Jun 26th, 3:40 PM - 5:00 PM

Data-Driven Hybrid Approach to Short-Term Predictions of Hydrological Events

Marina G. Erechtchoukova Prof.
York University, marina@yorku.ca

Peter A. Khaiter Prof.
York University, pkhaiter@yorku.ca

Dennis Khaiter
York University, dkhaiter1@gmail.com

iEMSs International Congress on Environmental Modelling and Software

9th International Congress on Environmental Modelling and Software
Fort Collins, Colorado, USA, Mazdak Arabi, Olaf David, Jack Carlson, Daniel P. Ames (Eds.)
https://scholarsarchive.byu.edu/iemssconference/2018/

Data-Driven Hybrid Approach to Short-Term Predictions of Hydrological Events

Marina G. Erechtchoukova[a], Peter A Khaiter[a], Dennis P. Khaiter[b]
[a]*School of Information Technology, Faculty of Liberal Arts and Professional Studies, York University, Canada* (*marina@yorku.ca*, *pkhaiter@yorku.ca*)
[b]*Faculty of Health, York University, Canada* (*dkhaiter1@gmail.com*)

Abstract: Application of supervised classification to short-term forecasting of hydrological events demonstrated that combining outputs of several individual learners into a final judgement improves the accuracy of predictions and creates more robust models. Given that predictions are generated as categorical values corresponding to a class label of a future hydrological event, the ensembles perform better when they incorporate black box models which disagree on the same subsets of data. To further extend the 'diversity of opinions' of ensemble members, black box models of a different type can be added to an ensemble of classifiers. Given that regression methods are aimed at accurate calculation of future magnitudes of hydrological characteristics as opposed to determining a class label denoting future hydrological conditions, the extension of the ensemble approach to regression and hybrid models looks promising to further increase the lead time of reliable predictions. The study investigated regression models applied for short-term predictions of hydrological events, such as flash floods, at a highly urbanized small watershed and their inclusion into ensembles of classifiers. The predictions were generated solely on readily available data collected by stream and rain gauges. The heterogeneous measurements of water levels and precipitation were combined and transformed into phase spaces using time-delay embedding. The potential for developing a hybrid model incorporating both classification and regression approaches was analysed. The results of this study are presented in the paper.

***Keywords**:* Supervised machine learning; hydrological prediction; ensemble; hybrid model; flash flood.

1 INTRODUCTION

The approaches to prediction of watershed hydrological conditions represent a diverse set of techniques. There are several taxonomies of these techniques, which have evolved over the years following the increasing complexity of the addressed tasks and expansion of the set. Mount et al. (2016) presented the current state of data-driven approaches and their potential for hydrological science and applications within the context of integrated hydrology domain. The study highlights the advantages of hybrid modelling tools that incorporate process-based formalization and data-driven analysis. Data-driven quantitative approaches to water resource assessment stemmed from empirical hydrological models where analytical expressions were derived from observation data to calculate hydrological characteristics. These models were well-accepted by practitioners despite their inability to provide exact solutions and reliable extrapolation to the conditions not observed at a time when the models were developed (Nash and Sutcliffe,1970). The approach expanded significantly with application of stochastic and data mining tools to hydrological data and resulted in black box models (ASCE, 2000; Eagleson, 1972; Young, 2003).

The majority of modelling tools are aimed at estimating the magnitudes of hydrological characteristics at points of interest over time. Artificial Neural Networks (ANNs) undoubtedly dominate other data mining approaches used for this purpose. They are widely applied to stream flow modelling in a classical

form and in combination with other artificial intelligence approaches, providing such techniques as Bayesian artificial neural networks (e.g., Humphrey et al., 2016), neuro-fuzzy systems (Nayak et al., 2005), or deep learning tools, (e.g., Li et al., 2016). Other techniques utilize regression methods, such as M5 (Quinlan, 1992) or evolutionary polynomial regression algorithms, and genetic programming (Icaga, 2005). The latter was identified by Elshorbagy et al. (2010) as the most successful tool for predicting some of the hydrological parameters.

