Time Series Scenario Composition Framework in Supporting Environmental Simulation Tasks

Chi-Yu Li[1], Frank Molkenthin[2]

Abstract

To answer the impacts under specific what-if scenarios together with simulation tools has been demanding in different environmental problems. In this contribution, a general software framework for time series scenario composition is proposed to deal with this issue. It is done through providing an interface to process available raw time series data and to compose scenarios of interest. These composed scenarios can be further converted to a set of time series data, e.g. boundary conditions, for simulation tasks in order to investigate the impacts. This software framework contains four modules: data pre-processing, event identification, process identification, and scenario composition. These modules mainly involve Time Series Knowledge Ming (TSKM), fuzzy logic and Multivariate Adaptive Regression Splines (MARS) to extract features from the raw time series data and then interconnect them. These extracted features together with other statistical information form the most basic elements, MetaEvents, for the semi-automatic scenario composition. Besides, a software prototype with two application examples containing measured hydrological and hydrodynamic data are used to demonstrate the benefit of the concept. The results present the capability of reproducing similar time series patterns from specific scenarios comparing to the original ones as well as the capability of generating new artificial time series data from composed scenarios based on the interest of users for simulation tasks. Overall, the framework provides an approach to fill the gap between raw data and simulation tools in engineering suitable manner.

1. Introduction

Due to the rapid development of different computer and sensor technologies, scientists and engineers nowadays are able to collect, access, handle, etc. huge amount of data more easily than before. Moreover, there also exist a number of simulation tools focusing on different aspects of different environmental problems with different time and space scales. However, one tool, which fills the gap between available mass raw data and simulation tools, is still not commonly seen, especially in investigating the impacts from certain what-if scenarios for further decision-making.

Some tools and studies also try to fill the gap between mass raw data and simulation tools in answering the impacts from certain scenarios. For instance, the National Groundwater Modelling System (NGMS) [1] contains several predefined scenarios for groundwater problems together with MODFLOW as the simulation tool. Some studies, such as in [2], use Monte Carlo based approaches to generate reasonable inputs for further simulation tools for flood risk assessment. In this contribution, a different approach is proposed by providing a flexible interface for users to create the scenarios of their interest. The entire process is semi-automatic and comprises different modules from data pre-processing to scenario composition. One purpose of this framework is to form the most basic elements, MetaEvents, which represent specific features of a set of independent collected time series data and also contain corresponding statistical information. The information, which MetaEvents contain, is not isolated but complementary. They know the “natural

[1] Brandenburg University of Technology Cottbus-Senftenberg, 03046 Cottbus, Germany, Chi-Yu.Li@tu-cottbus.de, Chair of Environmental Informatics

[2] Brandenburg University of Technology Cottbus-Senftenberg, 03046 Cottbus, Germany, Frank.Molkenthin@tu-cottbus.de, Chair of Environmental Informatics

order" of the phenomena through the processing of the available data. With such information, MetaEvents can be used as LEGO® bricks to compose or build scenarios of interest in engineering suitable manner. Afterwards, these scenarios can be converted into a set of corresponding time series data for further simulation tasks, e.g. boundary conditions.

2. Framework Concept

To fill the gap between available mass raw time series data from different sources with different space and time scales and the simulation tools, a software framework, as shown in Fig. 1, is proposed to resolve this issue. As shown in Fig.1, this framework contains four modules: data pre-processing, event identification, process identification, and scenario composition.

