Dynamic Time Warping Algorithm for Model Calibration

University of Waterloo

Department of Civil and Environmental Engineering

By

Shi Jie Tan

Aug 2018

Summary:

Hydrological modeling is an essential tool for simulating hydrologic processes. To ensure the accuracy of a model, the input parameters are typically calibrated, either manually or automatically. One of the most common statistics used in the calibration processes is the Nash Sutcliffe Efficiency Coefficient (NSE). However, the metric is shown to be sensitive to outliers, magnitude bias, as well as time offset bias in time series data.

One method of accounting for this time offset bias is through a similarity search algorithm – dynamic time warping (DTW). DTW uses a nonlinear alignment method to optimally match he indices of the observed and simulated values. R, a statistical programming language and its “dtw” library are used for determining the normalized alignment distance between the two time series data. Four different step patterns were selected for analysis including symmetric, asymmetric, and Rabiner and Juang step pattern. The result from the sample test data - Irondequoit shows that DTW is able to capture the offset in the observed and simulated data but tend to perform some mismatches. Further experiments can be conducted to use the normalized alignment distance as a metric for calibration.

Table of Contents:

1.0. Introduction

Hydrological modelling is a powerful tool for simulating the effects of watershed processes. In general, a model is intended to represent a simplified version of the real-world system. It typically consists of various parameters along with its unique characteristics which define the model. The two most fundamental inputs to any hydrological models are rainfall data and drainage area. Other characteristics such as soil properties, soil moisture content and topography are also typically considered. (Gayathri, Ganasri, & S, 2015)

To ensure that the model is indeed accurate, calibration is typically conducted by adjusting the input parameters such that the measured and simulated values match closely. Several established statistics for model evaluation and calibration include the Nash-Sutcliffe Efficiency Coefficient (NSE), percent bias, as well as the root mean squared error & standard deviation ratio (RSR). A novel approach for model calibration is proposed in this report using dynamic time warping – a similarity search algorithm.

2.0. Background

To reduce uncertainty in model simulation, calibration is typically required when evaluating the model. This is achieved by adjusting the input parameters based on the evaluation of simulated and observed data. The procedure begins with a sensitivity analysis followed by a manual or automatic calibration. Standard techniques for model calibration typically include NSE or RSR. A calibration procedure utilizing those two metrics is shown in Figure 1.

