t, u_t, \varepsilon_{t+1}) + \mathbf{H}_{t+1}(s_{t+1})] \tag{2.8}$$

where the Laplace criterion is applied and a finite horizon is considered; to this formulation:

$$\mathbf{H}_t(\mathbf{s}_t) = \min_{u_t} Q_t^*(x_t, u_t) \tag{2.9}$$

where

$$Q_t^*(\mathrm{x}_t, u_t) = \underset{\varepsilon_{t+1}}{E} [\mathbf{g}_t(\mathbf{s}_t, u_t, \varepsilon_{t+1}) + \mathbf{H}_{t+1}^*(s_{t+1})] \tag{2.10}$$

The function $Q_t^*(\mathrm{x}_t, u_t)$ is known in the literature as a Q-factor and provides, given x_t, the optimal cost-to-go at time t assuming that the control u_t is applied at the first step and an optimal policy is adopted in the following steps. Then the optimal policy is obtained by the solutions of this equation:

$$m_t^*(s_t) = \arg\min_{u_t} Q_{t+1}^*(s_{t+1}, u_{t+1}) \tag{2.11}$$

Q-Learning allows to include any information $\mathbf{I}_t$ into the controller as far as this information is observable, even if a model is not available . Conceptually, any additional information is considered as an augmented state component, like in SDP, however a model is not required for it.

In this thesis we will use a Fixed Class Policy method, see Section (3.2.4).

Chapter 3

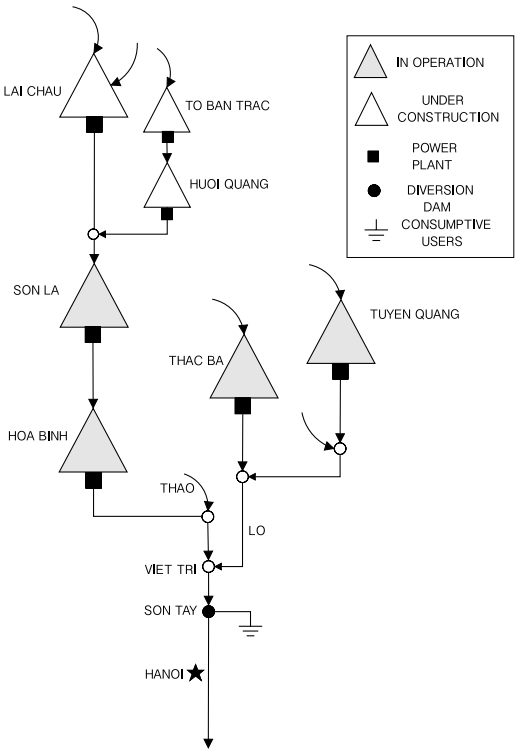
Case study: The Red River basin and the Hoabinh Reservoir

In this section the procedure, presented in the previous chapters, is applied to a real world case study: the Hoabinh reservoir in the Red River basin, Vietnam. This chapter investigates the use of exogenous information for the design of the optimal operation of the Hoabinh. Here follows a description of the study area, an analysis of hydrologic and climatic variables available, and finally the results from the application of the procedure described in Chapter 2 are analyzed.

Study Area

The Red River basin is the second largest river in Vietnam. It is located between 20^o00 N and 25^o30 N, and 100^o00 E and 107^o10 E. The total area of the basin is approximately 169,000 km^2, of which 81,240 km^2 (48%) in China, 86,600 km^2 (51.35%) in Vietnam, and the rest in Laos. Administratively, the Red River basin covers 26 provinces and cities in the northern region of Vietnam, with a total population of about 28 million.

The Red River, the main River downstream from Viettri, originates at the confluence point of three upstream tributaries: Da, Thao and Lo River (Figure 3.1). All these tributaries originate from China. Even though the catchment areas of the Da and Thao River



(a)

(b)

Figure 3.1: The Red River system. a) Schematic model of all the components of the system and b) Geographic map of the system.

basins are almost the same, the Da River contributes 42%, while the Thao River contributes only 19% of total flow to the Red River. The Lo River basin is the smallest one, but its contribution is 25.4%.

The whole basin is characterized by two distinguished seasons: rainy season from May to October and dry season from November to April. Annual rainfall varies from from 1,200 to 4,800 mm/year in Vietnam part, and about 80% of rainfall occurs in the rainy season. Because of uneven rainfall, flows through the basin are unevenly distributed in time, causing floods and water-logging in the rainy season and water shortages in the dry season. Among water sectors, irrigation is the biggest user, accounting for 90% of total used water. The agricultural land occupies approximately 1,874,100 ha and forestry land occupies 2,570,775 ha. Potential area for future agriculture and forestry development is estimated at about 3,919,500 ha, with an associated increasing of potential demand for the future.

Several reservoirs have been built and operated since 1970s. Thacba reservoir (Figure 3.1) is located on Chay River, a tributary of the Lo, starting its regulation in 1971. The main objective of this reservoir is hydropower generation. However, it increases the flow in the dry season from about 100 to 200 m^3/s. The Tuyenquang reservoir (Figure 3.1) on the Gam River (belonged to the Lo) and the SonLa reservoir (Figure 3.1), upstream of Hoabinh reservoir on the Da River, are under construction. Up to now, Hoabinh reservoir is the biggest one of Vietnam. So far it has been playing an important role in preventing and controlling flood, generating hydro electricity, and supplying water to irrigation, domestic use, industry, and other water users of the Red River Basin. The Hoabinh hydropower plant, at present covers approximately 15% of the electric production in the country. The hydropower plant has eight turbines with total design capacity of 1920 MW (maximum of 240 MW per turbine). Maximum release through the eight turbines reaches 2360 m^3/s. There are 12 bottom gates, each of size 10x6 m, located at the elevation of 56 m with possible maximum release of about 21,996 m^3/s. Six spillways with size of 15x15 m start from elevation of 102 m, allowing a possible maximum release of 14,100 m^3/s. Beside hydropower production, the main purposes of the reservoir operation are flood mitigation and water supply to the Red River Delta.

Due to rapid growth of population, quick development of the economy, and climate change, the basin has been facing many problems such as severe floods, water shortages, water pollution, and so on. The following is a brief description of the main problems

affecting this area.

Floods

In recent years, big floods have frequently plagued Vietnam and, in particular, the Red River Basin (see Table 3.1).

Table 3.1: Historical flood events at the main stations in the Red River Basin.

Station	River	Average flood flows (m^3/s)	Peak flows (m^3/s)	Date
Laichau	Da	7,242	14,2	Aug-1932
Sontay	Red	16,785	37,8	Aug-1971
Yenbai	Thao	5,143	10,3	Aug-1971
Chiemhoa	Gam	3,188	6,2	Aug-1971
Tuyenquang	Lo	5,156	11,7	Aug-1971
Vuquang	Lo	5,467	14	Aug-1971
Hamyen	Lo	2,897	5,7	Jul-1986
Hoabinh	Da	9,618	22,7	Aug-1996
Tabu	Da	9,919	22,7	Aug-1996

From Table 3.1 it is possible to see that the most of the big floods occurred in August and so that they are driven by the monsoon. In fact, the total rainfall in the rainy season accounts for nearly the 80% of yearly rainfall, especially, rainfall on August accounts for 21,5% of the yearly amount.

In general, the Red River floods are combinations of floods from the three upstream tributaries: Da, Thao and Lo River, and the percentage of simultaneous occurrence is rather high (72%). Flood of the Da River is the main source causing big floods of the Red River. On average, it contributes 49% of 8 day flood water (max is 69%) at Sontay. If only the flood peak is considered, there are 69% of flood peaks of the Da River constituting the flood peaks of the Red River. The Lo River flood is the second biggest source of Red River floods. One interesting note is that floods of the Lo and Da Rivers often coincided,

and there are about 34% of chance to form the big flood of the Red River.

Hydropower
So far hydropower has been the main energy source of Vietnam, in fact it contributes from 35% to 40% of total consumed energy . Hence it is important to guarantee the right amount of water for the hydropower production over all the year.

The Hoabinh reservoir is also used as a water supply for the irrigation district located downstream. Therefore it is importante that the implemented management ensures a sufficient amount of water to meet the irrigation demand especially in the dry season. In the PhD thesis of Quach (2011) it has been demonstrated that the water deficit is negligible for most of the year. So, in this work, it has not been considered as an objective to optimize.

3.1 Modeling the Hoabinh water system

Usually, the management of a reservoir must meet multiple objectives, often belonging to different stakeholders with conflicting interests. This is the case even for the Hoabinh. The operation of the Hoabinh must balance flood control, hydropower production and water supply for irrigation. As anticipated we will concentrate on the first two objectives and we won't consider the water supply.
For implementing an optimal control policy, it is necessary to define the structure of the optimal control problem and this one is composed by several parameters that must be specified; they are: the objectives, the evaluation horizon, the decision time step and the model of the system.

Objectives and indicators

The objectives considered in this study are the minimization of flood events and the maximization of the hydropower generation. The optimization of only 2 objectives, instead of 3, is considered because our target is not to find an optimal management policy for this reservoir (for a complete discussion about this topic see Quach, 2011), but to

demonstrate that the inclusion of exogenous variables in the structure of the final policy itself improves its performances. The procedure and the results shown here are also applicable in case of studies more complex and with more than two objectives.
The objectives are modeled through physical indicators that quantify the evaluation criteria the relevant stakeholders adopt in judging and comparing alternative operating policies. Indicators are expressed as the aggregation, through an appropriate operator, of step indicators (see Soncini-Sessa, 2007).

Flood

The flood mitigation is evaluated through an indicator measuring the water level in Hanoi. Prolonged periods of high water levels at the Hanoi Station correspond to high risk of dike break (Vorogushyn and Apel, 2010), and consequently high potential damage. For the sake of simplicity, a proxy indicator is used in designing policies, that is the positive difference between the water level and a threshold given value of 9,5 m that is the 1^{st} alarm flood level (information from Hansson and Ekenberg, 2002). Moreover, due to the high frequency and magnitude of flood in August (see Table 3.1), in this month a higher important weight is assigned:

$$g_{t+1}^{flo} = \begin{cases} 0 & if\, h_{t+1}^{Hn} \leq \bar{h} \\ \delta_t (h_{t+1}^{Hn} - \bar{h})^m & otherwise \end{cases} \tag{3.1}$$

where g^{flo} is the step cost of flood, h^{Hn} is the water level in Hanoi station and $\bar{h}$ is the flood threshold. δ_t is the seasonal coefficient (equals 2 in August and 1 otherwise), and m is a coefficient reflecting risk aversion, here assumed equal to 2. The objective considered in the optimization is:

$$J^f = \frac{1}{h} \sum_{t=0}^{h-1} [\max\{0, (h_{t+1}^{Hn} - \bar{h})\}]^2 \tag{3.2}$$

where h is the evaluation horizon.

Hydropower

Hydropower generation is a function of the daily energy production, P_{t+1}, defined as:

$$P_{t+1} = \varphi g \gamma \eta (H_{t+1}) r_{t+1}^t H_{t+1} \tag{3.3}$$

where φ is a coefficient of dimensional conversion, g is the gravitational acceleration (equal to 9,81 m/s^2), γ is water density (equal to 1000 kg/m^3), η is turbine efficiency, which is a function of the hydraulic head H_{t+1}, r^t_{t+1} is release through turbines. The hydraulic head is the difference between the water level upstream and downstream, i.e.

$$H_{t+1} = h^{up}_{t+1} - h^{do}_{t+1} \tag{3.4}$$

Finally, to formulate the immediate cost, the daily energy production (see Equation 3.5) is filtered by a time-varying dimensionless coefficient α_t, to taking account for the seasonal variability, i.e.

$$g^{hyd}_{t+1} = -\alpha_t P_{t+1} \tag{3.5}$$

where α_t is assumed equal to 2 two from April to June and 1 in the other months. The negative values of the production are considered, in this way the indicator is formulated as a cost to be minimized. Hence to maximize the hydropower generation the measure of this step cost must decrease. Then, the objective related to the hydropower energy issue, used in the optimization stage, is the averages of the previous step indicators over the whole evaluation horizon:

$$J^h = -(\frac{1}{h}\sum_{t=0}^{h-1} \alpha_t P_{t+1}) \tag{3.6}$$

Evaluation Horizon

The evaluation horizon considered for the Hoabinh reservoir covers the period 1994-2005. We start from 1994 because even if the Hoabinh reservoir was completed in 1989, it is fully operative since 1994.

Decision Time Step

Decision time step, t, is the time between one decision and the next.
Since the historical operation of the Hoabinh reservoir is based on a 1 day decision step, we decide to use this same step. To justify this decision we analyze the translation time

of the whole Red River system, i.e. to estimate times of water conveying from the most upstream points to the downstream points (in this case the downstream point considered is the city of Hanoi). This can be done by studying the cross-correlogram of the whitened series in question. Figure (3.2) suggests that the translation time between the upstream stations to the downstream one is approximately one day, i.e. the output of the next day is highly dependent upon inputs of previous day.