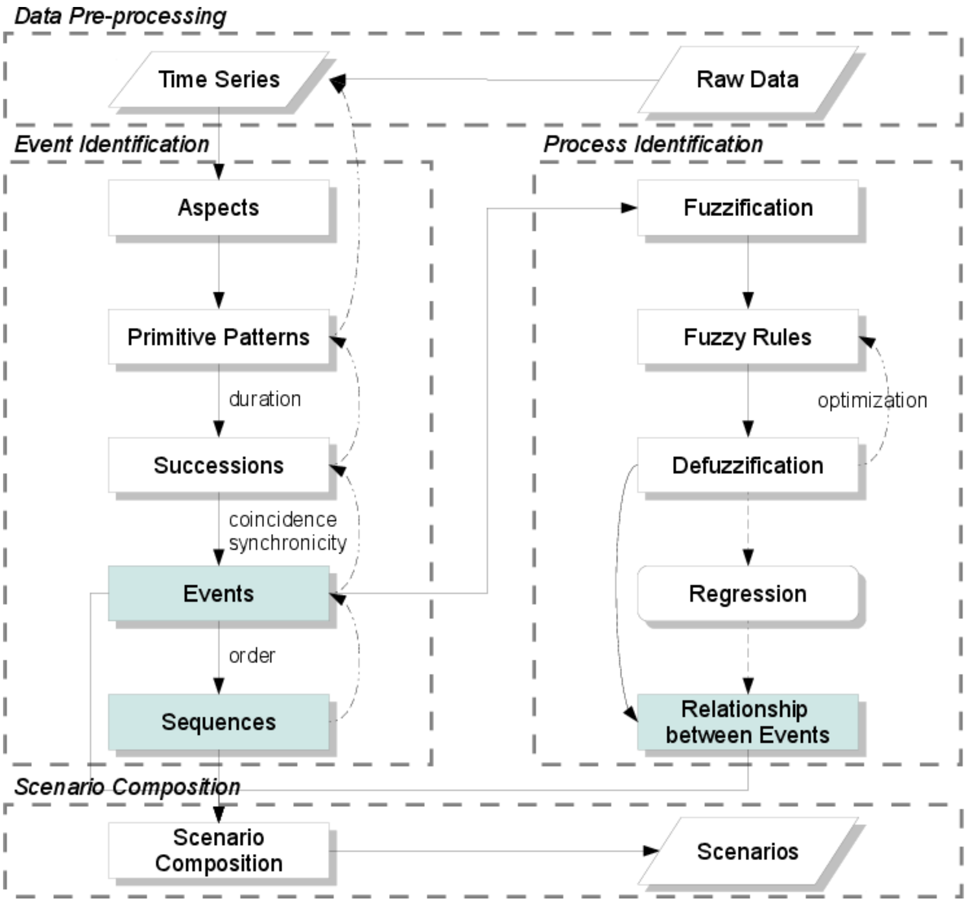

Figure 1: Framework of Scenario Composition

2.1. Data Pre-processing

Data pre-processing is a crucial step in any data-driven approach and so is in this framework, since it is also data-driven based. It prepares the raw data into the necessary time series sets, Aspects, for the subsequent steps in this framework. It is also problem- and domain-specific and can contain different techniques, such as gap-filling, noise-removing, etc. At the end of the data pre-processing, a set of time series data are grouped into different Aspects with different semantics, e.g. different physical state variables, different time series patterns, etc. For instance, a collection of air temperature and soil moisture time series data can be regarded as two Aspects due to the reason of different physical state variables; two precipitation time series data sets collected from different meteorological stations, e.g. in a plain and on a hill, can be considered two Aspects if they have different patterns.

2.2. Event Identification

The purpose of event identification is to extract features, Events, within a certain time interval among the entire time series data set based on the Time Series Knowledge Mining (TSKM) proposed by Mörchen [3]. These extracted Events will be in human-readable format. For instance, a hydrological data set containing the information of air temperature, soil moisture and precipitation can be identified an Event as *aridity* when the air temperature is *high*, the soil moisture is *low* and the precipitation is *low* if dry season occurs.

Event identification contains a sequence of steps, as shown in Fig. 1, and different techniques, such as the methods of clustering, segmentation, rule generation, etc., to extract Events as it is based on TSKM. These techniques are also problem- and domain-specific. The sequence of steps is semi-automatic and the knowledge and experiences of domain experts can be also taken into consideration while extracting Events.

2.3. Process Identification

What process identification does is to provide a way to identify and describe physical relationships among different state variables since the Events derived from event identification are descriptive and loosely-connected. This process identification is built upon Mamdani-type fuzzy inference system with Multivariate Adaptive Regression Splines (MARS) for better interpretation [4], and the later serves as a second mapping function to ensure better descriptions of the phenomena of interest.

2.4. Scenario Composition

As shown in Fig. 1, the information from both event identification and process identification will be merged in the module of scenario composition. Such information will form the basic component, MetaEvent, in scenario composition. The MetaEvent is the Event together with its metadata derived from the collected data, e.g. statistical information, describing the characteristics of the extracted features. A scenario is composed by a sequence of MetaEvents and it describes a situation which could possibly happen.