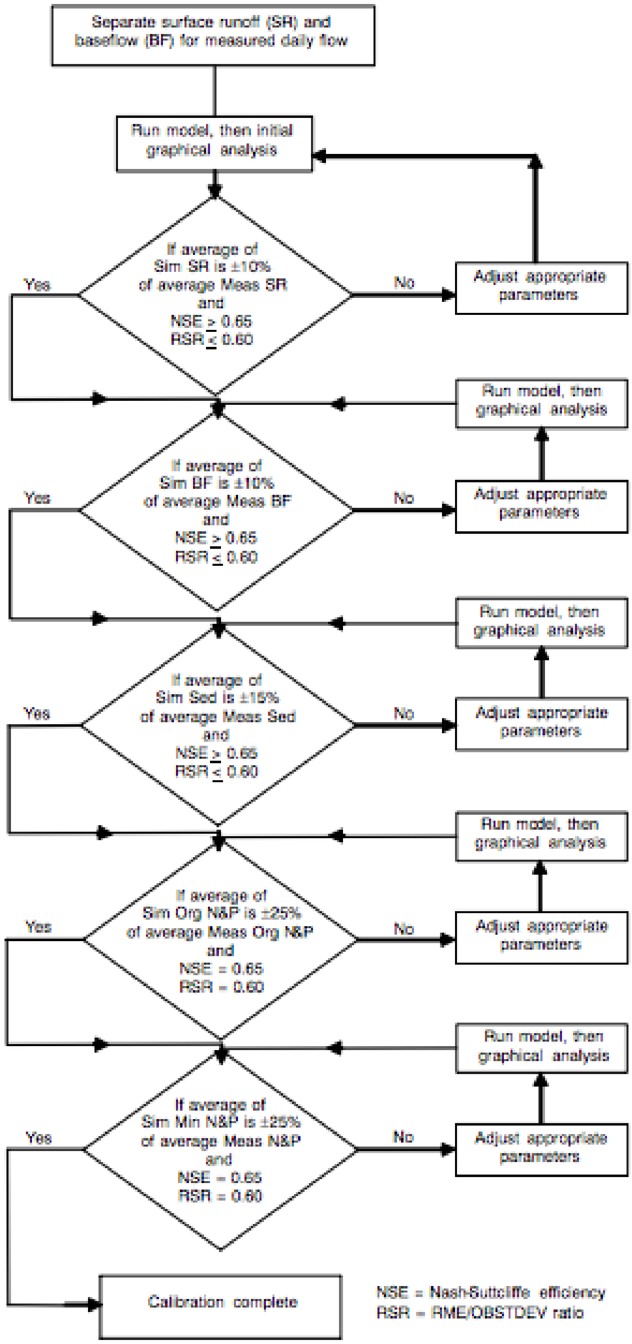

Figure 1. Calibration Procedure (Moriasi, et al., 2007)

3.0. Nash-Sutcliffe Coefficient

NSE is a common regression statistics used for calibrating hydrological models. A formulation and background of the NSE will be provided in this section.

3.1. Formulations

NSE is specifically adapted for evaluating the predictive power of a simulated hydrological model. The formulation is shown in the equation below.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Y_i^{obs} - Y_i^{sim})^2}{\sum_{i=1}^{n}(Y_i^{obs} - Y^{mean})^2}$$

Where Y_i^{obs} corresponds to the *i*th observed value, Y_i^{sim} corresponds to the *i*th simulated value, and $Y_{mean}$is the mean of all the observed data. The summation is used for accounting for all the observed and simulated data where n is the total number of observations.

If the predictions of a linear model are unbiased, the NSE value ranges from 0 to 1. If the predictions of a linear model are biased however, the NSE value can become negative.

A value of 1 corresponds to a perfect match between the modeled and observed data whereas a value of 0 indicates that the model prediction is as accurate as the mean of the observed data. A value of less than 0 indicates that the residual variance is larger than the data variance. Generally, values between 0 and 1 are acceptable (McCuen, Knight, & Gillian, 2006).

Based on the analysis of NSE conducted by McCuen, it was determined that outliers can largely influence the sample values of NSE. Furthermore, the time-offset bias and magnitude bias can negatively affect the accuracy of the metric.

3.2. Time Offset Bias

Time offset bias can arise in hydrological time series data when the rainfall and runoff data are not synchronized. This can be a result of the rain gauge located outside of the watershed or when the rainfall hyetograph is offset from the runoff hyetograph.

The NSE is sensitive to the time offset error. As a result, the numerator of the NSE increases and inflates. To assess the effects of this time offset error, a gamma distribution can be used with a shape parameter of 4.7. The shape parameter is "translated on the time axis to reflect a model that has not been properly calibrated to fit in the time domain but reproduces the magnitudes of the measured hydrograph" (McCuen, Knight, & Gillian, 2006). The result shows that as the offset interval increases, the value of the NSE decreases, meaning the less accurate the metric becomes. Similarly, the smaller the time interval, the less significant the offset is on the NSE.

4.0. Dynamic Time Warping

One method of accounting for the time bias in the time series data as seen in NSE is using a nonlinear alignment metric instead. Dynamic time warping is a type of sequence similarity search algorithms that finds the best match of the two time series using a nonlinear alignment. Figure 2 illustrates the idea of a nonlinear alignment where the index of one series can be matched to a different index in the other series.

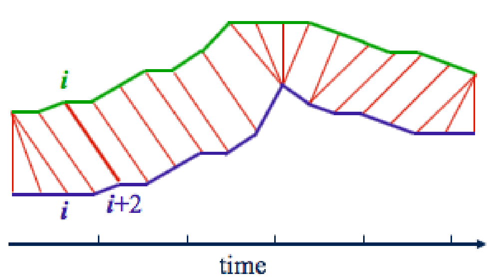

Figure 2. Non-linear Time Warping (Tsiporkova)

4.1. Algorithmic Definitions

The definition of the algorithm is as follows. Given two time series represented in the form of two vectors: $X = (x_1, \ldots, x_N)$ and $Y = (y_1, \ldots, \ y_M)$, there exists a local dissimilarity function f between the pairwise elements x_i and y_j.

$$d(i,j) = f\left(x_i, \ \ y_j\right) \geq 0$$

Where d is the cross-distance matrix between the vectors X and Y. The cross-distance matrix between the two vectors are used as the only input for the algorithm.