There are many problems in applied integrated hydrology, however, where categorical labels associated with subsets of data have to be identified. Such problems can be successfully supported by classification algorithms which produce rule-based models or decision trees. These algorithms are used as major tools or in combinations with other data mining algorithms. A method for constructing a decision tree that provides good estimates of flood frequency and supports sound extrapolation of observation data was proposed by Eagleson (1972). The Classical C4.5 algorithm was used as a benchmark in evaluation of local hydrological predictive models (Hewett, 2003). Decision tree classifiers provide relatively clear interpretations that are attractive for potential users. Prediction of hydrological events based on time series data mining was described by Damle and Yalcin (2007), Erehtchoukova et al. (2016b), McCulloch et al. (2008), and Segretier et al. (2013).

The distinguishing feature of data-driven techniques applied in hydrological modelling is the requirement for observation data collected at the real watersheds. The latter makes the developed models and their results site-specific. To improve the robustness of these models, a hybrid multi-model approach can be suggested. The approach requires a well-defined rule for the aggregation of predictions generated by individual models into a final judgement. Aggregation rules for black box models of the same type have been used successfully. An example is various combination rules utilized in ensembles of classifiers (Erechtchoukova et al., 2017). The paper presents the results of exploratory computations for investigation of the data-driven hybrid multi-model approach incorporating categorical predictions with estimated magnitudes of hydrological characteristics.

2 METHODOLOGY

2.1 Background

Rapid changes in hydrological conditions of small watersheds in response to heavy precipitation may result in fast water inundation of submerged territories, causing flash floods. These events, although short in their duration, are dangerous for highly populated areas and require thorough monitoring and accurate short-term forecasts for issuing alerts and undertaking preventive measures. The study was a part of a larger project on the development of an operational decision support tool for flood management in small highly urbanized watersheds. The tool is expected to generate predictions of hydrological conditions at a cross section of interest, utilizing meteorological and hydrological data routinely collected on watersheds.

The methodology for applying supervised classification algorithms to short-term prediction of hydrological events using time-delay embedding of readily available series of observations on water level and precipitation conducted at a watershed was one of the techniques proposed for this tool (Erechtchoukova et al., 2016b). The methodology produces a model which takes data available on a watershed as input variables and generates a label which determines the class of a future event, namely as a “high flow event” or “low flow event”, at the cross section of interest. The black box model is constructed as a heterogenous ensemble of classifiers. To apply the methodology, the time series of data collected from all observation sites of a small watershed are to be transformed into the phase space following the formula:

$$X(t+j\tau) = Y_1(t-1), Y_1(t-2), \dots, Y_1(t-K\tau), \dots Y_M(t-1), Y_M(t-2), \dots, Y_M(t-K\tau), Class(t+j\tau), \quad (1)$$

where $X(t+j\tau)$ is the element of the phase space built to generate predictions at time t with $j\tau$ lead time, τ is the time interval between two subsequent observations, $K\tau$ is the delay time, $Y_i(t)$ is the instantaneous measurement from i-th gauge at the time t, $i = 1,\dots,M$, $Class(t)$ is the class label of an event at the investigated cross section at the time t. The class label is assigned by the event characterization function formulated according to existing business rules:

$$f_M(X(t)) = \begin{cases} 'high', & Y_M(t) \geq H_{thresh} \\ 'low', & Y_M(t) < H_{thresh} \end{cases} \quad (2)$$

where H_{thresh} is the threshold used in the practice of operational flood management to determine the emergency event and is site-specific.

The predictions are generated as categorical values -- class labels of future events at the cross section of interest. Further extension of this methodology to incorporate data of varying granularity was previously described by Erechtchoukova and Khaiter (2017). Methodological aspects of the ensemble development for the given problem, including suitable diversity measures and combination rules were investigated by Erechtchoukova et al. (2016a, 2017).

2.2 Experiment settings

Although models developed using supervised classification and the proposed approach generated reliable predictions for reasonably long lead times, major directions for further improvement of the methodology were considered to maintain the robustness of constructed models and to extend the lead time of reliable forecasts. The analysis of misclassified tuples from testing sets revealed that all flood events observed over the investigated period of time were identified correctly and error occurred either at the very beginning or the very end of some flood events. This means that errors corresponded to the parts of hydrograph in the close proximity to the threshold used in the even characterization function (2). One of the possible ways to improve the performance of developed ensembles is to increase their diversity by introducing models of different types, namely, regression machine learning algorithms. It was first necessary to investigate the performance of regression models on the same sets of data to ensure that they may mitigate, not amplify the classifier errors.