The object of scenario composition is to provide a manner to compose synthetic scenarios of interest and, then, to generate the corresponding time series data based on these user-defined scenarios. To achieve this object of scenario composition, the framework has to offer users the necessary information of each MetaEvent, e.g. duration, mean values, next possible MetaEvents, etc., and an interface, e.g. Graphical User Interface (GUI), to compose the scenarios of interest.

3. Prototype and Application Examples

3.1. Prototype Design and Implementation

In order to demonstrate the concept of this framework, a prototype is designed and implemented in R and Java software technologies, and the information between R and Java environment is implemented with the help of Java Native Interface (JNI). Due to the reasons that lots of techniques involved in the framework is problem- and domain-specific, only the general algorithms are implemented in this prototype and the following examples is also carried out with this generalized prototype. Apart from being a data management/generation tool, it also keeps track of the related metadata to ensure its traceability. Besides, this implementation of prototype can work as a standalone application as well as serve as an add-on to other simulation tools.

3.2. BinghamTrib Data Set

Since the framework offers an interface for users to compose any scenario of interest, it also implies that it is difficult to "validate" the results with the traditional sense of validation. However, it demonstrates the capability of this framework to reproduce similar time series patterns comparing to the original ones from a specific scenario in this application example.

In this application example, the hydrological data BinghamTrib data set from the R package `hydromad` [5] is used for this demonstration. This data set contains three physical state variables: rainfall (mm/day), temperature (°C), and streamflow (mm/day). Among them, rainfall and streamflow data are collected for the Bingham River Trib at Ernies Catchment (2.68 km^2) by Department of Water, Water Information Provision section, Perth Western Australia. Temperature data are collected by Bureau of Meteorology, Australia. The collected time series data are on a daily basis and from 1974-05-18 to 2008-11-02 as shown in Fig. 2.