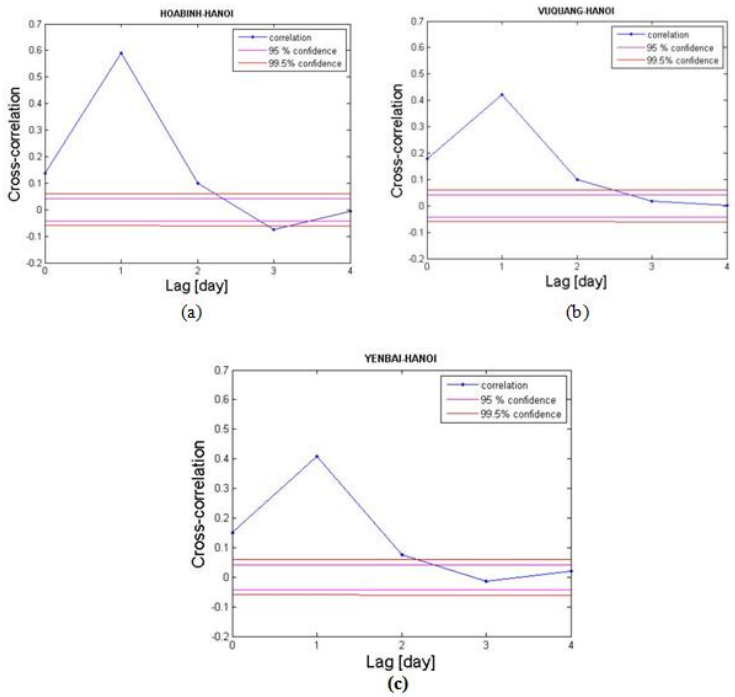


Figure 3.2: Translation time from upstream stations to downstream station. Stations are located: a)on the Da river, b)on the Thao river, c)on the Lo river.

Therefore, for the study purpose, the decision step can be assumed as constant and equal to a day.

Model of the Hoabinh system

The HoaBinh reservoir is an artificial reservoir with a storage capacity of 9.8 billion m^3 and an active storage of 6 billion m^3, corresponding to a level operational range of 37 m. The reservoir dynamics is modeled by daily mass balance equation considering

inflow from the Da River catchment, evaporation and release:

$$s_{t+1} = s_t + q_{t+1}^{\mathrm{HB}} - e_{t+1} S(s_t) - r_{t+1} \tag{3.7}$$

where s_t is the storage on day t, q_{t+1}^{HB} is the inflow to the HoaBinh reservoir (i.e. outflow of the Da catchment); e_{t+1} is the unitary surface evaporation (which follows a yearly pattern); $S(\cdot)$ is the reservoir surface computed as a function of the storage; and r_{t+1} is the release. The actual release r_{t+1} coincides with the release decision u_t only if the latter is feasible, i.e. included between the minimum and maximum feasible release that can be obtained when all the gates are completely closed or open, respectively. Such values are computed by integration of the continuous-time mass balance equation using the instantaneous minimum and maximum stage-discharge relation (see Castelletti, 2008) as given by the rating curves of the turbines, bottom gates, and spillways.

3.2 Application of the Procedure to the Hoabinh case study

In this section the application of the procedure 4 steps, presented in the Chapter 2, is shown.

3.2.1 First Step - Deterministic Optimization

As a first step it is necessary to evaluate if there is an effective space for improvement from the solutions offered by the traditional stochastic optimal control problem. To assess if this space actually exists a simulation experiment assuming perfect information system, that is, full knowledge of all future flows from the upper Da River and the tributaries Lo and Thao must be run. The associated upper bound of performances can be derived by solving a deterministic optimal control problem, i.e. finding the trajectory of release decisions (release scheduling) $\mathbf{u}^* = |u_0^*, u_1^*, \ldots, u_{h-1}^*|$ that minimizes the average aggregate cost under historical flow pattern of the Da, Thao and Lo River. The (single-objective) deterministic control problem to be solved, is

$$\min_{\mathbf{u}} \left(\lambda_1 \frac{1}{h} \sum_{t=0}^{h-1} g_{t+1}^{\mathrm{hyd}} + \lambda_2 \frac{1}{h} \sum_{t=0}^{h-1} g_{t+1}^{\mathrm{flo}} \right) \tag{3.8}$$

where $t = 0$ and $t = h-1$ are the first and last day in the optimization horizon ($1^s t$ January 1994 - $31^s t$ December 2005); g_{t+1}^{hyd} and g_{t+1}^{flo} are the immediate costs defined in Sect. 3.1,

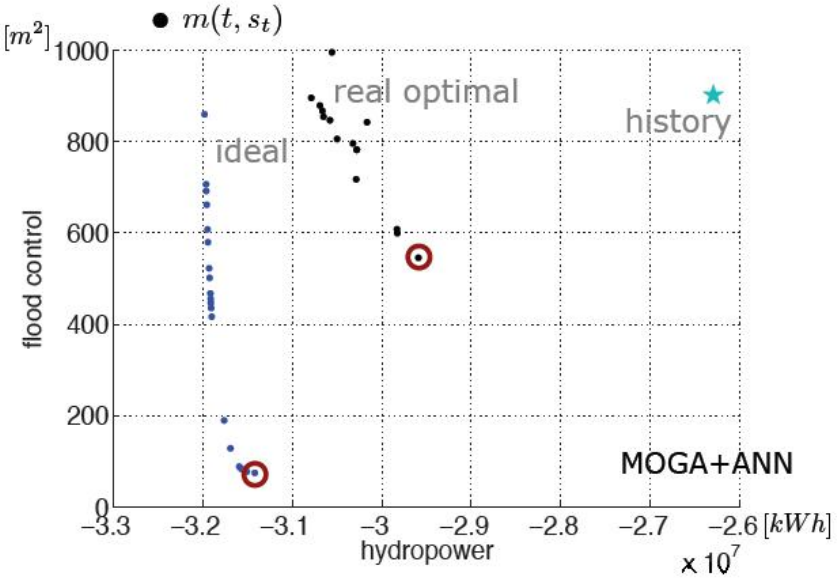


Figure 3.3: Pareto Frontiers for the horizon 1994-2005. The blue one is the solutions of a DDP problem; the black one is the solution of MOGA-ANN problem; the star indicates the historical operating rule.

whose value is computed as a function of the release scheduling $\mathbf{u}$ by simulation of the model described in Sect. 3.1; and λ_1, λ_2 are the aggregation weights. For a given combination of weights, the associated single-objective problem (Eq. 3.8) can be solved by Deterministic Dynamic Programming (DDP). By changing the weight values, different tradeoffs between the objectives are defined and the Pareto-optimal solutions are found, as shown in Figure (3.3).

Figure (3.3) shows that there is an effective space for improvements from the ideal frontier to the real optimal one. The latter was formulated by the application of a Multi-Objective Genetic Algorithm (MOGA) over an Artificial Neural Network (ANN) function family. This means that an ANN function family was selected for the operating rule (since it guarantee high flexibility and low complexity) and a MOGA was applied to determine the function parameters that minimize the average value of the immediate costs shown in Eq. (3.1) and (3.5). As shown in Figure (3.3) the release decision deriving from the black frontier is defined as a function of time and storage:

$$u_t = m(t, s_t) \tag{3.9}$$

We will concentrate on the possible improvement we can reach on the flood control

objective. In Figure (3.3) the points referred to an operating rule exclusively based on the optimization of the flood objective are circled in red. Indeed these points represent the solution of the DDP problem and of the MOGA-ANN problem respectively, where only the flood control optimization is considered.

3.2.2 Second Step - Candidate Variable Selection

The second step is to identify all the exogenous hydro-meteorological variables that, potentially, could improve the operating policy by giving more information about the physical functioning of the system. As introduced in Chapter 1 we must find the most valuable information, able to play as a surrogate of the perfect knowledge of the future inflow to the reservoir, to fulfill the input vector I_t, belonging to the equation:

$$u_t = m(t, s_t, I_t) \tag{3.10}$$

where the subscript t of I_t means that the information contained into this vector, are available at time t when the decision is made. Hence, all the available data for the specific system must be identified. However before proceeding towards collecting all the available data for the reservoir under exam, it is important to analyze the final goal of our management problem; this means to analyze the objectives that must be optimized and their dynamics. Indeed, as anticipated in Section (2.2), before collecting the variables, we must discriminate between the high frequency ones and the low frequency ones and choose regarding to our objectives. Since the objective, that we decided to optimize in this thesis, is the flood control, we concentrated on the high frequency variables like the precipitation data.

Before choosing what variables should be used we present a brief description of all the available data for the case of study.

Along the Da River basin are placed 18 meteorological stations and 4 hydrological stations (Figure 3.4), along the Thao River basin there are 3 meteorological stations and 1 hydrological stations (Figure 3.5), and, finally, along the Lo River basin there are 3 meteorological stations and 1 hydrological station (Figure 3.5).

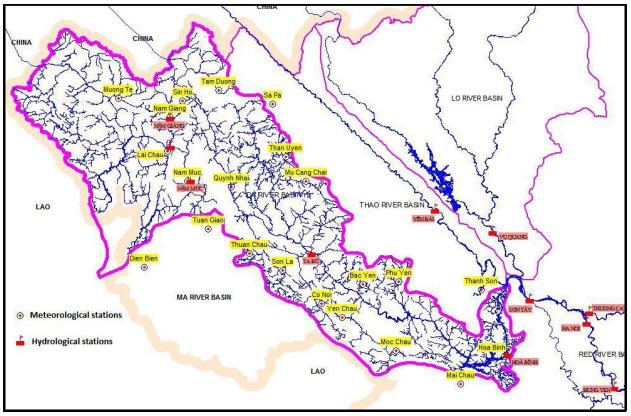


Figure 3.4: Da River Basin and location of available stations.

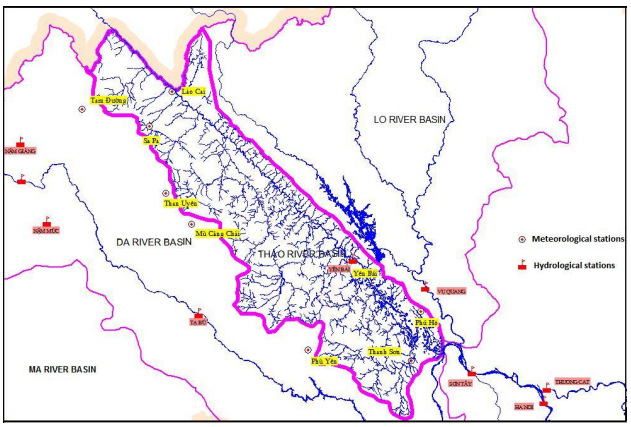


Figure 3.5: Thao River Basin and location of available stations.

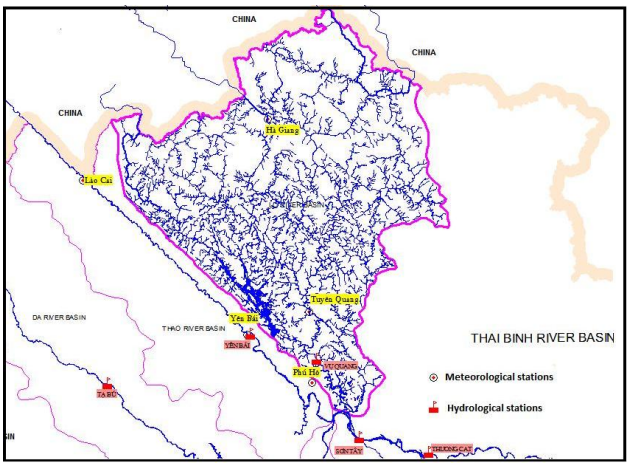


Figure 3.6: Lo River Basin and location of available stations.

At the Hoabinh reservoir, the daily historical time series were recorded from 1st January, 1989, including instantaneous water levels (h_t^{up}, h_t^{do}) at time t upstream and downstream of the dam, and interval releases $(r_{t+1}^t, r_{t+1}^b, r_{t+1}^s)$ from t to $t+1$ through turbines, bottom gates and spillways, respectively. So the total release of the reservoir r_{t+1} is the sum of all the above component releases. Other three related gauging stations are selected. They are Yenbai station on the Thao River, Vuquang station on the Lo River and Hanoi station on the Red River. Water levels at these stations are usually measured twice per day, at 7 o'clock in the morning and 19 o'clock in the afternoon. Here, only the first measurement is used as the level of the day in question. All these daily time series are available from 1958.

The collected information is from different sources in Vietnam, but the main source is in the Hydrological Division - Institute of Water Resources Planning. A serious problem is the lack of information on China's catchment area. So the data mentioned in this thesis are only data in Vietnam's catchment part.

The data available are shown in Table (3.2).

3.2.3 Third Step - Input Variable Selection

From the previous section it is possible to see that there are a lot of exogenous information available for this reservoir, thanks to the application of the Input Variable Selection (IVS) procedure is now possible to define who are the most valuable variables among all the previous mentioned.