To remap the time indices of the two vectors X and Y, warping functions $\emptyset_x$ and $\emptyset_y$ are used to "pick the deformation of the time axes such that the two time series are as close as possible to each other". (Giorgino, 1-25)

$$D(X,Y) = \min d_\emptyset \, (X,Y)$$

Similarly, the average accumulated distortion between the warped time series X and Y can be determined as follows:

$$d_\emptyset(X,Y) = \sum_{k=1}^{T} d \, (\emptyset_x(k), \emptyset_y(k)) \, m_\emptyset(k)/M_\emptyset$$

Where $m_\emptyset(k)$ is a per-step weighting coefficient and $M_\emptyset(k)$ is the corresponding normalization constant.

4.2. Warping Constraints

Constraints are typically imposed on the warping functions, including monotonicity, continuity, boundary conditions, warping window, and slope constraint.

Monotonicity is set to preserve the ordering of the alignment path such that the features are not repeated. Expressed in terms of the i and j axes, the monotonicity can be formulated as $i_{s-1} \leq i_s$ and $j_{s-1} \leq j_s$. Figure 3 illustrates the idea of monotonicity constraint.

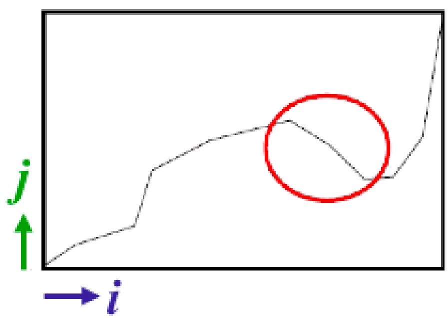

Figure 3. Monotonicity Constraint (Tsiporkova)

The second constraint is continuity such that the alignment path does not jump in time index and that no points are omitted in either of the time series. It can be expressed in terms of $i_s - i_{s-1} \leq 1$ and $j_s - j_{s-1} \leq 1$. Figure 4 illustrates the continuity constraint.

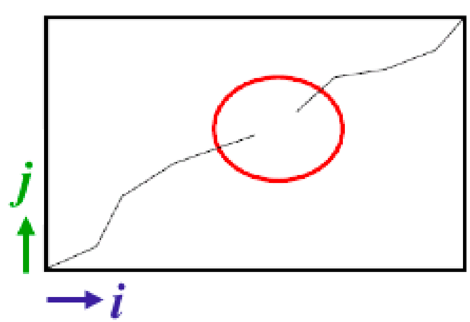

Figure 4. Continuity Constraint (Tsiporkova)

The third constraint is the boundary constraint such that the alignment starts at the bottom left and ends at the top right of the alignment path. It ensures that the alignment does not partially consider one series. It can be formulated as $i_1 = 1$, $i_k = n$, $j_1 = 1$, $j_k = m$. Figure 5 illustrates the idea of boundary constraint.

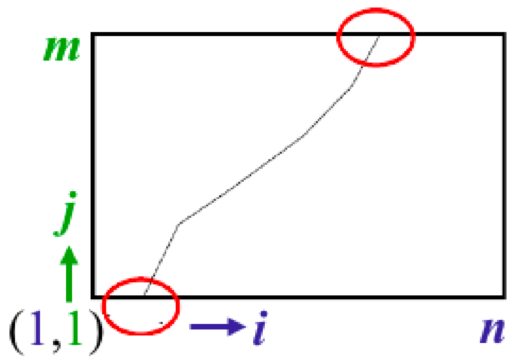

Figure 5. Boundary Constraint (Tsiporkova)

The forth constraint is the warping window such that the alignment does not skip different features or stop at similar features. A window band r is typically defined and expressed as $|i_s - j_s| \leq r$. Figure 6 illustrate this constraint.

