POLITECNICO
MILANO 1863

School of Civil, Environmental and Land Engineering
Course of Natural Resources Management

Project work: design of a forecast model of the Danube River streamflow and impacts of damming up in Gonyu

Group 9:

Sara D'Alessandro
Francesca Rampinelli
Giorgia Sica
Federico Staffa

Professors:
Andrea Francesco Castelletti
Matteo Giuliani
Matteo Sangiorgio

INDICE

1 INTRODUCTION

1.1 CASE STUDY

The Danube is one of the longest rivers in Europe, with its 2860 km it flows through ten countries of Central Eastern Europe. In this case of study we focus on the location of Gonyu *[Figure 1]*, placed in north-western Hungary and on the border with Slovak Republic.

This area plays a crucial role for agriculture and industries, demanding a large amount of water.

Figure 1: Location of case study

The purposes of the project are:

- to build a one day ahead forecast model of the river streamflow, trying a wide arrangement of modelling techniques;
- to discuss the impacts of damming up the river.

1.2 PRELIMINARY ANALYSIS

The available data is comprised of 27 years long historical time series of respectively the daily stream-flow of the river [m^3/s], daily precipitation [mm] and daily temperature [°C] *[Figure 2]*.

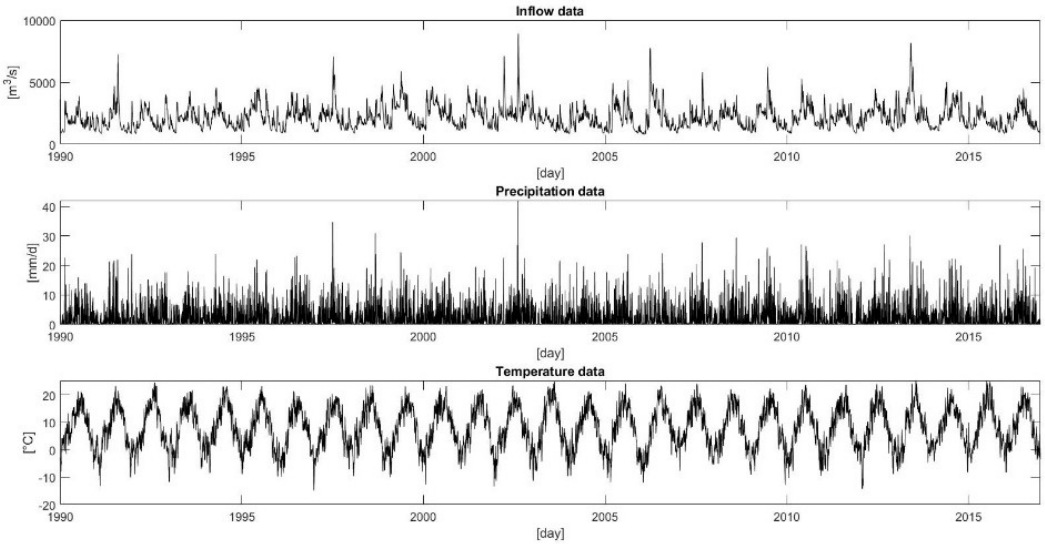


Figure 2: Data's time series

The natural inflow follows a repeating pattern along every year, which can be noted through the cyclo-stationary mean, reported in *Figure 3*.

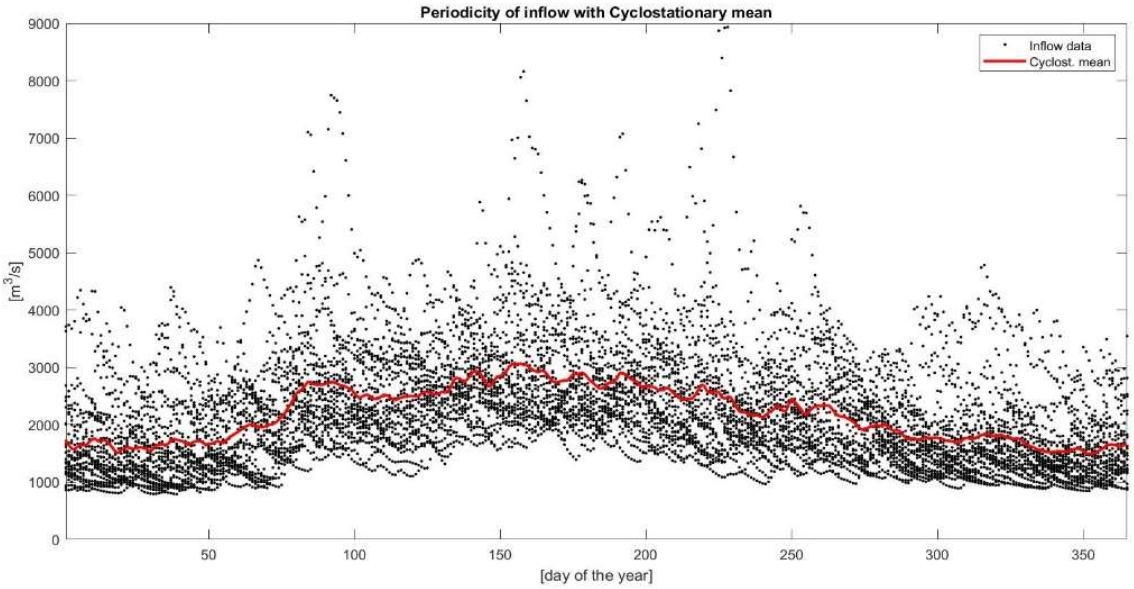


Figure 3: Cyclo-stationary mean trend

A more elaborate statistic is provided by the cyclo-stationary mean with moving average, where the daily mean is computed considering a window of a few days around the particular day. This yields a smoother representation. [*Figure 4*]