Generally, in a multi-objective context the final selection of $\mathbf{I}_t$ depends upon the relative importance of each objective in each ideal optimal policy which can be found by DDP, i.e. result of IVS are different from one intended policy to the other. Indeed the IVS chooses the variables that better explain the sequence of ideal release decisions defined by the resolution of the DDP problem, and each of these sequences, is created by applying a different combination of weights to the different objectives that the model must optimize. For instance, an extreme reservoir operating policy which favors hydropower production tends to keep the reservoir water level as high as possible and vice verse for a policy which prioritizes upstream flood mitigation. As anticipated, in this thesis to run the IVS algorithm was used a control policy which favors flood mitigation instead of the

Table 3.2: Available data in Red River system up to Hanoi

No	Stations	Time step	Elements	Observation time		Catchment	
1	Hoa binh	daily	Precipitation	1957	2004	Da river	
			Discharge	1956	2004	Da river	
			Temperature	1961	2007	Da river	
			Water level	1956	2004	Da river	
2	Thanh son	daily	Precipitation	1960	2004		Thao river
3	Mai chau	daily	Precipitation	1961	2004	Da river	
4	Moc chau	daily	Precipitation	1961	2004	Da river	
			Evaporation	1982	2002	Da river	
			Temperature	1961	2006	Da river	
5	Yen Chau	daily	Precipitation	1961	2004	Da river	
6	Phu yen	daily	Precipitation	1961	2004	Da river	
			Evaporation	1982	2002	Da river	
7	Bac yen	daily	Precipitation	1973	2004	Da river	
			Evaporation	1982	2002	Da river	
8	Co noi	daily	Precipitation	1964	2004	Da river	
			Evaporation	1982	2002	Da river	
9	Son la	daily	Precipitation	1961	2004	Da river	
			Evaporation	1965	2002	Da river	
			Temperature	1961	2006	Da river	
10	Thuan chau	daily	Precipitation	1960	2004	Da river	
11	Mu cang chai	daily	Precipitation				
			Evaporation	1982	2002	Da river	
			Temperature	1961	2006	Da river	
12	Quynh nhai	daily	Precipitation	1960	2004	Da river	
13	Tuan giao	daily	Precipitation	1958	2004		Ma river
			Evaporation	1982	2002		Ma river
			Temperature	1961	2006		Ma river
14	Than uyen	daily	Precipitation	1961	2004	Da river	
			Evaporation	1961	2002	Da river	
15	Nam muc	daily	Precipitation	1964	2004	Da river	
			Discharge	1960	2004	Da river	
16	Lai chau	daily	Precipitation	1957	2004	Da river	
			Discharge	1957	2004	Da river	
			Evaporation	1982	2002	Da river	
			Temperature	1961	2006	Da river	
17	Nam giang	daily	Precipitation	1974	2004	Da river	
			Discharge	1965	2004	Da river	
18	Sa pa	daily	Precipitation	1961	2004	Da river	
19	Dien bien	daily	Precipitation	1963	2004		Ma river
20	Sin ho	daily	Precipitation	1961	2004	Da river	
			Evaporation	1982	2002	Da river	
			Temperature	1961	2006	Da river	
21	Tam duong	daily	Precipitation	1970	2004	Da river	
			Evaporation	1982	2002	Da river	
22	Muong te	daily	Precipitation	1961	2004	Da river	
			Evaporation	1982	2002	Da river	
			Temperature	1961	2006	Da river	
23	Pha din	daily	Evaporation	1982	2002	Da river	
24	Yen Bai	daily	Precipitation				Thao river
			Discharge	1956	2004		Thao river
			Water level	1956	2004		Thao river
25	Lao Cai		Precipitation				Thao river
26	Vu Quang	daily	Precipitation				Lo river
			Discharge	1956	2004		Lo river
			Water level	1956	2004		Lo river
27	Phu Ho		Precipitation	1960	2004		Lo river
28	Tuyen Quang		Precipitation	1960	2004		Lo river
29	Ha Giang		Precipitation	1957	2004		Lo river
30	Son Tay	daily	Discharge	1956	2004		Red river
			Water level				Red river
31	Ha Noi	daily	Discharge	1956	2004		Red river
			Water level				Red river
32	Hung Yen	daiy	Water level	1956	2004		Red river
33	Thuong Cat	daily	Discharge	1956	2004		Duong river

hydropower production.
In Section (2.3) the formulation of the IVS algorithm has been explained. Basically given a vector of output (a sequence of release decisions) associated to an optimal policy resulted from applying DDP, and an in-time corresponding set $\mathbf{I}_t$ of related exogenous variables, the target of IVS is to find the most relevant inputs that explain the output by using the Iterative Input Selection (IIS) algorithm.
There are a few parameters that must a priori fixed on the basis of the problem specifics, and by empirical or trail-and-errors evaluations: n_{min}, $ScoreTh$, the number of trees M, the n_{folds} number and the maximum number p of the highest rank variables we want to be shown. We made a few trials to train the IIS algorithm and the best results are showed hereafter. For every input set we run the algorithm with this set of parameters:

N_{folds}	M	p
10	200	5

And with these three different termination criteria:

Termination Criteria		
1)	$n_{min} = 2$	$Tscore = 0,98$
2)	$n_{min} = 100$	$Tscore = 0,98$
3)	$n_{min} = 75$	$Tscore = 0$

We use three different Termination Criteria to understand what combination can actually guarantee the best performances. The results show that the better one is the first: $n_{min} = 2$, $Tscore = 0,98$. The graphic in Figure (3.7) demonstrates that the first combination of termination criteria is the one resulting in better performance for every input set.

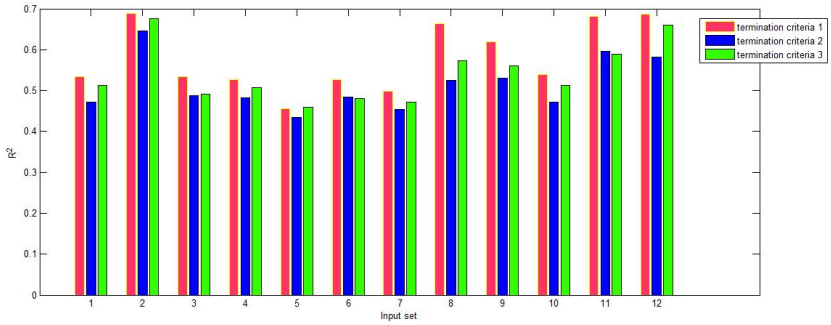


Figure 3.7: Performances of the 3 different combination of termination criteria.

All the data sets processed by the IIS algorithm, are generically composed in a matrix of the form:

$$F = \{< \mathbf{s}_t, \mathbf{t}, \tilde{\mathbf{I}}_t, \mathbf{u}_t >, t = 0, ..., h\} \tag{3.11}$$

where the variable $\mathbf{u}_t$ is the sequence of optimal release decisions obtained through the DDP and here represents the output that must be modeled, and $\tilde{\mathbf{I}}_t$ represents the set of all the collected variables we want to analyze. To this data-set we applied a random shuffle function in a way to mix randomly all the rows of that matrix. After applying this function the performances of the algorithm (measured by the coefficient R^2) improved significantly. Indeed by randomly mixing the matrix F, we eliminate the temporal correlation, and so the process can be assumed as a Marcov process, i.e. each row is assumed independent from the previous and from the subsequent row. After running the IIS, the set of the variables that better explains the optimal sequence of release decisions, is obtained, as shown schematically in Figure (3.8).

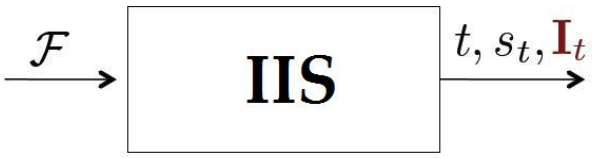


Figure 3.8: Generic scheme of the IIS algorithm functioning.

The value of the vector $\mathbf{I}_t$ of Eq. (3.10) has been proved in several trials, with different initial sets of the available variables. Every set differs from the other in terms of dimen-

sion, of included variables and of the step-time index in which they were considered. In these input vectors $\tilde{\mathbf{I}}_t$ we tried to use different associations of hydrologic (punctual or areal rainfall data, soil evapotranspiration and flow from different stations in the basin) and meteorological (temperature) information.

Between all the trial input sets tested, only the one with the highest coefficient of determination, R^2, is presented. This input set contains 30 variables, as shown in Table (3.3), and it is composed by storage values, time values, 3 inflows values and 25 rainfall data. The rainfall data are the daily, punctual measurement in all the stations on the Da, Thao and Lo rivers considered at the time instant $t-1$, $t-2$ (or in the time interval $[t-1;t)$, $[t-2;t-1)$).

Table 3.3: Variables contained into the Input vector.

N. and Name			
0	Time	15	p_{t-2}^{TU}
1	Storage	16	p_{t-2}^{SH}
2	q_t^{HB}	17	p_{t-2}^{SL}
3	q_t^{YB}	18	p_{t-2}^{LH}
4	q_t^{VQ}	19	p_{t-2}^{PY}
5	p_{t-1}^{MT}	20	p_{t-2}^{TC}
6	p_{t-1}^{SH}	21	p_{t-2}^{YC}
7	p_{t-1}^{TU}	22	p_{t-1}^{HG}
8	p_{t-1}^{SL}	23	p_{t-1}^{L2}
9	p_{t-1}^{LC}	24	p_{t-1}^{T1}
10	p_{t-1}^{PY}	25	p_{t-1}^{LC}
11	p_{t-1}^{MO}	26	p_{t-2}^{HG}
12	p_{t-1}^{TC}	27	p_{t-2}^{L2}
13	p_{t-1}^{YC}	28	p_{t-2}^{T1}
14	p_{t-2}^{MT}	29	p_{t-2}^{LC}

For a complete explanation about the abbreviations, see Appendix 1.
The result of running the IIS algorithm on this data set are resumed in Table (3.5) and in Figure (3.9).

Table 3.4: Result of the IIS algorithm on the candidate input vector.

Variable	R^2	Increase R^2
Storage	0,3154	31,5%
q_t^{HB}	0,3935	7,8%
Time	0,4666	7,3%
p_{t-2}^{LC}	0,4964	2,9%

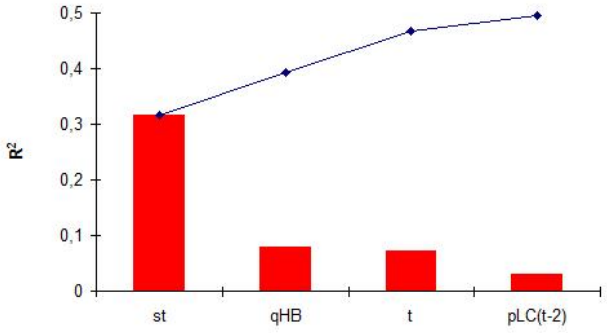


Figure 3.9: The variables selected by running the IIS algorithm on the candidate input data-set with the associated relative contribution to the overall performance of the underlying tree-based model.

The IIS algorithm on this data set reaches an explained variance of almost $R^2 = 0,5$. Unsurprisingly, most of the release signal is explained by the storage. However, the cumulative effect of the other three variables contributes for another 18% to the underlying model performance.
Considering the results shown in the previous Table, the input vector will be:

$$\mathbf{I}_t = (\mathbf{q}_t^{HB}, \mathbf{p}_{t-2}^{LC}) \tag{3.12}$$

With this value of $\mathbf{I}_t$ we can now proceed towards the fourth step of our procedure, the optimization, to create an operating policy able to exploit the value of this selected

hydroclimatic information.

3.2.4 Fourth Step - Optimization

Once the hydroclimatic information vector $\mathbf{I}_t$ has been selected, see Eq. (3.12), the next step in our procedure is the identification of an optimal operating rule conditioned upon this information, i.e.

$$\mathbf{u}_t = m(\mathbf{t}, \mathbf{s}_t, \mathbf{q}_t^{HB}, \mathbf{p}_{t-2}^{LC}) \tag{3.13}$$

To identify the operating rule in the infinite-dimensional space of functions $\mathbf{m}(\cdot)$ a stochastic optimization problem must be solved. Stochastic Dynamic Programming is by far the most widely used method to solve such a problem. However we already described that its application is subject to the limitations imposed by the so called curse of modelling and curse of dimensionality (see Section (1.1.1)). To overcome this problem in this work we used a parametrization-simulation-optimization approach. First, a prescribed function family $\hat{\mathbf{m}}(\cdot)$ is selected for the operating rule of Eq. (3.13) and then, the optimal parametrization θ^* is identified. In other terms, the stochastic optimization problem is traced back to the following non-linear programming problem:

$$\min_{\theta} J(\mathbf{x}_0, \theta, \mathbf{q}) \qquad \text{s.t. } (\mathbf{x}_0, \mathbf{q}) \text{ given and } \mathbf{u}_t = \hat{\mathbf{m}}(\mathbf{x}_t, t, \mathbf{I}_t; \theta) \tag{3.14}$$

In this study, we use an ANN as family function and we solve problem (3.14) by the use of a Multi-Objective Genetic Algorithm (MOGA), (for more information see Castelletti, 2008). The optimization aim is to find an appropriate set of parameters θ^* that minimize the average value of the immediate costs of Eq. (3.1) and Eq. (3.5). A set of these parameters constitute an "individual" in the MOGA. MOGA starts from a randomly selected population of N "individuals". The "fitness" (average value of the immediate costs) of each individual is tested by simulation of the system under historical flows of the upper Da, Thao and Lo River and the operating policy defined by the parametrization under exam. Then, a new population is generated by selection, crossover and mutation, and the process is repeated for a prescribed number of iterations. In this study, selection, crossover and mutation are performed according to the Non-dominated Sorting Genetic Algorithm NSGA II (see Deb, 2002).