In the next step, the extension of the ensembles was considered. Given that regression models produce magnitudes of dependent variables as opposed to classifiers generating categorical values, the issue of aggregation of predictions of individual members into a final judgement becomes important. In general, two approaches can be considered. In the first approach, the water level magnitudes predicted by regression models can be supplied to the event characterization function (2) to determine the predicted class of an event and, after that, a combination rule can be applied to aggregate the individual predictions into a final judgement. The second approach is based on the uncontested advantage of the classification algorithms to process both numerical and categorical values of independent variables. This advantage allows to utilize stacking without conversion of predicted magnitudes into class labels. The current study investigated the first approach, using majority vote as a combination rule, and leaving the alternative for future research.

Two regression algorithms were considered in the initial steps of the study, namely, REPTree (Witten and Frank, 2000) and M5Base (Quinlan, 1992), both available in WEKA software (Hall et al., 2009). The heterogenous ensemble of classifiers comprised C4.5 (Quinlan, 1993) implemented as J48, CART (Breiman et al., 1984), the NBTree algorithm (Kohavi, 1996), the Ridor classification algorithm (Gaines and Compton, 1995), and Jrip, suggested by Cohen (1995).

Regression models were constructed on the same dataset as the ensembles of classifiers with the only difference in representation of the target variable as the magnitude of water level at a cross section of interest at any given moment of time. After that, regression model predictions were converted into categorical values, determining the type of a hydrological event. To apply regression models, time series of observation data were transformed into a phase space using the time-delay embedding according to the formula:

$$X(t + j\tau) = Y_1(t-1), Y_1(t-2), \dots, Y_1(t-K\tau), \dots Y_M(t-1), Y_M(t-2), \dots, Y_M(t-K\tau), Y_1(t+j\tau), \quad (3)$$

where $Y_1(t + j\tau)$ is the target variable. The phase space was split into two subsets – training with approximately two thirds of the tuples and a testing set with the rest of the elements.

2.3 Model performance measures

There are several estimates of generalization error which are used to evaluate model performance. From the perspective of flood management, correct prediction of high-flow events is more important than prediction of low-flow events, and misclassification of low-flow events is less dangerous than misclassification of high flow events. This implies that recall and precision can provide more information than weighted averaged indicators for classification models. The recall is described by the ratio of the number of correctly identified high flow tuples to the total number of high flow tuples in the test set. The precision is calculated as the number of correctly identified high flow events tuples divided by the total number of high flow labels assigned by a classifier on the test set.

Given the necessity to compare and interpret models of different types, the mean absolute error, the maximum absolute error, and the relative absolute errors were considered for regression models. The mean and maximum absolute errors were calculated as the average and maximum absolute deviations of predicted magnitudes from the corresponding observed values of water level. The relative absolute error was calculated as the ratio of the mean absolute error to the error of ZeroR classifier, which predicts the most frequent value. All estimates of the selected measures were calculated in WEKA software.

3 EXPLORATORY COMPUTATIONS

3.1 Datasets

Data used in the study were collected by the Toronto and Region Conservation Authority (TRCA) on a watershed of Spring Creek, Ontario, Canada – a small stream flowing in highly urbanized area with up to 70% of impervious surface. This small watershed is characterized by a "flashy" response to intense precipitation during the warm season from April to November. The stream daily average baseflow estimates are close to 0.20m^3/s. The study was conducted on data collected during the dry year, 2014. The total amount of precipitation the watershed received during the investigated warm season was estimated at 539mm. Spring Creek's average instantaneous water discharge was approximately 0.47m^3/s, which totalled to an annual water discharge of 0.015km^3.

