US Army Corps of Engineers®

Field Data Recovery in Tidal System Using Artificial Neural Networks (ANNs)

by Bernard B. Hsieh and Thad C. Pratt

PURPOSE: The field data collection program consumes a major portion of a modeling budget. However, due to instrumentation adjustment and failure, the obtained data could be incomplete or producing abnormal recording curves. For instance, complete boundary condition data are often critical to the numerical modeling effort. The data may be unavailable at appropriate points along the computational domain when the modeling design work changes. In addition, the key locations, which usually have high gradient variation in the numerical model, could be partially missing. Therefore, the judgment of engineering design will lose its reliability if sufficient measurement is not available for those points. The problem of estimation of temporal and spatial variation as described requires more advanced techniques to solve both time-delay and nonlinearity features. In this Coastal and Hydraulics Engineering Technical Note (CHETN), Artificial Neural Networks (ANNs) are used to address the missing data recovery problem for the data collection activities for a tidal lagoon, Biscayne Bay, FL.

BACKGROUND: The use of modern computing techniques including soft computing and numerical models and their integration has become commonplace in managing water resources projects. While the latter methodology has been popular to address the physical phenomena, the former technique is paid less attention by the researchers. The main advantages of using numerical models are based on their capability of prescribing the physical laws in the modeling domain. However, their accurate usage often requires extensive computational resources and validation using extensive field measurements, and many system parameters need to be estimated, particularly for large-scale and complex systems. Hsieh (1997) has proposed a framework design of flow model validation using the integration method of numerical model, stochastic filter, and system simulation techniques. This CHETN presents an application to address the missing data recovery problem in that design.

ANNs modeling techniques to solve tidal hydraulic problems are a relatively new area (Dibike and Abbott 1999; Tsai and Lee 1999). ANNs are able to solve problems in a way that resembles human intelligence (Khonker et al. 1998). It learns by examples. In the sense that observations provide knowledge, they are able to capture the knowledge within a data set. Unlike traditional artificial intelligence and statistical solution approaches, ANNs are able to solve problems without any prior assumptions. As long as enough data are available, a neural network will extract any regularities or patterns that may exist and use it to form a relationship between input and output. ANNs have probably become the most efficient tools for generalization problems. The technique is also able to provide a map from one multivariable space to another through training, even when given a set of data with noise. These properties make ANNs well suited to problems of estimation and prediction for flow phenomena. Usually, the data set is divided into training, cross-validation, and testing portions. The training part is used to identify the optimal weights to bridge the input/output series while the cross-validation is used to monitor the training

process to avoid overtraining. The testing part is used to examine the performance of the ANNs so it is not used in the training process.

The most popular ANNs algorithm is the classical multilayer perceptron (MLP) model. MLPs (Figure 1) are feed-forward neural networks trained with a standard backpropagation algorithm. This is a topology of ANNs with eight inputs, one hidden layer (three nodes), and one output system. They are supervised networks, so they must be trained for the desired response. They can learn how to transform the input data into the desired response if sufficient patterns are present in the training data set. With one or two hidden layers, an MLP can approximate the performance of optimal statistical classifiers in difficult problems. Two other two algorithms, namely time-lagged neural networks (TDNN) and recurrent neural networks (RNN) are more powerful algorithms to solve time series forecasting and prediction problems requiring the capability of addressing time delay problems.

