

Gap Repair in Water Level Measurement Data Using Neural Networks

P. van der Veer, J. Cser, O. Schleider & E. Kerckhoffs

Delft University of Technology, Faculties of Civil Engineering and Informatics, Postbus 5048, 2600 GA Delft, The Netherlands

Email: p.vdveer@ct.tudelft.nl

Abstract

The paper presents a method for completing missing values in time series by applying neural network computation. Conventional methods to complete fragmentary time series like linear interpolation are chosen for small amounts of missing values; problems occur when there are larger intervals of missing values. So far such gaps are repaired by estimating techniques that use parameter estimation of mathematical models. However such an approach fails when there is not enough information for calibrating the model or when the model is too simplified for reliably completing the data series. Neural networks with their generalisation and memory properties are predestined for this category of problems. In the proposed method a two-layer feed-forward network is used to build spatial patterns out of time series intervals with some variations on one hand and on the other hand directly model processing is used for producing time series information as a function of additional and external information. A representative values approach combines characteristics of both the spatial and the direct approach. Parameters in the approach are the number of representative values in each input pattern and the sampling distance of representative values in the original time series. The discussed example refers to water level data that contains a basic tidal oscillation (which is computable by astronomical components) and additional set-up which is mainly a function of local meteorological conditions like wind speed and wind direction.

1 Introduction

Time series representing measurement data are produced by measuring devices and transmitted to its destination in order to be processed further, e.g. by simulation programs. Sometimes the flow of information in a measuring system can be interrupted. However, when the communication failure is only temporary often a measuring system stays working and will produce data again when the interrupt is over. The data series that has been produced appears then to have gaps. Generally simulation programs that process measurement data for calibration require continuous, complete data series. So the impact of a

measurement device failure is: data that are unsatisfactory as input for simulation models. If such gaps in data series occur, their maximum acceptable size depends on the data behaviour in the gap environment and of course on the requirements of specific application that will use the data. Gap repairing by methods only considering the raw time series without further information may be suitable for small intervals of missing values. In such cases, for example, linear interpolation may be used between the values at the start and end points of the gap. For larger intervals of missing values, additional information about the process that produced the data is necessary. That information, representing characteristics of the system, may come from other measurement data or from a simplified mathematical model that could describe the phenomena during a small period (the length of the gap), and that has been calibrated using other available measurement data.

2 Measurement data series gap repair

In this paper water level measurement data in a coastal area are considered. In many of such areas the water system is complex. There are many parameters in the system: water level, wind velocity, wind direction, wave energy, water flow velocity and water flow direction distributions, data about regional topology, etc. When we focus on the water level we see a periodic phenomenon that, as a rough approximation, can be described by a Fourier series. In this series, representing the astronomical influences on the water level, there are basic components and location-specific components. The series represent only a very rough description of the water level behaviour. Other components have to be included for a more accurate description. These components, together forming the water level setup, (which is defined as the difference between the astronomical water level and the actual water level), represent a dynamical system with many degrees of freedom. (In the examples section of this paper, the wind conditions are used as most relevant for the water level setup).

For such a complex system, gap repair in measurement data series is not simple. Only for small gaps simple interpolation techniques may be appropriate. For larger intervals of missing values, a conventional method is the use of a mathematical model. Using different measurement data the parameters of the model are estimated. Then the model is used to repair the gap. An inherent problem to this approach is the complexity of modeling, especially with respect to the effect of the meteorological conditions on the set-up. And not only the modeling itself is a problem. Also the available measurement data about meteorological data are generally dynamical, location-dependent, less accurate and less reliable than the water related data. For such effects probabilistic methods would be necessary.

For gap repair there is as a practical alternative: start with computing a first approximation of the expected water level from astronomical components, and then use some additional information about the water level setup at start and end points of the gap. For relatively small gaps such a method could be useful. This method is elegant because of its simplicity from a mathematical point of

view. Only basic mathematical formulations are used, without the necessity of model calibrations at every occurring gap.