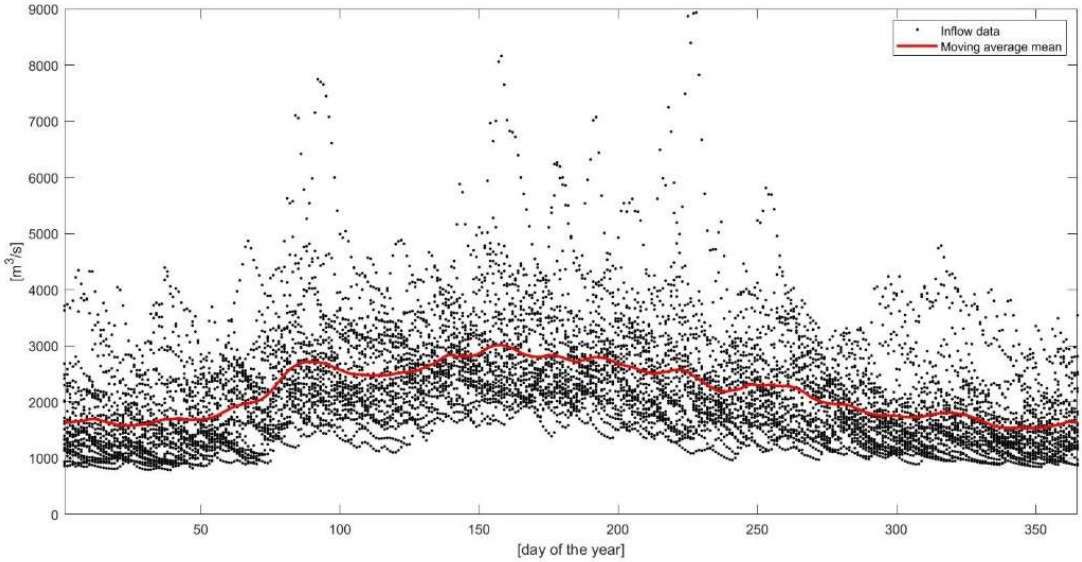


Figure 4: Moving average mean

When a phenomena is characterised by such a strong cyclo-stationarity, it is important to remove the seasonal trend, so the forecast model can correctly capture the variability of the data.

We then proceeded to remove this trend from all the data [*Figure 5*] with the following formula:

$$x_t = \frac{q_t - m_q}{s_q}$$

where x_t is the deseasonalized streamflow at time t, q_t is the observed streamflow at time t and m_q and s_q are respectively the cyclo-stationary mean and cyclo-stationary standard deviation.

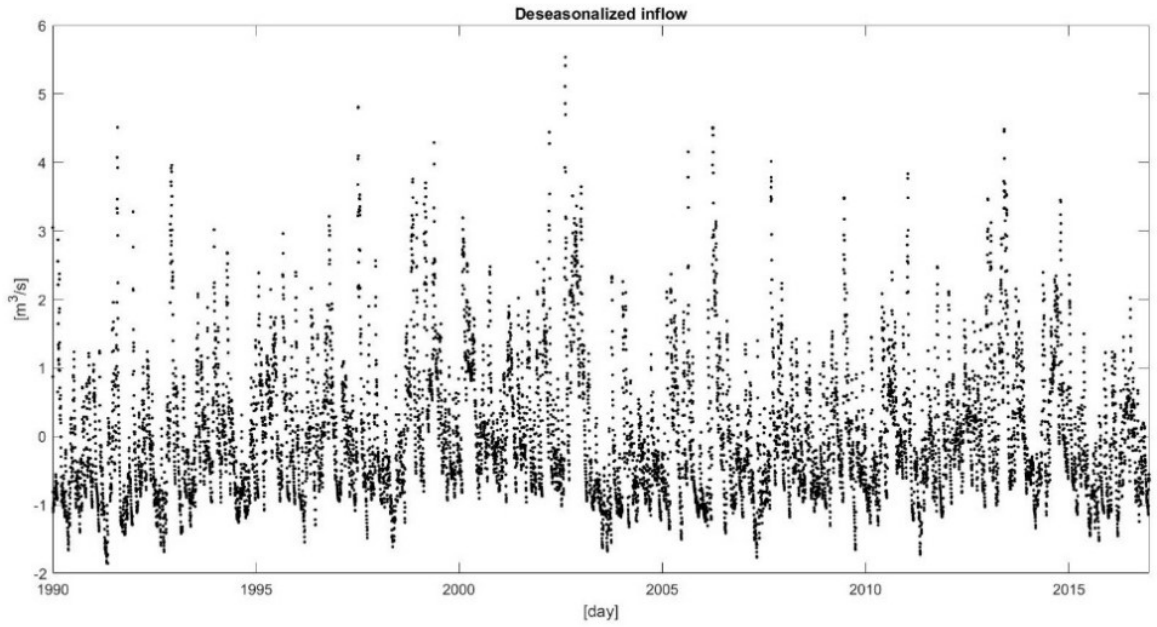


Figure 5: De-seasonalized inflow data, to be estimated

Finally, we evaluated the autocorrelation of the inflow [*Figure 6*].