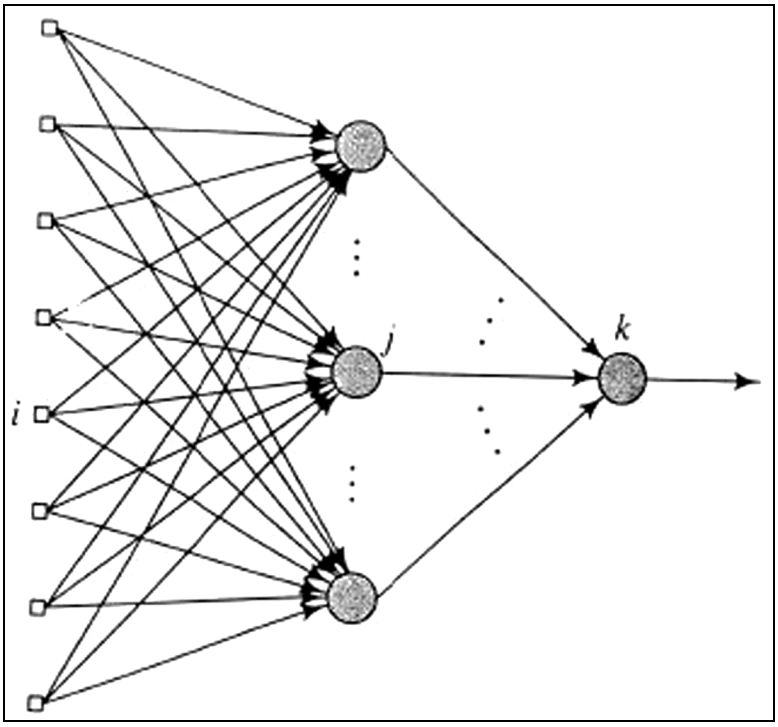


Figure 1. Fully connected feed-forward network
with one hidden layer and output layer

DATA RECOVERY SYSTEM (DRS): The DRS for missing data is based on the transfer function (response function) approach. The identification of system response is constructed by the training and cross-validation processes of learning from a common period between input and output series for ANNs. The testing portion (performance), which is not involved in the training, is used to compute the simulated output from additional input series. The simulated output using optimal weights from the best fit activation function (transfer function) can generate a recovered data series. This series is called the missing window. Three types of DRS are defined as follows:

a. *Self-recovery.* This type of recovery is based on a single time series itself. In this situation no other series can be used as the reference to create the response bridge. The method is

to break a long time series into two portions. The first part of the data is considered as the input, and the second part is regarded as the output function and contains the missing window. Longer time series training data sets that contain more significant patterns are critical for output performance.

b. *Neighboring station recovery.* This is the most typical recovery case. Obviously, the local recovery should have better performance than the remote recovery. If the involvement between input and output functions is a different parameter, this recovery is classified as the different parameter recovery. Otherwise, it is called the same parameter recovery.

c. *Multivariate parameter recovery.* Since the system response from input to output could involve more than one variable and receive different time delay, this more complex system requires physical cause and effect to identify the system structure. For example, the salinity variation for a particular location could be caused by the source tide, local wind, and nearby freshwater inflow for an estuary system.

STUDY AREA AND FIELD DATA COLLECTION PROGRAM: Pratt et al. (in preparation) summarizes the field data collection program for Biscayne Bay, a shallow, subtropical marine lagoon located on the southeast coast of Florida. It covers approximately 100 km from north to south and varies from less than 1.6 km to 13 km in width. It is bordered on the west by the south Florida mainland and on the east by a series of barrier islands and shallow, vegetated mud banks. The developed data sets and numerical models that can aid in the study and management of Biscayne Bay include circulation, salinity, and water quality. Bathymetry and geometry of the navigation channels, interconnecting canals and inlets, astronomical tide-induced currents, wind-induced currents, and freshwater inflow are major factors that determine circulation patterns.

The purpose of the field data collection program was to provide hydrodynamic results including velocities, flow distributions, circulation patterns, water levels, salinities, and meteorological measurements during long-term monitoring and short-term intensive surveys. The long-term monitoring equipment used to collect the data consisted of five bottom-mounted Acoustic Doppler Profiler (ADP) velocity meters, 12 water-level and salinity recorders, and one meteorological station within the study area.

KNOWLEDGE BASE AND ANNs MODELING: To perform the data recovery system, a number of stations with 15-min intervals during February 1998 were used to conduct the analysis. To identify the performance of ANNs, a week of data were purposely hidden to compare the simulation results. This recovery information is called the missing window in the system.

For the hydraulic engineering applications, the back propagation networks, the time-delayed networks, and the recurrent networks are used to perform the comparisons. The best performance was found to be partially recurrent networks. This Tech Note, unless indicated otherwise, will use the recurrent network to demonstrate the results. The software used for this study is NeuroSolutions (Version 3.02). The data set is divided into training (2 weeks), cross validation (1 week), and testing (1 week) portions. The performance analysis is represented by several

quantity statistical numbers, including mean square error, normalized mean square error (NMSE), mean/maximum/minimum absolute errors, and correlation coefficient (CC).