The TRCA Flood monitoring network has two gauges installed on the stream and two rain gauges collecting data on the hydrological conditions of the watershed. Stream gauges generate time series of water levels with 15-minute intervals, while rain gauges record data every five minutes. The time series of precipitation were aggregated to 15-minute granularity and synchronized with water level data. The duration of intense precipitation observed in recent years and the duration of high-flow events pre-determined the time-delay interval corresponding to the longest lag between a high-flow event and high intensity precipitation registered at the watershed. Subsequently, a phase space of tuples $X(t+j\tau)$ was re-constructed using a three-hour time-delay embedding following (3).

The predictions were generated for the most southern cross section equipped with the stream gauge. The water level threshold for this location is set to 172.75m. This magnitude was used in (2) to identify the class of an observed event.

3.2 Regression model evaluation

Models of both types were built for four lead time intervals ranging from 15 minutes up to one hour. Model's generalization error was evaluated on tuples from the testing set unseen at the training step corresponding to model development. This approach allows for pessimistic estimates. The results of the model evaluation are presented in Table 1.

The average estimates of the models' performance look very attractive, supporting predictions with one-hour lead time. According to these estimates, the M5Base algorithm produced the model which outperforms the one developed using the REPTree inducer. The maximum deviation of predicted magnitudes from corresponding observation data is significantly less for the model constructed by training the REPTree algorithm. For practical applications, however, it is necessary to consider the errors the models make when predicting high flow events. The maximum absolute error of the model

produced by M5Base algorithms was threefold greater than those made by REPTree based model, consistently for all considered lead time intervals.

Table 1. Regression model performance for different lead time intervals.

Lead time	Inducer					
	M5Base			REPTree		
	Mean abs(error), m	Maximum abs(error), m	Relative abs(error), %	Mean abs(error), m	Maxim abs(error), m	Relative abs(error), %
15 min	0.0029	0.534	4.6504	0.0046	0.151	7.3087
30 min	0.0044	0.475	7.0291	0.0070	0.151	11.1623
45 min	0.0059	1.528	9.3771	0.0067	0.443	10.6543
60 min	0.0079	1.819	12.5668	0.0083	0.551	13.1779

The predicted magnitude of water level was used to assign a class label to a corresponding tuple. This label was compared with the label of the corresponding tuple from the testing set to identify those tuples whose errors lead to misclassification of a future condition at the cross section of interest. Analysis of potentially misclassified tuples demonstrated that errors of regression models occur in close proximity of extreme values of water level, both maximal and minimal, so that generated predictions are behind actual values (Figure 1). The models performed well outside of these intervals.

One of the four general principles for the creation of an ensemble which outperforms its members is to ensure the 'diversity of opinion'. Given that subsets of tuples from the phase space, on which classifiers and regression models fail to predict high-flow events are not the same, it was expected that combining corresponding predictions from all models may improve the performance of the prediction tool.

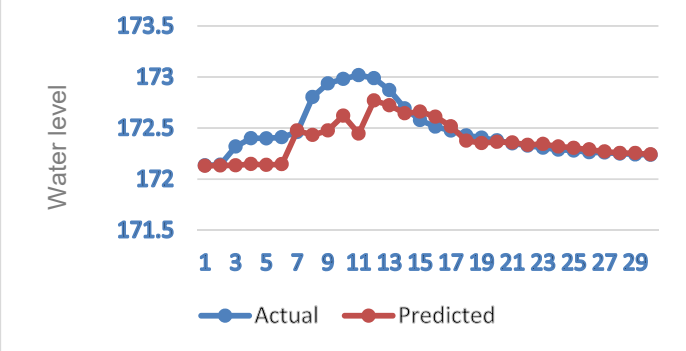

Figure 1. Pick flow modelling with 60 minutes lead time using M5Base inducer.

3.3 Ensembles evaluation

To test the data-driven hybrid approach, two ensembles: a heterogenous one consisting of the five classifiers and a hybrid one comprising all seven investigated models, were developed. The individual classifiers were constructed using methodology presented by Erechtchoukova et al. (2016b) by applying the following inducers: C4.5, CART, Jrip, NBTree, and Ridor. The same time series of observation data were transformed into a phase space following (1) and (2). The classification algorithms were chosen to ensure 'diversity of opinion' according to the analysis presented by Erechtchoukova et al. (2016a). Majority vote was applied to combine individual predictions of classifiers into a final judgement. The predictions of two regression models were converted into corresponding class labels and a hybrid ensemble was created with the same aggregation rule, i.e., majority vote. The performance of both ensembles was evaluated for four lead time intervals (see Figure 2).