However, if we have large gaps, for example a three hours gap while the measurement frequency is ten minutes, we meet problems with respect to data accuracy and data reliability. And especially in our coastal situation where a time series contains basic periodical components, simple interpolation techniques for the setup effects will still cause significant deviations from the actual course when the gaps are in unfavourable position, i.e. when they are larger than one cycle and the nearest known values are not near local maxima or minima. For such situation interpolation methods are needed that can handle complex systems. A neural network could be very useful in such an approach. Especially a combination with basic mathematical formulations could be promising. The basic information like the astronomical effects on the water level is computed from a well-known Fourier series and the neural network could then compete with the accuracy of the set-up interpolation. For this interpolation data from nearby nodes in the measurement system or from measurements before and after the gap may be used.

Water level measurement data are generally stored in a database. The quality of data is crucial for further processing of the data, e.g. in water level predictions. The absence of gaps in the series is an important quality factor of the database. Generally the quality enhancement of the database, by performing gap repair, is made off-line. This type of off-line gap repair is often done as a kind of preprocessing for other calculations that require complete data series. Of course this is not the most cost effective way of handling the measurement data. It could be much more desirable to process the gap repair on-line at the measurement location. Then complete data series would be transferred to the database and no additional processing would be necessary before the data can be used further. The suggested way of using neural networks for gap repair could be implemented at the measurement location. So this approach could contribute to data quality enhancement as well as to cost effectiveness of measurements.

3 Feed-forward neural networks

Feed-forward multilayer neural networks are general applicable and able to compete with many conventional techniques for a broad spectrum of tasks. An important application of them is prediction of a non-linear process from trained examples, where the neural network implicitly learns an internal parameter identification of the process.

Neural network response to presented inputs can be seen from two different points of view: First a classification viewpoint. If a feed-forward network is used in classification problems, the network associates new input with previously generated output of a training set input. It looks for training set input which has a minimal spatial difference to the new input. Then both input vectors are considered to belong to the same class. Because somewhat erroneous input patterns will be recognised as their correct counterparts this is fault tolerant behaviour of the neural network. Secondly a generalization

viewpoint. This is one of the most important aspects of neural networks. It can be illustrated by a simple example. As mentioned, a trained neural network performs a mapping which it has learned from examples represented in the training pattern set. The example network shown in Fig.1 has been trained to produce the average of the input values at the output by applying the training pattern set which consists of instances of the triple $(a, b, 0.5(a+b))$.

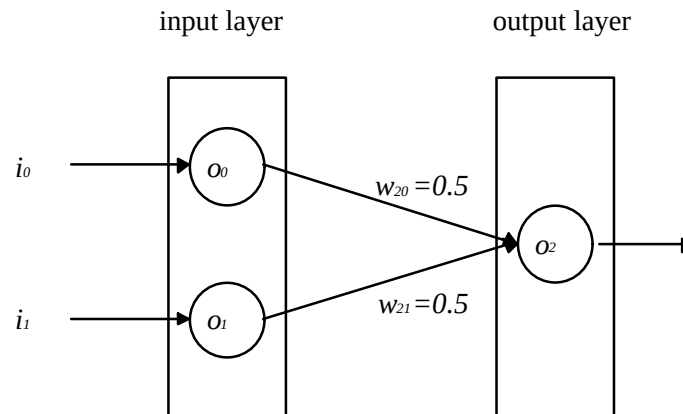


Fig. 1:

Simple example network to compute the average of the two input values. Obviously, it is correct for any input values.

But obviously, because the weights are set to 0.5 each in this constellation it performs the desired mapping not only for the training input but for any real value input. Here we see a form of the generalization properties of neural networks. However, in practice we often meet the problem that the network finds a weight constellation which satisfies the training set, but where in some cases, if the input is too far from the training set, the network output is not accurate. Therefore neural networks are very well capable for handling interpolation problems, but one has to be careful when a feed-forward neural network is used for extrapolation problems, especially when the data are far from the training set. Then the combination with different methods can be suitable; or the application of an other type of neural network may be useful.