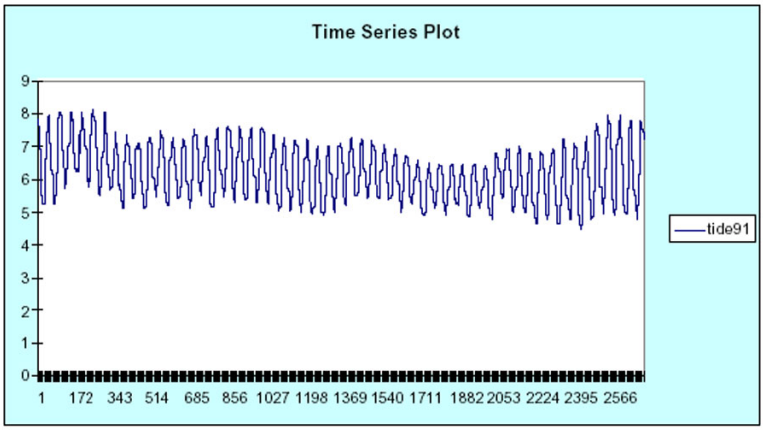
The following parameters are used to perform most of the recurrent networks:

Input layer
  AF=TDNN axon
  Depth in samples=10
  HL PE's= 2
  AF=linear axon
  Learn rule=momentum
  Step size=1.0
  Momentum factor=0.7
Hidden layer (HL)=1
Output layer
  AF=linear axon
  Learn rule=momentum
  Step size=0.1
  Momentum factor=0.7
Maximum epoches=1000

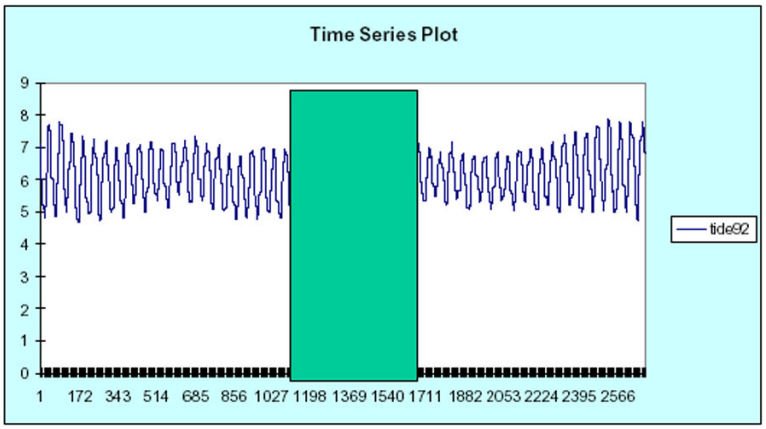
RESULT DEMONSTRATION:

a. *Self- recovery of surface elevations.* Tidal stations in the bay entrance are sometimes used to serve as the boundary condition for the numerical modeling study. The experience shows this application can avoid the iteration process for numerical modeling when the boundary condition is not available. The worst condition for the data recovery is that no other reference data set, such as neighboring station, can be used to construct the response function. Semidiurnal, diurnal, and neap-spring components dominate the harmonic constituents in the tidal system. It seems 2 months of surface elevations (two lunar cycles) are sufficient to construct the self-recovery scheme. The record can be divided into two parts: the first month's data are assumed as the input series (Figure 2a) and the second month's data with missing window (Figure 2b) are used as the output series. Shaded portions of the figures represent missing windows. The testing process of ANNs modeling creates the estimation of the missing window. Very good agreement (Figure 2c) was found from this missing window recovery result (CC=0.9716 and NMSE=0.0996).