This is an index of correlation at different time steps: we can see that even after 20 time steps it doesn't fall within a confident interval around 0, meaning the process is not totally random and can be estimated with autoregressive *(AR)* models.

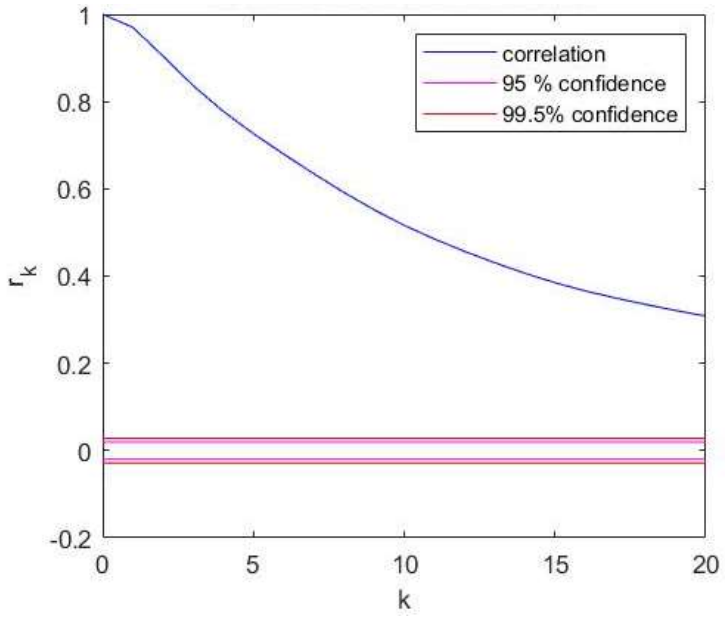


Figure 6: Autocorrelation of inflow data

2 DATA-DRIVEN FORECASTING MODEL

2.1 LINEAR MODELING

We started with the simplest models, these being *AR* and *ARX* models.

Being linear, they don't necessitate of big computational times to calibrate, even on fairly large datasets as the one available to us, although they generally have medium-low performances in reproducing the data.

We also decided to implement a basic cross validation procedure to scrutinize whether or not different splitting of the dataset in calibration and validation datasets could influence performances.

2.1.1 Autoregressive AR models

Firstly we investigated the effect of different splitting by checking performances of an *AR(1)* model changing a parameter *k* at every run.

Particularly, the dataset is divided in *k* sub-datasets of equal length, then the calibration/validation procedure is repeated *k* times using every time one of the different sub-sets as a validation dataset and the other *k-1* together as a calibration dataset. We've tried some different k values and we obtained the best performances with k = 3.

Once identified *k*'s value, the particular best instance of the calibration/validation dataset split is also chosen and kept constant for all the successive identifications done in the project.

And based on the R^2 and MSE values [*Figure 7*], we took split #3 as the adopted instance of calibration and validation datasets.

Split #2 was disqualified because even if it had objectively the highest R^2 score, higher of the R^2 scored in calibration, it also had the highest MSE across the splits.

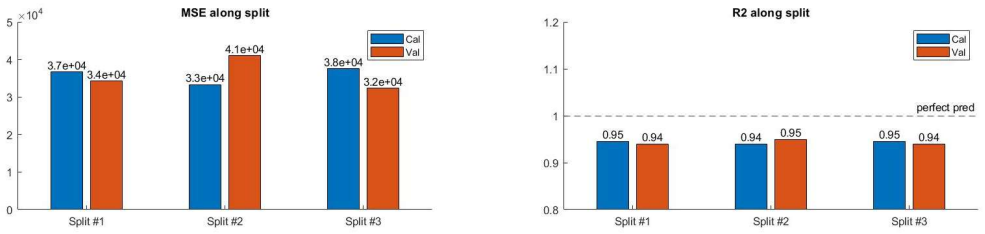


Figure 7: Scores of MSE and R^2 along the split

Next up, for the iterative identification of *AR(i)* models we defined max model order *i_max* = 10.

We can see in *Figure 8* that even after *i* = 3 we have a clear settlement of performances, thus a reasonable and optimal choice for the model order of ARs would be of *i* = 3.

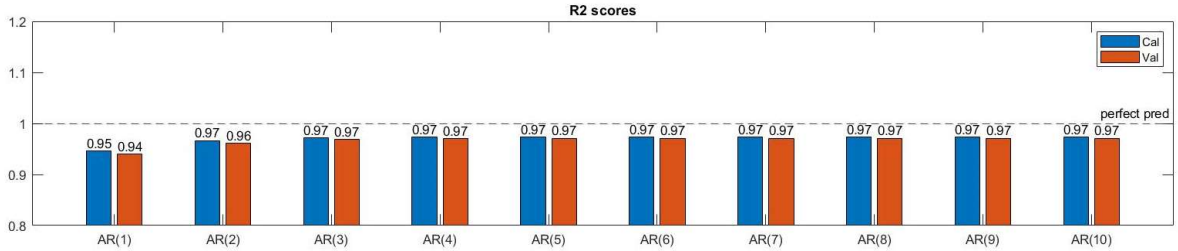


Figure 8: Scores of AR(i)

2.1.2 Autoregressive ARX models

Additionally, to the inflow data, we were provided two exogenous datasets regarding the temperature and the precipitation in the case study's area, these could prove useful in a more accurate modelling through ARX models.