4 DISCUSSION AND CONCLUSIONS

The two selected measures of the ensembles' performance exhibited different dynamics with increasing lead time interval. The recall, which reflects how well a model predicts high-flow events without consideration of low-flows, declined with the extended forecast horizon. The precision accounting for both: accurately predicted high-flow tuples as well as misclassified low-flow tuples from the testing sets - showed notable improvement for extended lead times. However, this was achieved mainly due to

more accurate predictions of low-flow events because the increase in recall values was not that significant.

Both ensembles demonstrated the same predictive ability on the short lead times of up to two observation intervals. The improvement in performance of the hybrid ensemble was revealed at longer lead time intervals as it was identified by both selected measures. Adding regression models with subsequent classification of their results to the ensemble increased the ensemble's *True-positive* rate from 56% to 64%, which makes an approach promising for short-term hydrological predictions.

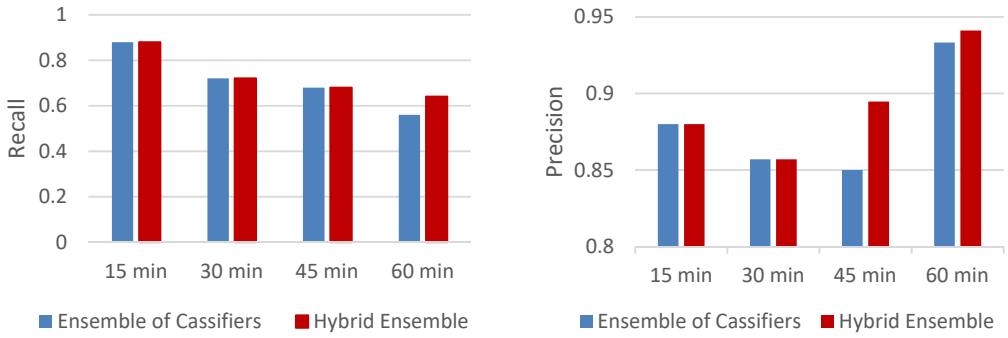

Figure 2. Comparison of heterogenous and hybrid ensembles.

The exploratory computations were conducted using default settings of employed machine learning algorithms. The aggregation rule assigned equal weights to all individual judgements. These experimental settings imply that the estimates obtained at this stage of the study are pessimistic in a sense that better performance of both ensembles can be achieved. Given that such fine-tuning of inducers is site-specific, it can be conducted after the framework for short-term predictions using the data-driven hybrid approach is complete. The current results confirm the advantage of expanding an ensemble to include models of different types and suggest further investigation of the hybrid approach. First of all, it is worthwhile to explore other ways to aggregate the results of individual ensemble members, starting from the stacking meta-modelling approach. It is also necessary to investigate ensemble membership and the extent to which balance between models of both types affects the 'diversity' of members' opinion. The conducted study is important for the hybrid modelling approach which incorporates data-driven techniques with process-based models as it provides insights on model selection based on their predictive abilities and means for aggregation of obtained decisions.

ACKNOWLEDGMENTS

Data used in the study was provided by the Toronto and Region Conservation Authority. The authors are thankful to TRCA and especially J. Duncan and J. Cao for providing data and necessary clarifications. The authors are grateful to editors and anonymous reviewers for their thoughtful suggestions and helpful comments on the manuscript.

REFERENCES

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000. Artificial Neural Networks in Hydrology. I: preliminary concepts. Journal of Hydrologic Engineering 5 (2), 115-123.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. Wadsworth International Group, Belmont, California.

Cohen, W.W., 1995. Fast Effective Rule Induction. In: Twelfth International Conference on Machine Learning, pp. 115-123.

Damle, C., Yalcin, A., 2007. Flood predicting using time series data mining. Journal of Hydrology 333, 305-316.

Eagleson, P.S., 1972. Dynamics of flood frequency. Water Resource Research 8 (4), 878-898.

