An Ensemble Method for Time Series Learning

Francesco Masulli - University of Pisa, Italy

in collaboration with
Daniela Baratta - INFM Genoa, Italy
Giovambattista Cicioni - CNR Rome, Italy
Léonard Studer - Université de Lausanne, Switzerland

Sept 2002

Outline

1. Time series learning
2. Hints from Dynamical Systems Theory
3. Ensemble Method based on Singular Spectrum Analysis Decomposition
4. Application to Rainfall Forecasting
5. Conclusions

Time series learning

Learning a mapping on the basis of a (possibly small) data set of examples is an ill-posed inverse problem (Haykin99).

Concerning temporal time series learning, noise, ambiguity of the mapping, and discontinuity of the signal affect the generalization performance of the learning machines.

A popular way to reduce ill-posedness in temporal data learning consists in assuming an input scale (Dietterich90) suitable to alleviate the mapping ambiguity problem.

To this aim we should find the optimal dimension of the input vector and the time lag between its elements.

After the setting of the mapping input vector and of other design issues, the temporal data can be learned by a machine.

Accurate learning of a continuous mapping is supported by the Universal Function Approximation property holding for some classes of learning machines including, e.g., Multi Layer Perceptrons, Radial Basis Functions Nets, and Fuzzy Basis Functions Nets (Cybenko89, Poggio90, Wang92).

However, for small data set, simple learning machines exhibit better generalization capabilities (Vapnik95).

In this talk: constructive framework for the design of time series learning machines, proposed in (Masulli95, Masulli99, Haykin98)

In particular:

- apply results and prescriptions related to the delay-embedding theorem (Takens81, Mane81) to the design of learning machines of continuous mappings of temporal data.
- decompositive ensemble method based on the Singular-Spectrum Analysis (SSA) (Vautard92) in order to extend the constructive approach to the learning of discontinuous and/or intermittent signals (Masulli99, Masulli2000).

Successful applications to the design of learning machines for:

- simulated non-linear and chaotic signal prediction (Masulli95)
- system identification (Masulli99)
- daily rainfall forecasting (Masulli2000, Masulli02).

Hints from Dynamical Systems Theory

State Space

A deterministic dynamical system is described by a set of differential equations.

Its evolution is represented by the trajectory in state space (of dimension n) of the vector $\mathbf{Q} = (x, \dot{x}, y, \dot{y}, z, \dot{z}, \ldots)^{\top}$

where $x, \dot{x}, y, \dot{y}, z, \dot{z}, \ldots$ are the variables of the system and their derivatives.

The figure made in state space by $\mathbf{Q}$ is the attractor of the system.

For non-linear systems, the dynamical variables $(x, y, z \ldots)$ are coupled.

The evolution of one variable (let say x) is not independent of all the other ones $(y, z \ldots)$.

Except for few simple phenomena, the set of differential equations is unknown.

Even, often the whole set of relevant effective dynamical variables is not always well defined.

But, as the variables are interdependent, the observation of only one of those brings information — maybe in an implicit way — on the other ones and consequently on the complete dynamical system.

This is the reason why time series of nonlinear dynamic systems are so useful.

Embedding Theorem

The question is now: "How to reconstruct the complete dynamical system with only the one-variable time series $(s_1, s_2, s_3, \ldots)$?"

The *Embedding Theorem* proposed independently in 1981 Takens and Mañé gives an answer to the above question.

In the Takens-Mañé theorem we consider an augmented vector **S** built with *d* elements of the time series.

The dimension of the vector *d* has to be greater than two times the box-counting dimension D_0 of the attractor of the system:

$$d > 2D_0 \tag{1}$$

A vector **S** satisfying the Takens-Mañé bound cited in the previous paragraph will evolve in a reconstructed state space, and its evolution will be in a diffeomorphic relation with the original **Q** state space point (a diffeomorphism is a smooth one-to-one relation).

In other words, for every practical purposes the evolution of **S** is a fair copy of the evolution of **Q**.

It is worth noting that: there is a distinction between the *order of the differential equation* (n) which is the dimension of the state space where live the true state vector **Q** and the *sufficient dimension of a reconstructed state space* (d) where the reconstructed vector **S** lives.

An Example

In order to elucidate the Embedding Theorem,

let consider a sine wave $s_t = A\, sin(t)$.

In d=1 (i.e. the s_t space) this wave oscillates in the interval $(-A, +A)$.

Two points which are close in the sense of Euclidean (or other distance) may have quite different values of $\dot{s}(t)$.

In this way two ”close” points may move in opposite directions along the single spatial axis.

In a two dimensional space (s_t, s_{t+T}), where T is a time lag, the ambiguity of the dynamics of points is resolved.

The system evolves on a figure (in general an ellipse) that is topologically equivalent to a circle.

If we draw the sine wave in the three dimensions (s_t, s_{t+T}, s_{t+2T}), no further unfolding occurs and the sine is represented as a new ellipse.

The Method of Embedding

In order to reconstruct the dynamical system we can use the *time delay embedding method* (Abarbanel96).

This method consists in building d-dimensional state vectors $\mathbf{S}_i = \left(s_i, s_{i+T}, \ \ldots \ , s_{i+(d-1)T}\right)$.

In principle, it suffices that $d \geq n$.

But, the *effective* dimension d is not directly related to the dynamical dimension n $-$ as in the case of weak coupled variables.

We must choose:

-> T (time lag)
-> d

Choosing the time delay

The time delay T (or *time lag*) used in the embedding has to be chosen carefully.

If it is too long, the samples $s_i, s_{i+T}, \ldots, s_{i+(d-1)T}$ are not correlated and then, in general, the dynamical system can not be reconstructed.

This happens in particular for chaotic systems, for which even two initially close chaotic trajectories will diverge exponentially in time.

If time delay T is too short, every sample is essentially a copy of the previous one, bringing very little information on the dynamical system.

First Minimum of Average Mutual Information

We use the Shannon's mutual information to quantify the amount of information shared by two samples in order to get an useful estimation of the time lag T.

Let's defined the *average mutual information* between measurements a_i drawn from the set A and measurements b_i drawn from set B.

The set of measurements A is made of the values of the observable s_i and the set B is made of the values s_{i+t} (t is a time interval).

Average mutual information is then :

$$I(t) = \sum_{s_i \in A, s_{i+t} \in B} P(s_i, s_{i+t}) \times log_2 \frac{P(s_i, s_{i+t})}{P(s_i)P(s_{i+t})}, \qquad (2)$$

where $P(\ldots)$ are probabilities distributions based on frequency observations.

It has been suggested (Fraser86, Fraser89, Vastano89, Abarbanel96) to take the time T, where the first minimum of I(t) occurs, as the value to use at the time delay in the phase space reconstruction.

In this way the values of s_n and s_{n+T} are the most independent of each other in an information-theoretic sense.

Moreover the first minimum of average mutual information is a good candidate for the interval between the components of the state vectors that will be input to the neural network model of the non-linear dynamical process.

Evaluating the Global Embedding Dimension

From the Embedding Theorem, the box counting dimension D_0 should be evaluated.

In principle, it can be estimated directly from the time series itself, but this task is very sensitive to the noise and needs large set of data points (order of 10^{D_0} data points) (Abarbanel96).

In order to avoid those problems, we can estimate the *embedding dimension* d_E, defined as the lowest (integer) dimension which unfolds the attractor,

i.e. the minimal dimension for which foldings due to the projection of the attractor in a lower dimensional space are avoided.

The embedding dimension is a *global* dimension and in general is different from the local dimension of the underlying dynamics.

The Embedding Theorem guarantees that if the dimension of the attractor is D_0, then we can unfold the attractor in a space of dimension d_E ($d_E > 2D_o$).

It is worth noting that d_E is not a necessary condition for unfolding, but is sufficient.

The dimension of input layer of the Multi-Layer Perceptron will be then of dimension high enough in order that the deterministic part of the dynamics of the system is unfold.

Global False Nearest Neighbors

In practice, the method of *Global False Nearest Neighbors* proposed by Abarbanel (1996), can be used to evaluate the embedding dimension d_E.

Given a data space reconstruction in dimension d, with data vectors $\mathbf{S}_i = (s_i, s_{i+T}, \ldots, s_{i+(d-1)T})$,

where the time delay T is the first minimum of average mutual information(Eq. 2).

Let be $\mathbf{S}_i^{NN} = (s_i^{NN}, s_{i+T}^{NN}, \ldots, s_{i+(d-1)T}^{NN})$, the nearest neighbor vector in phase space.

If the vector $\mathbf{S}_i^{NN}$ is a *false* neighbor (FNN) of $\mathbf{S}_i$, having arrived in its neighborhood by projection from a higher dimension because the present dimension d does not unfold the attractor, then by going to the next dimension $d+1$ we may move this point out of the neighborhood of $\mathbf{S}_i$.

We define the distance ξ between points when seen in dimension $d+1$ relative to the distance in dimension d as

$$\xi_i \equiv \sqrt{\frac{R_{d-1}^2(i) - R_d^2(i)}{R_d^2(i)}}, \tag{3}$$

then

$$\xi_i = \frac{\left| s_{i+dT} - s_{i+dT}^{NN} \right|}{R_d(i)}. \tag{4}$$

As suggested by Abarbanel(1996), $\mathbf{S}_i^{NN}$ and $\mathbf{S}_i$ can be classified as a false neighbor if ξ_i is a number greater than a threshold θ ($\xi_i \geq \theta$).

In many applications a good value for θ is 15.

In case of clean data from a dynamical system, we expect that the percentage of FNNs will drop from nearly 100% in dimension one close to zero when d_E is reached.

It is worth noting that, as we go to higher dimensional spaces the volume available for data grows as the distance to the power of dimension, and no near neighbor will be classified close neighbor.