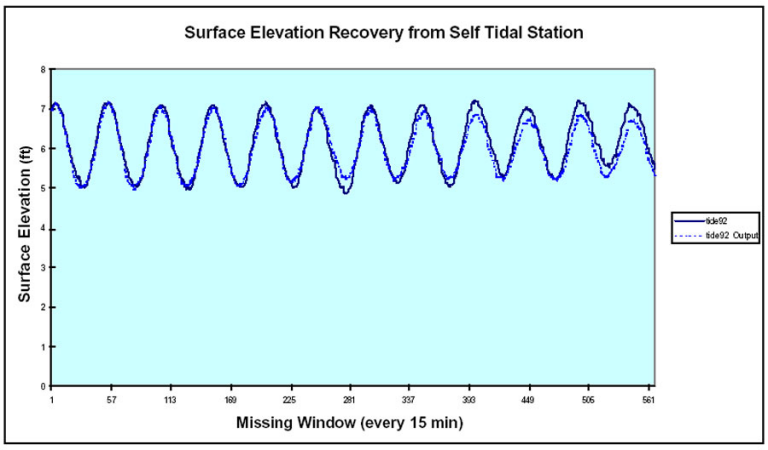
b. *Neighboring station recovery.* The first demonstration of this recovery was to use the tidal station (tide 8) (Figure 3a) to recover the partial missing record in a tidal station (tide 9) (Figure 3b) which is 9.6 km away. This is the most typical tidal signal propagation problem due to the friction effect. The excellent performance (Figure 3c) was obtained by using the recurrent ANNs (CC=0.9906 and NMSE=0.0205). Using the surface elevation to recover the salinity concentration at the same location (sta 11) was the second application. The poor results (CC=0.5576 and NMSE=0.6950) were due to other forcing factors, such as wind stress, freshwater inflow, and the local effect.



a. Input series for surface elevation at lagoon entrance (ft)

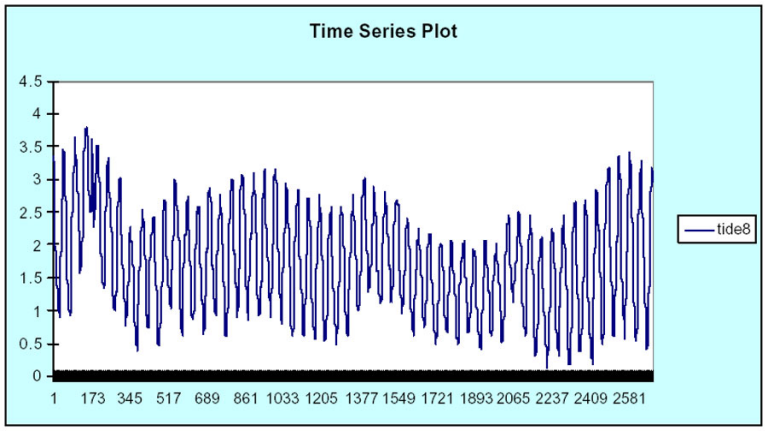


b. Output series for surface elevation at lagoon entrance (ft)

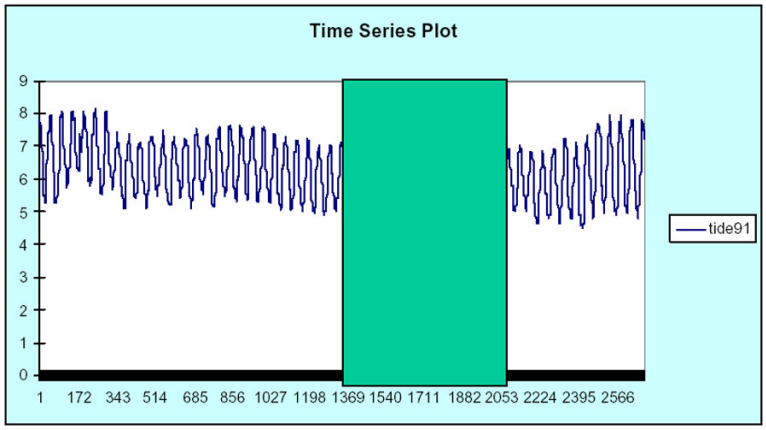


c. Self-recovery for surface elevation at lagoon entrance (ft)
(missing window recovery (dashed line))

