The Fifth Japan-Taiwan International Workshop
on Hydrological and Geochemical Research for Earthquake Prediction

Comparison of Several Anomaly Detection Methods on the Seismic Groundwater Level Series

**Tzong-Yeang Lee, Shu-Chen Lin,
Wei-Chia Chen, Feng-Sheng Chiu,Tzu-Cheng Chiu**

10-12 October, 2006, Tsukuba, Japan

Acknowledgement (I)

The first author would like to express the deep thanks for the invitation and financial support of *Tectono-hydrology Research Group, Institute of Geology and Geoinformation, Geological Survey of Japan National Institute of Advanced Industrial Science and Technology (AIST)* and be sure of good discussions on the topic of the earthquake-related groundwater changes.

Acknowledgement (II)

This work was supported in part by the Water Resources Agency (WRA), Ministry of Economic Affairs.

The authors would like to thank the Disaster Protection Research Center (DPRC) of National Cheng-Kung University (NCKU) for kindly permitting us to participate the “Planning of Groundwater Anomalies Associated with the Earthquake” (’01-’05) project and the “Development of Tectono-hydrology Monitoring System and Application of the Research Results” (’06-’09) project.

AGENDA

- □ Introduction
- □ Motive and Purpose
- □ Strategy (Methods and Procedures):

  Factors (Noises) Filtering Model

  - BAYTAP-G
  - TFM

  Methods of Anomaly Detection

  - anomaly announcement form (AAF)
  - outlier analysis (OA)
  - the variation of grey-window shifting (Di)
  - the measure of grey variation information series (Es)
  - the cutting series of grey progressive sliding (Em)

  Based on the grey theory
- □ Case Studies
- □ Concluding Remarks

Introduction *[1/5]*

- The earthquake event will often react out through the interface of the environment; the groundwater is a comparatively apparent one in a great deal of variables.
- The groundwater level (GWL) is apt to receive influences of the environmental factors, like as rainfall, tide, atmospheric pressure, river water-level and artificial pumping.
- These factors increase the difficulties to analyze the variability of GWL induced by the earthquake.

Introduction *[2/5]*

- □ To analyze these effects objectively, the noises to affect the GWL must be filtered out in advance.
- □ The development of factors (or noises) filtering model is needed and expected that it is more convenient to explore, interpret and analyze the physical (e.g. abnormal) phenomena caused by the earthquake event.
- □ In this study, there are two filtering models to be selected for this purpose. One is the BAYTAP-G and the other one is TFM. (The details will be described later)

Introduction *[3/5]*

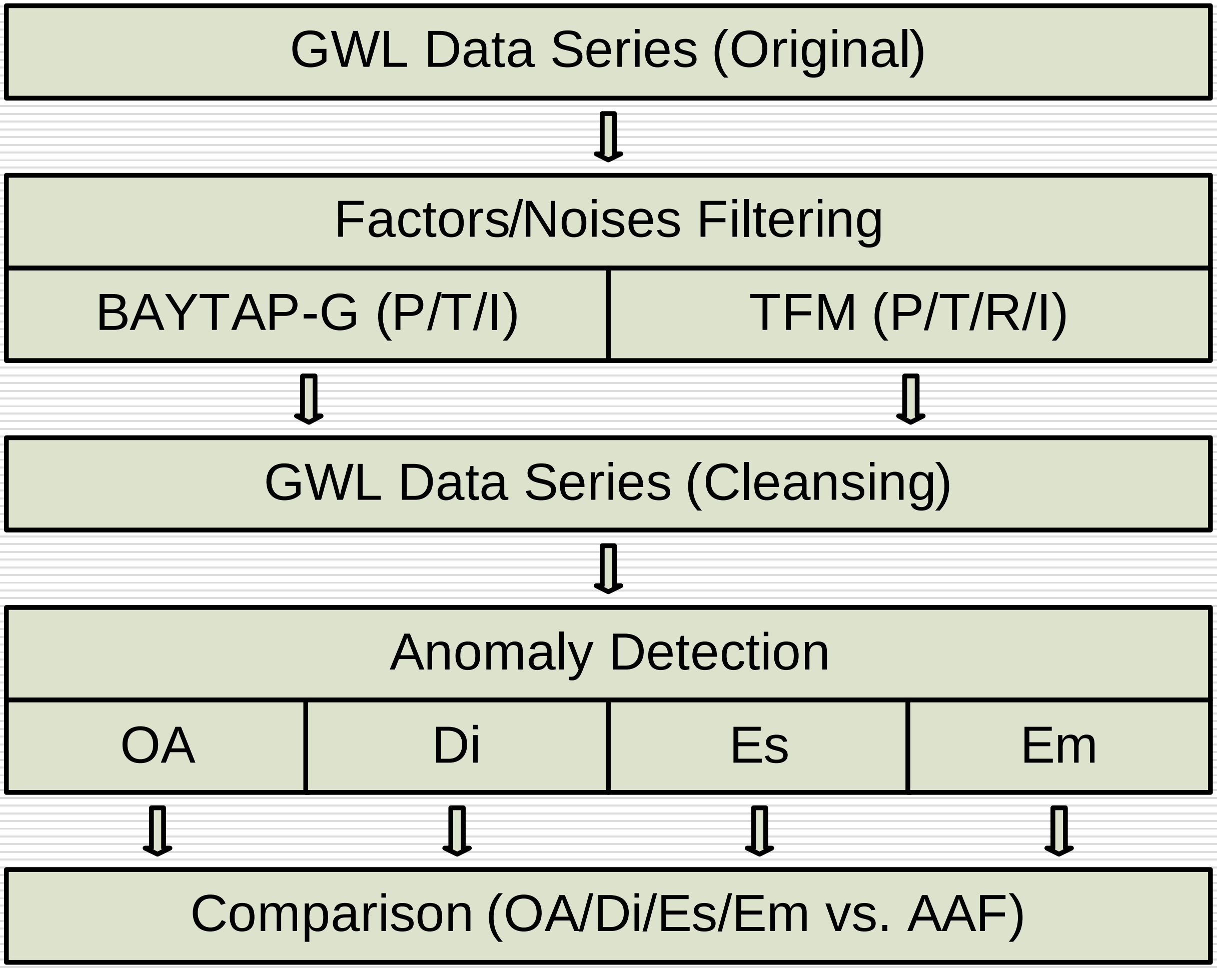
- □ If the BAYTAP-G or TFM is used to filter out the influences of affecting the original GWL data series, including the atmospheric pressure, tide, rainfall and irregular signal. After this procedure, the data can be taken as the “cleansing” data.
- □ Next, one thing is important. It is how to explore or decide the anomaly of the cleansing data.
- □ In this study, four detection methods are selected to check or test the cleansing data. The first one is based on the statistical theory (OA) and the others are based on the grey theory (Di, Es, and Em). (The details will be described later)

