京 都 大 学 防 災 研 究 所 年 報　第48号 B　平成17年4月
Annuals of Disas. Prev. Res. Inst., Kyoto Univ., No. 48 B, 2005

Spatial Interpolation of Runoff between Catchment Observation Stations using Local Linear Models

Paul James SMITH*, Katsuyoshi SEKII*, and Toshiharu KOJIRI

* Graduate School of Engineering, Kyoto University

Synopsis

A system capable of providing short-term runoff forecasts for all locations across a target watershed is introduced. Current runoff forecasting technology is generally capable of producing accurate short-term forecasts at only those locations in a watershed where runoff monitoring stations provide observation data. An interpolation method based on a database of distributed hydrological simulation results and using local linear models (LLM) and global linear models (GLM) is introduced to extend this ability to allow flood forecasts to be made for all locations in a watershed, not just those locations where runoff observations are available.

Keywords: flood forecasting, distributed hydrological model, local linear model, global linear model, interpolation

1. Introduction

It has long been the goal of flood forecasting to provide timely and accurate estimates of future discharge conditions at specific watershed locations. As such, topics involving design and calibration of hydrological models, real-time filtering of runoff estimates, and data-driven techniques for inferring hydrological time series patterns have received considerable attention. While accurate point-forecasts are ideal for predicting reservoir inflow, or for providing warnings to communities situated in highly concentrated areas, there also exists the as yet unfulfilled need to provide such forecasts in a distributed manner. While a number of different schemes have been successfully applied in recent years for using real-time runoff observation data to reduce forecast errors at point locations, a major difficulty in developing distributed flood forecasting systems is the lack of a real-time filtering or data assimilation strategy capable of reducing forecast error at all points within a watershed using real-time runoff observation data available at only a handful of observation stations.

A scheme is developed here which uses a distributed rainfall-runoff model to build a database of runoff data for various historical runoff events. Data mining techniques are then employed to search through the database to establish relationships based on the runoff rates between watershed locations. Knowledge of these relationships can be used in real-time to infer the future runoff rates at all non-observation locations in a watershed from the filtered predictions made at locations where real-time runoff observations are available.

Distributed flood forecasts allow location-specific decisions to be made for all watershed locations, and provide the opportunity for decision making that takes into account the future distributed, rather than point, watershed conditions. There are a wide range of beneficial applications of such a distributed flood forecasting system.

2. Distributed Flood Prediction Approach

The distributed flood prediction approach proposed here involves the following steps:

[1] Prediction of future discharge rates several hours ahead at watershed locations that contain discharge observation stations.

[2] Interpolation and extrapolation of these forecasts across the watershed based on an understanding of spatial and temporal relationships between the hydrographs at each watershed location.

3. Point Prediction

The distributed flood forecasting system proposed here is designed to work with a generic point forecasting prediction scheme that can make use of real-time runoff observations for reducing forecast error. Appropriate forecast systems include those based on a combination of the state-space Kalman filter with lumped-parameter runoff models (Kitanidis and Bras, 1980; Puente and Bras, 1987), data-driven models such as Artificial Neural Networks (Karunanithi et al., 1994; Lorrai and Sechi, 1995; Campolo et al., 1990), Genetic Programming (Khu et al., 2001; Liong et al., 2002), Support Vector Machines (Liong and Sivapragasam, 2002), and combinations of both physically-based and data-driven strategies (Babovic and Bojkov, 2001; Smith and Kojiri, 2004).

4. Proposed Interpolation Strategies

Local Linear Modeling (LLM) and Global Linear Modeling (GLM) are investigated for their application to interpolation and extrapolation of runoff rates along river channels.

Because the interpolation system developed in this research must be used to identify hydrological patterns for 100s of unique combinations of watershed locations under a variety of different hydrological conditions, it is essential to use a flexible strategy capable of adjusting itself to each different task in real-time. Additionally, in consideration of global climate change, it is desirable that the system as a whole can grow and adapt to changing hydrological conditions. For these reasons, both strategies use a database containing numerous precipitation-driven rainfall-runoff simulation results from a distributed hydrological model calibrated to the target watershed of interest. The simulated hourly discharge rates at each watershed location (1km spatial resolution) stored in this database can then be accessed in real-time to recognize spatial and temporal patterns between hydrographs at different locations in the watershed, thus removing the need for the development of numerous pre-defined models. In this way the most probable discharge rates at various unguaged locations in a watershed can be estimated based on observations or predictions of discharge rates at each available discharge observation station. The system can be automatically updated following each new observed precipitation event simply by performing a hydrological simulation and adding the results to the database, thus increasing the diversity of the knowledge in the database through the inclusion of new hydrological phenomena.