We started by checking performances of *ARX(1,1)* models with either precipitation or temperature data singularly, also either proper or improper, in order to assess their statistical significance in fitting the streamflow data [*Table1*].

Table 1: proper/improper ARX(1,1) tests with the 2 different exogenous datasets

	ARX(1,1) temperature pro	ARX(1,1) temperature imp	ARX(1,1) precipitation pro	ARX(1,1) precipitation imp
R2 calibration	0,946	0,947	0,963	0,949
R2 validation	0,940	0,941	0,959	0,941

We found that the most indicative exogenous dataset is the precipitation one, while the temperature one influences less the performances.

Next, we decided to implement both the temperature and precipitation datasets in an iterative investigation of the best model order *i* of proper/improper *ARX(i,2)* models. [*Figure 9*]

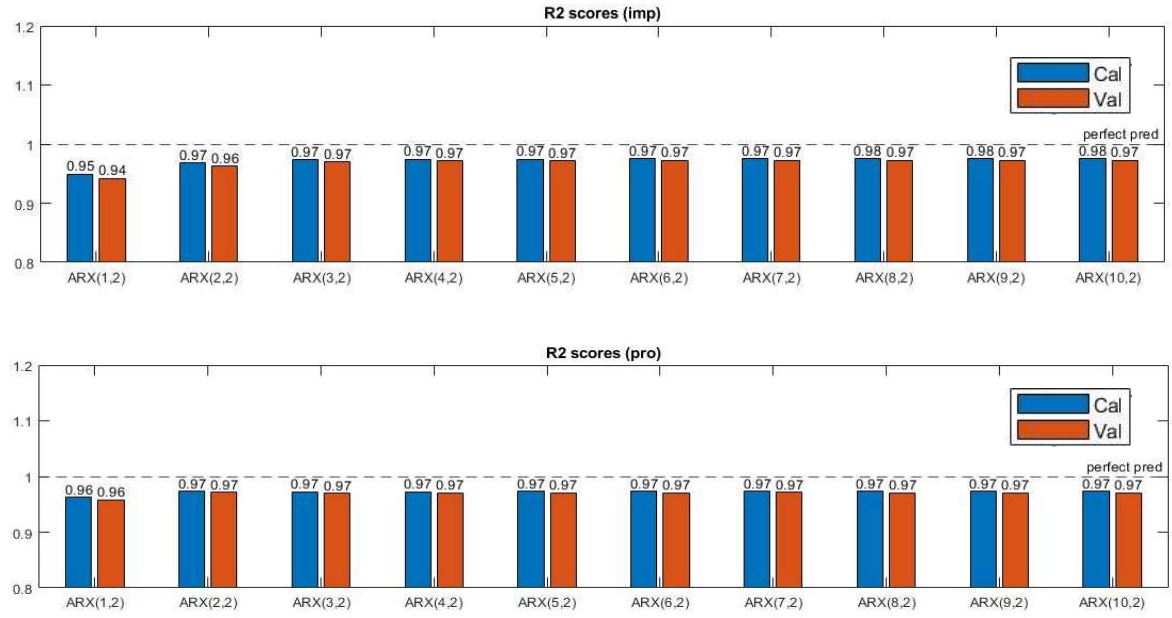


Figure 9: R^2 of ARX(i,2) proper

In this case the improper and proper models have similar performances particularly after i=3, but in the first two orders the proper one seems better than the improper one.

Furthermore, performances mostly stagnate after i = 2 for the proper model and after i = 3 for the improper one, so an ideal choice of model would be ARX(2,2) proper.

2.2 NON LINEAR MODELING

Non-linear models are a powerful tool to fit data generally much better than linear models, this however comes at the cost of increasing complexity and thus requiring much more computational time.

We investigated Artificial Neural Networks (*ANN*), Classification And Regression Trees (*CART*) and Random Forests (RF).

2.2.1 ANNs

We implemented networks with different neuron distribution along layers, these being:

- *n* neurons on one single layer → *Shallow ANN*
- *n* neurons equally distributed on *l* layers → *Deep ANN*

Particularly, we used *n* = 20 and *l* = 2.

Additionally, since the training of ANNs is a non-linear optimization process, we repeat the procedure N times to counteract any possible "bad" initializations, that would lead to bad final performances. However, we found that different initializations have not a large influence on the performances, so we chose arbitrarily one for each ANN model category. As we can see in *Table 2*, an ideal choice of model would be ANN deep proper.