Introduction [4/5]

□ Two models are used for filtering the original GWL data and four methods are applied to detect the anomaly of the cleansing data in this study.

□ All the results are compared with the “Anomaly Announcement Form (AAF)” established by the Disaster Protection Research Center, National Cheng-Kung University.

Introduction [5/5]



GWL Data Series (Original)

Factors/Noises Filtering

BAYTAP-G (P/T/I)

TFM (P/T/R/I)

GWL Data Series (Cleansing)

Anomaly Detection

OA

Di

Es

Em

Comparison (OA/Di/Es/Em vs. AAF)

P = atmospheric pressure
T = tide
R = rainfall
I = irregular signal

OA = outlier analysis
Di = the variation of the grey-window shifting
Es = the measure of the grey variation information series
Em = the cutting series of the grey progressive sliding

The Flowchart of Data Analysis

Motive and Purpose

- One of objective in the project is to offer the (computer) tools for exploring the groundwater micro-behavior and explaining the interrelation of earthquake and groundwater.
- In this study, we focus more attentions on the development of the automatic procedures to achieve the goal described above.
- The automation of data analysis is necessary for the project, but the performance of the anomaly detection should be more concerned.

Factors (Noises) Filtering – BAYTAP-G

- □ The BAYTAP-G model is developed by the Institute of Statistical Mathematics and National Astronomical Observatory in Japan.
- □ The model can be used to filter the influences of affecting the GWL, including the atmospheric pressure, tide and irregular signal.
- □ It uses the Akaike's Bayesian information criterion (ABIC) to obtain the adequate model, but the detail is neglected in here.

Factors (Noises) Filtering – TFM *[1/3]*

- □ The transfer function model (TFM) is developed by the Disaster Protection Research Center, National Cheng-Kung University in Taiwan.
- □ The model can be used to filter the influences of affecting the GWL, including the atmospheric pressure, tide, rainfall and irregular signal.
- □ Regression analysis is known to a statistical method used in modeling relationships that exist between variables.
- □ The TFM is an extension of the linear regression model: regression with serially correlated errors.
- □ It uses the Bayesian information criterion (BIC) to obtain the adequate model.

Factors (Noises) Filtering – TFM *[2/3]*

□ The full equation of transfer function model includes:

1. incorporate the "memory" of its past by lagged (dynamic) regression.
2. incorporate the serial (cross) correlations by the general regression.

$$y_t = \underbrace{\overbrace{\sum_{i=1}^{p} a_i y_{t\text{-}i}}^{\text{memory effect}} + \overbrace{\sum_{m=1}^{M}\sum_{j=1}^{q_m} b_{i,m} x_{t-j,m}}^{\text{memory effect}}}_{\text{cross correlation effect}} + e_t$$

Anomaly Detection - OA *[1/4]*

- □ Time series observations are sometimes influenced by interruptive, unexpected, uncontrolled events, or even unnoticed errors of typing and recording. The consequences of these interruptive events create spurious observations that are inconsistent with the rest of time series. Such observations are usually referred to as ***outliers***.
- □ The main references in this study are Chen et al. (1990) and the SCA statistical system (2000).

Anomaly Detection - OA *[2/4]*

- The full equation of modeling the effects of outliers includes:
  1. modeling the noise effects by ARIMA.
  2. modeling the input effects by dynamic regression.
  3. modeling the outlier effects by specific function.

$$\boxed{Y_t = C_0 + \overbrace{\sum_{j=1}^{k} v_j(B)X_{jt}}^{\text{input effect}} + \overbrace{\omega L(B) I_t(t_1)}^{\text{outlier effect}} + \overbrace{N_t}^{\text{noise effect}}}$$

Anomaly Detection - OA [4/4]

- □ There are four types (L(B)) of outliers:
  (1) additive outlier (AO): an event that affects a series for one time period only.
  (2) innovational outlier (IO): an event whose effect is propagated according to the ARIMA model of the process.
  (3) level shift (LS): an event that affects a series at a given time, and whose effect becomes permanent.
  (4) temporary change (TC): an event having such an initial impact and whose effect decays exponentially.
- □ At present, it is not mainly concerned on the type of outlier but pays close attention to the time-point and statistical significance of outlier.