4.1 Local linear modeling

(1) Introduction

Local Linear Modeling is used here to approximate the relationship of future runoff states at watershed locations without discharge observation stations using the filtered predictions (and recent observations) of future runoff states at observation station locations. This method provides an effective tool for finding an estimate or prediction for a query vector **x** by fitting a parametric function in the neighborhood of **x**. Unlike global models such as Artificial Neural Networks which seek to fit a single global model to all of the training data, local models use only those training samples that are most similar to the query vector **x** to obtain a locally parametric model suitable for estimating $f(\mathbf{x})$ in the vicinity of **x**. As linear regression is only used in the vicinity of the query, the LLM strategy is capable of modeling solution spaces that are globally non-linear.

A local regression model is used to approximate a relationship between the query vector and output vector by drawing upon database simulation data and embedding it into a suitably-determined state space. This state space is searched for the *k* nearest neighbors closest to the query vector. A regression is then performed on the neighborhood, from which an estimate of the state of the non-observation location can then be made.

Regressions of polynomial degree zero and one are respectively referred to here as Local Averaging Models (LAM) and Local Linear Models (LLM). Regressions of higher polynomial degree are possible, however only those of degree one are considered here.

Atkeson et al. (1997) give the following linear model, which assumes that the constant 1 has been appended to all the input vectors **x** to include a constant term.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d + \varepsilon \qquad (1)$$

Here β_i are the set of model parameters requiring identification, x_i are the model inputs, d is the dimensionality of the training data and ε is an error term to be minimized. The training examples are collected in matrix **X** and the model parameters are collected in matrix $\boldsymbol{\beta}$.

$$\mathbf{y} = \boldsymbol{\beta}\mathbf{X} \qquad (2)$$

The model is determined through estimation of the parameters β_i using a regression which minimizes

$$\sum_i \left(x_i^{\mathrm{T}}\beta - y_i\right)^2 \qquad (3)$$

through solution of the normal equations

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\beta = \mathbf{X}^{\mathrm{T}}\mathbf{y} \qquad (4)$$

with the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ inverted for β:

$$\beta = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \qquad (5)$$

(2) Nearest neighbors search

An exhaustive search strategy is used to find the k nearest neighbors to the query vector, which requires that the Euclidean distance d_{E} between the query vector **q** and each data point **x** in the database be calculated for every query made.

$$\begin{aligned} d_{\mathrm{E}}(\mathbf{x},\mathbf{q}) &= \sqrt{\sum_j \left(\mathbf{x}_j - \mathbf{q}_j\right)^2} \\ &= \sqrt{\left(\mathbf{x}-\mathbf{q}\right)^{\mathrm{T}}\left(\mathbf{x}-\mathbf{q}\right)} \end{aligned} \qquad (6)$$

Efficient nearest point search algorithms are available to speed the nearest neighbor lookup process, such as the k-d trees scheme (Bentley, 1980; Moore, 1991) which creates a data structure for storing the set of training points taken from a d-dimensional space, to allow for rapid subsequent lookup. In the case of this research, the system is designed to be flexible to allow for changes in database size, data quality, and hydrological and climate change. The query vector has a variable form to allow for the unique requirements of each location within a watershed and for changes in the temporal correlation between discharge rates at spatially separated locations. For this reason and as database search time is negligible, an efficient search algorithm is deemed unnecessary.

An option to prevent a given regression from being dominated by data points all taken from the same simulation is included in the system, whereby the maximum fraction of nearest neighbors that may be chosen from a given simulation event i is restricted to be

$$k_i \le 1/\left(a * n\_sim + b\right), \quad i = 1,\ldots,n\_sim \qquad (7)$$

where a and b are chosen by the user such that their sum is unity (a=0.05, b=0.95 is used in this research) and *n_sim* is the number of simulations stored in the database.