Table 2:R^2 scores for proper/improper Shallow and Deep ANN tests

	ANN shallow pro	**ANN shallow imp**	**ANN deep pro**	**ANN deep imp**
Calibration	0,964	0,948	0,965	0,949
Validation	0,957	0,938	0,955	0,939

2.2.2 CARTs and Random Forests

CART models implement a greedy algorithm in the identification procedure, so they tend to overfit to some degree, this issue is counteracted by:

- Acting on the termination criterion of the identification, tuning identification parameters like *MinLeafSize*, *MaxNumSplits*, *MinSplitSize* and *MaxDepth* etc…
- Implementing a more advanced architecture, like Random Forests (*RF*), which are essentially an ensemble of CARTs.

So, we chose to try different values of Min Leaf Size for both CART and Random Forests models. As we can see in *Table 3* an ideal choice of model would be the CART with min leaf size 50.

Table 3: R^2 scores for CART and RF tests with different Min Leaf Size

	CART mls 10	**CART mls 50**	**CART mls 100**	**RF mls 10**	**RF mls 50**	**RF mls 100**
Calibration	0,971	0,958	0,948	0,957	0,918	0,883
Validation	0,948	0,950	0,942	0,938	0,915	0,888

2.3 FINAL CONCLUSIONS

In *Table 4* we summarized the value of the best performances for each model category.

Table 4: R^2 scores for the best options for each model category

	AR(3)	**ARX(1,1) Precipitation proper**	**ARX(2,2) proper**	**ANN Deep proper**	**CART mlf 50**
Calibration	0,972	0,963	0,973	0,965	0,958
Validation	0,969	0,959	0,972	0,955	0,950

From what emerged, an optimal choice of a data-driven forecasting model for the case study would be a proper *ARX(2,2)*. We found more efficient a data-driven model with both the exogenous information available, given that they improve the performances without increasing the model order.

Additionally, we discovered that neural networks, CARTs and Random Forests lead to performances comparable to those obtained with linear models, despite the theoretical better learning capabilities. The performances of non-linear models would have to be way higher to justify the introduction of more complicated data structures, and the related much longer computational times, so we rejected this type of models.

3 DAMMED ALTERNATIVE

3.1 DAM DESIGN

We proceeded with the second purposes of the project, discuss the impacts of damming up the river,

firstly defining the dimension of the reservoir.

We used the Sequent Peak Analysis method (SPA) to compute the optimal storage of the dam, by means of searching the maximum total storage needed to satisfy a specified desired release along the temporal domain of the available data [*Figure 10*]. This is a more realistic procedure than the Rippl method as it exploits the reservoir's dynamics simulation for the computation.

However, during the simulation of the natural case the release has been considered equal to the inflow, given the lack of a natural basin.

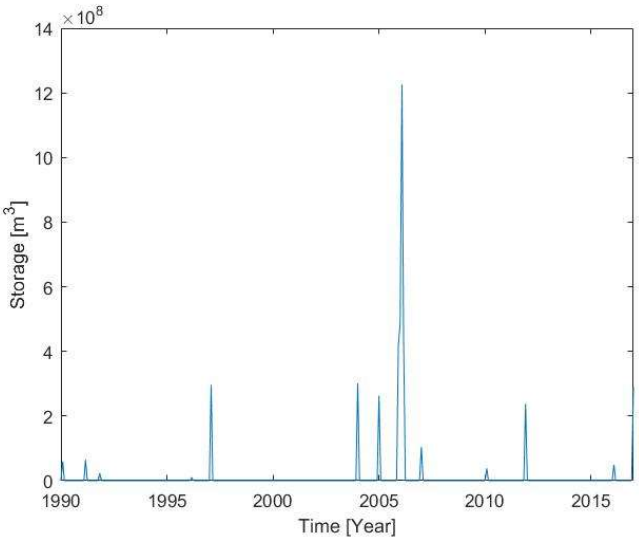


Figure 10: Storage needed for supporting the demand

The water demand downstream of the dam is arbitrarily set to the *10th percentile* of the historical sequence of the inflows, meaning 1170 m^3/s. We found that the maximum storage needed to support this demand is equal to $1{,}23 \times 10^9$ m^3 (1,23 Mm^3) which is set to be the capacity of the dam itself.

The surface was arbitrarily assumed as 50 km^2, based on an existing dam in the same area. This is an approximative assumption justify by the difficulty in finding structures with similar capacity and physical characteristics in the domain considered.

Then, under the hypothesis of cylindrical lake, the corresponding level of the active capacity is 24,5 m. [*Table 5*]

Table 5: Dam's identified features

Capacity [Mm³]	1,23
Surface [km²]	50
Heigh [m]	24,5

3.2 INDICATORS

They are a number associated to every stakeholder influenced by the management of the dam allowing to compare the degree of the stakeholders satisfaction of every adopted control policy.

The first indicator concerns the farmers and it is evaluated as the average of the squared daily water deficit, where the power of 2 is implemented to consider the risk aversion of the stakeholders.

$$I_{irr} = \frac{1}{H}\sum_{t=1}^{H}[(w - x_t)^+]^2$$

Where:

$(w - x_t)^+ = \begin{cases} w - x_t & \text{if } w > x_t \\ 0 & otherwise \end{cases}$

w = Water demand

x_t = Release at time t

H = Number of observations

The second indicator concerns the riparians and it is evaluated as the annual average number of the days with the occurrence of the dam flooding in the area, considering the flooding level at 24 [m].

$$I_{flo} = \frac{1}{N}\sum_{t=1}^{H}\Gamma(h_t \geq h_F)$$

Where:

$\Gamma(h_t \geq h_F) = \begin{cases} 1 & if\ h_t \geq h_F \\ 0 & otherwise \end{cases}$

N = Number of years of the domain

H = Number of observations

h_t = Height of the lake at time t

h_f = Height of the lake over which a flood event happens

In the natural case (i.e. the alternative 0) this indicator is considered equal to zero as there's no actual basin, furthermore, zero is the ideal value of the indicator.