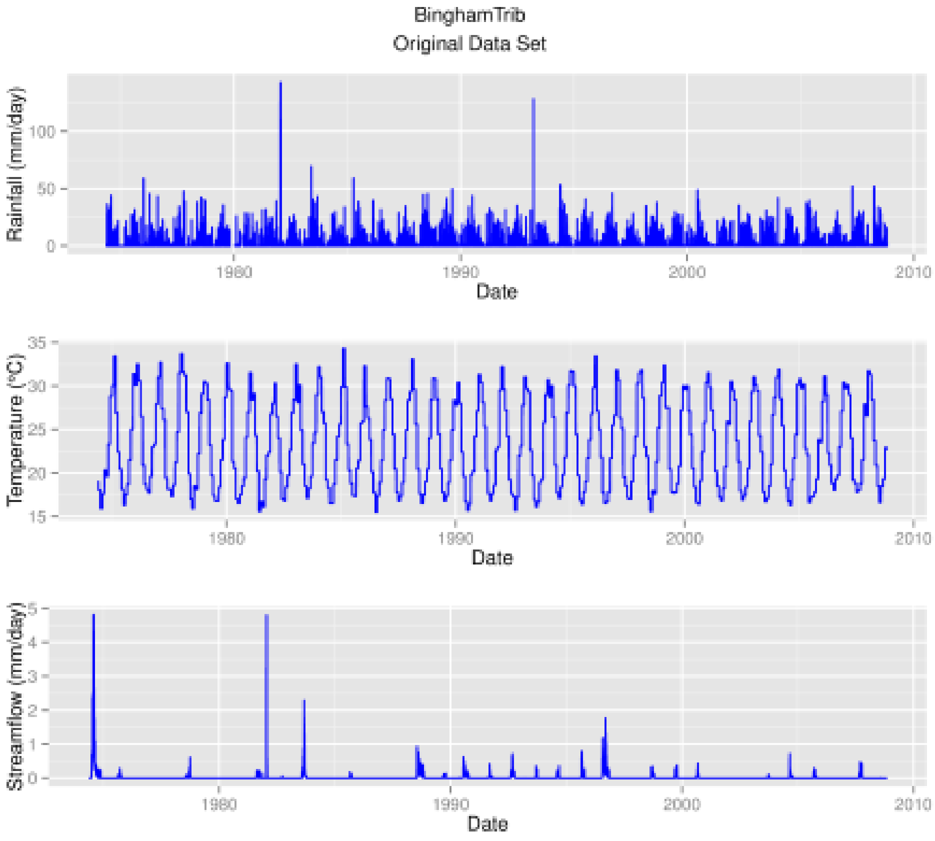

Figure 2: Original BinghamTrib Data Set from R package `hydromad`

This data set is almost complete, and only some rainfall records are missing. Through inspecting the history of rainfall and the neighboring values of the gaps, these gaps are filled with the value of zero. Moreover, to derive MetaEvents, the time series set has to be grouped into meaningful categories according to the concept of TSKM. This can be done manually with the help of expert's knowledge, regulations, or through different possible techniques, e.g. clustering. Here, for the demonstration purpose, the general purpose *k*-means clustering is applied on this data set. Depending on the type of data and through trial and error, these rainfall, temperature and streamflow data are grouped into five, five and three categories. These grouping together with the default settings of the prototype, there are total 42 MetaEvents derived. With these derived 42 MetaEvents as basis, a specific scenario is composed to reproduce the original data and the comparison between the original time series data and the scenario-converted ones in the range of

10 years from 1990 to 2000 is shown in Fig. 3. As shown in Fig. 3, with the specific scenario, it is possible to reproduce original time series patterns. Two derived MetaEvents used to compose this scenario, MetaEvent 21 and MetaEvent 41, are also indicated in Fig. 3, and MetaEvent 41 is able to capture the rainfall peak value in year 1993. While looking into the generated time series data, stepwise patterns can be observed due to the reason that the current prototype implementation uses the mean value of each physical state variable within each MetaEvent as a default value for time series data generation. This reason also leads to the fact that these generated default values are not able to catch extreme values.