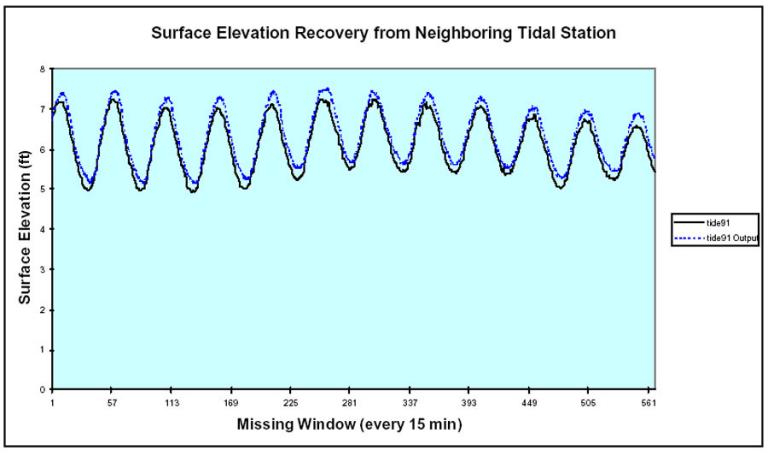
Figure 2. Self-recovery of surface elevations
(To convert feet to meters, multiply by 0.3048)



a. Input series for surface elevation within lagoon (ft)



b. Output series for surface elevation at lagoon entrance (ft)



c. Neighboring station recovery for surface elevation at lagoon entrance (ft)
(missing window recovery (dashed line))

Figure 3. Neighboring station recovery
(To convert feet to meters, multiply by 0.3048)

c. *Multivariate parameter recovery.* The tidal current is a very important parameter in the tidal hydrodynamic system and is costly to collect. The x-component wind stress (Figure 4a) and the x-component of current in sta 3 (Figure 4b) receive signals from the ocean tide (sta 9 from 6.4 km away). The data recovery for this case is shown in Figure 4c. Except for the small short periodic variation, the results show very good pattern match.

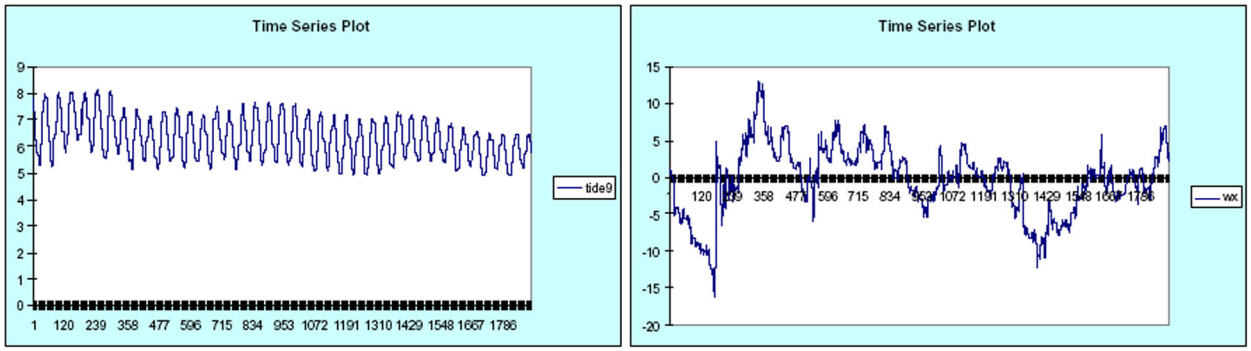
RECOVERY RELIABILITY:

a. *Physical forcings as input.* As in the previous analysis, the reliability of data recovery also depends on the selected parameter and how well the forcing functions are included as the model input. A comparison (Table 1) addresses the reliability of tidal current recovery due to physical processes, namely, tidal forcing, surface slope, and related physical parameters. While tidal current due to the surface slope between two neighboring surface elevation shows the best results, two other approaches also obtain very high correlation. The main source of errors comes from small-scale variation. This could be caused by any other local effect or other physical parameters not addressed well enough. The results show that the wind stress contributes only very minor improvement for the analysis. This is probably because the effects of the wind stress are longer duration (sample depth) than the tidal forcing. A further analysis using mixture networks to separate the input influence could be the alternative approach.