The third indicator concerns the environment and it is evaluated as the difference of days with level lower than a Minimum Environmental Flow (MEF) in the regulated and natural case.

$$I_{env} = \left|(I_{env_{reg}} - I_{env_{nat}})\right|$$

Where:

$I_{env_{reg}} = \frac{1}{N}\sum_{t=1}^{H}\Gamma(x_{t\_reg} < MEF)$

$I_{env_{nat}} = \frac{1}{N}\sum_{t=1}^{H}\Gamma(x_{t\_nat} < MEF)$

$\Gamma(x_{t\_reg} < MEF) = \begin{cases} 1\ if\ x_{t\_reg} < MEF \\ 0\ otherwise \end{cases}$

$\Gamma(x_{t\_nat} < MEF) = \begin{cases} 1\ if\ x_{t\_nat} < MEF \\ 0\ otherwise \end{cases}$

H = Number of observations

N = Number of years of the domain

x_{t_reg} = Regulated release at time t

x_{t_nat} = Natural release at time t

MEF = 25th percentile of the inflows time series, set as threshold

As in the previous indicator, it is considered equal to zero for the natural case. Furthermore, zero is the ideal value of the indicator because low level events caused by the dam should be as similar as possible to those happening naturally.

All the indicators must be minimized.

3.3 POLICY SHAPE

The regulation policy of the dam is designed to satisfy the requirements of all the stakeholders, it's a piecewise linear function, defined by 4 parameters [*h1, h2, m1, m2*]

- When the lake level is below the level *h1* the release is a linear function passing through the origin of the system and with gradient *m1*. The release is lower than the demand *w* because of the necessity to save water.
- Between the levels *h1* and *h2* the release is equal to the water demand of the agriculture sector.
- When the lake level is above the level *h2* the release is a linear function with gradient *m2*. The release is higher than the demand to prevent flood events.
- When the lake level exceeds the height of the dam the release is equivalent to the maximum natural release.

The maximum natural release is modelled as a line interpolating 0 and the point C, identified by the dam's height h = 24,5 m and the maximum historical inflow 8932,1 m^3/s [*Figure 11*].

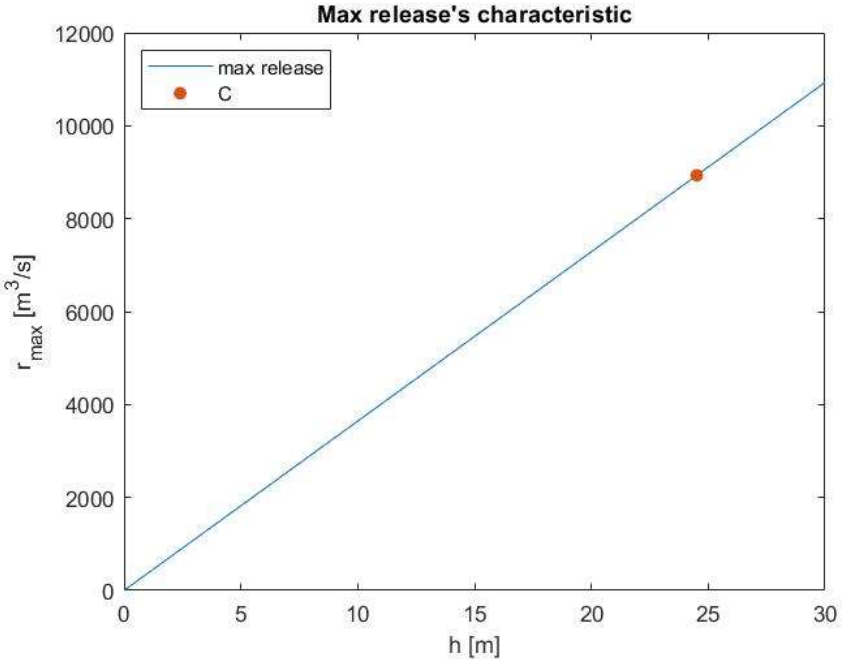


Figure 11: Dam's maximum release

3.4 EMODPS

In order to obtain the optimal policy, we used NSGAII, a genetic algorithm that searches for the optimum by mimicking the process of natural selection. Starting from an initial population, the algorithm selects the best solutions according to the pareto dominance criterion, then a new population is generated, and the process is iterated for a fixed number of generations.

The solution found is suboptimal because of the iterative nature of the algorithm.

NSGAII is initialized by setting a population of 40 individuals and the number of generations at 20. The fixed parameters of the policy are given as input with the ranges of the parameters to be optimized, arbitrarily set as:

- *h1* [0 – 10 m]
- *h2* [10 – 24 m]
- *m1* [100 – 900 m^2/s]
- *m2* [500 – 1000 m^2/s]

The algorithm generates a three-dimensional Pareto frontier [*Figure 12*].