In this case we can modify the Eq. 4 as

$$\xi_i = \frac{\left| s_{i+dT} - s^{NN}_{i+dT} \right|}{R_A}, \tag{5}$$

where A is the nominal “radius” of the attractor defined as the Root Mean Square (RMS) error value of data about its mean, e.g.:

$$R_A = \frac{1}{N} \sum_{i=1}^{N} \left| s_i - s_{av} \right|, \tag{6}$$

$$s_{av} = \frac{1}{N} \sum_{i=1}^{N} s_i. \tag{7}$$

In (Montarsolo98) a very efficient implementation of FNN algorithm is presented. This algorithm is based on the work by Nene and Nayar (Nene97).

It is worth noting that there are two main arguments that can suggest to size the input layer of a predictor based on MLPs smaller than the evaluation obtained using the FNN method.

In fact this evaluation is still an upper bound, and moreover for an assigned size of the training set, a limitation of the complexity of the learning machine can lead to better generalization.

Bells Whistles and Pitfalls of FNN

- The global FNN calculation is simple and fast.
- The FNN calculation applied to signals coming from two different outputs of the same dynamical system gives, in general, two different values of d_E. Then from each signal we will obtain different reconstructed coordinate systems, but both consistent with the original dynamical system.
- FNN method is valid even if the signal of interest results from a filtered output of a dynamical system (Abarbanel96, Dave97).

- If the signal is contamined by noise (assumed to be generated by an high dimensional system), it may be that the contamination will dominate the signal of interest and FNN will show the dimension required to unfold the contamination. Here, a simple byproduct of FNN calculation is an indication of noise level in a signal.

Ensemble Method based on Singular Spectrum Analysis Decomposition

Singular Spectrum Analysis

The methodology described in the previous section has been successfully applied in the design of Multi-Layer Perceptrons and Neuro-Fuzzy systems to

- forecasting of simulated non-linear and chaotic systems (Masulli97, Masulli97)
- real world problem such as the modeling of the vibration dynamic of a real system consisting in a 150 MW Siemens steam turbine (Masulli99).

The proposed methodology can not be directly applied to forecasting discontinuous or intermittent signals, as the universal function approximation theorems for neural networks (Cybenko89) and fuzzy systems (Wang92b) require the continuity of the function to be approximate.

In order to avoid the effect of discontinuities of a signal we can apply the Singular-Spectrum Analysis (SSA) (Kumaresan80, Pike84, Vautard92, Lisi95) to the signal to be forecasted.

In SSA the state vector $\mathbf{S}_i = (s_i, s_{i+1}, \ldots, s_{i+M-1})$ is a temporal window (augmented vector) of the series s, made up by a given number of samples M.

The cornerstone of SSA is the Karhunen-Loève expansion or Principal Component Analysis (PCA) that is based on the eigenvalues problem of the lagged covariance matrix Z_s.

Z_s has a Toeplitz structure, i.e. constant diagonals corresponding to equal lags:

$$\begin{pmatrix} c(0) & c(1) & . & . & . & c(M-1) \\ c(1) & c(0) & c(1) & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & c(1) \\ c(M-1) & . & . & . & c(1) & c(0) \end{pmatrix} \tag{8}$$

In absence of prior information about the signal it has been suggest (Vautard92) to use the following estimate for Z_s:

$$c(j) = \frac{1}{N-j} \sum_{i=1}^{N-j} s_i s_{i+j} \tag{9}$$

The original series can be expanded with respect to the orthonormal basis corresponding to the eigenvectors of Z_s

$$s_{i+j} = \sum_{k=1}^{M} p_i^k u_j^k, \quad 1 \leq j \leq M, \quad 0 \leq i \leq N - M \tag{10}$$

where p_i^k are called *principal components* (PCs) and the eigenvectors u_j^k are called the *empirical orthogonal functions* (EOFs), and the orthornomality property

$$\sum_{k=1}^{M} u_j^k u_l^k = \delta_{jl}, \quad 1 \leq j \leq M, \quad 1 \leq l \leq M \tag{11}$$

holds.

It is worth noting that SSA does not resolve periods longer than the window length M.

Hence, if we want to reconstruct a strange attractor, whose spectrum includes periods of arbitrary length, the large M the better, avoiding to exceeding $M = \frac{N}{3}$ (otherwise statistical errors could dominate the last values of the autocovariance function).

In (Vautard89, Ghil91, Vautard92, Keppenne93, Lisi95, Ghil97) many applications of Singular Spectrum Analysis have been presented, including noise reduction, detrending, spectral estimate, and prediction.

Concerning the application of SSA to prediction, that is the main interest of the present paper, it is supported by the following argument:

Since the PCs are filtered version of the signal and typically band-limited, their behavior is more regular than that of the raw series s, and hence more predictable.

Vautard and Ghil (1992) fit an autoregressive (AR) model for each individual PC using the AR coefficient estimate of Burg (1978), while Lisi, Nicolis and Sandri (1995) used Multi-Layer Perceptrons in order to estimate filtered version of the raw signal using obtained using SSA.

In order to reduce the computational costs we decompose the raw series *s* in reconstructed waves corresponding to SSA subspaces equivalent to similar explained variance and we predict them using Multi-Layer Perceptrons combined with independent evaluation of time lag using the first minimum of mutual information and embedding dimension using False Nearest Neighbors method.

Reconstructed components and reconstructed waves

Following Vautard and Ghil (1992), suppose we want to reconstruct the original signal s_i starting from a SSA subspace A of k eigenvectors.

By analogy with Eq. 10, the problem can be formalized as the search for a series $\hat{s}$ of length N, such that the quantity

$$H_A(\hat{s}) = \sum_{i=0}^{N-M} \sum_{j=1}^{M} (\hat{s}_{i+j} - \sum_{k \in A} p_i^k u_j^k)^2 \qquad (12)$$

is minimized.

In other words, the optimal series $\hat{s}$ is the one whose augmented version $\hat{S}$ is the closest, in the least-squares sense, to the projection of the augmented series S onto EOFs with indices belonging to A.

The solution of the least-squares problem of Eq. 12 is given by

$$\hat{s}_i = \begin{cases} \frac{1}{M}\sum_{j=1}^{M}\sum_{k\in A} p_{i-j}^k u_j^k & \text{for } M \le i \le N-M+1 \\ \frac{1}{i}\sum_{j=1}^{i}\sum_{k\in A} p_{i-j}^k u_j^k & \text{for } 1 \le i \le M-1 \\ \frac{1}{N-i+1}\sum_{j=i-N+M}^{M}\sum_{k\in A} p_{i-j}^k u_j^k & \text{for } N-M+2 \le i \le N. \end{cases} \qquad (13)$$

When A consists on a single index k, the series $\hat{s}$ is called the kth RC, and will be denoted by $\hat{s}^k$.

RCs have additive properties, i.e.

$$\hat{s} = \sum_{k\in A} \hat{s}^k \qquad (14)$$

In particular the series s can be expanded as the sum of its RCs:

$$s = \sum_{k=1}^{M} \hat{s}^k \qquad (15)$$

Note that, despite its linear aspect, the transform changing the series s into $\hat{s}^k$ is, in fact, non-linear, since the eigenvectors u^k depend non-linearly on s.

If we truncate this sum to an assigned number of RCs, the explained variance of the related augmented vector $\hat{S}$ is the sum of the eigenvalues associated to those RCs, while the estimation of the resulting reconstruction error is the sum of the eigenvalues corresponding to the remaining RCs.

As a consequence, it is suitable to order the RCs following the value of the eigenvalues.

Let be $A_1, A_2, \ldots, A_L$ L disjoint subspaces, then a *reconstructed wave* (RW) Ω_l ($l = 1, \ldots, L$) is defined as

$$\Omega_l = \sum_{k \in A_l} \hat{s}^k, \;\; 1 \leq l \leq L. \tag{16}$$

Then, from Eq.s 15 and 16, one can obtain:

$$s = \sum_{l=1}^{L} \Omega_l, \tag{17}$$

that says that the original series s can be recovered as the sum of all the individual RWs.

Ensemble Method

In order to design a predictor for complex signals, such as discontinuous and/or intermittent signals, we can apply the following approach that combines an unsupervised step and one supervised one, building-up an such a way an ensemble of learning machines:

- Unsupervised decomposition: Using the Singular Spectrum Analysis, decomposes the original signal S in reconstructed waves (RWs), corresponding to subspaces with equal explained variance;
- Supervised learning: Prepares a predictor for each RW using the methodology
- Operational Phase: The prediction of the original signal S is then obtained as the sum of the predictions of individual RWs, i.e. using Eq. 17.

Note1:

It is worth noting that, sometime the most complex waves (in general those corresponding the the low eigenvalues) cannot satisfactory predicted, using the available data.

Following the criteria of the *best prediction* (Lisi95) in the Eq. 17 we can excluded them if, when if enclosed in the sum, make worse the overall prediction.

Note 2:

The proposed ensemble methods is a additive, but each machine learns a different component of the signal

1 Application to Rainfall Forecasting

2 Data Set and Methods

Forecasting of daily rainfall intensities series of 3652 samples each, collected by 135 stations located in the Tiber river basin in the period 01/01/1958 - 12/31/ 1967.