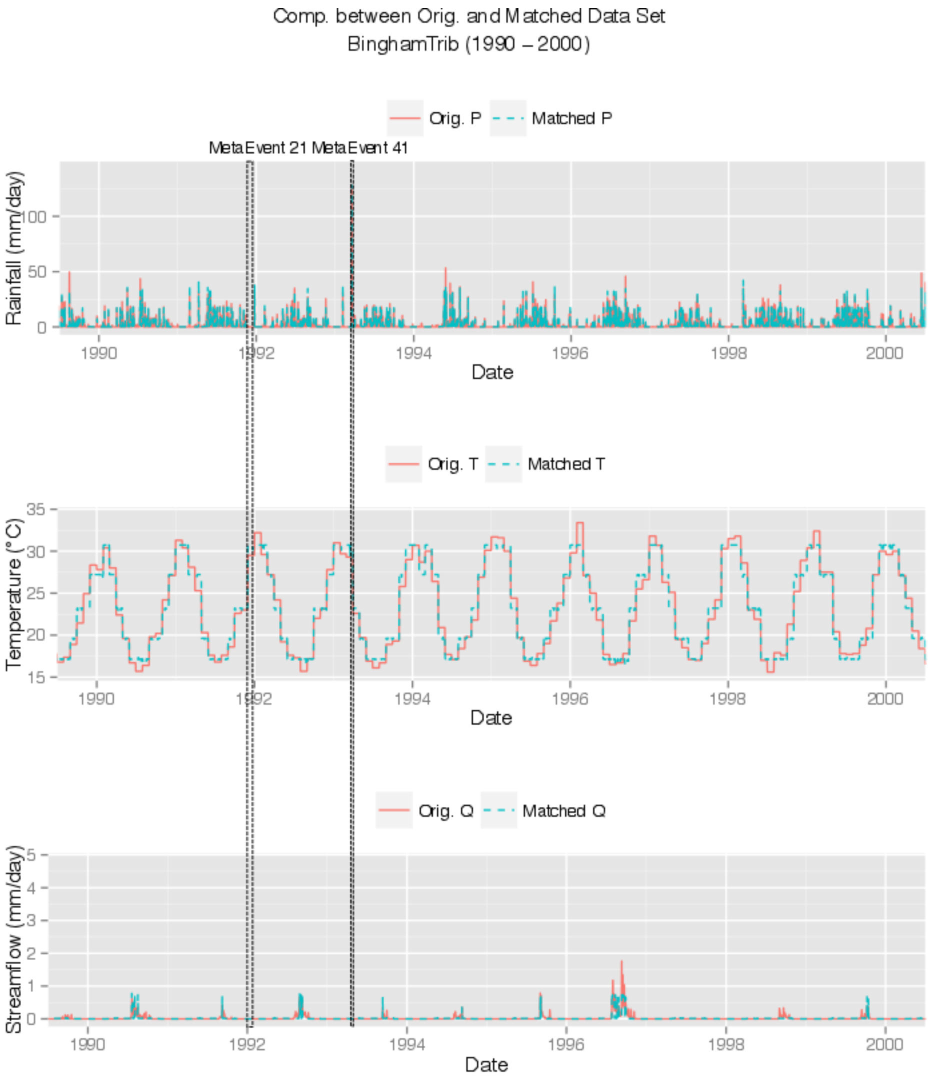

Figure 3: 10-year Comparison of Original BinghamTrib Data Set with the Matched MetaEvent Default Values (BinghamTrib Data Set 1990 – 2000)

3.3. Oder River Data

In this application example, the data set describing the 1997 Oder Flood in Germany [6] is chosen to demonstrate how the framework can be used with other simulation tools. For this purpose, the focus of this application example is to generate time series data from a specific user-defined scenario, and then to apply these data as boundary conditions in the selected simulation tool. The study area of this application example focuses on the part of the Oder river, starting from the town of Eisenhüttenstadt to the city of Frankfurt (Oder), on the border between Germany and Poland, and the distance of this river section is about 30 km long.

For the purpose of the demonstration, this simulation is simplified and only the flow condition of the river from the town of Eisenhüttenstadt to the city of Frankfurt (Oder) without dam breach is considered. With this simplification, DHI Mike 11 is chosen as the 1-D simulation tool for this demonstration and time series data collected for this application with different resolution and time span are as follows:

- discharge data (m^3 /s) at daily intervals from 1996-01-11 to 1997-11-01 at Eisenhüttenstadt
- water level data (m) at 15-minute intervals from 1996-11-01 to 1997-11-02 at Eisenhüttenstadt
- water level (m) at 15-minute intervals from 1996-11-01 to 1997-11-02 at Frankfurt (Oder)

Before this demonstration, a simple pre-processing is carried out to keep all time series data have the same time span and resolution. This is done by reducing the time span from 1996-11-01 12:00:00 to 1997-11-01 12:00 and converting the daily discharge data at Eisenhüttenstadt to the ones at 15-minute intervals with the help of spline interpolation as shown in Fig. 4.