For our task, the gap repairing problem, we have to find a suitable neural network method that uses the mentioned properties. In the experiments we will use complete series of measurement data. We make artificial gaps in the series to be able to evaluate later the results of our approach for gap repair. We will use multilayer feed-forward networks using backpropagation, a non-feedback learning paradigm. In such a method the neural network does not refer to its previous computation results automatically. The approaches that can be followed are:

- Direct presentation of the water level values, where the network has to construct a water level estimation value from information about both water

level and wind condition values of the time step prior to the current one. Thus this gap repairing is seen as a prediction task.

- Spatial presentation of series fragments. Here the time component is handled as a component in a spacial pattern. This obviously enforces rather large input series (and thus rather large networks) as well as experiments on determining an appropriate structure of the network, e.g. with varying number of layers. There is a number of factors influencing the quality of results: position of the gap in the series, size of the gap and the number of composed patterns for training.
- Spatial presentation of series fragments with reduced complexity. On one hand this reduction is realized by reducing the amount of input data. The data series is replaced by a smaller series of *representative values*. This series is used for the pattern recognition. On the other hand this approach uses a pre- and postprocessing method: before the neural network data processing first well known aspects like the astronomical influences on the water level are filtered out from the input data. These influences are described by simple mathematical formulations (Fourier series). So there is no need for involving them in the pattern recognition process. After the neural network data processing the results are corrected for the aspects that were filtered out previously.
- An improvement of the representative values approach is also applicable. Basically it uses the *representative values* approach: but now the missing values are approximated from patterns from both sides of the gap. So from an extrapolating technique it becomes an interpolating technique. Therefore it may be useful for large gaps.

4 Spatial patterns experiments

Generally multilayer feed-forward networks are not the standard architecture for temporal processing and there does not seem to be an obvious way to deal with the time component. Some experiments should show whether a feed-forward neural network approach is suitable at all for the task of gap repair in water level data series. For example, it is not possible for a feed-forward network to produce a dynamical output by receiving a constant input (what can be done with recurrent network under certain conditions), because it can only be a model for (continuous) functions without time component, thus one input value will always have the same output value.

Given a feed-forward network, the most promising issue appears to be finding an optimal network input data representation and, especially in our case of water level modelling, reducing the amount of input information by the introduction of a representative values approach. The introduction of additional information about related phenomena shows to improve the results. For example: not only measured water level data are used; also measured data meteorological parameters is included.

The following aspects have to be investigated:

- Training an archive of instances of a water level time series in a suitable way so that it can be generalised to other input patterns.

- Adding data from related phenomena, like meteorological information. Special attention has to be paid to an appropriate representation of the phenomena in the input patterns.
- Investigations about the general representation of gaps (missing values). If gaps are present in the patterns, the way of representation both in the training and in the test patterns is in question.
- Considerations about the network environment. For example, constellation details are what will be pre-processed, what will be done by the network itself and what will be post-processed.
- Loss of generality is to be suspected when particular solutions are developed. Thus, an interesting question is to what extent the solutions are valid for other fields of application.

Experiments to give an answer about the applicability of feed-forward networks for our task of gap repair concern especially the first mentioned aspect. The other aspects refer to quality enhancements of the method. Our preliminary experiments deal with the pattern recognition properties of feed-forward networks, applied to relatively large patterns. An interval of k values of the time series has been taken both as input and output pattern. In this way the neural network was trained to perform the same data behaviour as the physical system. According to the general paradigm learning from examples, the gap repairing property should be trained from a variety of example gap constellations. In practice, this makes training complicated.

In the next experiments, an input pattern consists exactly of the known part of the time series, whereas the corresponding output pattern provides only the missing values. Here, the network performs a multidimensional mapping. It produces an m -dimensional identification of an n -dimensional input, if the experiment is compared to pattern recognition. If the intervals are chosen arbitrarily, the network can not produce similar output to similar input, which makes the network task complicated. The size of the output layer determines a fixed size of the gap, and by taking the known part of the time series as input pattern, only one gap is allowed per training pattern.