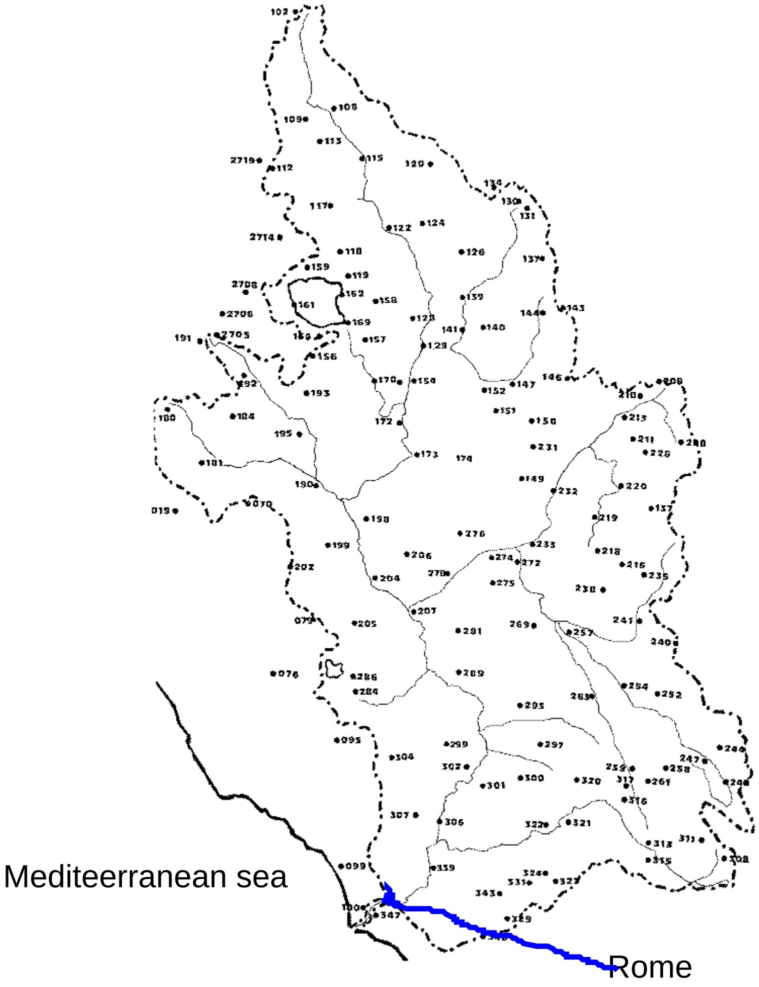


Figure 1: Distribution of the 135 stations in the Tiber river basin.

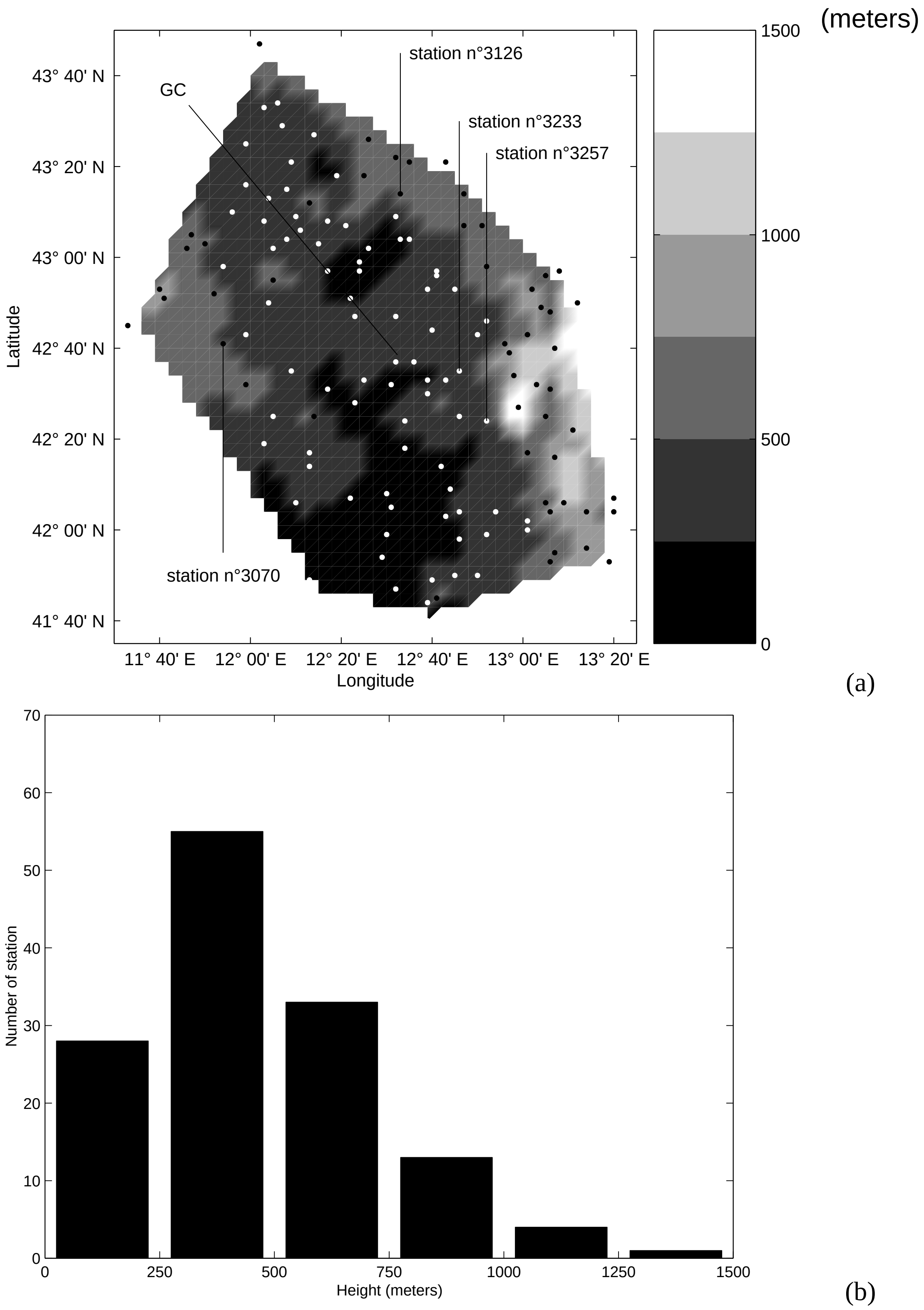


Figure 2: (a) Height map (in meters on the sea level) of the 135 stations on the Tiber river basin. The Geographic Center (GC) of the 135 stations, the two stations more correlated to MS, and the two stations less correlated to it (see Tab. 4) are shown on the map. (b) Histogram of station's height.

The data processing started by considering the series of the Mean Station (MS), defined as the average of all 135 rainfall intensity series (Fig. 3).

In Fig 4 a window on the period 07/01/66 - 12/30/66 is presented in order to better show the discontinuity and intermittence of the studied signal.

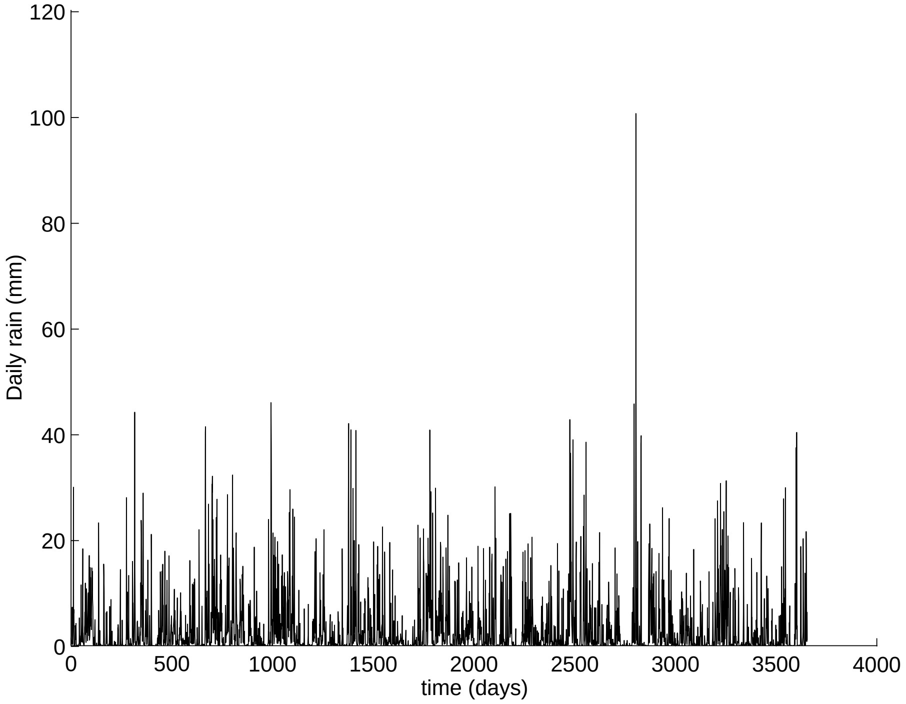


Fig. 3 Mean Station: Daily rain millimeters. Period 01/01/1958 - 12/31/1967.

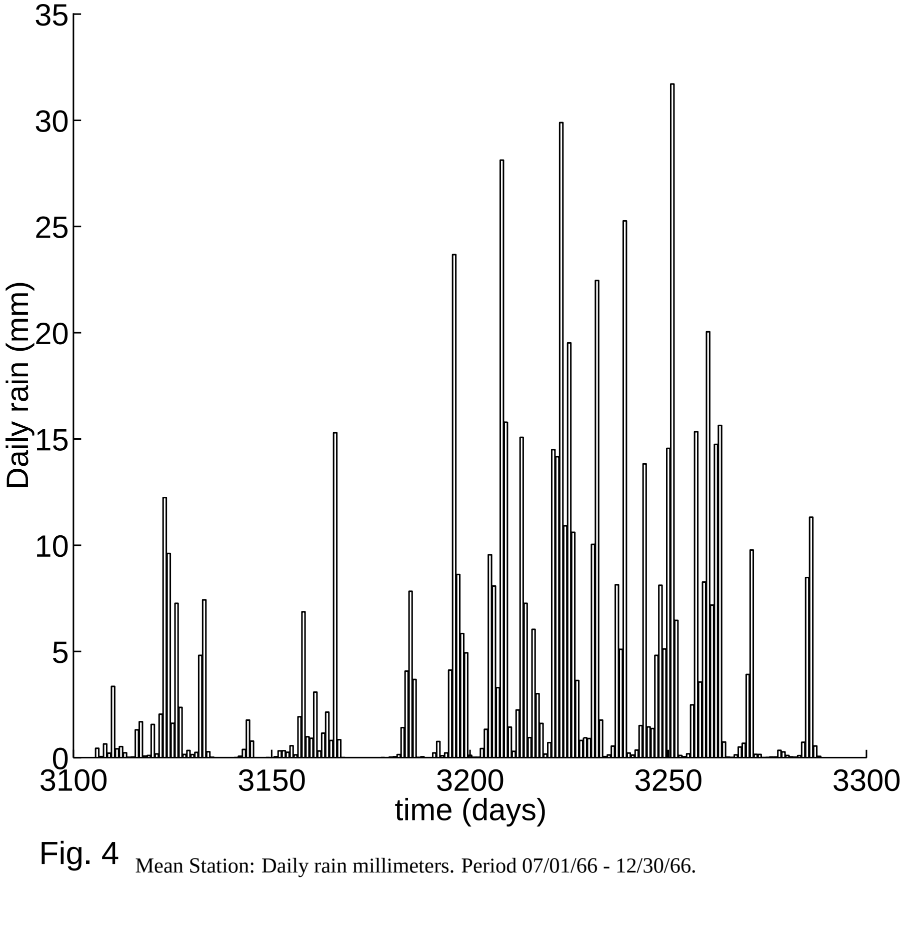


Fig. 4 Mean Station: Daily rain millimeters. Period 07/01/66 - 12/30/66.