Anomaly Detection - Di *[1/3]*

- □ The variation of grey-window shifting (Di) is based on the grey system theory.
- □ According to the grey system theory, the GM (1,1) model is defined as

the order of differential equation

the number of variable

$$\frac{dx^{(1)}}{dt} + ax^{(1)} = b$$

where

(1) a and b are coefficients

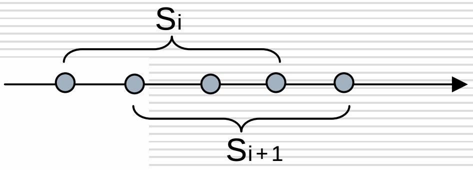
(2) $x^{(1)}(k) = \sum_{k=1}^{n} x^{(0)}(k)$

- □ The solution of GM(1,1) is

$$x^{(1)}(k) = \left(x^{(0)}(1) - \frac{b}{a} \right) e^{-a(k-1)} + \frac{b}{a}$$

Anomaly Detection - Di [2/3]

- The window S_i and shifting of this window S_{i+1} are used for GM(1,1) modeling, then the predicted value is created for individual model.



$$y^{(1)}_{S_i}(k+1) = (x^{(0)}_{S_i}(1) - \frac{b_{S_i}}{a_{S_i}}) \times e^{-(a_{S_i} \times k)} + \frac{b_{S_i}}{a_{S_i}}$$

$$y^{(0)}_{S_i}(k+1) = y^{(1)}_{S_i}(k+1) - y^{(1)}_{S_i}(k) \quad k = 0, 1, \ldots, n_1 - 1$$

the predicted value of window S_i

$$w^{(1)}_{S_{i+1}}(k+1) = (x^{(0)}_{S_i}(1) - \frac{b_{S_i}}{a_{S_i}}) \times e^{-(a_{S_i} \times (k+n_2))} + \frac{b_{S_i}}{a_{S_i}}$$

$$w^{(0)}_{S_{i+1}}(k+1) = w^{(1)}_{S_{i+1}}(k+1) - w^{(1)}_{S_{i+1}}(k) \quad k = 0, 1, \ldots, n_1 - 1$$

the predicted value of window S_{i+1}

- The predicted absolute error of window S_i and S_{i+1} is

$$e_{S_i} = \sum_{k=1}^{n_1} \left| y^{(0)}_{S_i}(k) - x^{(0)}_{S_i}(k) \right| \qquad e'_{S_{i+1}} = \sum_{k=1}^{n_1} \left| w^{(0)}_{S_{i+1}}(k) - x^{(0)}_{S_{i+1}}(k) \right|$$

Anomaly Detection - Di [3/3]

- For window S_{i+1}, calculate the absolute variation of $e_{S_{i+1}}$ and $e'_{S_{i+1}}$.

$$d_{S_{i+1}} = \left| e'_{S_{i+1}} - e_{S_{i+1}} \right|$$

- When the window is shifted, the $d_{S_{i+1}}$ is used to check the change of data structure.
- The threshold value needs to be assigned for testing the anomaly. The <mean+2*st.dev.> is suggested in this study.

Anomaly Detection - Es [1/2]

- The measure of grey variation information series (Es) is based on the grey system theory and information entropy.
- The calculation steps of the Es method are described in brief as follows:
  (1) Normalize the data series

$$y_i = f(x_j) = \left(\frac{1}{1+x_j}\right) \Bigg/ \left(\sum_{i=1}^{S} \frac{1}{1+x_i}\right)$$

  (2) Calculate the information entropy

$$I(X) = -K \sum_{j=1}^{S} y_j \ln y_j \quad (j \in J;\ K = 1/\ln 2)$$

Anomaly Detection - Es [2/2]

(3) Define the relative measure of variation information

$$I_a(X) = \frac{I_d(X)}{\max I_d(X)} \times 100\% = \frac{I_{\max}(X) - I(X)}{I_{\max}(X) - I_{\min}(X)} \times 100\%$$

- The threshold value needs to be assigned for testing the anomaly. The <mean+2*st.dev.> is suggested in this study.
  (It is the same of Di.)

Anomaly Detection - Em [1/2]

- The cutting series of grey progressive sliding (Em) is based on the Es method. According to the basis of Es method, the time-point and magnitude of variation in time series are concerned.
- The calculation steps of the Em method are described in brief as follows:

  (1) Re-arrange the data series:

  $$X_j = [x(1), x(2), \ldots, x(j), \bar{x}(j), \ldots, \bar{x}(j)] \quad (j = 1, 2, \cdots, N)$$

  where $\bar{x}(j) = \sum_{k=1}^{j} x(k)/j$

  (2) Calculate the $I_a(X_j)$ (the Es method)

Anomaly Detection - Em [2/2]

(3) Define the measure of cutting series of grey progressive sliding

$$\Delta I_a(X_j) = [I_a(X_j) - I_a(X_{j-1}) / I_a(X_j)]$$

- □ The threshold value needs to be assigned for testing the anomaly. The <mean+2*st.dev.> is suggested in this study.
  (It is the same of Di and Es.)

Comparison of Di, Es and Em

- The Di method:
  1. The minimum data number of GM(1,1) modeling is 4. (we take 4 for window size)
  2. If the data value is continuously the same, this method fails and needs to use the Es or Em method.
- The Es method:
  1. Calculate the information entropy of data
  2. To compared with the max-minimum of information entropy in whole period.
- The Em method:
  1. Calculate the information entropy of data
  2. To compared with the information entropy of previous window.