To make a fair comparison with historical operation, in the optimization process the system simulation uses historical data over the period 1961-1978 (optimization horizon) and the final population is then re-simulated over the period 1994-2005 (evaluation horizon). The results obtained by the simulation of the operating rule of Eq. (3.13) to the Hoabinh reservoir over the horizon 1994-2005, are shown in the following graphic:

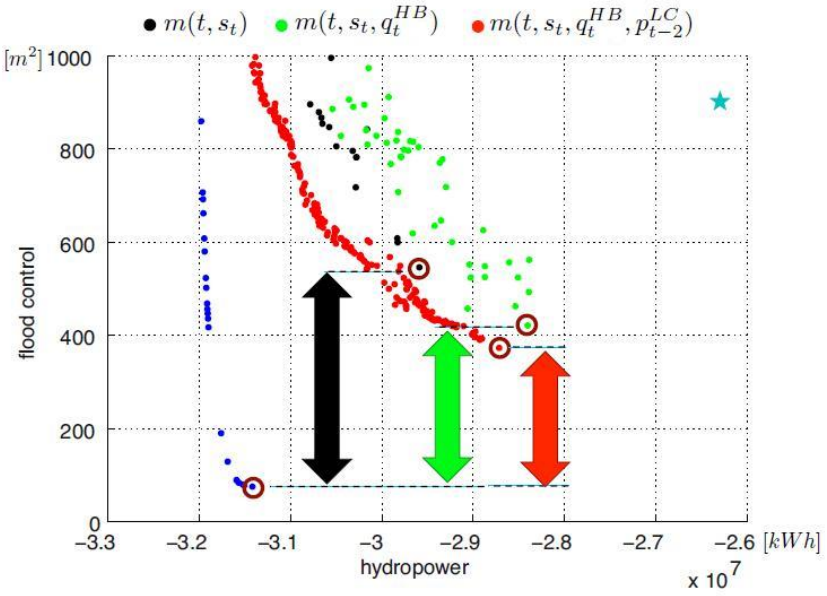


Figure 3.10: Pareto Frontiers for the horizon 1994-2005. The distance measurement between the different Frontiers performance is based on the points circled in red, they are the points representing the optimization of only the flood objective.

From Figure (3.10) it is possible to see that there is an actual improvement of about 20% in including the selected hydroclimatic information contained in Eq. (3.12). This means that the exploiting of some relevant hydroclimatic variables can improve the management of this reservoir. To analyze deeply the obtained result we studied the sequence of water level measured in the Hoabinh reservoir, simulated with the different operating policy, and the peak-flow event of August 1996, shown respectively in Figure (3.11) and (3.12).

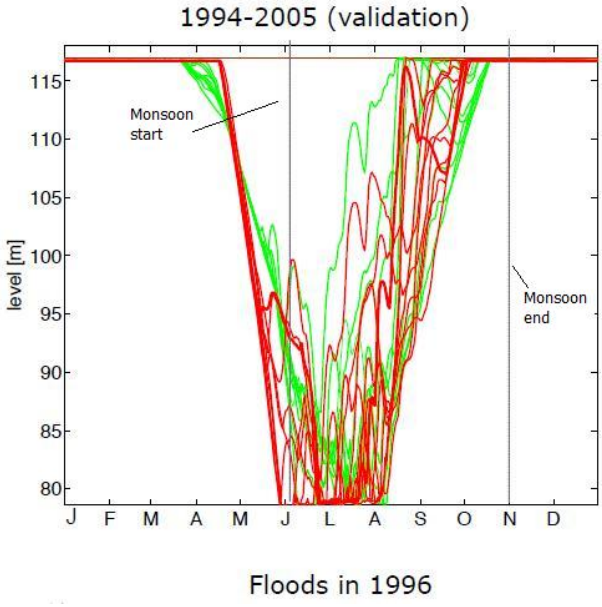


Figure 3.11: The Hoabinh water level produced by the operating policy exploiting the hydroclimatic information over the evaluation horizon 1994-2005. The red one is the operating policy exploiting the $\mathbf{I}_t$ vector.

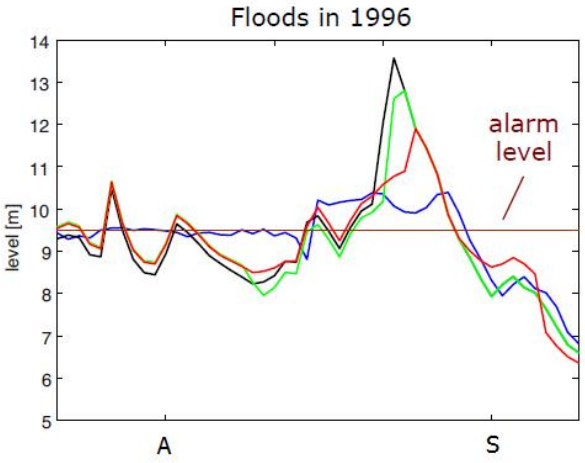


Figure 3.12: Details of the peak-flow event occurred in August 1996, water level measured in the Hanoi station.The red one is the operating policy exploiting the $\mathbf{I}_t$ vector, the blue one is the operating policy obtained with DDP, the black one is the historical operation and the green one is the operating policy exploiting only one variables of the $\mathbf{I}_t$ vector, q_t^{HB}

From Figure (3.12) it is clear that there is an improvement in the reservoir management by using the operating policy with the hydroclimatic information. Indeed it is shown

that the operating policy with the $\mathbf{I}_t$ vector (the red one), is able to reduce the peak-flow water level measured in Hanoi, during the flood of August 1996.

3.3 Further improvement: Large-Scale Atmospheric Circulation Phenomena

Recently, a lot of studies have been based on the analysis of some climatologic indicators able to explain the inter-annual variability in some areas of the world and able to improve the future forecasts of the hydro-climatologic conditions (see e.g. Makkearson (2008) and Maity, (2006)). Indeed, nowadays, it is well recognized that the time series of hydrologic variables, such as rainfall and inflow are significantly influenced by various large-scale atmospheric circulation patterns measured with these climatologic indices. Furthermore, in literature, is everyday more studied the possibility to improve basin-scale inflow forecast using the information of large-scale atmospheric circulation phenomena such as the ENSO indices. For these reason, to further improve our analysis, we decide to test if the inclusion of these atmospheric circulation phenomena will improve the results we obtained.

Large-scale Atmospheric Circulation Phenomena

Almost all of these climatologic indicators are related to the phenomenon known as El-Niño Southern Oscillation (ENSO). This phenomenon is a large-scale, coupled ocean-atmospheric, process and it is currently considered as one of the most significative factors influencing the hydro-climatic global variability (for more information see Kahya and Dracup (1993) or Allan (2000)). The so called El-Niño is a large-scale anomalous warming of sea surface temperature (SST) over the central and eastern Pacific Ocean with associated change in pressure field. In normal years, SST of the western part of the equatorial Pacific Ocean remains warmer than that of the eastern part, and, pressure at the eastern part of the Pacific Ocean is higher than that of the western part. During anomalous years, SST of the eastern part of the equatorial Pacific Ocean becomes warmer-than-normal and the pressure field is reversed, i.e. the anomalous pressure in the eastern part of the Pacific

becomes lower than that in the western part. Instead, anomalous cooling of the SST over the eastern part of the Pacific is known as La-Niña. Whereas the anomalous sea-saw variation of the pressure field, between the eastern and the western parts of this Ocean, is the so called Southern Oscillation as discussed by Maity (2006). The El-Niño is an almost periodic phenomenon; i.e. it is repeating over time, following an irregular period of 4 to 5 years.

There is an increasing number of studies investigating the relationship between inflow and ENSO. In literature, indeed, it is every day more solid the hypotheses that the relation between the ENSO phenomenon and inflow is stronger than the relation between ENSO and the precipitation data, since the variability of rainfall is reflected within the hydrologic runoff process and, moreover, inflow integrates spatial in addition to temporal information; for more information see Chiew (1998).

To quantify and to analyze the ENSO phenomenon a few different indicators can be used:

- The Southern Oscillation Index (SOI)
- The Sea Surface Temperature (SST)
- The Multivariate ENSO index (MEI)

The SOI index measures the strenght of the Southern Oscillation and it is the most used indicator to study the ENSO phenomenon. The SOI is computed from fluctuations in the surface air pressure difference between Tahiti and Darwin, Australia. El-Niño episodes are associated with negative values of the SOI, meaning that the pressure difference between Tahiti and Darwin is relatively small.

There are a few different methods of how to calculate the SOI. The method used by the Australian Bureau of Meteorology is the Troup SOI (Troup, 1965) which is the standardised anomaly of the Mean Sea Level Pressure (MSLP) difference between Tahiti and Darwin. It is calculated as follows:

$$SOI = 10\frac{[Pdiff - Pdiffav]}{SD(Pdiff)} \tag{3.15}$$

where

$Pdiff$ = (average Tahiti MSLP for the month) - (average Darwin MSLP for the month),

$Pdiffav$ = long term average of $Pdiff$ for the month in question, and
$SD(Pdiff)$ = long term standard deviation of $Pdiff$ for the month in question.
The multiplication by 10 is a convention. The SOI is usually computed on a monthly basis. Daily or weekly values of the SOI do not convey much in the way of useful information about the current state of the climate, and accordingly the Bureau of Meteorology does not issue them. All this information and the SOI monthly data from 1876 until now, are available on the Australian Bureau of meteorology.
The MEI index is the latest indicator created to analyze the ENSO phenomenon. It is computed from surface marine data filtered through spatial cluster analysis and based on six different observational fields: sea level pressure, zonal and meridional wind component, sea surface temperatures, near-surface air temperatures, and total cloudiness as described by Wolter and Timlin (1993). It is more complex than the other indices, but it is also more complete; in fact it is based on multiple different variables related to both ocean and atmospheric systems. It is able to capture more information about ENSO since it is a coupled ocean-atmospheric phenomenon. Moreover the MEI index has less vulnerability to errors in single variable fields because it combine information from many different fields. In Figure (3.13) the monthly standardize MEI data are shown; the positive value are related to the presence of El-Niño episodes, while the negative ones represent La-Niña episodes.

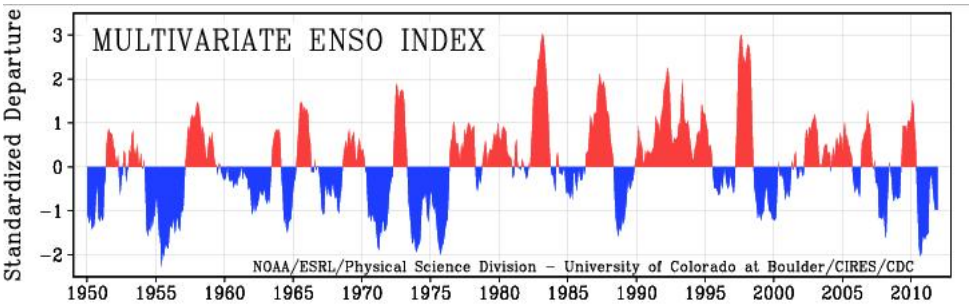


Figure 3.13: Standardized monthly data of MEI index from 1950 until nowdays. The positive values represent El-Niño episodes, while the negative represent La-Niña episodes .

Instead the SST index have been a primary expression of global climate anomalies for several decades. In fact, El-Niño Southern Oscillation produces a sea surface oscillation in the Pacific Ocean. Precisely, it is measured as a mean, of the superficial layer of the Ocean, on four specific regions of the Pacific, see Figure (3.14). The ”superficial layer” definition changes between the different measurement technologies (it varies from a few

millimeters up to 20 meters under the sea level). The four different regions, above which, the SST index is calculated, are all disposed in the Pacific Ocean, as shown in Figure (3.14):

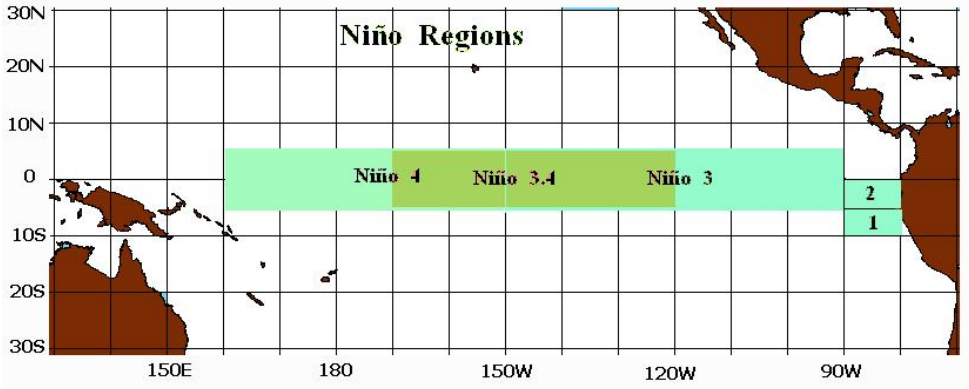


Figure 3.14: The 4 different regions of SST measurement.

All the information and the data about SST, SOI and MEI index are available on the National Oceanic and Atmospheric Administration (NOAA) website.

3.3.1 Iterative Input Selection with the ENSO indices

We tried to include the previously described ENSO indices, SST1·2, SST3, SST4, SST3·4, MEI, SOI, into the input vector described in Section (3.2.3). We ran the IIS algorithm on this new data-set to see if the overall performance might improve.
We discovered that the IIS performance improves with the inclusion of these climatic indices. The final coefficient of determination gets to a value of $R^2 = 0,68$ and, moreover, the selected final variables includes one of these indices, specifically the SST12, as shown in Figure (3.15) and in Table (3.5).