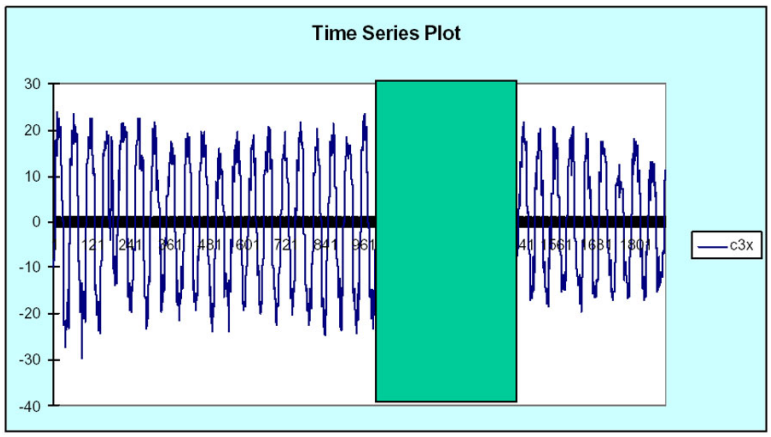
Table 1
Reliability of Tidal Current Recovery Due to Physical Forcing Parameters (Correlation Coefficient and NMSE (cm/sec))

Input/Output	Training	Cross-Validation	Missing Window
Tide/tidal current	0.933 (0.131)	0.935 (0.126)	0.967 (0.080)
Surface slope/current	0.944 (0.110)	0.943 (0.111)	0.972 (0.069)
Tide; wind/current	0.936 (0.126)	0.939 (0.118)	0.968 (0.077)
Wind/current	0.105 (0.988)	0.104 (0.994)	0.100 (0.995)

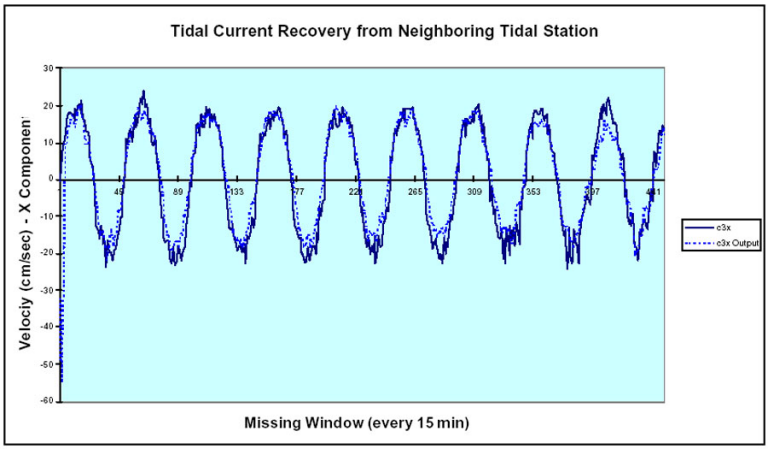
b. *Missing window size.* An important objective for simulating the missing DRS is to determine how the performance might be related to the size of the missing window. A comparison was conducted by using a tidal current simulation example with missing window sizes of 100, 200, 300, and 600 values. While the training data used the same length of record, the cross-validation used less information when the missing window enlarged. No significant differences (Table 2) were found from both training and cross-validation (CC and NMSE). The reliability of missing data recovery gets lower as the window size gets smaller. This unexpected result is due mainly to the initial simulation having larger errors than the following time-steps. When the window size gets smaller, these errors contribute a higher percentage of total error to the overall performance. This suggests that, when the missing window gets very small, the simulation window could enlarge the window in the beginning end (about 20 more time-steps from this case).



a. Input series for surface elevation at lagoon entrance (ft) and wind stress (m/sec)
(To convert feet to meters, multiply by 0.3048)



b. Output series for tidal current in sta 3 (cm/sec)



c. Multivariate parameter recovery for tidal current using surface elevation and wind stress as inputs (recovery – dashed line)

Figure 4. Multivariate parameter recovery

Table 2
Recovery Reliability of Tidal Current (Tidal Forcing and Wind Stress as Inputs) Due to Missing Window Size (Correlation Coefficient and NMSE (cm/sec))