3 Learning the Mean Station

Fig. 5 shows the graph of the mutual information of the MS's time series. Its first minimum gives $T = 7$.

This value has been used as the time lag for the computation of Global False Nearest Neighbors. The graph of FNN is shown in Fig. 6. Till $d = 6$ the curve decreases with the growing of dimension, and then reaches a plateau of 20%. The embedding dimension is then $d_E = 6$.

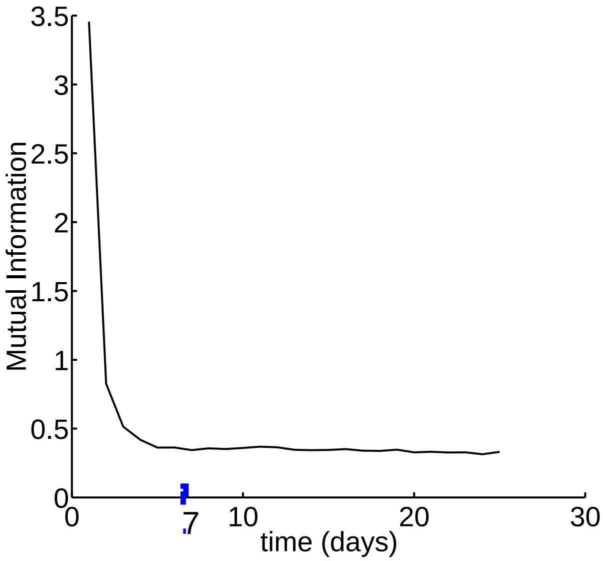


Figure 5: Mean Station: Mutual Information. The first minimum is for $t = 7$.

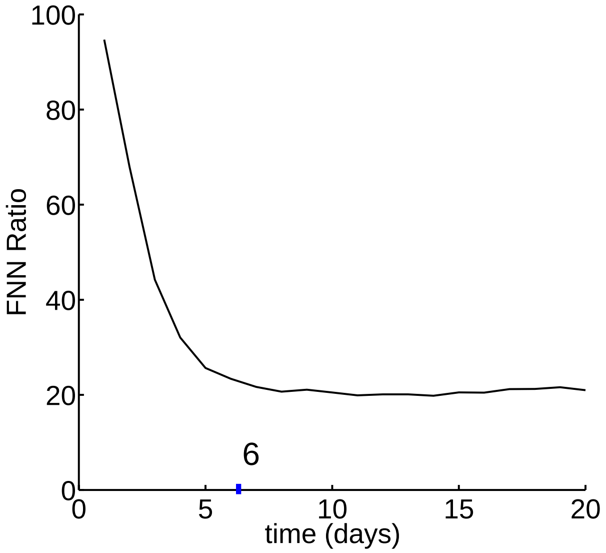


Figure 6: Mean Station: Global False Nearest Neighbors.

Following the constructive approach, we designed a predictor based on a Multi-Layer Perceptron.

The MLP was made up by two hidden layers of 5 units, an input layer of 6 inputs spaced by a time lag of 7 days.

The results obtained by such a way are poor, due to the discontinuity of the hydrological variable.

In order to reduce the effects of the discontinuities, we used the SSA decomposition ensemble method.

We applied the Singular-Spectrum Analysis (SSA) to a signal corresponding to the first 3000 samples of MS series. The window width used for the SSA was $M = 182$, i.e. 6 months, that is a period sufficient to take in account seasonal periodicities of the related physical phenomena.

Fig. 7 shows the ordered list of eigenvalues and the explained variance of the reconstructed signal using an increasing number of RCs.

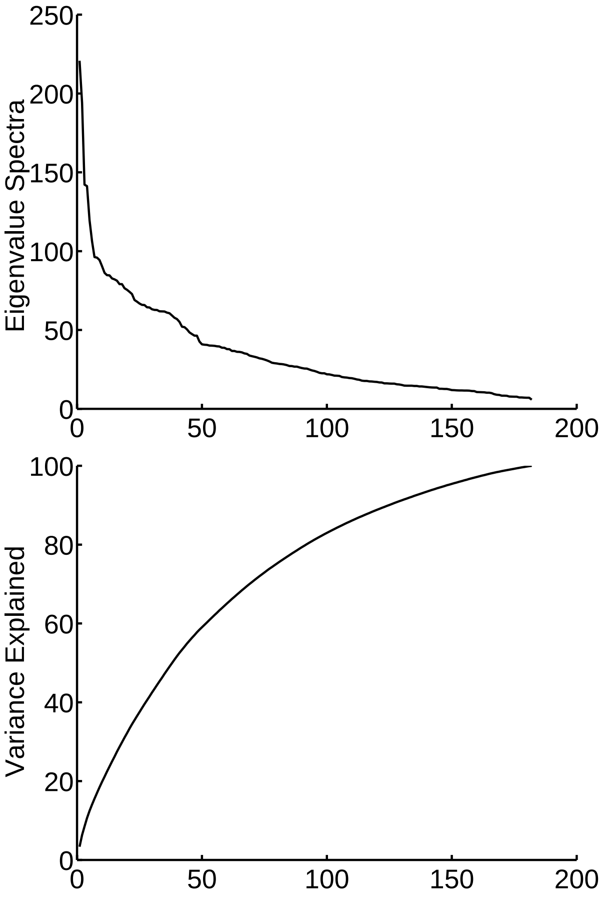


Figure 7: Mean Station: Eigenvalues spectrum (up) and explained variance of the augmented vectors related to an increasing number of RCs (down).

Table 1: Reconstructed waves (RWs) from disjoint SSA subspaces (each of them explaining 10% of the variance) and corresponding reconstructed components (RCs). The SSA is performed using using a window of 182 days.

RW	RCs
Ω_1	1-4
Ω_2	5-11
Ω_3	12-19
Ω_4	20-28
Ω_5	29-39
Ω_6	40-52
Ω_7	53-70
Ω_8	71-93
Ω_9	94-126
Ω_{10}	127-182

Then, from the original MS series we obtained 10 waves $\Omega_1, \dots, \Omega_{10}$ reconstructed from 10 disjoint sub-spaces, each of them representing a 10% of the explained variance (Tab 1).

Waves $\Omega_1, \ldots, \Omega_6$ (corresponding to the first 52 RCs), are more regular than the remaining waves (corresponding to subspaces with low eigenvalues) are more complex (Fig. 8).

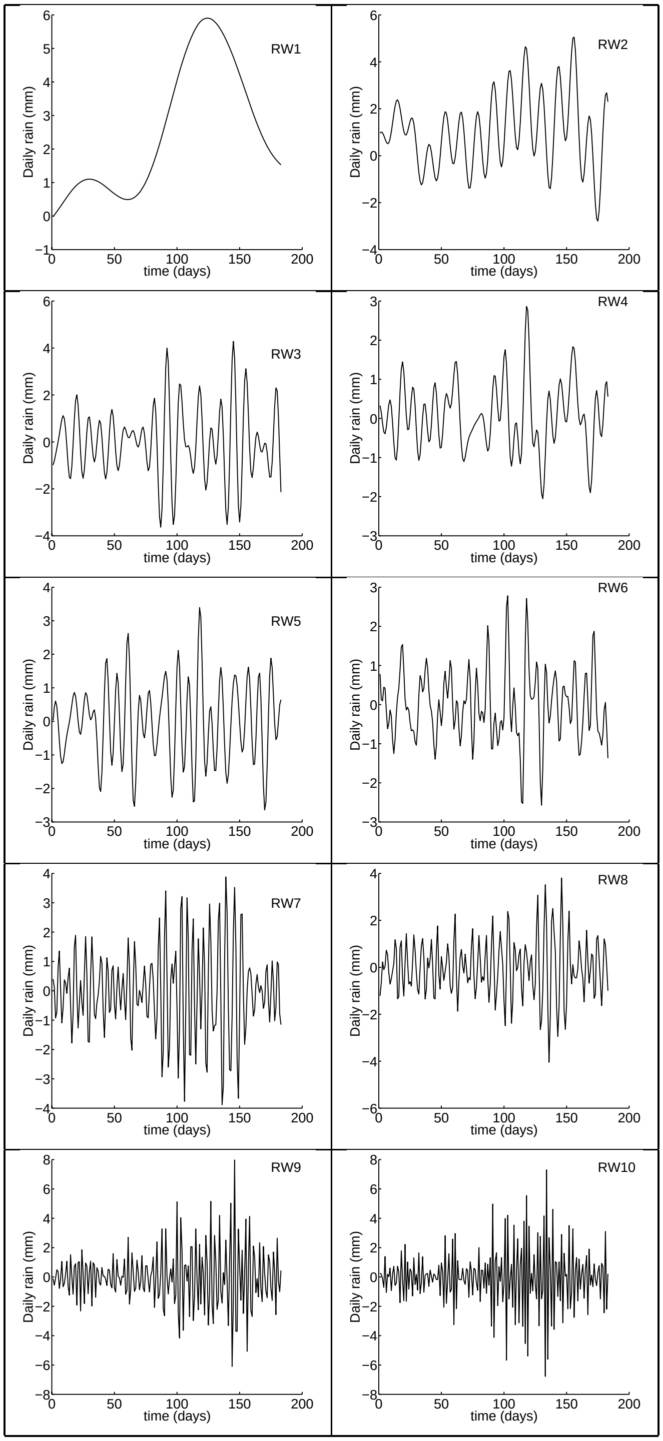


Figure 8: Reconstructed Waves. Period 07/01/1966 - 12/30/1966.