Table 3.5: Result of the IIS procedure with the candidate input vector containing ENSO indices.

Variable	R^2	Increase R^2
Time	0,3601	36%
Storage	0,6119	26%
q_t^{VQ}	0,6438	2,4%
$SST12$	0,6742	3%
q_t^{YB}	0,6893	1,5%

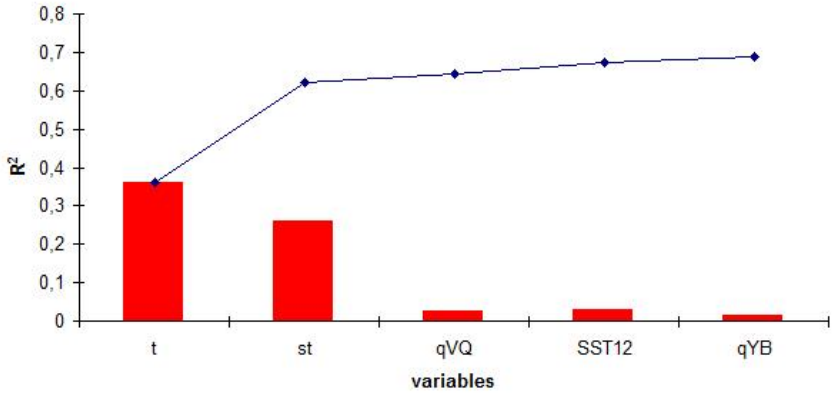


Figure 3.15: The variables selected by running the IIS algorithm on the candidate input data-set with the associated relative contribution to the overall performance of the underlying tree-based model.

From Table (3.5) it is possible to see that there is an improvement in the coefficient of determination of the 3% thanks to the SST12 variables. This indicates that actually exists a relationship between the climatologic index and the inflow to the Hoabinh.

Chapter 4

An Inflow Forecasting Model: Artificial Neural Network

The traditional way of exploiting hydroclimatic information in the reservoirs management has been, so far, represented by the construction of inflow forecasting models. In the previous chapters we presented an innovative procedure to use these information based on the selection of the most relevant hydroclimatic variables between all the available ones. In this Chapter we want to analyze if those most significant selected variables, can actually improve also the performance of a classic inflow forecasting model. We decide to use an Artificial Neural Network to implement this inflow forecasting model, because it is an excellent tool, able to capture the non-linear relationship between two time series, if any, and does not depend on the distributional form of the data set.

4.1 Architecture and training of the ANN

An artificial neural network is a mathematical structure designed to mimic the information processing functions of a network of neurons in the brain (Hinton, 1992; Jensen, 1994). ANNs are highly parallel systems that process information through many interconnected units that respond to inputs through modifiable weights, thresholds, and mathematical transfer functions. Each unit processes the pattern of activity it receives from other units, then broadcasts its response to still other units. ANNs are particularly

well suited for problems in which large data-sets contain complicated nonlinear relations among many different inputs. ANNs are able to find and identify complex patterns in data-sets that may not be well described by a set of known processes or simple mathematical formulae. Unlike a process-based model, it is not necessary to know exactly how those variables interact, the nature of the physical processes that cause those patterns, or any mathematical representation of those processes before applying an ANN. Hence, the ANN is an information processing systems trying to simulate, within an informatics system, the functioning of a biological nervous structure that are composed by a big quantity of nervous cells (or neurons) connected in a complex network. Some of these units receive information from the external environment, others emit results again to the external environment, while others (if there are) communicate only with the other units inside the network: they are respectively defined as input units, output units and hidden units, as shown in Figure (4.1):

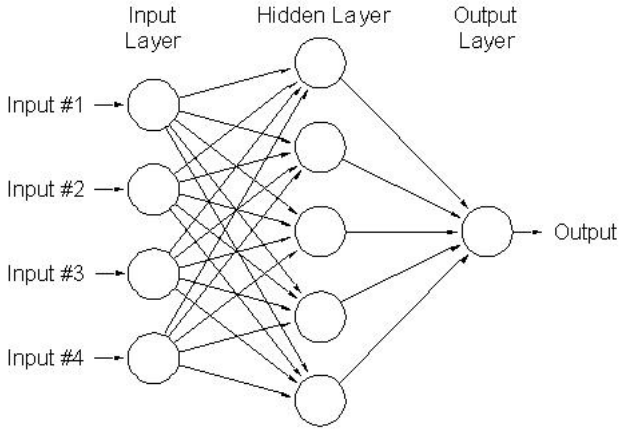


Figure 4.1: A representation of a simple 3-layer feed-forward artificial neural network with 4 inputs, 5 hidden nodes, and 1 output.

The input-output relation, i.e the transfer function of the network, does not need to be programmed but it is inferred by a learning process based on a training data set including some examples of input with its corresponding output: in this way the network can learn that relation. In fact, in the context of prediction, the network training consists of providing input-output examples to the network, and minimizing an error function with respect to the network weights. The network is trained with an appropriate method (typically the back-propagation), that uses these data for the purpose of modifying the weight and other parameters of the network itself in the way to minimize the prediction error function

related to the training set. If this training is successful, the network learn the unknown input-output relation and so it is able to make forecast even where the output is unknown. For this reason it is used for making inflow forecast trying to exploit the possible relation between the inflow formation upon a specific basin and the global circulation phenomena or the soil moisture levels.
The training algorithm uses a mean-squared error objective function, which tries to minimize the average squared error between the network's output, and the target value over all the training data pairs. This research of the minimum of the objective function is commonly pursued by using a gradient descent algorithm.
One important aspect of the ANN methodology is the design of the network architecture. In most of the studies, the network architecture is decided in an heuristic way (trials and errors). So we did in our case. This architecture is based on some parameters that must be initialized, specifically they are:

- Layers number
  Neural networks can be divided in two categories: Single-layer perceptron and Multi-layer perceptron. The first one is the simplest and the earliest kind of feed-forward network and it consists in a structure with a single layer of output nodes fed directly by the input data via a series of weights. While in the Multi-layer perceptron structure between the input and the output layer there is the presence of one or more hidden layers, whose nodes are called hidden neurons; each neuron in one layer has direct connections to the neurons of the subsequent layer.

- Hidden neurons number
  These are neither in the input layer nor the output layer. These neurons are essentially hidden from view, and their number and organization can typically be treated as a black box to people who are interfacing with the system. As a rule the number of hidden neurons must be greater than or equal to the number of input variables.

- Type of hidden neurons
  The hidden neurons are differentiated on the basis of the activation function. This function describes the output behavior of a neuron, it 'connects' weighted sums of the units in one layer to the values of the units in the next layer; it must be differentiable and monotonically increasing. There are several genres of activation functions

such as the sigmoid function, the hyperbolic tangent function or piecewise-linear function.

In this thesis a multi-layer network with one single hidden layer is used, i.e. the structure showed in Figure (4.1). The Levenberg-Marquardt back- propagation algorithm is applied (LMBP) (Hagan and Menhaj, 1994), and the activation function chosen for the hidden neurons is an hyperbolic tangent.
The parameters described previously are the theoretical base for the construction of an ANN, afterwards there are several other more technical parameters that need to be fixed. First of all the data set available to train the network must be divided in Calibration horizon and Validation horizon, where the former is the data set over which the network is trained and the latter is the set for testing it. Another parameter to be fixed is the 'Number of total iteration' i.e. the times the training process is repeated, each time starting from a different initialization; after all these different running the best network is selected. Finally there is one more parameter, the 'Number of epochs' that represents the maximum number of trials the gradient descent algorithm can do in the research of the minimum value of the gradient.
We made a few trials to train the ANN for this study and the best results are showed in hereafter.

4.2 Inflow forecasting on the Da River using rainfall data

Here the most typical example of hydrological ANN is applied: the rainfall-runoff model.
Downstream of the Dá River is located the Hoabinh reservoir. Therefore a good forecast of the future inflows along this river is essential for a smart management of the reservoir. In the studies previously made on the Hoabinh management was carried out that to be able to empty the reservoir without creating any damage to the capital city Hanoi, it is necessary to know the approaching of a peak inflow with, at least, 5 days of warning (see Quach, 2011). For this reason it was imposed to the ANN an aggregated inflow on a 5 days step as output. In this way it was implicitly required to the ANN to forecast the cumulated inflow of the 5 next coming days from the time instant t.

We tried to use several different input configurations, by referring to the results obtained in the previous Chapter (see Sections (3.2.2) and (3.2.3)). We analyzed the cross-correlation of every of those variables with the output we want to forecast. Only a few of them resulted having a weak correlation with the output; we ran the forecasting model only using the more correlated ones. Hence the input configurations tested are:

1. q_t,p_t
2. q_t, P_5^{WR}
3. q_t,q_{t-1},p_t
4. q_t, q_{t-1}, P_5^{WR}
5. q_t,p_t,p_{t-1}
6. q_t,q_{t-1}

where:
q_t and q_{t-1} represent the inflow to the Hoabinh reservoir at the present time instant t and at the previous one respectively; p_t is the rainfall data of the farthest hydrological station situated on the Da River from the Hoabinh reservoir, the Muongte station (see Figure 3.4); while P_5^{WR} indicates the weighted areal rainfall (aggregated with the Thiessen Polygons method) cumulated over the interval [t-5,t).
All the previous input configurations have been analyzed both using all available data sequence on the entire evaluation horizon, and considering only the data included in the rainy season between June 1st and until September 30th of each year in question.
We studied the cross-correlation between the proposed variables, the results are shown in the graphics of Figure (4.2) and (4.3).

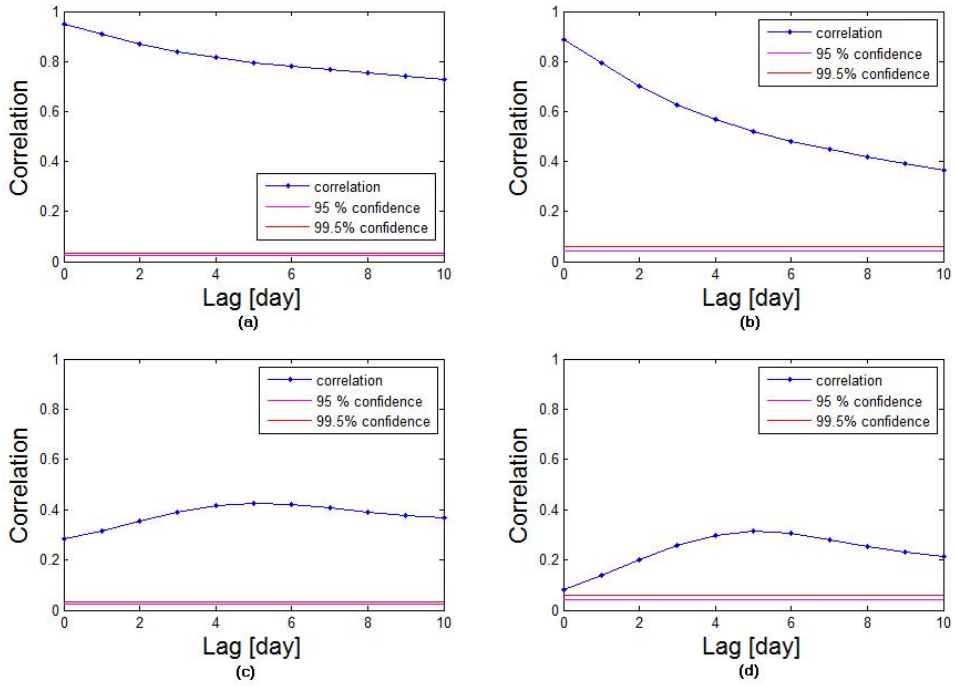


Figure 4.2: Cross-correlograms between the cumulated inflow of the 5 next coming days (the output that we want to forecast), the inflow itself q_t and the rainfall p_t registered at the Muongte Station: a) represents the cross-correlation of q_t with the output over the whole evaluation period; b)represents the cross-correlation of q_t with the output over the rainy season; c) represents the cross-correlation of p_t with the output over the whole period; d) represents the cross-correlation of p_t with the output over the rainy season.

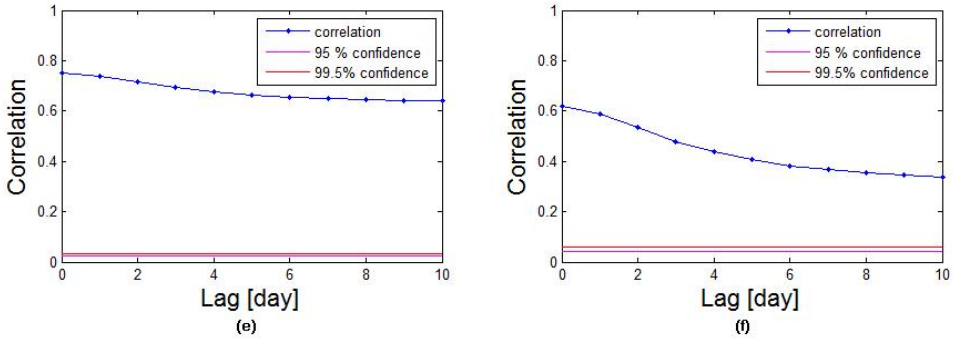


Figure 4.3: Cross-correlograms between the cumulated inflow of the 5 next coming days (the output that we want to forecast) and the areal rainfall P_5^{WR} over the Da River basin: e) represents the cross-correlation of the output with over the whole evaluation period; f)represents the cross-correlation of q_t with the areal rainfall over the rainy season.