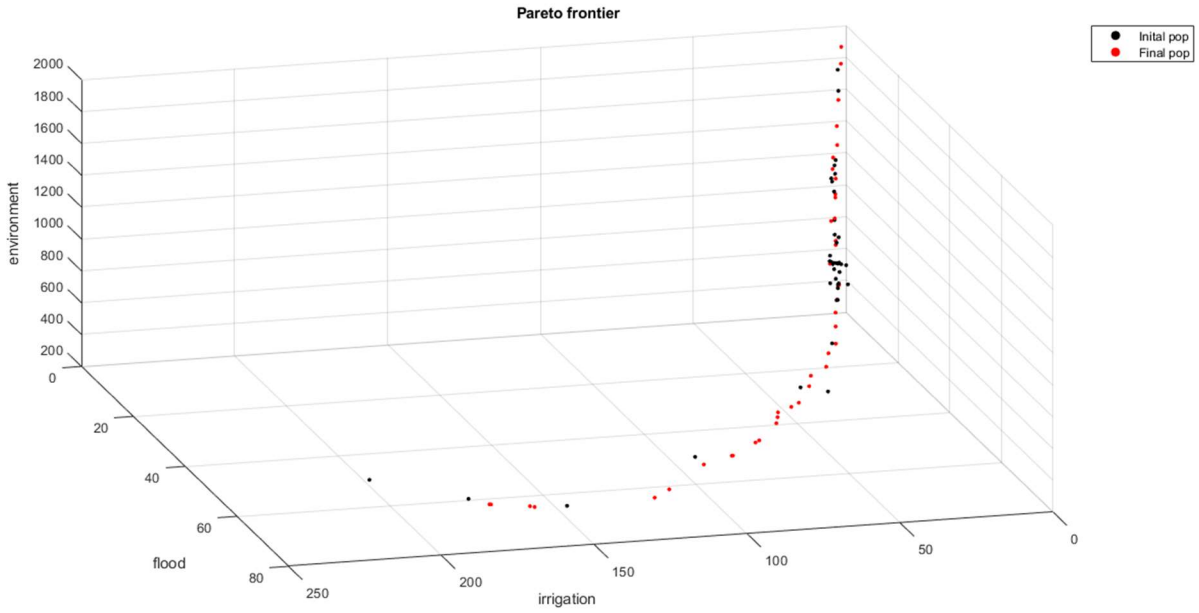


Figure 12: Pareto frontier generated with EMODPS

The black points are the initial population (selected randomly by the algorithm), while the red ones are the final population. The two population are not so clearly separated possibly because the number of generations of the algorithm isn't large enough.

It is also clear that the stakeholders related to irrigation and environment mostly agree, while the one related to floods is in strong opposition to them.

3.5 SELECTED ALTERNATIVES

To conclude our analysis, since the considered objectives are conflictual, meaning there's no clear optimal solution for all the stakeholders considered, we selected a few promising alternatives of dam operation.

- **A1:** best alternative for irrigation;
- **A2:** best alternative for flood mitigation;
- **A3:** best alternative for environment;
- **A4:** a possible solution of compromise between the three objectives.

These solutions are compared between each other and with the Business As Usual (**A0**), the alternative defined as the base case in which only the river is present and the reservoir is not built [*Table 6 and Figure 13*].

Table 6: Scores of the indicators for the different alternatives

Indicators	A0	A1	A2	A3	A4
I_{irr} $[(m^3/s)^2/days]$	2899.5	260.19	1177.79	1870.1	408.93
I_{flo} [days]	0	65.86	0.037	0.148	1.963
I_{env} [days]	0	171.80	4.306	1.713	3.973