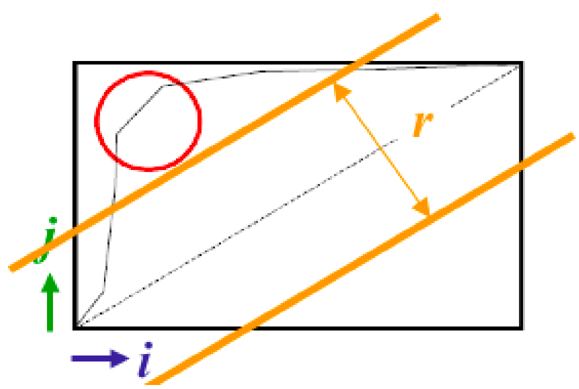

Figure 6. Window Constraint

Lastly, the sloping constraint can be implemented such that the alignment path does not become too deep or shallow. This effectively prevents short parts of the sequence matched with long parts of the sequence. The constraint can be formulated as $(j_{sp} - j_{s0}) / (i_{sp} - i_{s0}) \leq p$ and $(i_{sq} - i_{s0}) / (j_{sq} - j_{s0}) \leq q$. Figure 7 illustrates the slope constraint.

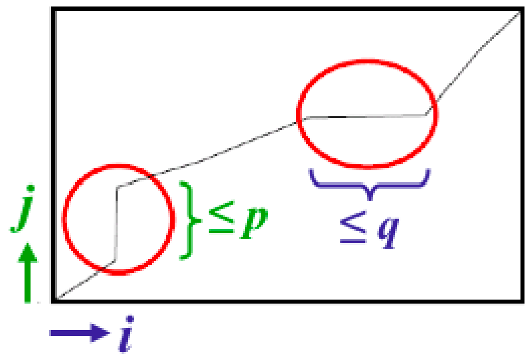

Figure 7. Slope Constraint (Tsiporkova)

4.3. Step Patterns

Step patterns compute the transition between the matched pairs and the corresponding weights. The step patterns specify $\emptyset(k+1)$ given the previous time steps $\emptyset(k)$ and $\emptyset(k-1)$ (Giorgino, 1-25).

The algorithms used in computing the step patterns are based on the method of dynamic programming where every time the algorithm is implemented, it references the solutions of subproblems that have already been computed. Four common step patterns can be used for computing the algorithms: symmetric 1, symmetric2, asymmetric and Rabiner and Juang's step pattern. The formulas for the step patterns are as follow:

Symmetric 1:

$$g[i,j] = \min\ (g[i-1,j-1] + d[i,j], g[i,j-1] + d[i,j], g[i-1,j] + d[i,j])$$

Symmetric 2:

$$g[i,j] = \min\left(g[i-1,j-1] + 2 \times d[i,j], g[i,j-1] + d[i,j], g[i-1,j] + d[i,j]\right)$$

Asymmetric:

$$g[i,j] = \min\left(g[i-1,j] + d[i,j], g[i-1,j-1] + d[i,j], g[i-1,j-2] + d[i,j]\right)$$

Rabiner and Juang:

$$g[i,j] = \min\left(g[i-2,j-1] + d[i-1,j] + d[i,j],\ g[i-2,j-2] + d\ [i-1,j] + d[i,j],\right.$$
$$\left. g[i-1,j-1] + d[i,j],\ g[i-1,j-2] + d[i,j]\right)$$

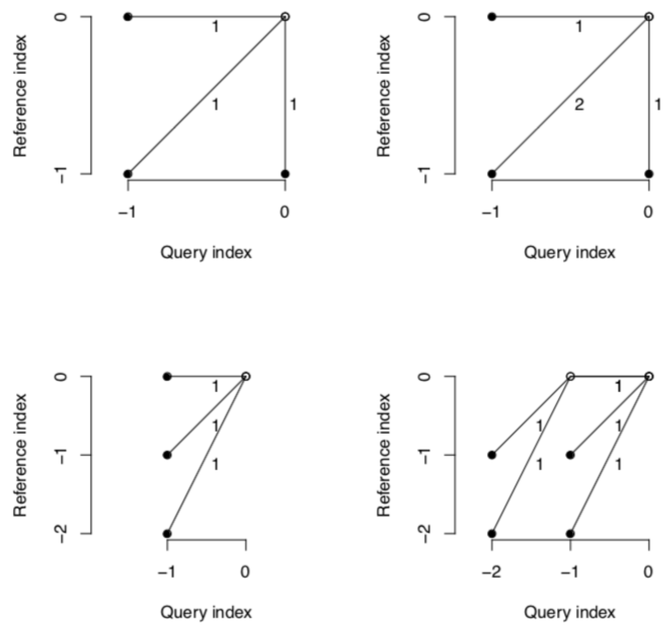

Figure 8. Four Step Patterns. From Left to Right and Top to Bottom: Symmetric 1, Symmetric 2, Asymmetric, Rabiner and Juang (Giorgino, 1-25)

Figure 8 shows the graphical form of the four step patterns where the horizontal axis represents the query index and the vertical axis represents the reference index.