From this first analysis it can be observed that the output has a very high correlation

with the inflow itself, this suggests that the system is highly auto-correlated (graphic a). Second we can see that for every variable the correlation during the rainy season is weaker than the one over the whole period (graphic b, d, e). Moreover it appears that the correlation between the output we want to forecast and the rainfall data p_t is very low (graphic c).
For all these different input configuration an ANN with a common structure was used; the technical parameters of the network have been fixed as shown in Table (4.1):

Table 4.1: Technical parameters of the network architecture

Calibration Horizon:	1961-1978
Validation Horizon:	1994-2004
Number of hidden neurons:	4
Number of total iteration:	120
Number of epochs:	1000

All the resulting models were compared between each others, using different graphical and mathematical tools. The highest explained variance R^2 was searched as discriminant between models. The models based only on data over the rainy season performs significantly worse than the models over the whole period; this confirms the observations made on the cross-correlation graphics. For this reason all the following analysis are made only on the result of the models based on the entire evaluation horizon.
We decide to shows later some graphics of the models with the highest R^2, i.e. the models based on the data sets 3 and 6 over the whole evaluation horizon. These models reached respectively a $R^2 = 0.9057$ and $R^2 = 0.9076$. From these results it is possible to see that there is a weak improvement in using also the rainfall information instead of using only the inflow data, this probably means that the system is highly autocorrelated and so the importance of the autoregressive part of this model is much higher with respect to the exogenous part represented here by the rainfall data. This result confirms the conclusions drawn from the previous analysis of the cross-correlation graphic. To deeply analyze this question, models based only on data over the rainy season were tested, to better understand if, at least in that period, the relative weight of the exogenous part was higher for

the purpose of prediction. But, as said before, the weak results obtained confirm that the system is highly autocorrelated.

The performances of these models were compared with those of a simple linear model, to understand if there is an effective improvement in the forecasting performances using a complex model, like the ANN, instead of a simpler one. In the case of the input configuration number 3 the linear model reached an explained variance of $R^2 = 0.8966$, while in the case of the configuration number 6 its explained variance is $R^2 = 0.8978$. Hence, from a preliminary analysis the ANN reaches better performances than the linear models. We used the input configurations number 3 and 6 to create the hydrographs showed in Figure (4.4) and in Figure (4.5) respectively. They display, both, two forecasted q_t, over the first year of the validation horizon (i.e. 1994), the red one obtained with the ANN model and the green one obtained with a simple linear model, while with the blue line is displayed the measured inflow for that year. For both is shown the detail of a peak flow event during the wettest months (July, August, September) in Figure (4.6) and Figure (4.7).

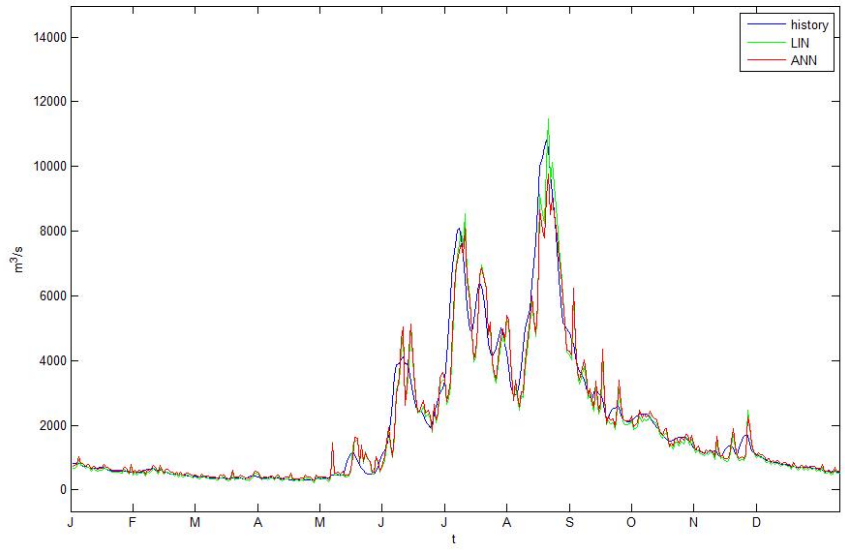


Figure 4.4: Hydrograph obtained with the input configuration number 3 with historic inflow to the Hoabinh for the year 1994(blue), forecasted inflow by ANN(red), forecasted inflow by a linear model(green).

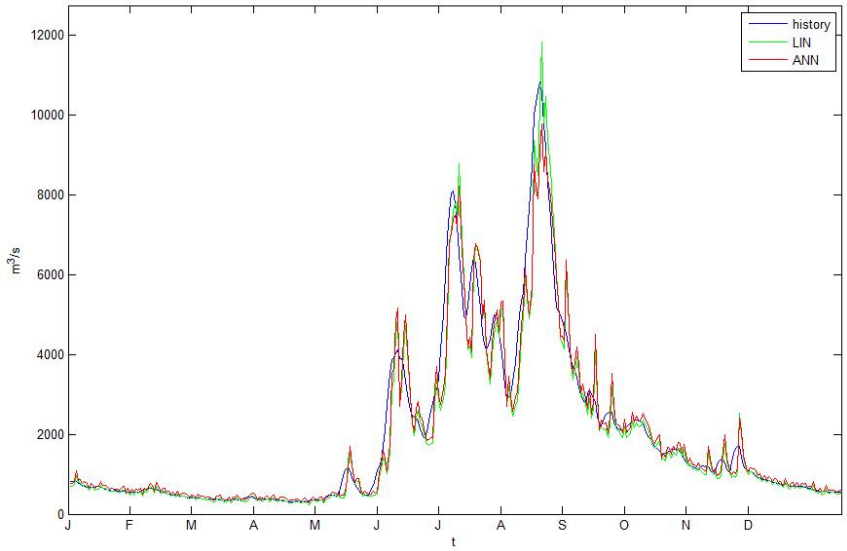


Figure 4.5: Hydrograph obtained with the input configuration number 6 with historic inflow to the Hoabinh for the year 1994(blue), forecasted inflow by ANN(red), forecasted inflow by a linear model(green).

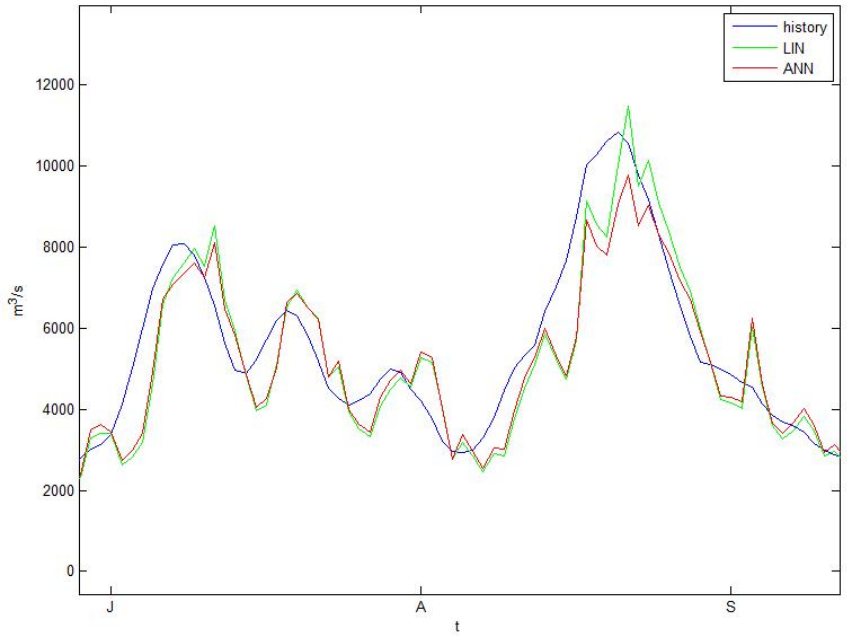


Figure 4.6: Detail of a peak flow for the input configuration 3.

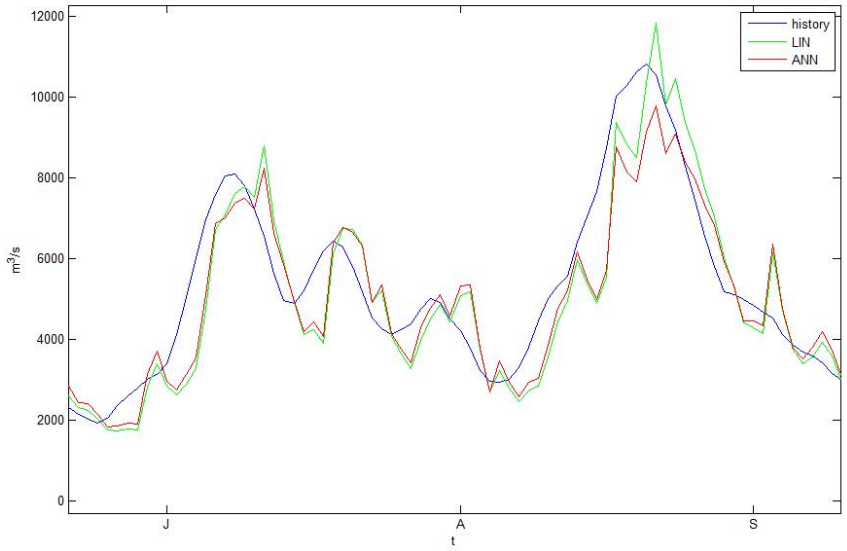


Figure 4.7: Detail of a peak flow for the input configuration 6.

From the previous hydrographs it is possible to see that the ANN and the linear model are very good in the reproduction of the low-flow events, while they are not so precise in the reproduction of the peak flows. Moreover it appears a delay between 1 and 3 days in the reproduction of the peak flow events in both models.
The hydrographs are very similar to each other, this suggests that the introduction of the rainfall data does not significantly improve the forecasting performance; the linear model are comparable to the ANN ones, as shown also by the scatter plots in Figure (4.8) and in Figure (4.9) .

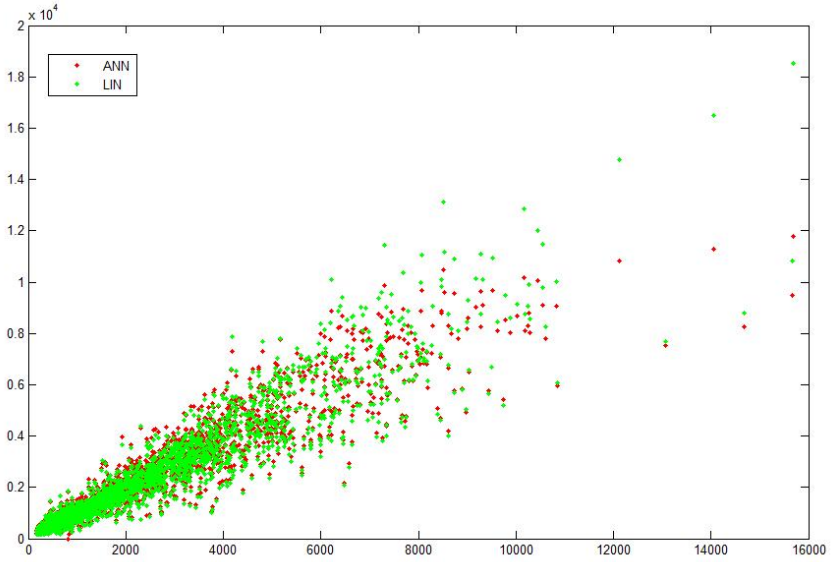


Figure 4.8: Scatter plot of the inflow predicted by the linear model(blue),and the inflow predicted by the ANN(red) obtained with the input configuration number 3.

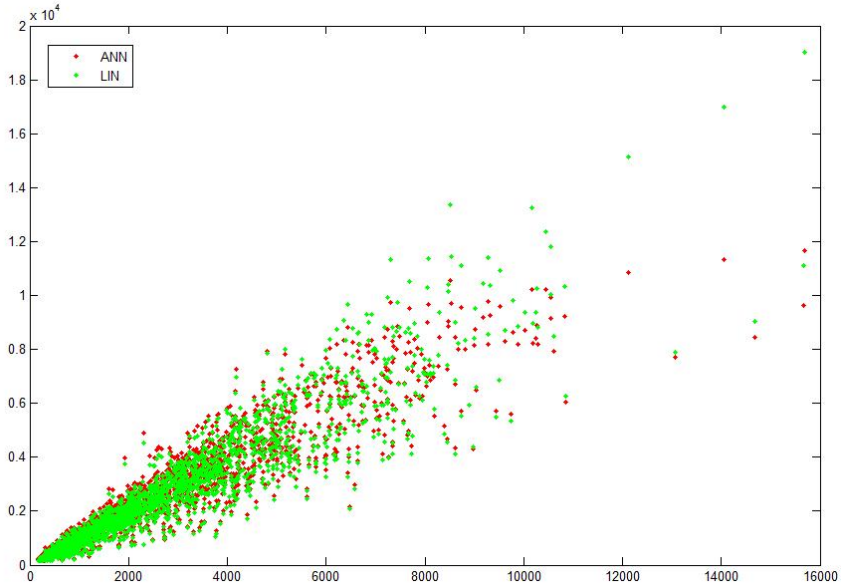


Figure 4.9: Scatter plot of the inflow predicted by the linear model(blue),and the inflow predicted by the ANN(red) obtained with the input configuration number 6.