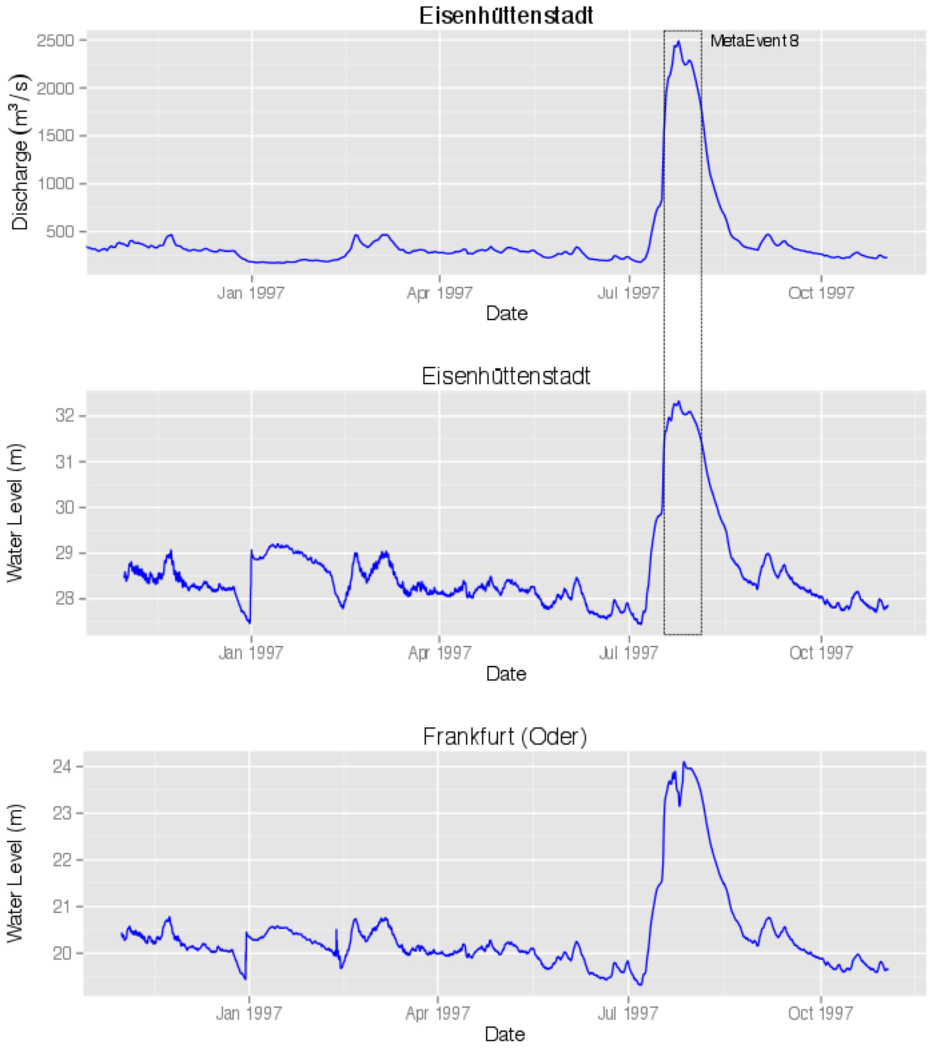

Figure 4: Measured Time Series Data at Eisenhüttenstadt (1996-11-01 to 1997-11-02) and a derived MetaEvent describing the "peak"

As the steps in the previous example, the discharge data at Eisenhüttenstadt and the water level data at Frankfurt (Oder) are grouped into five and three categories respectively, and total numbers of nine MetaEvents are derived. For instance, a derived MetaEvent 8 describing the "peak" is also shown in Fig. 4. This MetaEvent 8 contains some basic information, for instance: mean discharge value of 2239.40 m^3/s at Eisenhüttenstadt, mean water level of 23.72 m at Frankfurt (Oder), 4% of the event frequency, etc. Based on these derived MetaEvents, a scenario containing two peaks instead of the original one-peak scenario is composed to demonstrate the capability of composing scenarios of interest and the usage of the framework together with other simulation tools. The generated time series data for discharge at Eisenhüttenstadt and the water level at Frankfurt (Oder) based on this two-peaks scenario are as shown in Fig. 5. The first peak is composed by MetaEvent 6, MetaEvent 7 and MetaEvent 9, and the second peak is composed by MetaEvent 7, MetaEvent 8 and MetaEvent 7 as shown in Fig. 6. These generated discharge and water level time series data are served as upper and lower boundary conditions individually for a calibrated and validated 1-D Mike 11 model. As mentioned earlier, the default generated time series data appear stepwise patterns due to the settings of current prototype implementation. The water level at downstream based on this two-peaks scenario is simulated as shown in Fig. 6.