Furthermore, in recognizing that some observation stations will be more important than others in the regression stage, the elements of the query vector can be weighted during the nearest neighbor search to give priority to data elements from observation stations that have hydrographs that are highly correlated with the query location's hydrograph. These observation stations will often be those that are geographically closest to the query location. One approach towards choosing appropriate weights involves using the magnitude of the correlation vector $\boldsymbol{\varphi} = \left(\phi_1, \phi_2, \cdots, \phi_m\right)$, which is a measure of how highly correlated each query vector element is to the runoff at the target location. This correlation can be estimated from simulated data in the database, and assumes a linear relationship. These measures of correlation can be used to weight the elements of the query vector when searching for nearest neighbors: the higher the value of ϕ_j, the more influence the corresponding query vector element will have in determining suitable nearest neighbors for the regression.

This modified measure of distance between query point and data point is referred to here as the Dimensionally Weighted Euclidean Distance (DWED).

$$\begin{aligned} d_{\mathrm{DWED}}(\mathbf{x},\mathbf{q}) &= \sqrt{\sum_j \phi_j^2 \left(\mathbf{x}_j - \mathbf{q}_j\right)^2} \\ &= \sqrt{\left(\mathbf{x}-\mathbf{q}\right)^{\mathrm{T}}\boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{\varphi}\left(\mathbf{x}-\mathbf{q}\right)} \end{aligned} \qquad (8)$$

4.2 Global regression

As the number of nearest neighbors approaches *n_sim* the modeling approach moves from a local modeling strategy to a global regression strategy. This global regression approach can be considered as an extension of the local linear regression described above, using all available simulation data in searching for a relationship between the particular combination of locations under investigation.

4.3 Choice of query vector form

The proposed interpolation system is designed to exploit the correlation that exists between the discharge rates at different locations within the same watershed. It is therefore desirable to tailor the form of the query vector to suit each individual watershed location such that it maximizes the use of available correlated data. Since observations of discharge rates and filtered predictions of future discharge rates are available at observation stations, data from these locations form the basis of the query vector.

(1) Temporal correlation between elements

An estimate of the correlation between the hydrograph of the target non-observation point and the hydrographs from each observation station is determined. In most cases there will exist a given time lag at which the two hydrographs being compared have the highest correlation. For example a target location's present discharge rate will have a higher correlation with an upstream location's discharge rate from a number of time steps prior, compared with its present discharge rate. In other words, the influence of an upstream location's discharge takes some time to be felt by downstream locations. In the case of using the interpolation system in a prediction scenario, the optimal time lag for each combination of locations is chosen to be the non-positive time lag that shows the largest correlation. In the case where the observation station is downstream of the target location, the optimal time lag for that observation station will nearly always be zero, since positive time lags have no relevance in a prediction scenario.

The query vector for a given target location thus takes the following form:

$$\mathbf{q}(t) = \left(Q_1^{t+s1}, Q_2^{t+s2}, \cdots, Q_m^{t+sm} \right) \qquad (9)$$

where $s1$, $s2$,…, sm refer to the optimal lag of each of the m observation stations.

4.4 Additional elements

The inclusion of a number of additional query vector elements, which refer to other factors related to the hydrological dynamics in the watershed, may result in an increase in the accuracy of the interpolation method. Division of data points in the database into groups related to the stage of the hydrograph at the time of observation is considered. Here the hydrograph stage is simply described by one of the following four descriptors: (low flow / rising / peak / falling). A low flow level is set for each observation station based on hydrological records, with any discharge rate below or equal to that defined as 'low flow'. Any discharge rate above this level is then grouped based on the following rules:

- If the second derivative of the discharge rate time series is negative: 'peak'
- Else, if the first derivative of the discharge rate time series is positive: 'rising limb'
- Else, if the first derivative of the discharge rate time series is negative: 'falling limb'

In this way, the inflection points of the hydrograph are chosen as the transition points between rising limb, peak, and falling limb regions.

5. Application

This section presents the results of an application to test the ability of the local modeling scheme to faithfully model the temporal-spatial relationship between watershed locations based on the distributed rainfall-runoff simulation results.

The application is conducted for two typhoon events that occurred in the vicinity of the Nagara River watershed in Japan's Chubu region. This watershed is relatively steep and is prone to rapid flooding during typhoon periods. The vast majority of residences and facilities that require protection from flooding are located in the south of the watershed. Discharge observation stations exist within the watershed at the downstream locations of Chusetsu and Akutami, and the mid-stream locations of Mino and Shimohorado.