Anomaly Detection – AAF

□ The control and management procedure of data from the groundwater observation wells in this project is to go on according to the following seven steps:

(1) measurement of environmental information

(2) recording/storage of environmental information

(3) checking and processing of environmental information

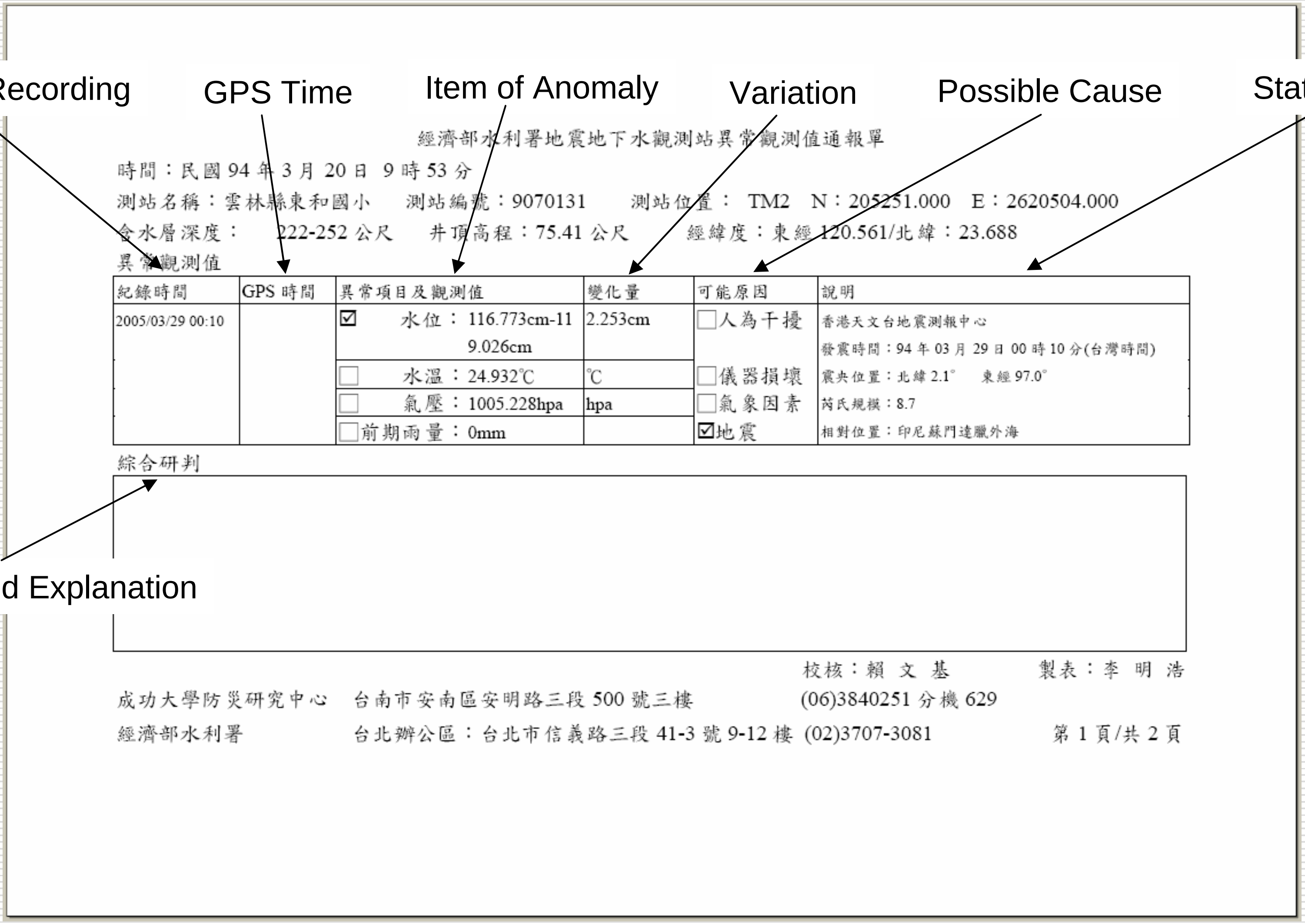
(4) noise filtering and data analysis ← ***By BAYTAP-G Model***

(5) identification/determination of anomaly

(6) data explanation and anomaly description

(7) making and proposing of the form

An Example of AAF [1/2]



Time of Recording

GPS Time

Item of Anomaly

Variation

Possible Cause

Statement

Integrated Explanation

經濟部水利署地震地下水觀測站異常觀測值通報單

時間：民國 94 年 3 月 20 日 9 時 53 分

測站名稱：雲林縣東和國小　測站編號：9070131　測站位置： TM2 N：205251.000 E：2620504.000

含水層深度：　222-252 公尺　井頂高程：75.41 公尺　經緯度：東經 120.561/北緯：23.688

異常觀測值

紀錄時間	GPS 時間	異常項目及觀測值	變化量	可能原因	說明
2005/03/29 00:10		☑ 水位：116.773cm-119.026cm	2.253cm	□人為干擾	香港天文台地震測報中心 發震時間：94 年 03 月 29 日 00 時 10 分(台灣時間)
		□ 水溫：24.932℃	℃	□儀器損壞	震央位置：北緯 2.1° 東經 97.0°
		□ 氣壓：1005.228hpa	hpa	□氣象因素	芮氏規模：8.7
		□前期雨量：0mm		☑地震	相對位置：印尼蘇門達臘外海