The conclusion of our experiments is that repairing gaps with feed-forward neural networks is possible. Some comments, however are relevant. For applicable results the network has to be trained to perform a mapping that can easily be generalised to test patterns that are not included in the training set. Or each input pattern has to match one of the trained patterns, at least "to some extent"; in this case test patterns must correspond to training patterns in their phase lengths and displacement. That correspondence appears to be an important issue for that approach.

One solution, however probably rather costly, is to have a variety of training patterns for one process: by training several patterns of the same process, for example a water level time series with a phase displacement for each pattern, the archive for looking up gap intervals is larger, so the network function is more complicated but the chance to find a matching pattern is better. By training several displaced instances of the same time series, more starting points for presenting fragmentary input patterns to the network are given. The

size of these shifted instances, which implies the number of input units of the network, can be chosen according to conditions of the particular application.

The optimal choice for the spatial pattern approach is not to train gaps at all, but to let the network perform mapping for the trained time series intervals such that it performs a data behaviour like the original phenomena. It shows that changes of the network structure like the number of hidden layers and hidden units as well as other parameters like the applied learning rule and activation function belong to the fine tuning stage of network design.

5 Direct presentation method

A basis for the representative values method and more in the direction of (partial) recurrent networks is the direct presentation method. Here, the time series is presented value-by-value at the input layer, together with additional information. That additional information may be from related phenomena that are measured or from water level measurement at nearby locations.

The idea is that a water level value is quite well-determined by the astronomical component of the water level, by the previous set-up values and by the previous wind conditions. Different interpretations may apply for these three value categories. For example, depending on the wind direction and the distance of measuring devices from each other, the effect on the set-up will be more or less delayed. A suitable combination of set-up and wind condition values has to be found. In this approach, only one single value instead of the complete gap is computed from one output pattern. Thus as many pattern presentations are needed as the gap size in time steps is. A very simple network results. There is a huge training set but each of the patterns is small. A characteristic aspect of this sequential approach of generating single values is that each network output is used in the next input pattern.

To illustrate the performance of this method an example is shown. Given a time interval T and the three time series water level f , wind direction w_r and wind speed w_s with values for each time step $t \in T$, the first step is to decompose respectively transform these values in a network-suitable form, where $f(t) = s(t) + a(t)$, $s(t)$ is the set-up time series and $a(t)$ are the astronomical calculated water level values. The wind condition vector with direction w_r and magnitude w_s is transformed to two-dimensional co-ordinates w_x and w_y . Finally, each vector (s, a, w_x, w_y) builds an input pattern p . The corresponding output pattern is the next set-up value $s(t+1)$.

For testing and using the network, the time series is sampled with the same parameters, then a first pattern is constructed with the input values that are closest to the gap, so that the output value is the first value of the gap interval. This output value forms the next input value, together with the wind conditions and the astronomical value for the next time step. This procedure is repeated until the whole gap is covered.

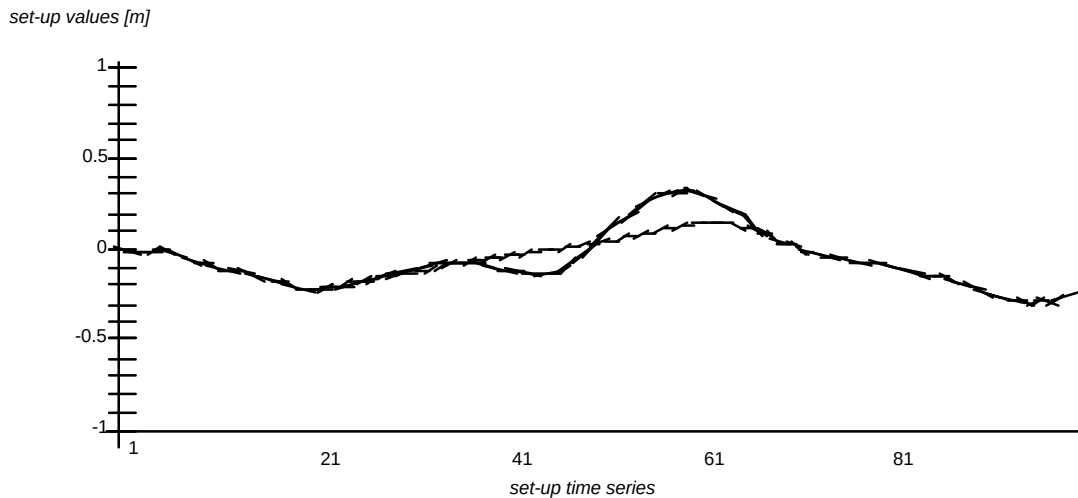


Fig. 2:

A direct presentation approach. The network is tested with patterns containing known additional information. Each output value is included in the next input pattern. The result is not accurate, but the reconstruction (from value 20 to 70) follows the general trend.

In Fig.2 is shown that the result follows the general trend, although the network did not follow the course of the set-up time series exactly. We see this direct presentation approach as a first test architecture for the re-use of network output values, which becomes necessary for approaches if the gap has larger dimensions.

6 A hybrid approach using representative values

A representative values approach was developed, based upon experiences of preceding experiments. It includes the following steps:

- Removal of any computable information that is not necessary for neural network processing. In our water level case, the astronomical water level which is seen as the basic oscillation can be simply computed from a Fourier series, independently from the neural network. If this reference is subtracted from the original time series, the result is the set-up time series.
- Limitation of time steps and related amplitude values. For the input patterns, only a few representative equidistant values of the water level set-up time series and values of the wind condition time series are used. This reduces the input space significantly. A suitable way of choosing the representative values has to be chosen so that the original time series can be reconstructed with a minimum loss of information. A representative value is defined as a value that is representative for the values of a part of the time series, such that the original series can be reconstructed from the representative values with sufficient accuracy. With a smaller sampling distance, less information is lost, but then patterns become more complicated and consequently it will

take more processing activity for the network to repair a gap. So here we have to find experimentally an optimum between accuracy and processing time.

- Normalisation of the pattern presentation. The information about the relative position of the gap will not be needed for the network. Basically only fragmentary parts of the time series are inputs for the network. The network will not have to learn where in the input the gap is, but only how the time series looks like in the neighbourhood.
- Reconstruction of the time series by reverse execution of the preceding steps after calculating the gap values. First the repaired values are shifted back to their original relative position in the time series. Post-processing generates a complete set-up time series from the representative values and finally, the astronomical water level is added to obtain the original water level time series, including the repaired gap.

The size of the gap is not restricted, but if it exceeds the range of one representative value, the next representative value has to be computed from previous ones, like in the direct presentation approach. So this representative values approach is a hybrid of a spatial pattern approach and a direct presentation approach. Like in the spatial pattern approach, more than one value is presented at once, and like the direct presentation approach output patterns are re-used (if the gap is larger than the covering distance of a representative value).

7 Examples of application

An experiment illustrates the hybrid representative values approach. Table 1 shows relevant practical data of the experiment.

Water level values of a measuring period of five days (720 values at a time step of 10 minutes) were reduced by the astronomical calculated water level values to obtain set-up values

$s(1), \dots, s(720)$.

Experiment values	
representative	
network type	multilayer feed forward
no. of hidden layers	1
no. of input / hidden / output units	5/6/1
links	fully connected
shortcuts	none
hidden and input layer activation	logistic
output layer activation	identity
no. of training patterns	33
training error	$< 10^{-5}$

Table 1:
Details of an experiment with representative values.

In the experiment:

$$T = \{20, 40, 60, \dots, 720\} \text{ and}$$

$$n = 36$$

representative values result, which were quantized to intervals of 5 cm where

$$S = \{-0.95, \dots, -0.1, -0.05, 0, 0.05, 0.1, \dots, 1\},$$

thus

$$m = 40$$

different values are allowed, assuming the set-up values do not exceed 1 m. For the gap representation, three values which represent the tidal cycle prior to the gap, and the two wind components which are computed from the original wind direction and speed values with

$$w_x = w_y = \{-0.5, -0.48, \dots, 0.48, 0.5\},$$

are input and the representative value of the gap location is output. Because three values are needed in each input pattern, with the available data set, only 33 patterns can be trained. This assumes a gap length of about 20 values for this constellation and per pattern. In the experiment, linear interpolation between the representative values was used to reconstruct the set-up time series, and after that the astronomical values were added to build the water level time series. The difference between actual and reconstructed water level is very small, as shown in Fig.3.