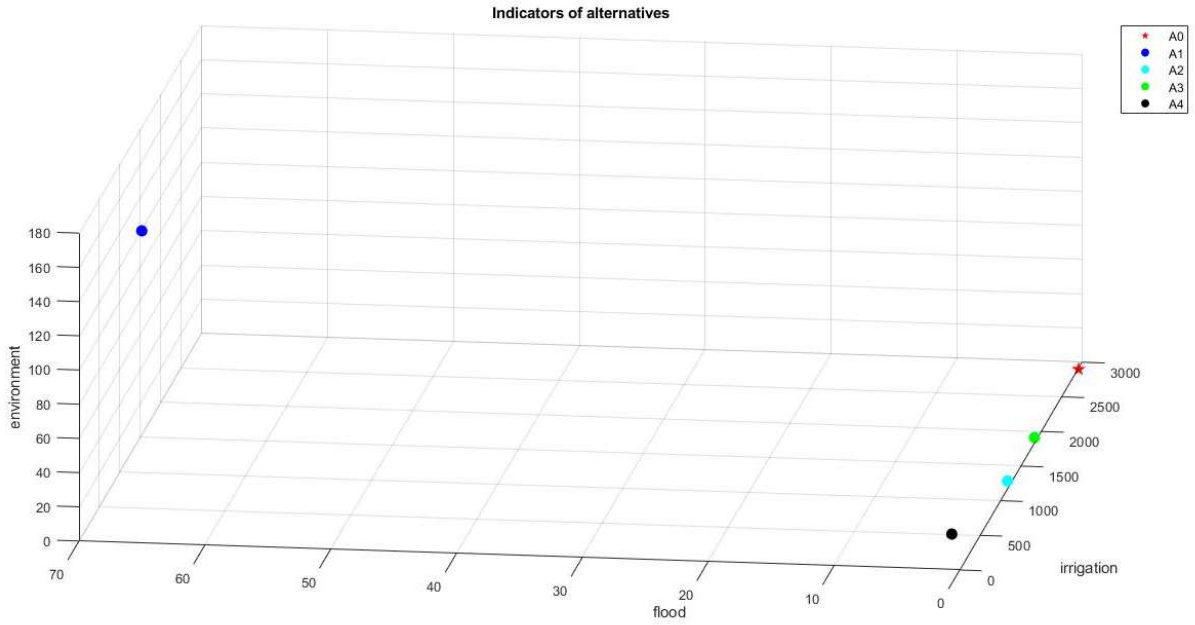
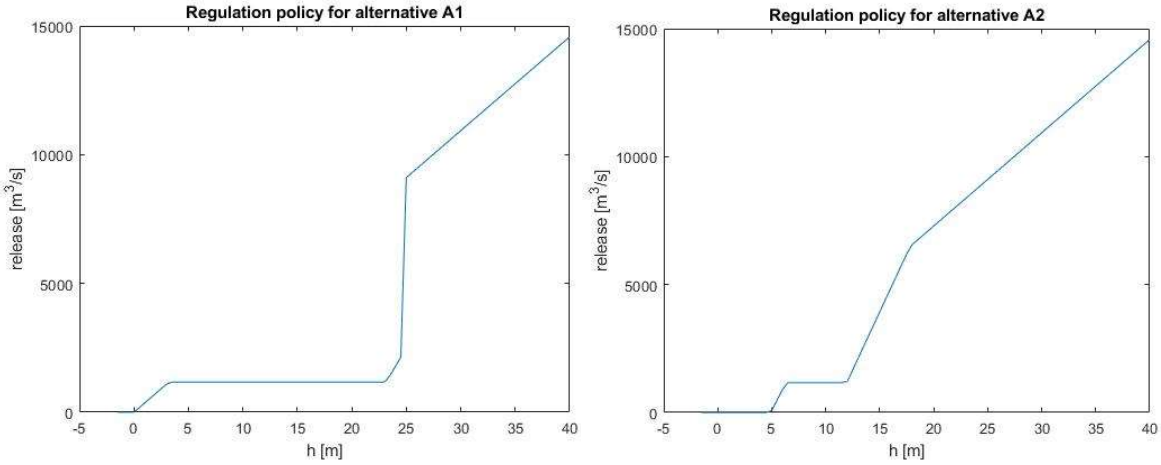


Figure 13: Indicators in 3D objective space

While *Figure 14* shows the regulation policies associated to the alternative considered.



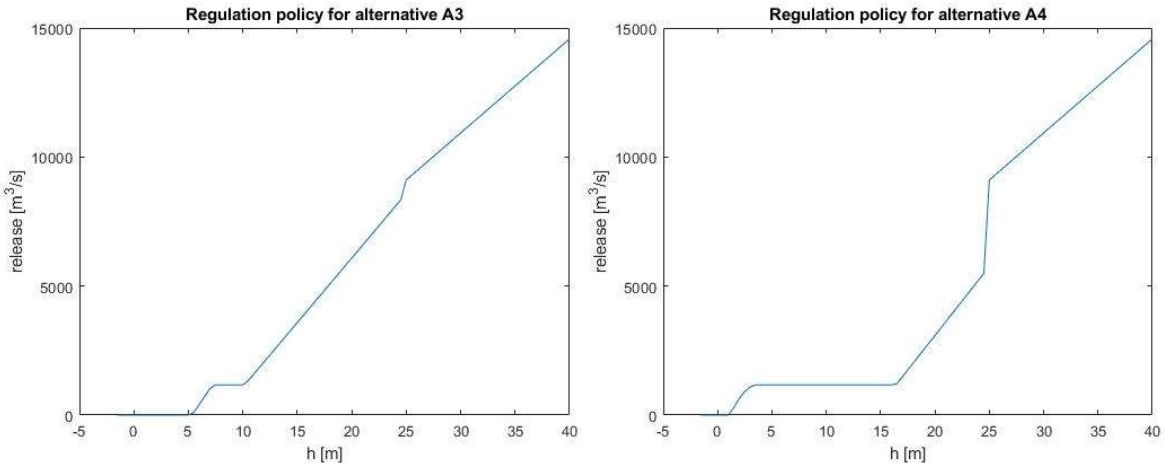


Figure 14: Regulation policy for each alternative

3.6 FINAL CONCLUSIONS

Ultimately, the selected compromise alternative A4 seems to be an ideal operating policy.

Comparing it with A0, the red star in *Figure 13*, we can see that we can effectively increase a lot the satisfaction of the irrigation stakeholders only slightly sacrificing performances of flooding and environment indicators.

In conclusion, from the perspective of the farmers, damming up the river is favourable while for the other stakeholders it is unprofitable. A reasonable solution could be offering a compensation/mitigation in order to make also them supportive on building the dam.