5.0. Dynamic Time Warping Implementation

The implementation of dynamic time warping using the sample hydrological data as well as the results will be discussed in this section. The script is provided in Appendix 1 for further reference.

5.1. Procedure

The "dtw" library from R is used for analysis. The basic syntax is as follows:

$$dtw(x, y, dist.method = "Euclidian", step.pattern = symmetric2)$$

Where a distance method, step pattern and window band can be selected.

All four types of step patterns were applied to the data sets and four alignment functions were defined: "alignment.rj", "alignment.as", "alignment.s1", "alignment.s2". To compute the average distance along the warping curve, the alignment distance was further normalized. Lastly, the step pattern with the lowest overall alignment distance was selected as the optimal alignment distance.

5.2. Data Analysis

A sample time series data set Irondequoit was used for analysis. The data set is provided in Appendix 2 for reference. The graph of the simulated vs observed hydrographs is plotted in Figure 9 as shown below. The simulated hydrograph corresponds to the black line and the observed hydrograph corresponds to the red line. As shown, the simulated hydrograph tends to follow the same trends as the observed data. The observed hydrograph however also shows much greater fluctuations than the simulated hydrograph.

Figure 10 shows the dynamic time warping effect on the simulated and observed plot. As expected, the peak of the simulated data also tends to match with the peak of the observed data. However, it can be observed in some instances, the peaks of the observed time series are matched to the farther peaks of the simulated time series instead of the closer peaks. According to human intuition, those are perceived as mismatches.

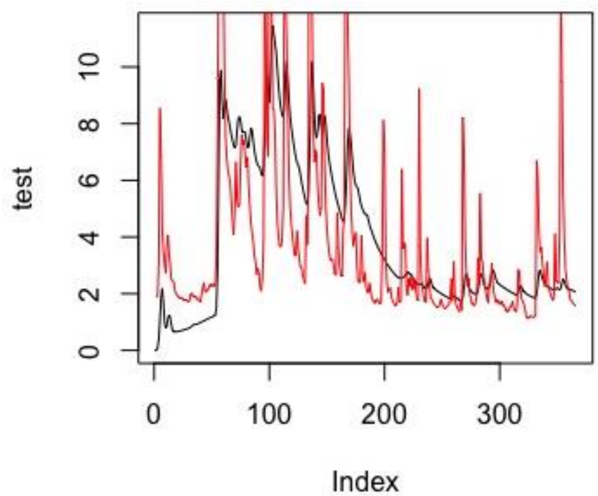

Figure 9. Simulated vs Observed Hydrographs (Observed and Simulated Data in Red and Black)

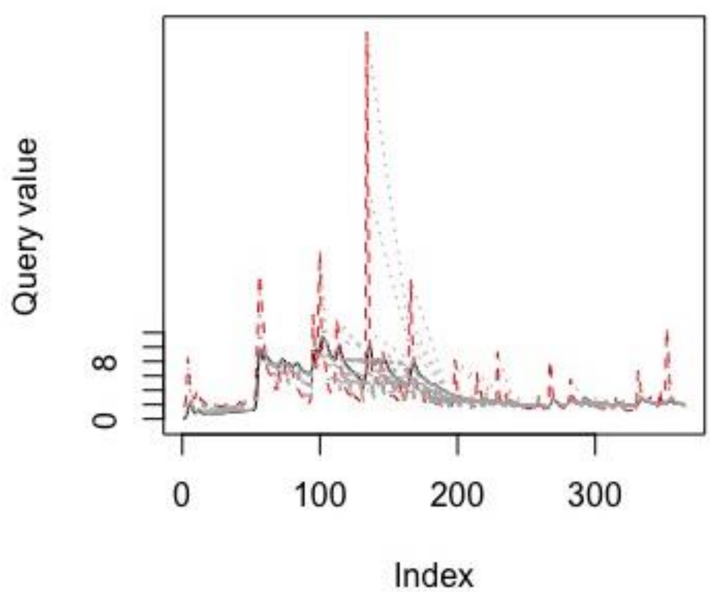

Figure 10. Simulated vs Observed Hydrographs with Dynamic Time Warping Effect

Figure 11 clearly shows the path of the alignment of the two time series data with the query index (simulated values) on the horizontal axis and the reference index (observed values) on the vertical axis. It can generally be observed that the algorithm performs well with the alignment path approximately linear. This indicates a relatively low alignment distance cost.