These scatter plots show that, with both the considered input configurations, the linear model and the ANN model is strong in the reproduction of low flows (the dots are concentrated on the bisector) while showing weak performance in the peak flow reproduction

(the dots are more dispersed). Again the scatter plots of the 2 configurations are really similar and this confirms the idea that the system is highly based on the autoregressive part.

Since all the previous models appear to have very similar performances another mathematical analysis have been made, in this way it is tried to addition information for better understanding the trend of every models. Some more indices, in addition to the R^2, have been taken into consideration.

Table 4.2: Summary of performance metrics measured on the validation horizon

Model	Input Configuration	R^2	RMSE	PDIFF	PEP	MAE	AME
Linear	q_t,p_t	0,8258	850,1	-1212,8	-7,7	431,4	7389,1
ANN	q_t,p_t	0,8437	798,8	6280,1	40,1	418,7	8293,5
Linear	q_t, P_{55}^{WR}	0,8381	819,6	383,4	2,4	442,7	7217,8
ANN	q_t, P_{55}^{WR}	0,8475	789,4	5978,4	38,1	424,2	8047,1
Linear	q_t,q_{t-1},p_t	0,8966	654,2	-2836,9	-18,1	329,1	5901,7
ANN	q_t,q_{t-1},p_t	0,9057	621,1	3917,1	24,9	319,2	6411,3
Linear	q_t,q_{t-1},P_5^{WR}	0,9044	630,1	-2479,8	-15,8	312,6	5697,6
ANN	q_t,q_{t-1},P_5^{WR}	0,1139	1900,3	1246,6	79,5	1369,4	1246,6
Linear	q_t, p_t, p_{t-1}	0,8276	845,3	-1045,9	-6,6	431,3	7486,9
ANN	q_t, p_t, p_{t-1}	0,8450	795,7	5454,4	34,7	418,5	8038,1
Linear	q_t,q_{t-1}	0,8978	651,7	-3324,2	-21,2	317,1	5638,4
ANN	q_t,q_{t-1}	0,9076	614,5	4245,6	27,1	314,7	6290,4

Where:

RMSE is the acronym of Root of Mean Squared Error and it is calculated with the following formula:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{q}-\hat{q})^2} \tag{4.1}$$

PDIFF is the acronym of Peak Difference and it is calculated with the following formula:

$$PDIFF = \max \bar{q} - \max \hat{q} \tag{4.2}$$

PEP is the acronym of Percent Error in Peak and it is calculated with the following formula:

$$PEP = \frac{(\max \bar{q} - \max \hat{q})}{\max \bar{q} * 100} \tag{4.3}$$

MAE is the acronym of Mean Absolute Error and it is calculated with the following formula:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} (|\bar{q} - \hat{q}|) \tag{4.4}$$

AME is the acronym of Absolute Maximum Error and it is calculated with the following formula:

$$AME = \max |\bar{q} - \hat{q}| \tag{4.5}$$

From the previous Table several observations can be made. First it is possible to notice that the linear models perform actually worse than the ANN, even if, they show a better prediction capacity of the peak flow events. This is possible to be deduced from the PDIFF and PEP indicators, whose values (absolute values) are always smaller than the ones associated with the ANN. Moreover another note can be done: the linear model of the configuration q_t, P_5^{WR} presents a PDIFF value of 2.4, this means that it is able to reproduce the peak flow events much better than any other models we tried. We have confirmation of this increased precision even if we look at the hydrograph produced from the corresponding model, see Figure (4.10). This could mean that the areal rainfall contains information more useful for the future inflow prediction. In fact, the areal weighted rainfall over the basin and accumulated on the previous 5 days, could represent a more complete and efficient information in creating a more accurate prediction of future inflows.

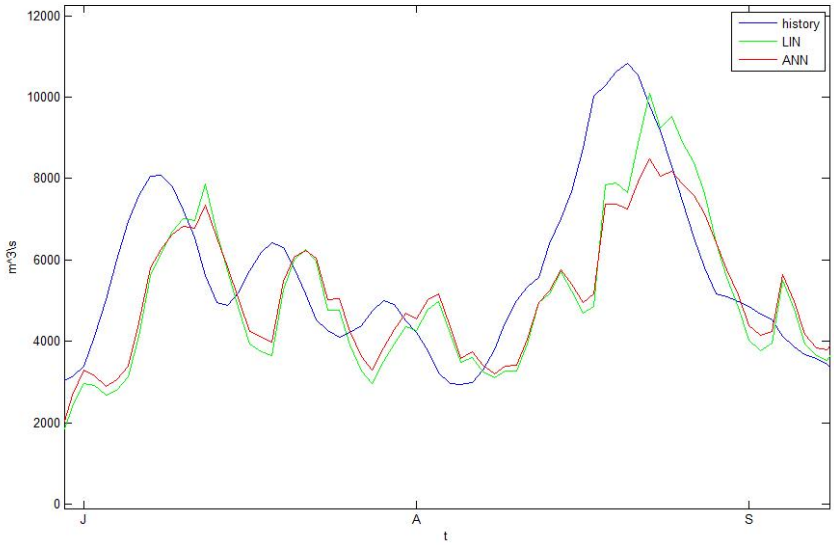


Figure 4.10: Detail of a peak flow for the inflow configuration q_t, P_5^{WR} with historic inflow to the Hoabinh for the year 1994(blue), forecasted inflow by ANN(red), forecasted inflow by a linear model(green).

However, even if this graphic shows better performance in the reproduction of the peak-flow respect to graphics of Figure (4.6) and (4.7), it shows that there is a bigger delay in the peak reproduction, i.e. there is 4 to 5 days of delay.

4.3 Inflow forecasting on the Da River using an ENSO index

In this trial the objective is to investigate the influence of large-scale atmospheric circulation phenomena on the basin-scale inflow variation and possible improvement of inflow prediction by incorporating the information of such large-scale atmospheric circulations. Also in this case an ANN approach is used to model the complex relationship between inflow and large-scale atmospheric circulations.
The ENSO index used in this study is the *Sea Surface Temperature* from El-Niño 1-2 region. We use this index because it is the only one selected by the IIS among all the other ENSO indices (see Section (3.3.1)).

The input configurations used are:

1. q_t,q_{t-1},SST_t

2. q_t,P_t^{WR},SST_t

where:

q_t, q_{t-1} are the inflow to the Hoabinh, P_t^{WR} is the areal rainfall aggregated with the Thiessen Polygons method and SST_t is Sea Surface Temperature from Niño 1-2 region.
Even this time as a first analysis we studied the cross-correlation between the SST Index and the output that we want to forecast, the results are shown in the graphics of Figure (4.11).

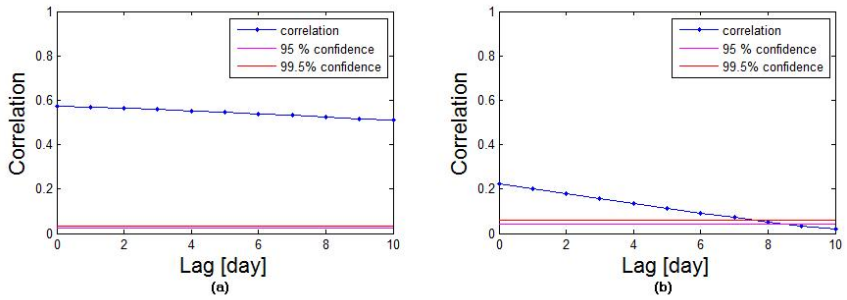


Figure 4.11: Cross-correlograms between the cumulated inflow of the 5 next coming days (the output that we want to forecast) and the SST Index values: a) represents the cross-correlation of the output with the SST over the whole evaluation period; b)represents the cross-correlation of the output with the SST over the rainy season.

From these graphics we can observe that the correlation over the rainy season is lower than the one over the whole period.
The technical structure of the ANN is the same of the previous case, see Table (4.1).
With this trial we obtained an overall R^2 similar to the previous trials: $R^2 = 0.9072$ for the first configuration and $R^2 = 0.8603$ for the second. If we analyze the hydrograph of the first configuration, showed in the Figure (4.12), we can see that the ANN is performing well on the low-flow events but again, not so well on the peak flows. As for the previous case a linear model is used to compare the performance of the ANN.

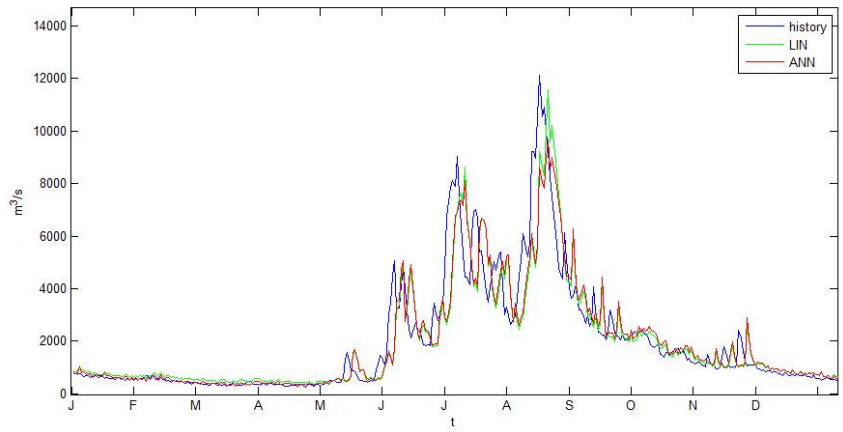


Figure 4.12: Hydrograph of the inflow configuration 1 with historic inflow to the Hoabinh for the year 1994(blue), forecasted inflow by ANN(red), forecasted inflow by a linear model(green).

From this hydrograph we can note that there is a delay in the reproduction of the peak flows of about 1 to 3 days.

Even this time, the scatter plot, shown in Figure (4.13)is very similar to the previous cases. It shows that the ANN and the linear model perform in a very similar way on the low-flow, but on the peak-flow the linear model has a better forecasting capacity.

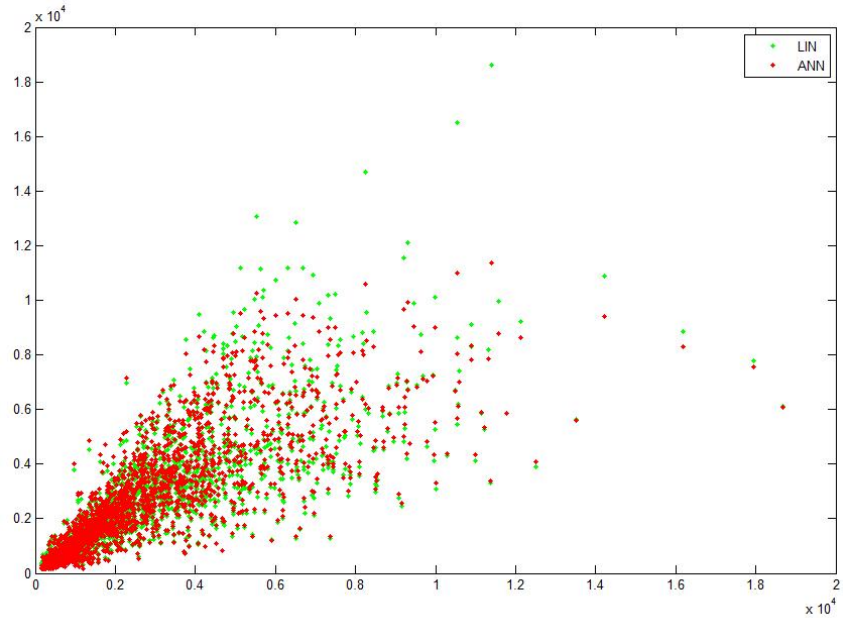


Figure 4.13: Scatter plot of the input configuration 1 with the inflow predicted by the linear model(blue),and the inflow predicted by the ANN(red).

As the previous case all the input configuration have been tested both over the whole evaluation period and then only over the rainy season (beginning of July - end of September) for each year. Even this time the models tested on the rainy season, perform significantly worse with respect to the models over the entire period, thus again, all the following analysis have been made only on the model based on the entire evaluation horizon.
Since the 2 input configurations performs very similarly some more performance metrics measured on the validation horizon have been analyzed, see Table (4.3):

Table 4.3: Summary of performance metrics measured on the validation horizon

Model	Input Configuration	R^2	RMSE	PDIFF	PEP	MAE	AME
Linear	q_t,q_{t-1},SST_t	0,899	642,5	-2927,7	-18,6	342,4	5818,9
ANN	q_t,q_{t-1},SST_t	0,9067	618,3	4180,9	26,6	317,2	6388,1
Linear	q_t, P_t^{WR}, SST_t	0,8486	789,5	-3280,7	-20,9	413,1	6969,2
ANN	q_t, P_t^{WR}, SST_t	0,8603	758,7	4721,1	30,1	408,1	7015,1

As a conclusion it is possible to say that there is no remarkable improvement in using the SST12 index to construct an inflow forecasting model.

In conclusion, we can compare the results shows in Table (4.3) and (4.2); all input configurations have, both in the case of the ANN and in the case of the linear, an explained variance very similar that changes in a very small percentage from the configuration q_t,q_{t-1}. This suggests again that there is no improvement in including exogenous information like p_t or SST_t in the input set because the system seams to be really autocorrelated. We can conclude that the "black-box" approach of the ANN probably is not suitable to deal with the hydroclimatic variables selected for the Da river system.