Fig. 9 shows the mutual information for each RW, while Fig. 10 shows the corresponding Global False Neighbors plots.

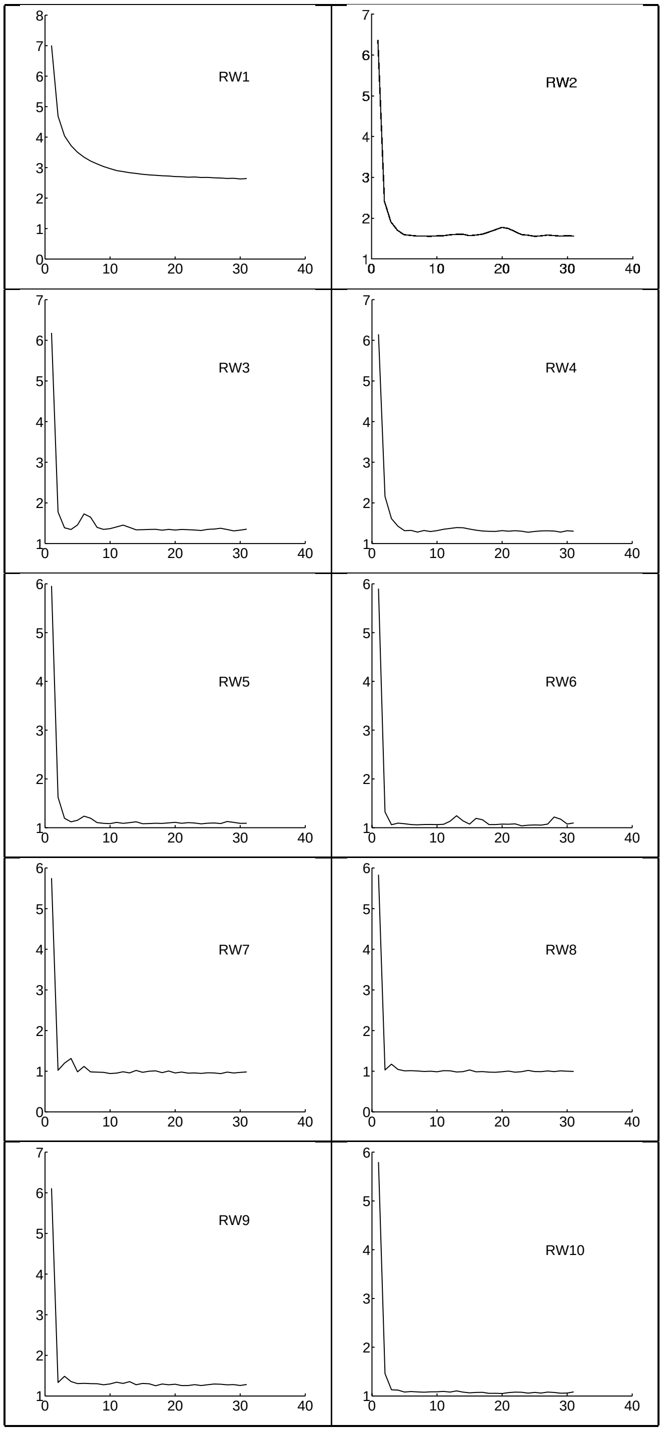


Figure 9: Reconstructed waves - Mutual Information.

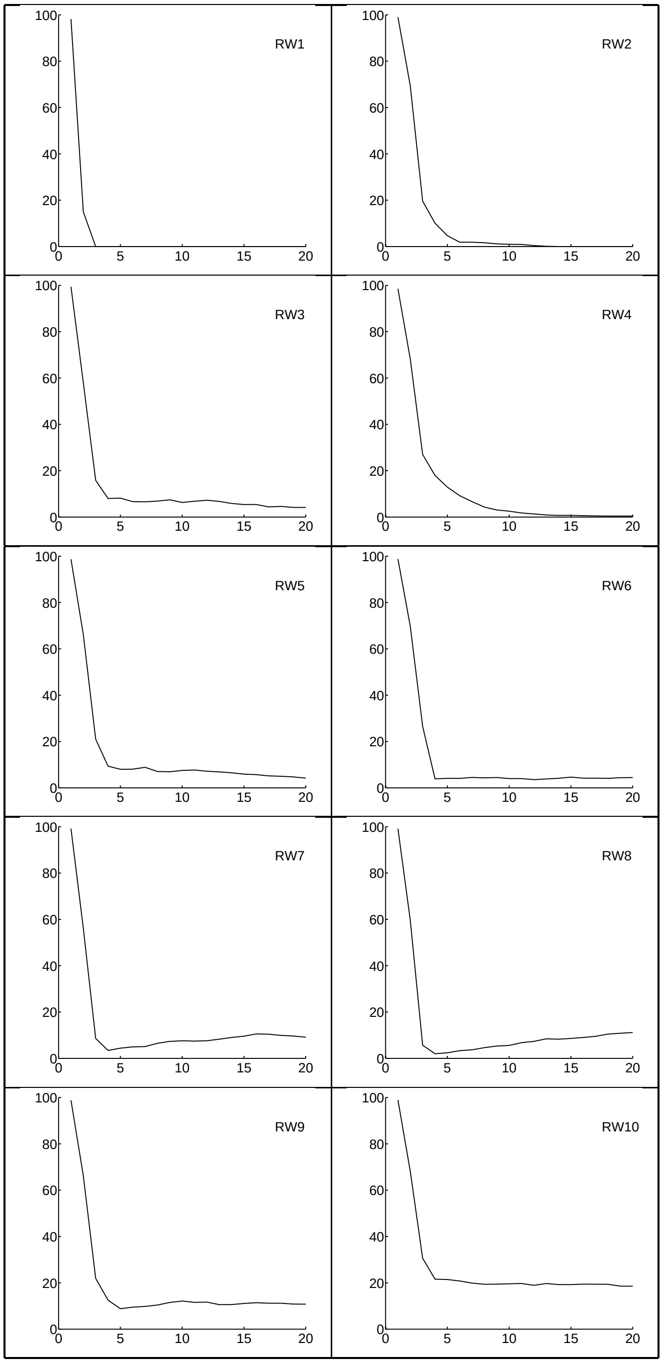


Figure 10: Reconstructed Waves - Global False Neighbors using T=7.

The evaluations of the first minimum of mutual information and of d_E for each RW are presented in Tab. 2.

Table 2: First minimum of Mutual Information (T) and embedding dimension (d_E) computed using T and other time lags for each reconstructed wave.

RW	T	$d_E(T)$	$d_E(7)$	$d_E(1)$
Ω_1	22	4	3	2
Ω_2	9	18	14	3
Ω_3	4	10	7	4
Ω_4	5	18	14	4
Ω_5	4	14	9	4
Ω_6	3	5	6	4
Ω_7	2	4	4	5
Ω_8	2	4	4	6
Ω_9	2	5	5	4
Ω_{10}	5	10	8	4

Then, we designed a neural predictor based on a MLP for each individual wave of the MS, following the constructive approach, implementing, in such a way, a SSA decomposition ensemble of learning machines.

The best results for each RW have been obtained using as inputs windows of 5 consecutive elements and two hidden layers with dimensions described in Tab. 3.

Table 3: Size of the hidden layers (L1 and L2), Root Mean Square (RMS) error and Maximum Absolute (MAXA) error for each reconstructed wave - Size of MLPs Input Layer=5.

RW	L1	L2	RMS	MAXA
Ω_1	6	4	.02	.05
Ω_2	8	5	.03	.12
Ω_3	6	4	.04	.15
Ω_4	8	4	.04	.11
Ω_5	8	5	.06	.14
Ω_6	8	4	.15	.40
Ω_7	4	4	.15	.38
Ω_8	6	4	.64	1.92
Ω_9	3	4	.75	2.40
Ω_{10}	3	4	.29	.90

As each wave contains 3652 daily samples, in our case for each wave we obtained a data set of 3646 associative couples, each of them consisting of a window of 5 consecutive elements, as input, and the next day rainfall intensity, as output.

Each MLP was trained using the first 2000 associative couples (*training set*), using the error back-propagation algorithm with momentum (Vogl88), and a batch presentation of samples.

The following 1000 associative couples (*validation set*) were used in order to implement an early stopping of the training procedure.

The remaining 646 were used for measuring the quality of the forecasting of the reconstructed wave (*test set*).

Results on the Mean Station

The prediction results for each reconstructed wave are presented in Tab. 3 and in Fig. 11.