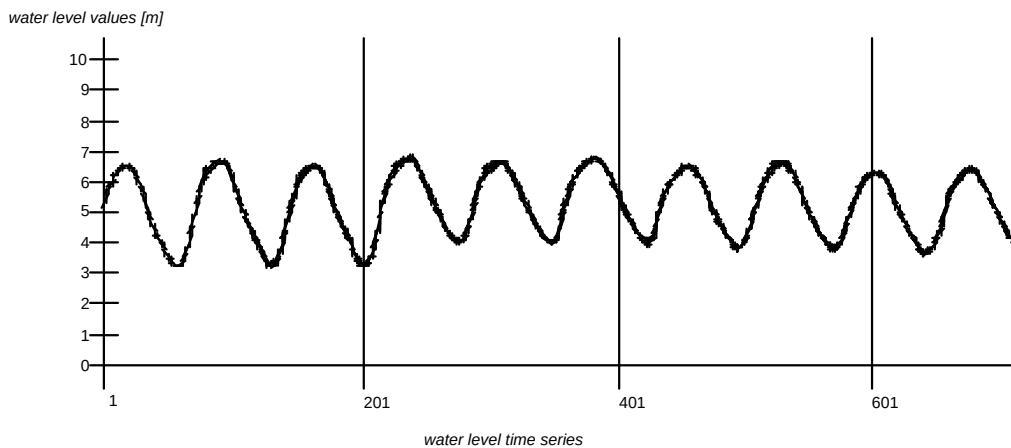


Fig.3:

The representative values approach. The original water level time series is reconstructed sufficiently enough by the trained representative values.

The advantage of this approach is the relatively small number of short patterns, which enables fast learning. In the example, all 33 possible three-value set-up/two-value wind component combinations were trained, which requires six hidden units.

A good way for a first test of the trained network is to shift the sampling interval so that new representative values for the same data set and thus new patterns which are independent from the training patterns are the result. In the corresponding example the same procedure as for preparing training patterns is applied. All parameters are the same as in the training phase, except that

$$S=\{30, 50, 70, \dots, 710\} \text{ and}$$

$$m=32$$

for the test. Network output are the representative values for possible gaps. Fig.4 shows the expected and actually reproduced representative values. Tolerating one quantization step of 0.05m, 53.125% of the patterns are recognised even with only 33 known patterns.

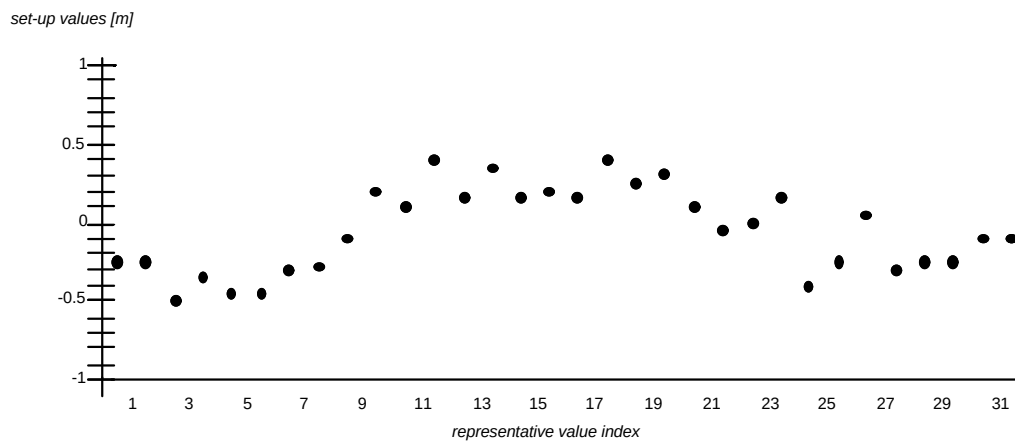


Fig.4:

The representative values approach: test result without re-using values, but with different sampling of the original time series. The network reconstructs each set-up value from known preceding values. Most of the values are sufficiently recognised (white: sampled desired values, black: network output from examples of a different sample)

For using the network appropriately, many more patterns have to be trained, because with the chosen parameters (e.g., the distance of two representative values), 33 patterns only represent about nine days of water level measurements. Depending on quantization and magnitude of the time series interval used for training the network, one may expect that patterns reoccur, even if they describe slightly different situations. Thus, the number of patterns to be trained is not merely a function of the considered interval length.

When the size of the gaps is larger than the space between two representative values, the network has to rely on its own results, because part of the input pattern is output from previous computations. Because it is a step-by-step reconstructing approach like direct presentation and values from both sides

of the gap are available, an improvement step may be considered, which defines two networks of the same type but with a different training set. One repairs the gap from the "right" side, one from the "left" side. Additional parameter is the weight vector for combining the two series to the result series, which will then be post-processed like in the previous approach.

Remember here, that a feed-forward network can only be used for prediction of a process when the complete behaviour of the process is known (is trained), which in particular means that there is a finite set of patterns that can describe this behaviour. This is, even approximately, not practically possible without reducing the network input to representative values of the process, as previous experiments showed.

8 Conclusion

This paper investigates repairing gaps in time series from two points of view. The first, a spatial pattern approach showed that a feed-forward network is able to reconstruct a part of a time series when it belongs to the training set. The size of this training set is a big problem when parts of a non-linear dynamical system like the water level are processed. The conclusion was to modify the pattern presentation to make the network task easier.

The second approach was to reconstruct parts of a time series by calculating the gap value by value. The output patterns consist only of one time series value, input patterns include the meteorological conditions and the astronomical water level component. This approach uses the output values to form the input pattern for the next calculation step.

The representative values approach contains aspects of both methods. Still a spatial pattern is presented to the network, but by abstracting to a few representative values and leaving the decomposition and reconstruction to pre- and post-processing methods the network has to deal with shorter and less patterns. Gap size and position are not restricted, thus if one representative value can not cover the gap, it has to be included in the input pattern to produce the next one.

Because it is a step-by-step reconstructing approach like direct presentation and values from both sides of the gap are available, an improvement step may be considered, which defines two networks of the same type but with a different training set. One repairs the gap from the "right" side, one from the "left" side.

The presented approaches use feed-forward networks and the inherent method is recognition and association of spatial patterns. Further research will concern the application of recurrent networks for time series forecasting. Basic ideas of the presented hybrid approach will be useful in that approach as well.

Literature

1. Funahashi, K.-I. On the Approximate Realization of Continuous Mappings by Neural Networks, *Neural Networks*, 1989, **2**, 183-192.
2. Hertz, J., Krogh, A., Palmer, R.G. *Introduction to the theory of neural computation*. Addison-Wesley, Redwood City, CA, 1991.

3. Hornik, K., Stinchcombe, M., White, H. Multilayer feedforward networks are universal approximators, *Neural Networks*, 1989, **2**, 359-366.
4. Hornik, K., Stinchcombe, M., White, H. Universal approximation of an Unknown Mapping and Its Derivatives Using Multilayer Feedforward Networks, *Neural Networks*, 1990, **3**, 551-560.
5. Narendra, K.S., Parthasarathy, K. Identification and Control of Dynamical Systems Using Neural Networks. *IEEE Transactions on Neural Networks*, 1990, **1**.
6. Rumelhardt, D.E., Hinton, G.E., Williams, R.J. Learning internal representations by error back propagation. In: *Parallel distributed processing: Explorations in the microstructure of cognition, Vol.1, Foundations*, Rumelhardt, D.E., McClelland, J.L., et al., MIT Press, Cambridge, MA, 1986