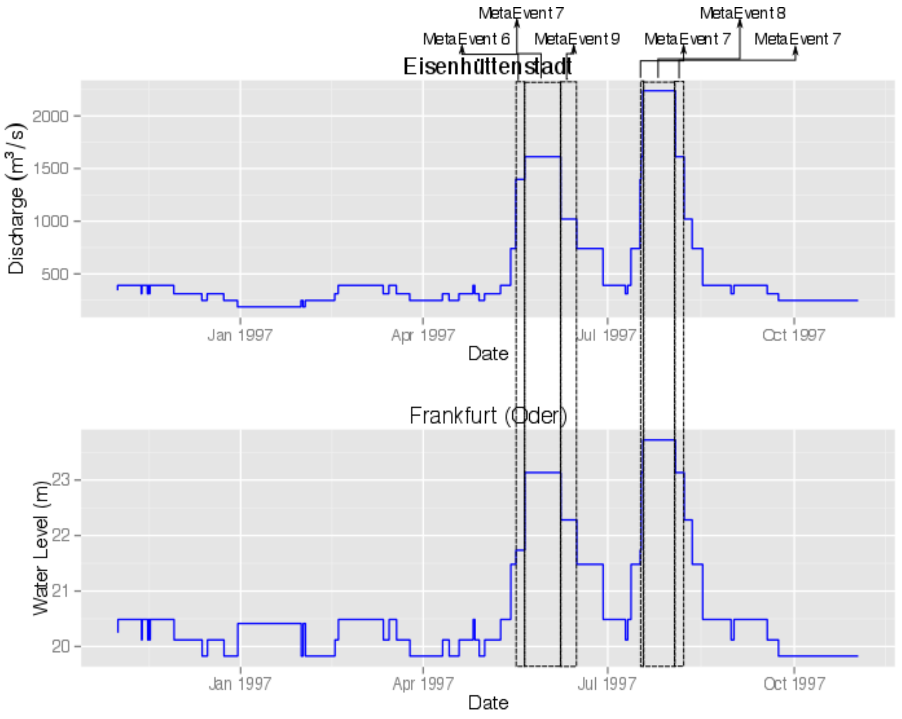

Figure 5: Generated Two-Peaks Time Series Data

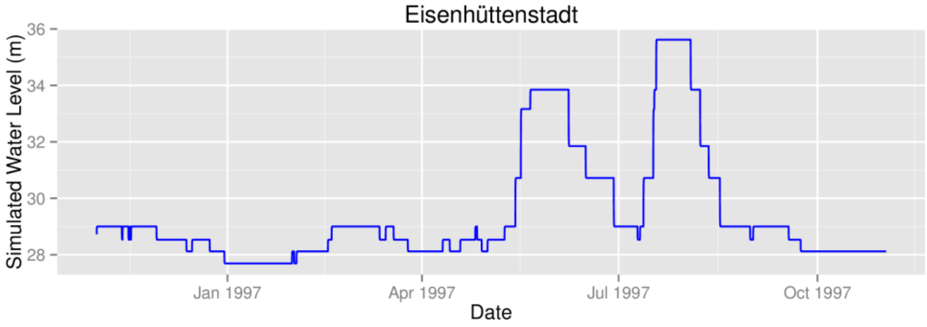

Figure 6: Simulated Water Level at Eisenhüttenstadt

4. Conclusions

In order to fill the gap between mass raw data and simulation tools, a framework of time series composition is proposed here to resolve this issue. With this framework, features of a set of time series data are extracted and form the most basic elements, MetaEvents, for scenario composition. With these MetaEvents, which contain not only features but also corresponding statistical information, users can compose the scenarios of interest with the help of information offered from the framework. The process to break a set of time series and to form MetaEvents is semi-automatic and the techniques required are problem- and domain-specific. In this contribution, a generalized prototype is introduced and this prototype with two different application examples also demonstrate the concept and the capability of the framework. Due to the generalization of the prototype, it still shows the room for improvement, especially in the generated time series data. Overall, it provides a general tool for users to be able to compose scenarios of their own interest and to convert the data for further simulation tasks to answer the impacts under such scenarios in engineering suitable manner.

References

[1] M. I. Whiteman, C. H. Maginness, R. P. Farrell, P. J. A. Gijsbers, and M. Ververs, “The National Groundwater Modelling System: providing wider access to groundwater models,” *Geol. Soc. Lond. Spec. Publ.*, vol. 364, no. 1, pp. 49–63, Jan. 2012.

[2] A. J. Kalyanapu, D. R. Judi, T. N. McPherson, and S. J. Burian, “Monte Carlo-based flood modelling framework for estimating probability weighted flood risk,” *J. Flood Risk Manag.*, vol. 5, no. 1, pp. 37–48, 2012.

[3] F. Mörchen, “Time Series Knowledge Mining,” Dept. of Mathematics and Computer Science, University of Marburg, Germany, 2006.

[4] M. B. Gorzałczany, Computational Intelligence Systems and Applications: Neuro-Fuzzy and Fuzzy Neural Synergisms. Springer, 2002.

[5] F. Andrews and J. Guillaume, hydromad: Hydrological Model Assessment and Development. 2012.

[6] Landesumweltamt Brandenburg (LUA), Ed., “Das Sommerhochwasser an der Oder 1997 - Fachbeiträge anläßlich der Brandenburger Ökologietage II,” in *Studien und Tagungsberichte, Schriftenreihe des Landesumweltamtes Brandenburg*, vol. 16, Potsdam, 1997.