A kinematic wave-based distributed rainfall-runoff model is prepared for the watershed comprising 1556 1km^2 mesh cells, and two sub-surface layers. The land use, surface slope and flow path (Figure 1), and channel characteristics are specified for each mesh cell. Model calibration and database preparation are performed using simulation results from 10 major precipitation events that occurred in 2000-2004.

Validation of the system is performed using two additional independent runoff events that occurred in 2003. Two scenarios are investigated here. The first scenario involves interpolating discharge rates for a location (Mino) that has observation stations located in both upstream (Shimohorado) and downstream

(Akutami, Chusetsu) locations. The second scenario involves extrapolation of discharge rates to a location (Shimohorado) that has no observation stations located upstream, and three observation stations located downstream (Mino, Akutami, Chusetsu). In each case the observed runoff at the target location is only used for verification, and as such these locations are assumed to be without observation stations.

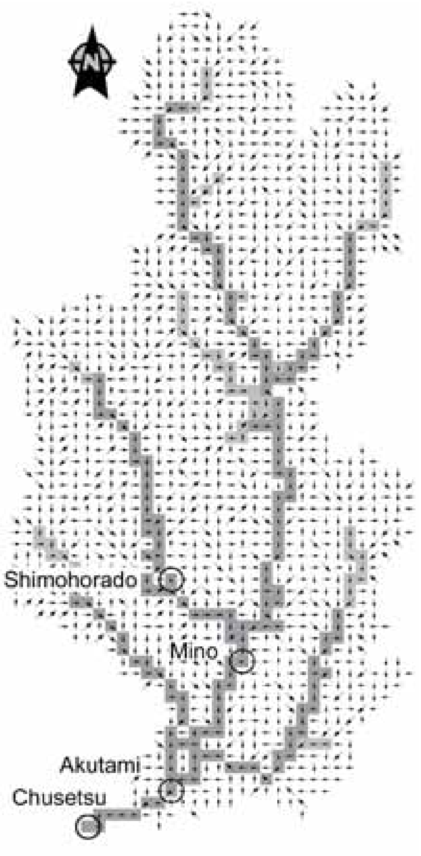

Figure 1 Nagara River watershed flow routing map

The observed discharge rates at the four locations for the two events used in this application are given in Figure 2 and Figure 3.

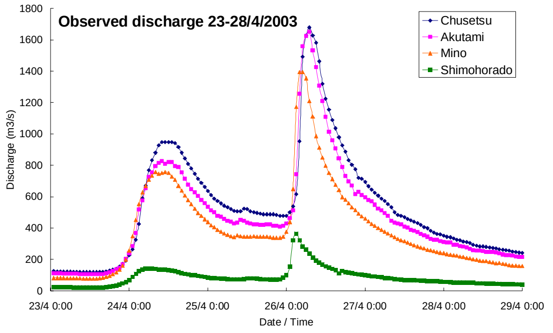

Figure 2 Observed discharge, Event 1: 23-28/4/2003

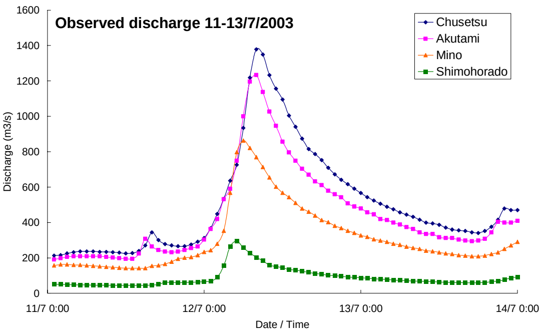

Figure 3 Observed discharge, Event 2: 11-13/7/2003

6. Results and Discussion

Correlations between hydrographs contained in the simulation database to determine optimal query vector forms suggest that for a scenario unrelated to prediction that the discharges at the two target locations are best described by functions of the following form, where superscripts refer to hourly time lag steps and subscripts refer to location names:

Mino: $Q^{t}_{M} = f(Q^{t+2}_{C}, Q^{t+1}_{A}, Q^{t-1}_{S})$

Shimohorado: $Q^{t}_{S} = f(Q^{t+3}_{C}, Q^{t+3}_{A}, Q^{t+1}_{M})$ (11)

Results using local linear modeling with a small number of nearest neighbors gave unstable results for both Mino and Shimohorado. It was found that stability and accuracy of the interpolation and extrapolation results improved as the number of nearest neighbors approached the number of data points in the database, equivalent to the global regression approach. The high linear correlation between hydrographs at each location also suggests that global regression is a valid approach.