綜合研判

校核：賴 文 基　　製表：李 明 浩

成功大學防災研究中心　台南市安南區安明路三段 500 號三樓　(06)3840251 分機 629

經濟部水利署　台北辦公區：台北市信義路三段 41-3 號 9-12 樓 (02)3707-3081　第 1 頁/共 2 頁

Case Studies

- □ Part I
  - Comparison of BAYTAP-G and TFM by OA
- □ Part II
  - Comparison of OA, Di, Es, Em and AAF

Data Acquisition and Research Scope

- The data come from the observation stations of Water Resource Agency, Ministry of Economic Affairs (the title of project: Planning of Groundwater Anomalies Associated with the Earthquake).
- There are 8 observation wells in Taiwan for the study.
- Data
  - 12 groups of time series (case c1 ~ c12)
  - time period: September, 2003 ~ May, 2004
  - data (GWL) recording by hourly time interval
  - data filtering by BAYTAP-G model or TFM

Just the results of case c1 and c2 are shown here.

Part I - Comparison of BAYTAP-G and TFM by OA

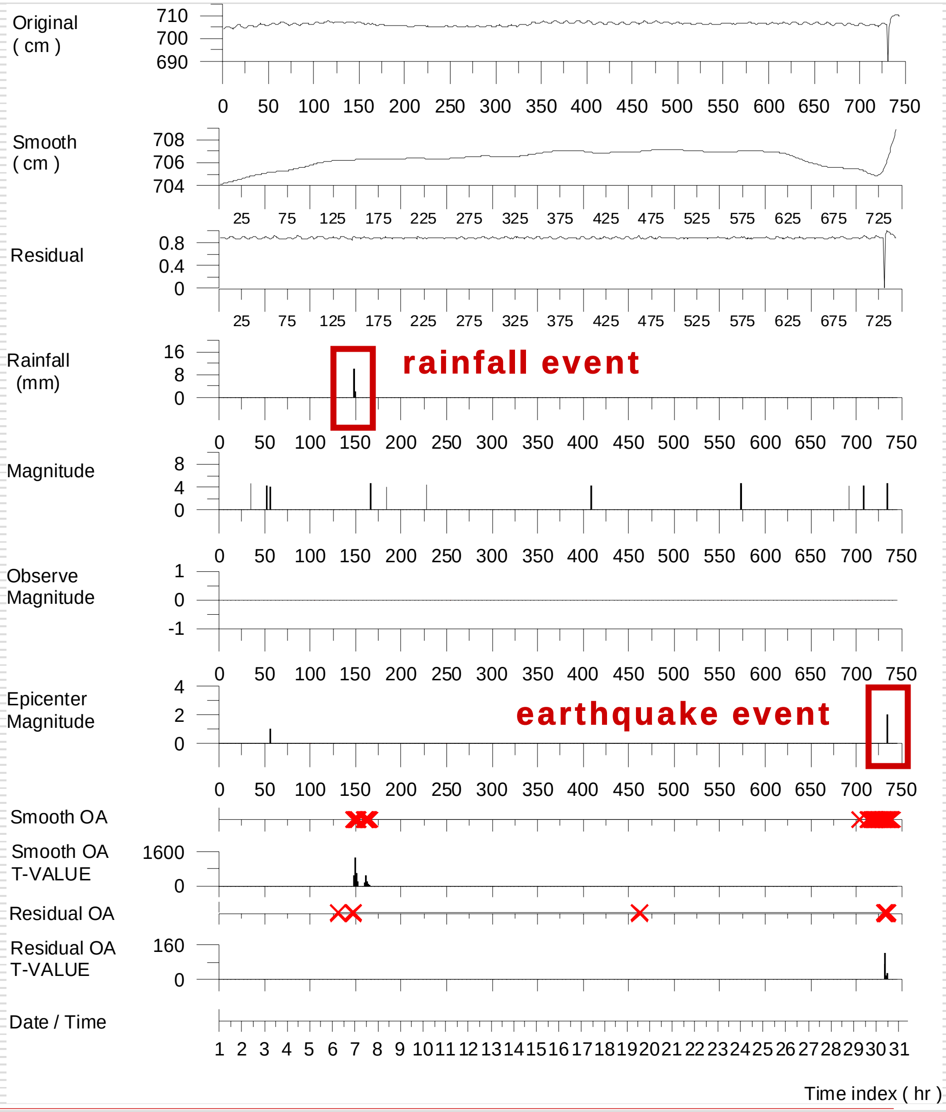
The original data

The cleansing data by BYATAP-G filtering

The cleansing data by TFM filtering

OA for BAYTAP-G

OA for TFM



The anomaly detection result of OA in case C1 from the BAYTAP-G and TFM filtering