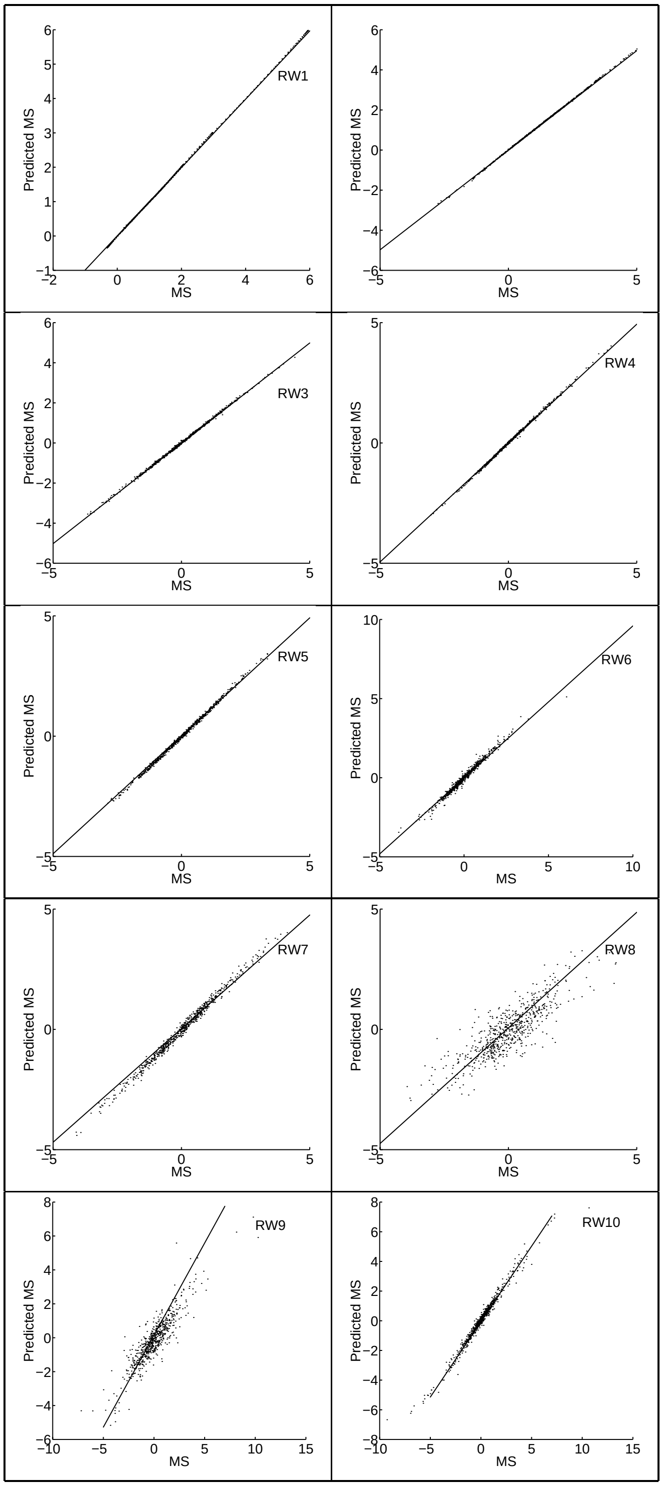


Figure 11: Reconstructed Waves - Scatter plots on the test set (using MLPs with 5 inputs).

The predictions obtained using the SSA decomposition ensemble of learning machines (i.e., the sum of the predictions of the 10 waves) at 1 day ahead are very satisfactory, as for the resulting MS prediction the Root Mean Square (RMS) error on the test set is .95 mm of rain, while the Maximum Absolute (MAXA) error is 6.47 mm, i.e., the predicted signal is substantially coincident with the measured MS rainfall intensity signal.

As shown in Figs. 12, 13, and 14, the predictions of the MS rainfall intensity signal are substantially coincident with the measured MS.

Note that in the comparison shown in Fig. 12 the predicted signal is clamped to zero.

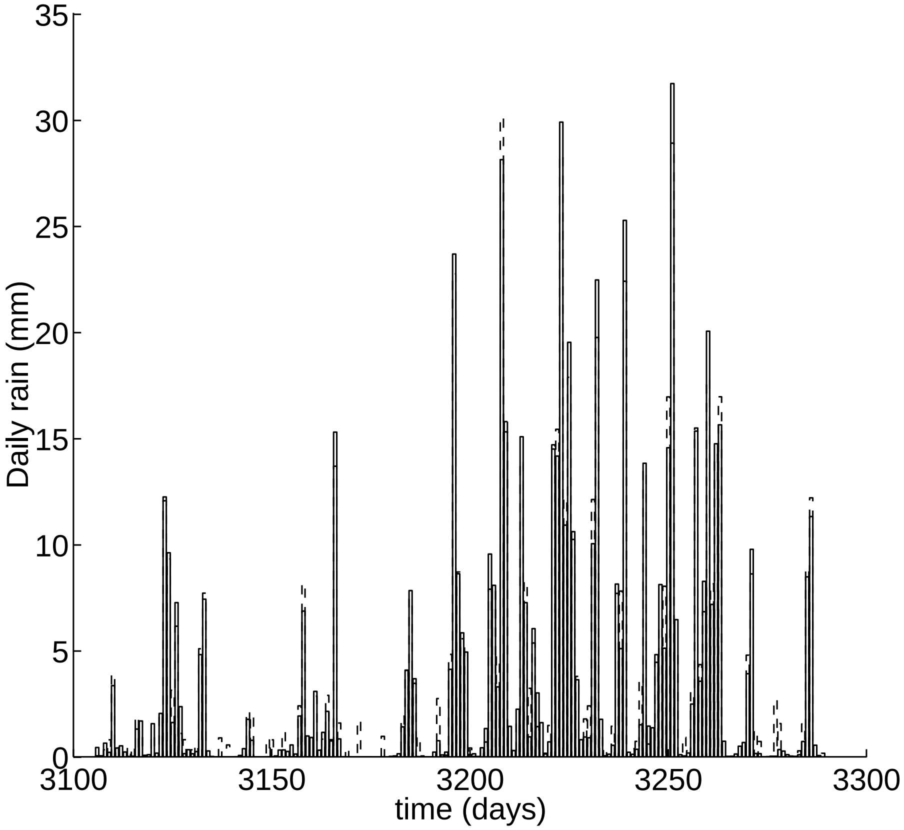


Fig.12 Mean Station: One day ahead forecasting in the period 07/01/66 - 12/30/66 of the test set using the ensemble of 10 MLPs with 5 inputs.

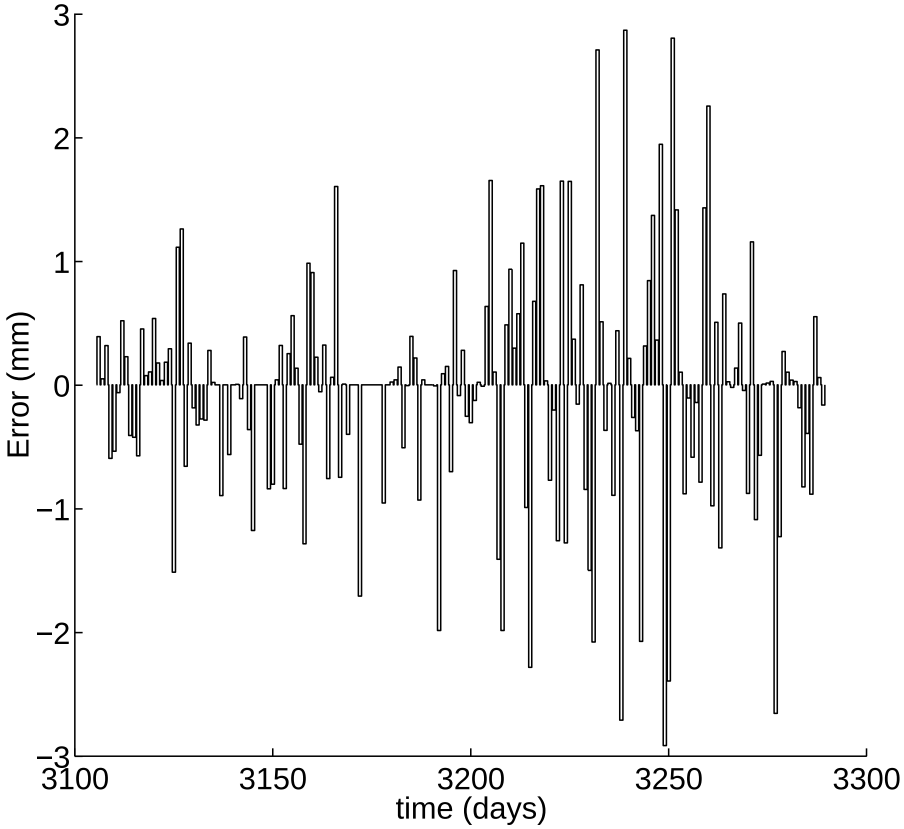


Fig.13 Mean Station: One day ahead forecasting. Errors in the period 07/01/66 - 12/30/66 of the test set using the ensemble of 10 MLPs with 5 inputs.

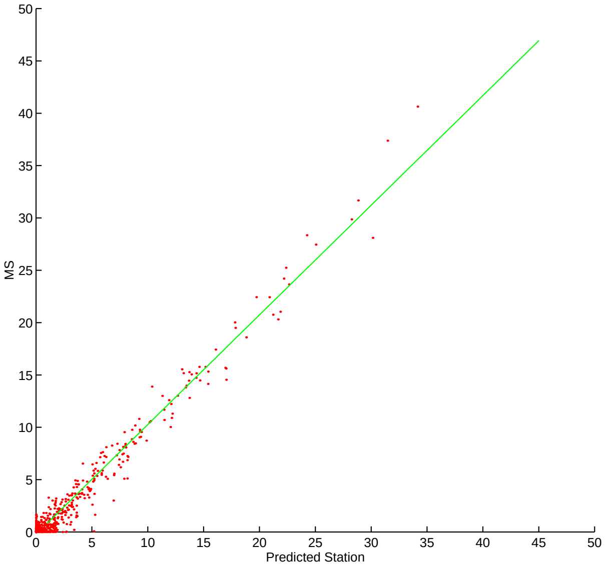


Figure 14: Mean Station: scatter plot of the 1 day haed forecasting on the test set using the ensemble of 10 MLPs with 5 inputs.

It is worth noting that the design of the ensemble learning machine is critical. Choosing a window $M = 182$ for the SSA, the best prediction results were obtained using MLPs with four or five inputs and two hidden layers.

Using MLPs predictors with four inputs we obtained results slight worse.In this case the RMS for MS is 1.05 mm and the MAXA is 8.05 mm for MLPs predictors using four inputs.

We notice that the Maximum Absolute error occurs the same day (11/05/1967) than for the architecture using MLPs with five inputs.

A different window for SSA can give results of inferior quality. E.g., using M=256 as the window for SSA we obtained good prediction performances only for for waves $\Omega_1, .., \Omega_6$, corresponding to 60% of the explained variance (first 76 RCs).

The resulting generalization of the SSA decomposion ensemble was poor, even leaving out in Eq. 17 the predictions of $\Omega_7, .., \Omega_{10}$ as, if enclosed in the addition, make worse the overall prediction.