Chapter 5

Conclusions and further research

In the introduction of this thesis we posed a question: Can, the direct inclusion of exogenous information in a reservoir operating policy, improves the performance of a reservoir management?
After all the analysis made along this work we can answer affirmatively. Indeed the aim of this thesis was to propose a new approach for exploiting exogenous information (in our case we always referred to hydroclimatic information) for improving reservoir management. The idea we developed, was to collect some relevant hydroclimatic variables able to give information about, and so to play as a surrogate of, the future inflow to the reservoir. In a way to implement an operating policy by directly including them into it, and so, without passing through a traditional inflow forecasting model. We tried to move from the model-based approach of exploiting exogenous information to a model-free one. This study was applied to a real world management problem, i.e. the operation of the Hoabinh reservoir in Vietnam.

With regard to the management of the Hoabinh reservoir we assessed that the procedure implemented, selecting the most relevant hydroclimatic information and including them directly into the operating policy, is actually performing well. The obtained results show that the information selected as the most informative allows to improve the policy performance of about 20%, on the flood control objective, with respect to an equivalent policy not conditioned upon any hydroclimatic data.
Moreover we applied our procedure (but this time without the optimization step) to an initial input set containing some ENSO indices. The Iterative Input Selection (IIS) algorithm

applied to this new data set, performs better than in the previous case, and it selects among the most valuable variables an ENSO index, i.e. the SST12. This demonstrates that the inclusion of these atmospheric circulation indices might be helpful in the inflow understanding and that they might improve the reservoir management.

Finally, in the last chapter of the thesis, we try to test if the most relevant hydroclimatic variables, selected with the application of the IIS algorithm, can actually improve also the performance of a classic inflow forecasting model. We use an Artificial Neural Networn (ANN) model and a linear model. The results obtained deny our hypothesis, i.e. both the ANN model and the linear model, ran on the Da river basin, show that there is no improvement in the inflow forecasting by using some relevant hydroclimatic variables. Probably this is due to the inadequacy of the "black box" approach to deal with these variables and, probably, a physically based model is more suitable.

In conclusion, the proposed approach gave promising results, both in analyzing the most informative variables among a wide set and in creating operating policy with higher performances with respect to traditional methods.
Although the results of this thesis seem to represent a good improvement over the typical approaches in the reservoir management, many aspects of the proposed approach require further investigations. Future directions of investigation might be the following:

- To extend the procedure created to multi-objective management problem, since in this thesis we concentrate on the optimization of only one objective.
- To deeply analyze the possible role of the ENSO indices, to study the relationship between them and the inflow formation and trends in the Vietnam geographical area.
- To improve the mathematical and statistical tools of the Iterative Input Selection algorithm.
- To deeply analyze the use of the hydroclimatic information, selected by the IIS algorithm, in inflow forecasting models.

Appendix A

Abbreviations

The following there is a legend of all the abbreviations used in this work.

Abbreviations

The very first letter of the input variable labels q, p, T and E denotes respectively inflow, precipitation temperature and evapotranspiration; the superscript stands for the name of stations (see Table(A.1)); the subscript denotes the values of data in time: e.g. $t, t-1, t-2, t-3$, etc, (or in the interval $[t-1;t)$, $[t-2;t-1)$, $[t-3;t-2)$, $[t-4;t-3)$, etc) respectively. The capital letter P indicates the cumulated sum of rainfall; for example p_2^{MT} means the cumulated rainfall at the station MT, that is equals to the cumulated rainfall in the interval $[t-2;t)$; the P_3^{MT} means the cumulated rainfall at the station MT, that is equals to the cumulated rainfall in the interval $[t-3;t)$ and so on.
For geographical references see Figure (A.1).

Table A.1: Stations.

Abbreviation	Name	River
MT	Muongte	Da
TU	Phadin	Da
SH	Phuyen	Da
SL	Sinho	Da
LH	Lai Chau	Da
PY	Tamduong	Da
MO	Thanuyen	Da
TC	Tuangiao	Da
YC	Yenhau	Da
HG	Ha Giamg	Lo
L2	T Quang	Lo
T1	Thanh Son	Thao
LC	Lao Cai	Thao

Table A.2: Other abbreviations.

Abbreviation	Meaning
WR	Weighted rainfall with the Thiessen Polygon procedure over the Da river
WD	Water Demand at Sontay station
SST12	Sea Surface Temperature over the Ocean Pacific region 1·2

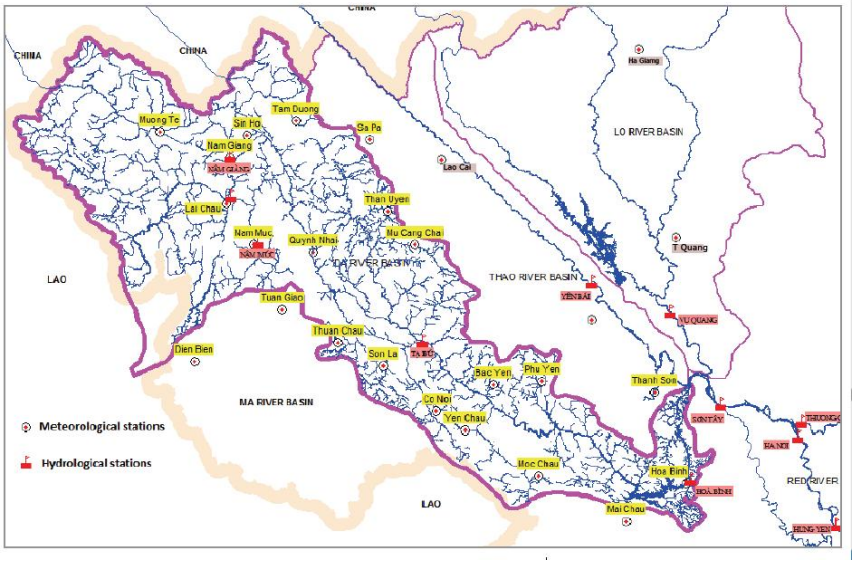


Figure A.1: Main stations on the Red River Basin.

Bibliography

[1] R.J. Allan. ENSO and climatic variability in the past 150 years. In: Diaz H.F., Markgraf V. (Eds.), *Multiscale variability and global and regional impacts*, Cambridge University Press, Cambridge, 3-55, 2000.

[2] R.E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5-32, 2001.

[4] A. Castelletti, S. Galelli, A. Salvetti, and A. Ventimiglia. Extremely Randomized Trees and Feature Ranking for daily streamflow prediction. In *Proceedings of the 9th International Conference on Hydroinformatics*, 7-11 September, Tianjin, CN, 2010.

[5] A. Castelletti, F. Pianosi, X. Quach, and R. Soncini-Sessa. Assessing water reservoirs management and development in Norhtern Vietnam. *Hydrol. Earth Syst. Sci.*, 16:189-199, 2012.

[6] A. Castelletti, and F. Pianosi. Improved reservoirs operation by hydroclimatic information. In *Proceedings 10th International Conference on Hydroinformatics*, Hamburg, Germany, 2012.

[7] A. Castelletti, S. Galelli, M. Restelli and R. Soncini-Sessa. Tree-based reinforcement learning for optimal water reservoir operation. *Water Resources Research*, 2010b.

[8] J. Chandimala, and L. Zubair. Predictability of stream flow and rainfall based on ENSO for water resources management in Sri Lanka. *Journal of Hydrology*, 335, 303-312, 2007.

[9] F.H.S. Chiew, T.C. Piechota, J.A. Dracup, and T.A. McMahon. El Niño/Southern Oscillation and Australian rainfall, streamflow and drought: links and potential for forecasting. *Journal of Hydrology*, 204: 138-149, 1998.

[10] A. Cutler and Z. Guohua. PERT - Perfect Random Trees Ensembles. *Computing Science and Statistics*, 33:490-497, 2001.

[11] A. Dai, K.E. Trenberth, and T. Qian. A global dataset of Palmer Drought Severity Index for 1870-2002: Relationship with soil moisture and effects of surface warming. *American Meteorological Society*, 2004.

[12] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182-197, 2002.

[13] S. Galelli. *Dealing with complexity and dimensionality in water resources management*. Phd Thesis, Politecnico di Milano, 2010.

[14] A.K. Gobena, and T.Y. Gan. Incorporation of seasonal climate forecasts in the ensemble stream flow prediction system. *Journal of Hydrology* 385, 336-352, 2010.

[15] G. Guariso, S. Rinaldi, and R. Soncini-Sessa. The Management of Lake Como: A Multiobjective Analysis. *Water Resources Research*, 1986.

[16] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 2006.

[17] K. Hansson and L. Ekenberg. Flood mitigation strategies for the red river delta. In *International Conference on Environmental Engineering, An International Perspective on Environmental Engineering*, Canada, 2002.

[18] M.T. Hagan, and M.B. Menhaj. Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks*, 5(6):989-993, 1994.

[19] M.I. Hejazi, X. Cai, and B. Ruddell. The role of hydrologic information in reservoir operation - Learning from historical releases. *Advances in Water Resources*, 31, 1636-1650, 2008.

[20] G.E. Hinton. How neural networks learn from experience, *Scientific American*, 267, 144-15, 1992.

[21] B.A. Jensen. Expert systems - neural networks. *Instrument Engineers' Handbook 3rd ed*, Radnor, Pennsylvania, 48-54, 1994.

[22] E. Kahya, and J.A. Dracup. *Influences of type 1 El-Nino and La-Nina events on streamflows in the Pacific southewest of the United States*. University of California, Los Angeles, 1998.

[23] E. Kahya, and J.A. Dracup. U.S. streamflow patterns in relation to the El Niño/Southern Oscillation. *Water Resources Research* 29(8): 2491-2503, 1993.

[24] S.S. Kashid, S. Ghosh, and R. Maity. Stream flow prediction using multi-site rainfall obtained from hydroclimatic teleconnection. *Journal of Hydrology*, 395, 23-38, 2010.

[25] S. Khan, A.R. Ganguly, S. Bandyopadhyay, S. Saigal, D.J. Erickson, V. Protopopescu, and G. Ostrouchov. Non-linear statistics reveals stronger ties between ENSO and the tropical hydrological cycle. *Geophysical Research Letters*, 2006.

[26] J.W. Labadie.Optimal operation of multireservoir systems: state-of-the-art review. *Water Resources Planning and Management*, 130(2):93111, 2004.

[27] A. Maas, M.M. Hufschmidt, R. Dorfam, H.A. Thomas, S.A. Marglin, and G.M. Fair. *Design of Water Resources Systems*. Harward University Press, Boston, USA, 1962.

[28] R. Maity, and N. Kumar. Basin-scale stream-flow forecasting using the information of large-scale atmospheric circulation phenomena. *Hydrological Processes*, 22, 643-650, 2008.

[29] A. Makkeasorn, N.B. Chang, and X. Zhou. Short-term streamflow forecasting with global climate change implications - A comparative study between genetic programming and neural network models. *Journal of Hydrology* 352, 336-354, 2008.

[30] K. Miettinen. Nonlinear Multiobjective Optimization. *Kluwer Academic Publishers*, Dordrecht, NL, 1999.

[31] G. Muluye, and P. Coulibaly. Seasonal reservoir inflow forecasting with low-frequency climatic indices: a comparison of data-driven methods. *Hydrological Sciences Journal*, 52:3, 508-522, 2010.

[32] M. Nguyen. *Tree-based input selection for hydrological modelling*. Master's thesis, Politecnico di Milano, 2010b.

[33] X. Quach. *Assessing and optimizing the operation of the Hoabinh reservoir in Vietnam by multi-objective optimal control techniques*. Phd Thesis, Politecnico di Milano, 2011.

[34] R. Soncini, A. Castelletti, and E. Webebr. *Integrated and participatory water resources management: Theory.*, Elsevier, Amsterdam, NL, 2007a.

[35] R. Soncini, A. Castelletti, and E. Webebr. *Integrated and participatory water resources management: Practice.*, Elsevier, Amsterdam, NL, 2007b.

[36] J.A. Tejada-Guibert, S.A. Johnson, and J.R. Stedinger. The value of hydrologic information in stochastic dynamic programming models of multireservoir system. *Water Resources Research*, 31(10): 2571-2579, 1995.

[37] A.J. Troup. The southern oscillation. *Quarterly Journal of the Royal Meteorological Society*, 91: 490-506, 1965.

[38] S. Vorogushyn, B. Merz, K.E. Lindenschmitdt, and H. Apel. A new methodology for flood hazard assessment considering dike breaches. *Water Resources Research*, 46(10):1-17, 2010.

[39] K. Wolter, M.S. Timlin Monitoring ENSO in COADS with a seasonally adjusted principal component index. In *Proceedings of the 17th Climatic Diagnostics Workshop*, Norman, OK (USA), 52-57, 1993.

[40] W. Yeh. Reservoir management and operations models: a state of the art review. *Water Resources Research*, 1985.

[41] Z. Zhao, S. Sharma, A. Anand, F. Morstatter, S. Alelyani, and H. Liu. Advancing Feature Selection Research - ASU Feature Selection Repository. *School of Com-*

puting, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, 2010.

[42] *http://www.cpc.ncep.noaa.gov* website of Climate Prediction Center database of the National Oceanic and Atmospheric Administration.