Results using global regression for Mino are given in Figure 4 and Figure 5, and for Shimohorado in Figure 6 and Figure 7. Table 1 gives the root mean square (RMS) error and mean absolute relative (MAR) error for the integration at Mino and the extrapolation at Shimohorado for the two events.

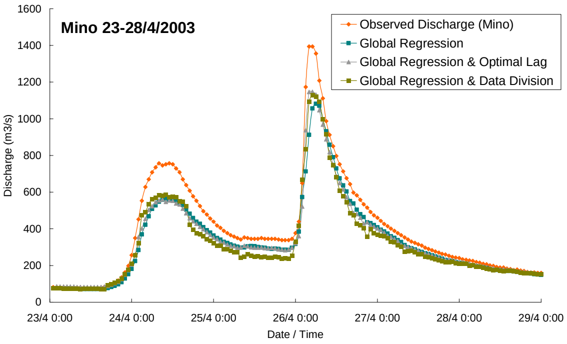

Figure 4 Interpolation for Mino, April 2003

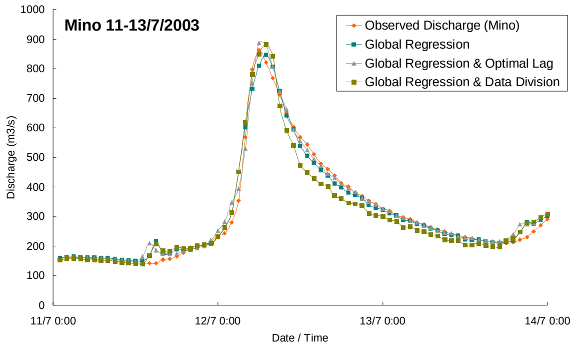

Figure 5 Interpolation for Mino, July 2003

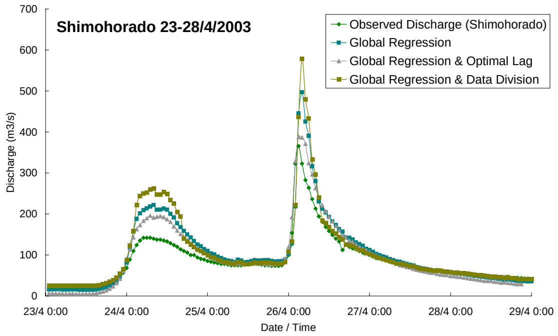

Figure 6 Extrapolation for Shimohorado, April 2003

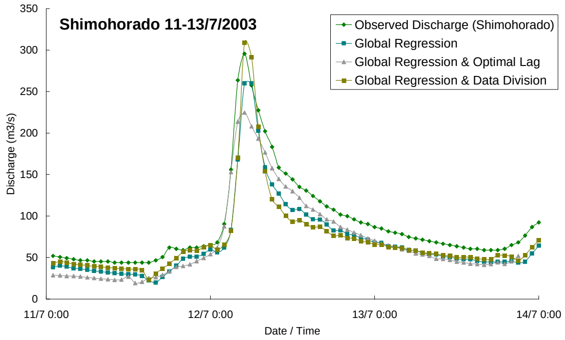

Figure 7 Extrapolation for Shimohorado, July 2003

Table 1 Results for Mino and Shimohorado

	GR[a]		GR / lag		GR / division	
	RMS[b]	MAR[c]	RMS	MAR	RMS	MAR
Mino						
E1[d]	101	0.140	89.9	0.128	99.8	0.165
E2[e]	23.8	0.0543	21.9	0.053	38.0	0.090
Shimohorado						
E1	36.9	0.201	24.5	0.243	48.5	0.176
E2	25.0	0.251	21.8	0.270	26.1	0.210

[a] Global regression
[b] Root mean square error (m^3/s)
[c] Mean absolute relative error
[d] E1: Event 23-28/4/2003
[e] E2: Event 11-13/7/2003

Application results indicate that the global regression strategy proposed here is capable of estimating hydrographs at distributed positions within a watershed based on knowledge of the hydrographs at positions located at a distance. As would be expected, hydrograph shape is estimated accurately, with rising and falling limbs, and hydrograph peaks timed well. For the unseen events, a mean absolute relative error in magnitude of the estimated runoff of the order of 0.05~0.15 was achieved for the two cases of interpolation for runoff at Mino, with less accurate results for extrapolation to the distant location of Shimohorado of the order of 0.20~0.25.