We underline that the dimension of the optimal input layer (i.e. 5) is smaller than the d_E evaluated with the FNN method (Tab. 2).

This choice is supported by the generalization trade-off due to complexity of the learning machine and limed size of the training set

Concerning the time lag between inputs, we investigated different values as the first minimum of the mutual information is only a prescription and not a theoretical result (Masulli97).

The plateau in the FNN plots of Fig. 5 is a symptom of the presence of high dimensional noise (Abarbanel96).

After the SSA decomposition we can notice that the noise is concentrated mainly in RW10 and also in RW3, RW5, and RW9, as shown in the plateaus in their FNN plots (Fig. 10.

Learning the Individual Stations

The daily rainfall series of individual stations are more discontinuous of the MS, but are well correlated to it.

In Fig. 15(a) we plot the correlation map of the daily rainfall series of the individual stations to the MS, while the corresponding histogram is presented in Fig. 15(b).

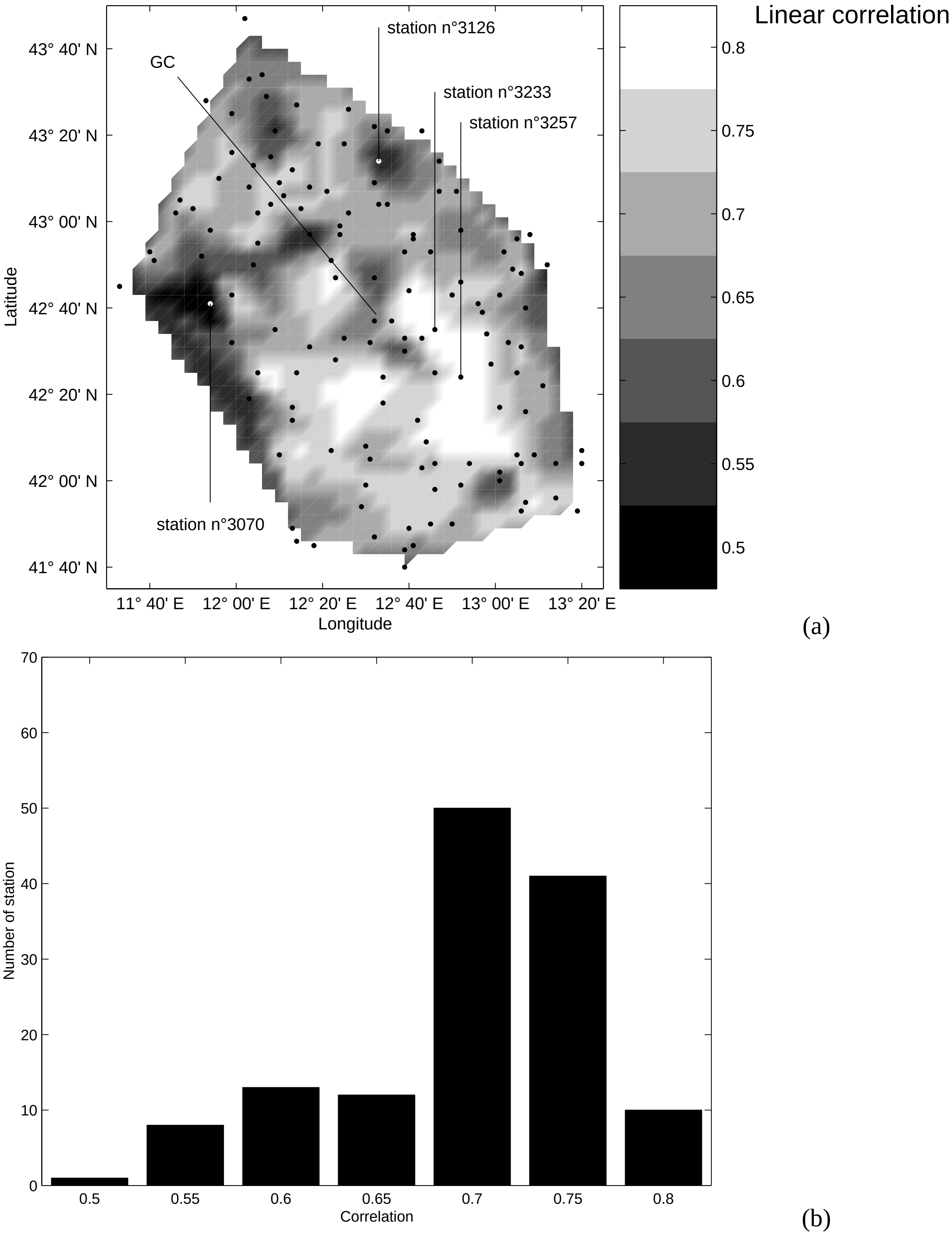


Figure 15: (a) Correlation to MS map of the 135 stations on the Tiber river basin. The Geographic Center (GC) of the 135 stations, the two stations more correlated to MS, and the two stations less correlated to it (see Tab. 4) are shown on the map. (b) Histogram of station's correlation to MS.

The average linear correlation coefficient is .7.

Moreover, using the Fisher-Snedecor test, we find a linear dependence at the 0.01 level of station's correlation versus the distance from Geographical Center (GC) of the 135 stations.

GC is defined as the average position of the 135 stations. Its longitude with respect to Greenwich is: 42 38' 88.8" E , its latitude is: 12 32' 51.0", and its Height is 473.1 meters.

To the aim of design efficient predictors of for the individual stations, we explored the following alternative approaches:

Approach A: Design of a single neural predictor for each station, sizing of its input layer using the measurement of the average mutual information and the method of Global False Nearest Neighbors.

Approach B: Implementing the unsupervised decompositive ensemble method based on SSA for each station, following the same approach previously presented for the MS.

Approach C: Decomposing the series of a station using the SSA already performed on the MS, calculating the RCs, aggregating the RCs in 10 RWs following Tab. 1, and then training one MLP for each RW. The prediction of the station's series will be the sum of the predictions of the 10 RWs.

Approach D: Decomposing the series of a station using the SSA already performed on the MS, calculating the RCs, aggregating the RCs in 10 RW following Tab. 1. The prediction of the station's series will be the sum of the predictions of the 10 RWs obtained using the MLPs trained for the MS (with hidden layers shown in Tab. 3).

Note that the Approaches B, C, and D are ensemble methods based on the SSA decomposition of the signal, with different flavors.

Results on Individual Stations

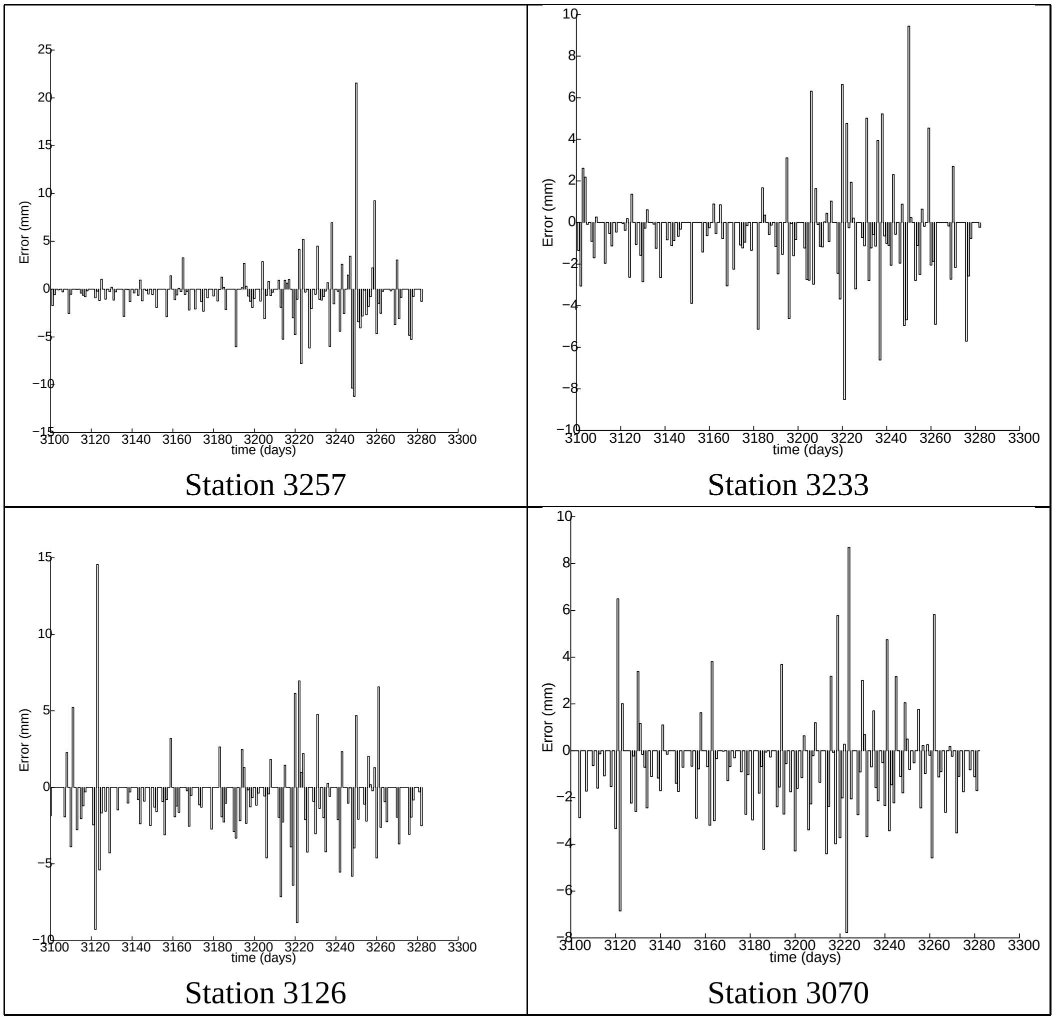
From our experimentation, the Approach A is unable to give useful results for any individual station, as well as for he MS station.