Part I - Comparison of BAYTAP-G and TFM by OA

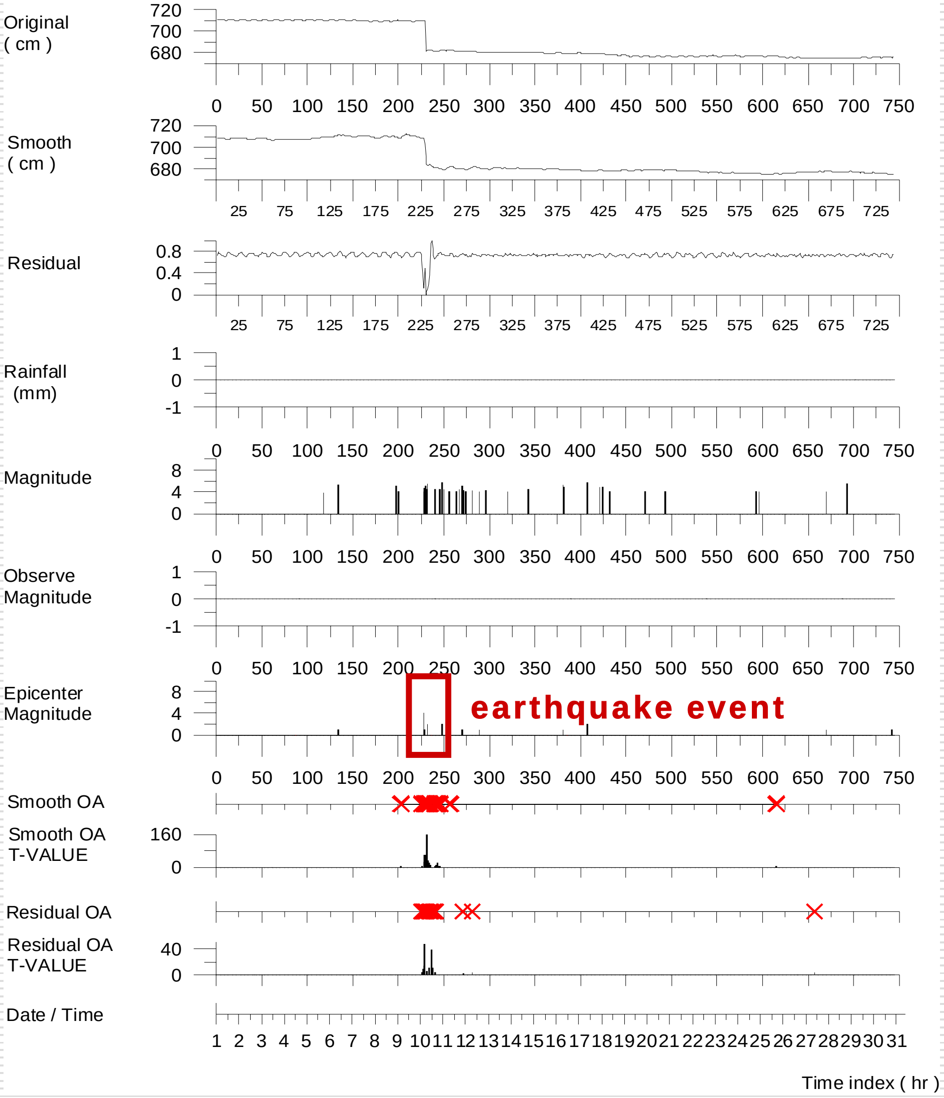
The original data

The cleansing data
by BYATAP-G filtering

The cleansing data
by TFM filtering

OA for BAYTAP-G

OA for TFM



The anomaly detection result of OA in case C2 from the BAYTAP-G and TFM filtering

Part I - Comparison of BAYTAP-G and TFM by OA

- □ The TFM cooperated with the BIC is efficient and automatic for filtering the environmental factors and obtaining the adequate model.
- □ To inspect the anomaly detection results of the OA method from the BAYTAP-G and TFM filtering, the TFM is similar to the BAYTAP-G.
- □ The TFM may be an alternative method for factors (noises) filtering, but it has many advantages and conveniences, such as
  (1) easy to increase the variables
  (2) systematic approach
  (3) fast (once) to estimate parameters
  (4) easy to update the model