Window Size	Training	Cross-validation	Missing Window (Testing)
100	0.955 (0.088)	0.967 (0.074)	0.880 (0.246)
200	0.955 (0.088)	0.966 (0.077)	0.926 (0.146)
300	0.953 (0.093)	0.968 (0.081)	0.946 (0.108)
600	0.954 (0.090)	0.967 (0.077)	0.957 (0.084)

c. *Missing window location.* The main feature using ANNs is to recognize and learn the historical patterns. Therefore, another critical issue for DRS is the reliability due to the missing window location from the entire data set. This is particularly important when the data length is not very long. This test is applied to the simulation of tidal current due to the surface slope between two neighboring surface elevations. The original data set was divided into four quarters. Two quarters were used to perform the training, one quarter was used to conduct the cross-validation, and the remaining quarter was used to generate the missing window (testing). Four combinations (Table 3) with the sequences of training, cross-validation, and testing (missing window) were investigated by checking the performance due to the location of the missing window. The highlighted correlations in Table 3 show very satisfactory performance. The analysis indicated that this is due primarily to the pattern similarity between the data from quarters 2 and 4. The pattern for data from quarter 1 is quite different from the other quarters. Therefore, the pattern similarities are still the major factor to assure the good performance for missing data recovery. It is not because of the order of data representation during the learning processes.

Table 3
Recovery Reliability of Tidal Current (Surface Slope) Due to Location of Missing Window (Correlation Coefficient and NMSE (cm))

Data Representation (quarters)			Training	Cross-validation	Missing Window
Tr	C-V	Te			
1, 2	3	4	0.930 (0.140)	0.924 (0.152)	**0.971 (0.065)**
1, 4	2	3	0.914 (0.171)	**0.972 (0.056)**	0.919 (0.158)
3, 4	1	2	0.952 (0.095)	0.898 (0.300)	**0.978 (0.055)**
2, 3	4	1	**0.976 (0.046)**	**0.967 (0.071)**	0.916 (0.219)

CONCLUSIONS: ANNs were used to simulate missing data recovery. The partially recurrent networks receive the best performance for a tidal lagoon system in the Biscayne Bay data collection program. The surface elevation is the easiest physical parameter for self-recovery, neighboring station recovery, and multivariate recovery, while the conservative parameters, such as salinity, are more difficult to recover due to the complex input system and their response speed. The mixture ANN approach may be the alternative to improve the solution. The performance due to missing window size is not only directly related the length of total data but also associated with the initial portion of the simulation. The data representation of assigning the

order of learning processes due to the missing window location is not significant. The degree of pattern similarity between the training data and the testing data determines the performance.

ADDITIONAL INFORMATION: For further information, contact Dr. Bernard B. Hsieh (Voice: 601-634-3679, e-mail: hsiehb@wes.army.mil or Mr. Thad C. Pratt (Voice: 601-634-2959, e-mail: Thad.C.Pratt@usace.army.mil), U.S. Army Engineer Research and Development Center, Coastal and Hydraulics Laboratory. For information about the Coastal Inlets Research Program, please contact Dr. Julie Rosati, Julie.D.Rosati@usace.army.mil . Any mention of a commercial product does not constitute an endorsement by the Federal government. This CHETN should be cited as follows:

> Hsieh, B. B. and Pratt, T. C. (2001) "Field data recovery in tidal system using artificial neural networks (ANNs)," Coastal and Hydraulics Engineering Technical Note CHETN-IV-38, U.S. Army Engineer Research and Development Center, Vicksburg, MS. http://chl.wes.army.mil/library/publications/chetn/

REFERENCES:

Dibike, Y., and Abbott, M. (1999). "Application of artificial neural networks to the simulation of a two dimensional flow," Journal of Hydraulic Research, 37(4), 435-446.

Hsieh, B. B. (1997). "An optimal design of field data collection for improving numerical model verification," Proceedings of the 27th Congress of the International Association of Hydraulic Research, San Francisco, 815-820.

Khonker, M., et al. (1998). "Application of neural networks in real time flash flood forecasting," Proceedings of Hydroinformatics 98, Babovic and Larsen (ed), Copenhagen, Denmark, August 24-26, 1998, 777-781.

McAdory, R., Pratt, T. C., Hebler, M. T., Fagerburg, T. L., Curry, R. (in preparation). "Biscayne Bay field data report," U.S. Army Engineer Research and Development Center, Coastal and Hydraulics Laboratory, Vicksburg, MS.

NeuroDimension, Inc. "NeuroSolutions, v3.02, developers level for windows," Gainesville, FL.

Tsai, C., and Lee, T. (1999). "Backpropagation neural network in tidal-level forecasting," Journal of Waterway, Port, Coastal, and Ocean Engineering, ASCE 125(4), 195-202.