The Approach B, while is the most computationally expensive, at the same time leads to poor results, that we could ascribe to ill-conditioning in the SSA due to the significant presence of noise in the series of an individual station.

The Approaches C and D give similar good results.

Using the Approach D (that is less computationally expensive than Approach C) the average RMS for all the stations is about 2.71 mm of rain

The results obtained with Approach C are often slight better than those of Approach D.



Rieti
lin corr = .82
rank=1
RMS-D = 2.40
RMS-C = 1.93

Station 3257

Arrone-Terni
lin corr =.81
rank=2
RMS-D = 1.72
RMS-C = 1.49

Station 3233

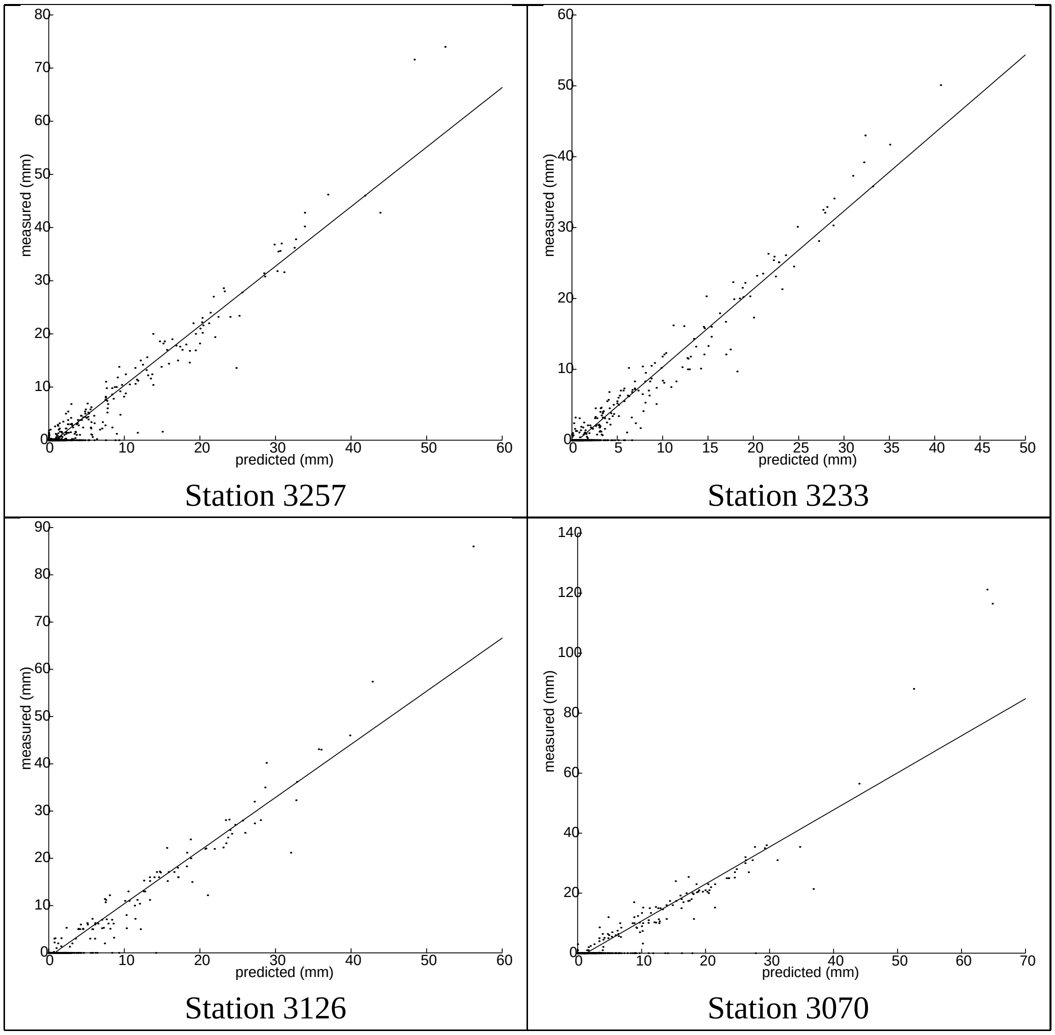
Scritto-Perugia
lin corr = .53
rank=134
RMS-D = 2.31
RMS-C = 1.56

Station 3126

San Lorenzo N.
- Viterbo
lin corr = .45
rank =135
RMS-D = 4.51
RMS-C = 2.35

Station 3070

Figure 16: Errors in the period 07/01/1966 - 12/30/1966 using ensembles of 10 MLPs with 5 inputs.The plots are relative to the two stations more correlated to MS, and to the two stations less correlated to it (see Tab. 4).



Station 3257

Station 3233

Station 3126

Station 3070

Figure 17: Scatter plots on the test sets using ensembles of 10 MLPs with 5 inputs. The plots are relative to the two stations more correlated to MS, and the two stations less correlated to it (see Tab. 4).

In Fig. 18 the one day haed prediction RMS for the 135 stations is presented in form of as a geographic map and as an histogram.

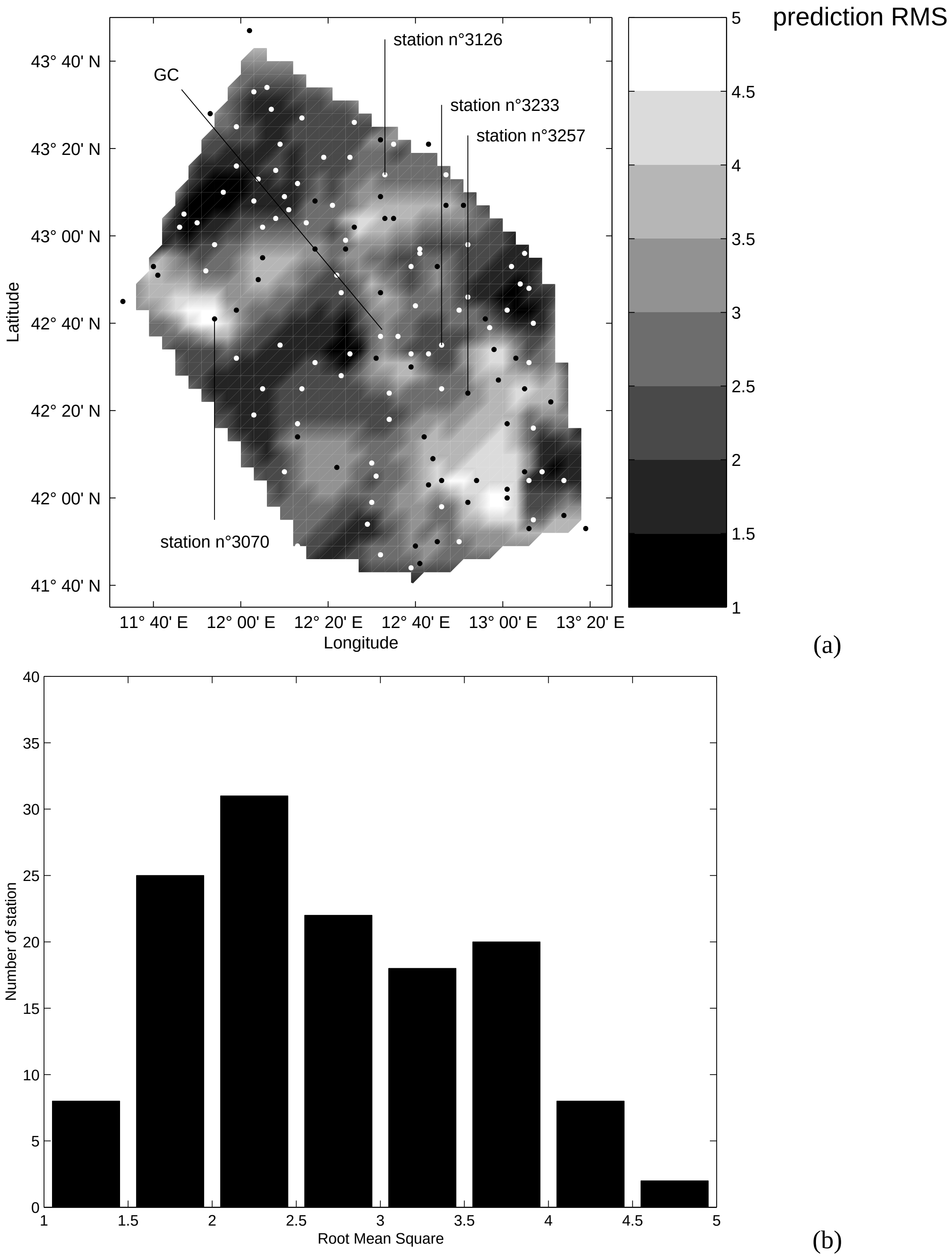


Figure 18: (a) One day haed prediction RMS map for the 135 stations on the Tiber river basin. The Geographic Center (GC) of the 135 stations, the two stations more correlated to MS, and the two stations less correlated to it (see Tab. 4) are shown on the map. (b) Histogram of station's one day haed prediction RMS.

Conclusions

Constructive methodology to the design of efficient predictors even for complex signals, such as discontinuous or intermittent signals.

Ensemble method that combines an unsupervised and a supervised step:

Unsupervised decomposition: The original signal is decomposed in reconstructed waves (RWs), using the Singular Spectrum Analysis.

Supervised learning: For each RW we design and train a MLP predictor using suggestions from dynamical systems theory.

In the operational phase the prediction of the original signal is obtained as the sum of the predictions of individual RWs.

The daily rainfall predictions of MS are very satisfactory, with a Root Mean Square error equal to .95 mm of rain.

Learning of individual stations. Steps:

1. Decompose the series of the station using the SSA already performed on the MS; calculate the RCs and aggregate the RCs in a number of RWs.
2. Train one MLP for each RW.

The prediction of the station’s series is the sum of the predictions of the 10 RWs.

It is possible to skip step 2 and using in prediction the MLPs already trained for the MS.

The daily rainfall predictions on individual station obtained with this latter approach show an average Root Mean Square errors of 2.71 mm of rain.