Part II - Comparison of OA, Di, Es, Em and AAF

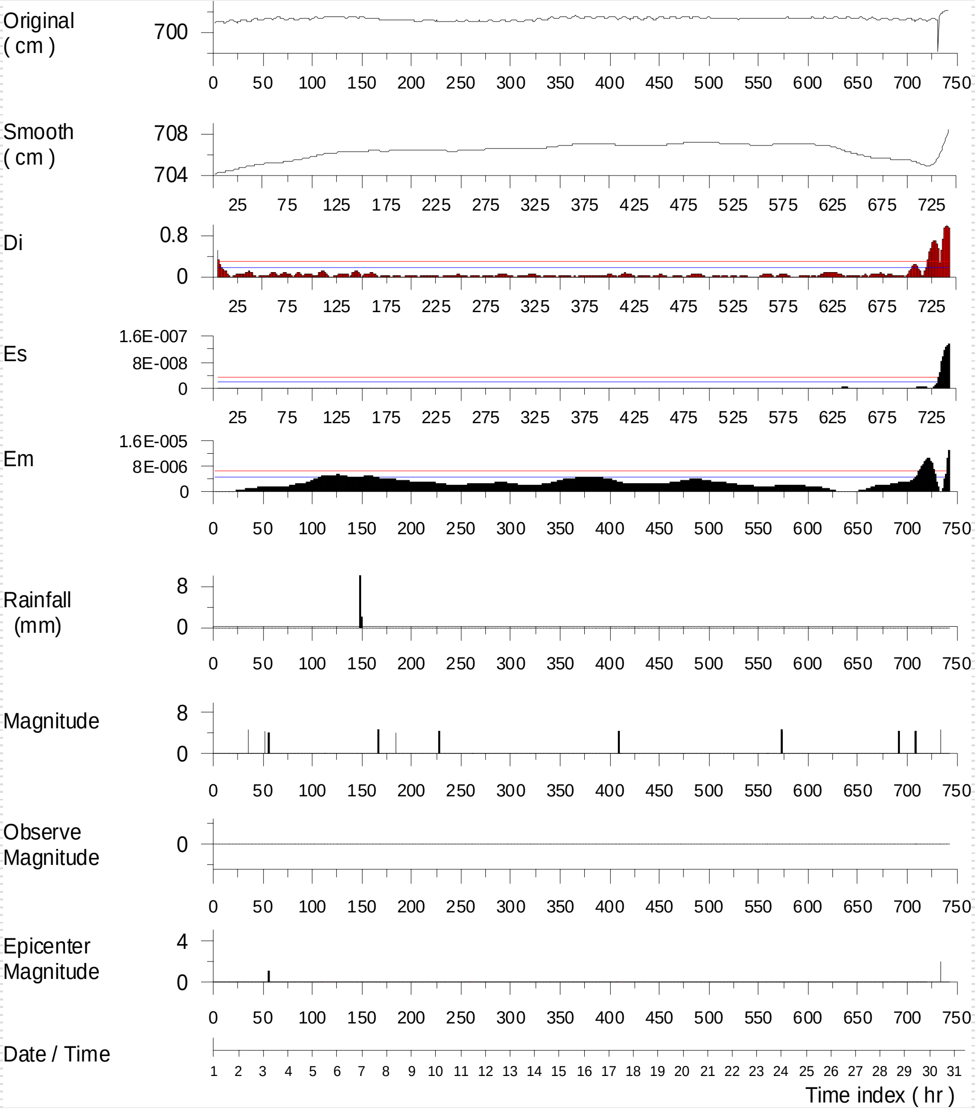
The original data

The cleansing data

The Di method

The Es method

The Em method



The anomaly detection time-series-plot of the Di, Es and Em in case C1

Part II - Comparison of OA, Di, Es, Em and AAF

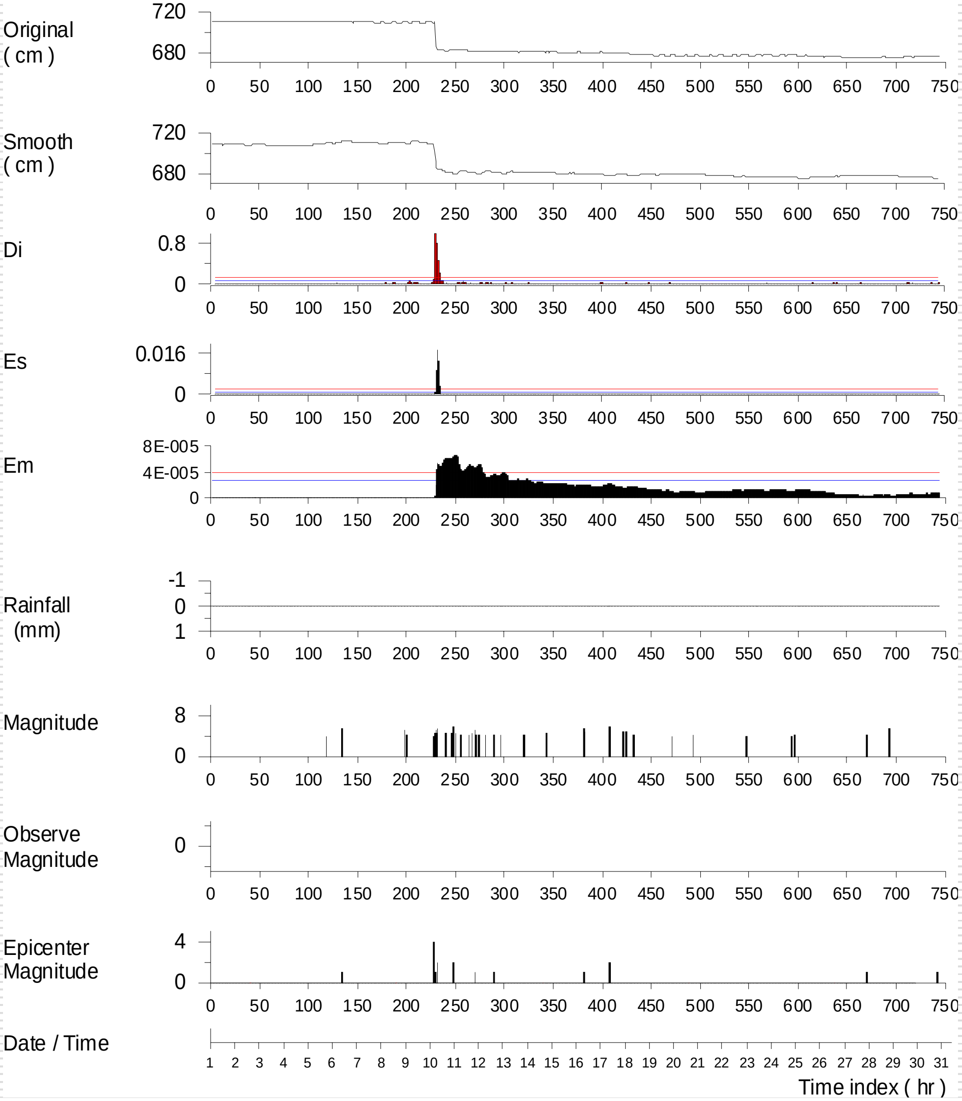
The original data ············▶

The cleansing data ············▶

The Di method ············▶

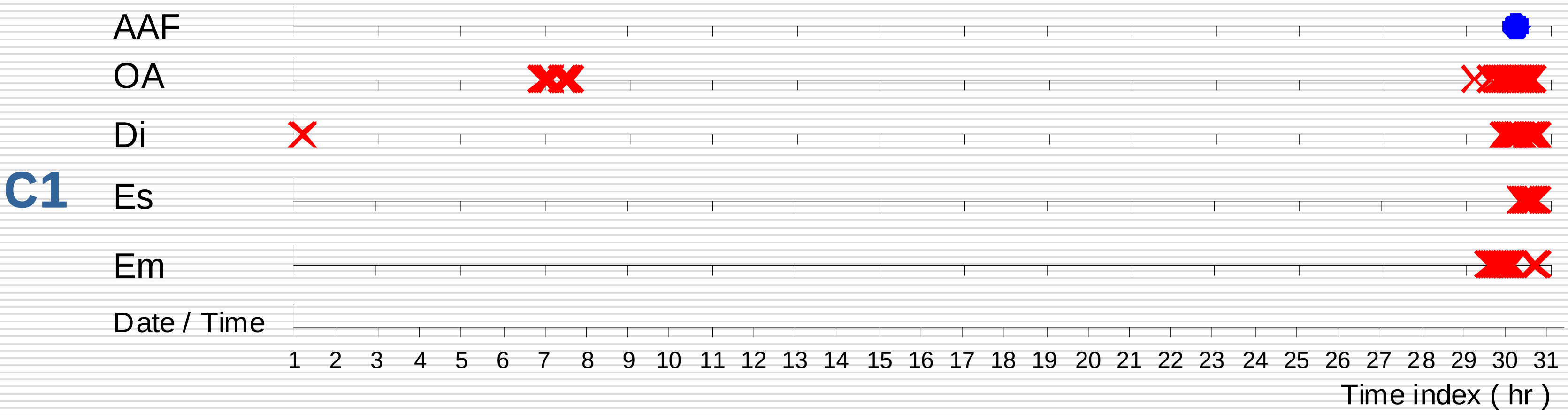
The Es method ············▶

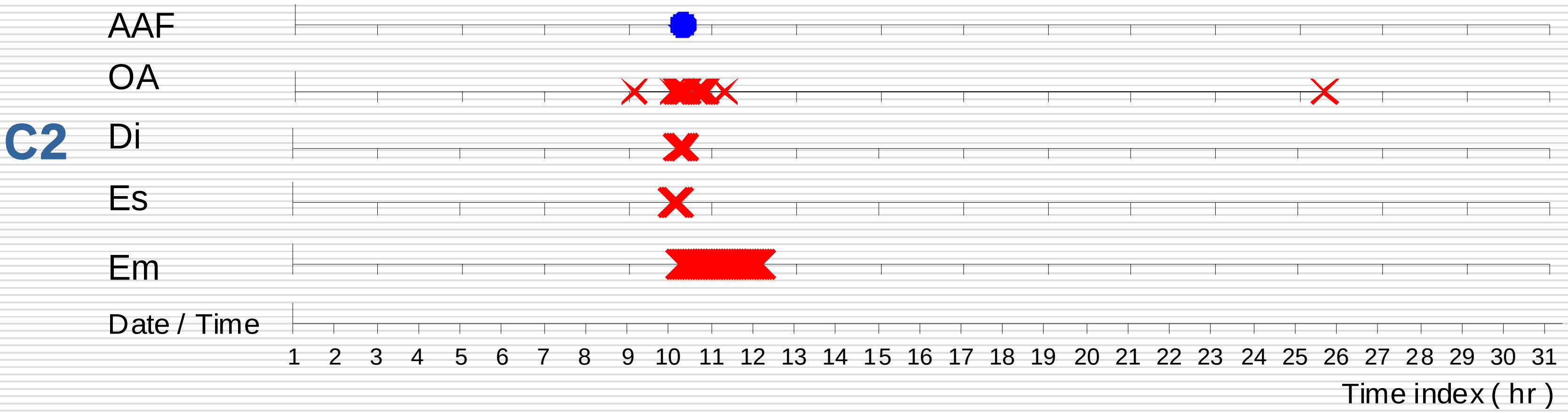
The Em method ············▶



The anomaly detection time-series-plot of the Di, Es and Em in case C2

Part II - Comparison of OA, Di, Es, Em and AAF





The anomaly detection result of the OA, Di, Es, Em and AAF in case C1 and C2

Part II - Comparison of OA, Di, Es, Em and AAF

- Three (Di, Es and Em) anomaly detection methods based on the grey system have the features:
  (1) the time of preparation is short
  (2) the data number is small for modeling
  (3) easy to model building
  (4) fast to estimate parameters
  (5) automatic to execute the procedure
  (6) flexible to adjust the model
- The time, period and intensity of the anomaly can be extracted by the Di, Es or Em method.
- The methods based on the grey system can be used for the real-time analysis. It is possible to provide the leading (pre-cursor) information.

Concluding Remarks [1/2]

- □ To compare the results of four detection methods to the AAF, the AAF with seven-step procedure is moderately subjective, but four detection methods with the standard operation procedure may be more objective.
- □ The OA method has the properties of rigorous theory, but the execution procedure is not easy to automatize. It is used as a quantitative method, in which the earthquake is regarded as an intervention event. The response function is established based on the changes of the GWL before and after the earthquake.

Concluding Remarks [2/2]

- □ Three methods (Di, Es and Em) based on the grey system theory have lots of merits, including the simple, fast and automatic, but the threshold value to test the anomaly needs to be set firstly from different observation stations.
- □ All four methods may offer the tools for exploring the groundwater micro-behavior and contribute to explaining the relationship of earthquake and groundwater.

Thanks for Your Attention and Cooperation