The results showed that a slight improvement in accuracy was gained for the interpolation at Mino through optimization of the query vector to consider the time lags at which the target location is optimally correlated. Division of the data points in the database to reflect their position in a hydrograph (baseflow, rising limb, peak, falling limb) to train separate regression models for each hydrograph stage showed mixed results with an increase in accuracy only for the extrapolation case at Shimohorado. These results are inconclusive regarding the benefit of employing lag optimization and data division.

7. Discussion and Conclusions

A strategy for interpolation and extrapolation of runoff rates across a watershed has been introduced. Results indicate that global regression can be used to estimate the shape, timing and magnitude of hydrographs separated from reference locations where runoff observations or predictions are available. Further investigation is required to determine the ability of the system to accurately extrapolate results to locations greatly separated from observation locations.

References

Atkeson, C.G., Moore, A.W., and Schaal, S. (1997): Locally weighted learning, Artificial Intelligence Review 11, pp. 11–73.

Babovic, V. and Bojkov, V. H. (2001): Runoff Modelling with Genetic Programming and Artificial Neural Networks, D2K Technical Report 0401-1. Denmark: Danish Hydraulics Institute.

Bentley, J. L. (1980): Multidimensional divide and conquer, Communications of the ACM, Vol. 23, No. 4, pp. 214-229.

Campolo, M., Andreussi, P. and Soldati, A. (1999): River flood forecasting with a neural network model, Water Resources Research, Vol. 35, No. 4, pp. 1191-1197.

Karunanithi, N., Grenney, W.J., Whitley, D. and Bovee, K. (1994): Neural networks for river flow prediction, Journal of Computing in Civil Engineering, Vol. 8, No. 2, pp. 201-220.

Khu, S.T., Liong, S.Y., Babovic, V., Madsen, H. and Muttil, N. (2001): Genetic programming and its application in real-time runoff forecasting, J. American Water Resources Association, Vol. 37, No. 2, pp. 439-451.

Kitanidis, P.K. and Bras, R.L. (1980): Real-time forecasting with a conceptual hydrological model: 1. analysis of uncertainty, Water Resources Research, Vol. 16, No. 6, pp. 1025-1033.

Liong, S.Y., Gautam, T.R., Khu, S.T., Babovic, V., Keijzer, M. and Muttil, N. (2002): Genetic Programming: A new paradigm in rainfall runoff modeling, J. American Water Resources Association, Vol. 38, No. 3, pp. 705-718.

Liong, S.Y. and Sivapragasam, C. (2002): Flood stage forecasting with support vector machines, J. American Water Resources Association, Vol. 38, No. 1, pp. 173-186.

Lorrai, M. and Sechi, G.M. (1995): Neural nets for modeling rainfall-runoff transformations, Water Resources Management, Vol. 9, pp. 299-313.

Moore, A. (1991): An Introductory Tutorial on kd-trees, Technical Report No. 209. Computer Laboratory, University of Cambridge, Robotics Institute, Carnegie Mellon University.

Puente, C.E. and Bras, R.L. (1987): Application of nonlinear filtering in the real time forecasting of river flows, Water Resources Research, Vol. 23, No. 4, pp. 675-682.

Smith, P.J. and Kojiri, T. (2004): Error correction of a rainfall-runoff model using genetic programming and a self-organizing map [Identeki puroguramingu to jiko soshiki mappu ni yoru ryushutsu ryuryo moderu no gosa shori]. Proceedings of 2004 Annual Conference, Japan Society of Hydrology and Water Resources. Muroran (Japan): Japan Society of Hydrology and Water Resources [In Japanese]

ローカル線形回帰分析を用いた流域観測所間の空間的流量内挿手法

Paul James SMITH*・関井勝善*・小尻利治

*京都大学工学研究科

要旨

本研究では，流域の各地点における実時間洪水予測システムを提案する。流量観測地点における流量予測を用いて，観測所がない地点での流量を推定する目的で，ローカル線形回帰分析とグローバル線形回帰分析を用いた空間的流量内挿手法を開発する。従来の地点における流量予測モデルに内挿手法を組み合わせたシステムを採用し，これらの全流域的な実時間洪水予測への適合性について検討する。

キーワード: 洪水予測，分布型流出モデル，ローカル線形回帰分析，グローバル線形回帰分析，内挿