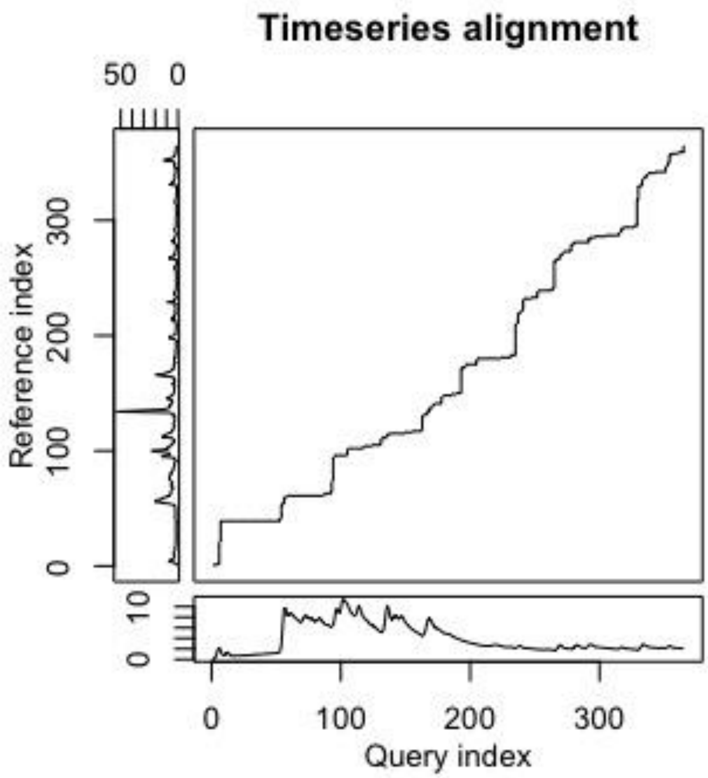

Figure 11. Time Series Alignment

6.0. Conclusion

Hydrological modelling is an essential tool for modelling hydrologic processes. Calibration is typically used for ensuring the accuracy of the model. Typically, the NSE and RSR are used during the calibration procedure. One shortcoming of NSE as discussed in this report is the time offset bias as a result of linear alignment. To potentially improve this, a similarity search algorithm –

dynamic time warping can be used for optimally matching the indices of the observed and simulated data based on the shortest distances. It can be observed from the results that dynamic time warping is able to capture the offset in the time series data but tend to perform some mismatches. Further experiments can be proceeded to use dynamic time warping’s normalized alignment distance as a metric for calibrating hydrological models.

References:

Gayathri, D., Ganasri, P., & S, D. (2015). A Review on Hydrological Models. *Elsevier*, 1001-1007.

Giorgino, T. (1-25). *Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package* . Padua: National Research Council of Italy.

McCuen, R., Knight, Z., & Gillian, C. (2006). Journal of Hydrologic Engineering. *Evaluation of the Nash-Sutcliffe Efficiency Index*.

Moriasi, D. N., Arnold, J. G., Liew, M. W., Bingner, R. L., Harmel, R. D., & & Veith, T. L. (2007). Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *2007 American Society of Agricultural and Biological Engineers*, 885-900.

Tsiporkova, E. (n.d.). *Dynamic Time Warping Algorithm for Gene Expression Time Series.*

Appendix A:

```
library("dtw")
# specify library path
.libPaths()

# load the sample data
setwd("/Applications/DTW/")
hy <- read.csv("Irondequoit_Hydrographs.csv")

# process the data
Hydro.sim <- hy[,c(5)]
Hydro.obs <- hy[,c(6)]

Hydro.length <- length(Hydro.sim)

ref <- window(Hydro.obs, start = 1, end = 366)
test <- window(Hydro.sim, start = 1, end = 366)

# plot observed vs simulated data
plot(test, col = 'black',type ='l')
lines(ref,col = 'red')

# step patterns
alignment <- dtw(test, ref, keep=TRUE)

# plot
dtwPlotTwoWay(alignment)
dtwPlotThreeWay(alignment)

# find normalized distance
ratio <- alignment$normalizedDistance
print(ratio)
```