Elshorbagy, A., Corzo, G., Srinivasulu, S., Solomatine, D.P., 2010. Experimental investigation of the predictive capabilities of data driven modelling techniques in hydrology - Part 2: Application. Hydrology and Earth System Science 14, 1943-1961, DOI:10.5194/hess-14-1943-1961.

Erechtchoukova, M.G., Khaiter, P.A., 2017. The effect of data granularity on prediction of extreme hydrological events in highly urbanized watersheds: A supervised classification approach. Environmental Modelling and Software 96, 232-238. DOI:10.1016/j.envsoft.2017.06.054.

Erechtchoukova M.G., Khaiter P.A., Ditmans, M., Khaiter, D., 2017. Role of a 'combination rule' in hybrid short-term prediction of hydrological events. In: Syme, G., Hatton MacDonald, D., Fulton, B. and Piantadosi, J. (Eds.) MODSIM2017, 22nd International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2017, pp. 866–872. ISBN: 978-0-9872143-7-9.

Erechtchoukova, M.G., Khaiter, P.A., Khaiter, D.P, 2016a. Measures of Diversity in Ensembles of Classifiers for Short-term Prediction of Hydrological Events. In: Sauvage, S.,Sánchez-Pérez, J.M., Rizzoli, A.E. (Eds.), 2016. Proceedings of the 8th International Congress on Environmental Modelling and Software, July 10-14, Toulouse, France, pp. 754-761. ISBN:978-88-9035-745-9.

Erechtchoukova, M.G., Khaiter, P.A., Saffarpour, S., 2016b. Short-term predictions of hydrological events on an urbanized watershed using supervised classification. Water Resources Management 30 (12), 4329-4343, DOI: 10.1007/s11269-016-1423-6.

Gaines, B.R., Compton, P., 1995. Induction of Ripple-Down Rules Applied to Modeling Large Databases. Journal of Intelligent Information Systems, 5(3), 211-228

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I, H., 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations 11 (1).

Hewett, R., 2003. Data mining for generating predictive models of local hydrology. Applied Intelligence 19, 157-170.

Humphrey, G.B, Gibbs, M.S., Dandy, G.C., Maier, H.R., 2016. A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network. Journal of Hydrology 540, 623-640.

Icaga, Y., 2005. Genetic algorithm usage in water quality monitoring networks optimization in Gediz (Turkey) river basin. Environmental Monitoring and Assessment 108, 261-277.

Kohavi, R., 1996. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In: Second International Conference on Knowledge Discovery and Data Mining, pp. 202-207.

Li, C., Bai, Y., Zeng, B., 2016. Deep feature learning architectures for daily reservoir inflow forecasting. Water Resources Management 30 (14), 5145-5161.

McCulloch, D.R., Lawry, J., Cluckie, I.D., 2008. Real-time forecasting using updateable linguistic decision trees. Fuzzy Systems, DOI: 10.1109/FUZZY.2008.4630634.

Mount, N.J., Maier, H.R., Toth, E., Elshorbagy, A., Solomatine, D., Chang F.-J., Abrahart, R.J., 2016. Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei Science Plan. Hyrological Science Journal 61 (7), 1192-1208.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models. Part I – a discussion of principles. Journal of Hydrology 10, 282-290.

Nayak, P.C., Sudheer, K.P., Rangan, D.P., Ramasastri, K.S., 2005. Short-term flood forecasting with a neurofuzzy model. Water Resources Research 41, W04004. DOI:10.1029/2004WR003562.

Quinlan, R., 1993. C4.5: programs for machine learning. Morgan Kaufmann Publishers, San Mateo, CA.

Quinlan, J. R., 1992. Learning with continuous classes. In: The Australian Joint Conference on Artificial Intelligence, pp. 343--348. World Scientific, Singapore.

Segretier, W., Collard, M., Clergue, M., 2013. Evolutionary predictive modelling for flash floods. In: Proc. Evolutionary Computation (CEC), 2013 IEEE Congress.

Witten, I.H., Frank, E., 2000. Data mining: practical machine learning tools and techniques with java implementations. Morgan Kaufmann, San Mateo, CA

Young, P., 2003. Top-down and data-based mechanistic modelling of rainfall–flow dynamics at the catchment scale. Hydrological Processes 17, 2195-2217.