Paul H. Whitfield, Environment Canada, 700-1200 West 73rd Avenue, Vancouver, BC, V6P 6H9, Canada

Reporting Scale and the Information Content of Streamflow Data

Abstract

Hydrometric records are collected on a nearly continuous basis in many hydrologic metering programs around the world. These measured signals are typically reported on a time scale longer than the original observations, for instance daily, monthly, or annual runoff. The analysis presented here examined records for several years from a number of watersheds in British Columbia. It addresses whether or not the reporting of daily mean discharge is appropriate for streams of different size scales and from different hydrologic regions of the province. Daily mean discharge is an appropriate reporting scale for watersheds of large area and low precipitation. However in smaller watersheds, and particularly those subjected to coastal rains, a finer time interval is more a more relevant scale.

Introduction

Records of streamflow are collected from networks of stations by many agencies around the world. Caselton and Husain (1980) suggest that the stations in such networks should be considered as transmitting hydrologic information. The hydrologic information is contained in the data records collected manually or automatically. The sampling interval is an important decision in the design of data acquisition systems (Gupta 1973). If there are significant serial correlations, redundant information will result if the sampling is too frequent (Whitfield 1983). However, infrequent sampling would lead to a loss of information (Gupta 1973).

Many gauging programs collect observations continuously and then report data at a time scale much reduced from the scale at which the original measurements were made. One example is that continuous analog records of water levels are converted into daily mean discharges. This daily data contains much information about central tendency, but often much of the hydrologic information contained in the original signal is lost.

Streamflow records have many properties, and many uses. Several authors (Matalas and Langbein 1962; Quimpo and Yang 1970; Gupta 1973) have described the information content of the mean of such records. The information about central tendency contained in such records is used in assessing economic and engineering aspects of water. However, much of the detailed information contained in such records has been lost. Historically, the information which was lost was of little consequence as the amount of detail was difficult, if not impossible, to use in hydrologic analysis. However, several authors have suggested that there be more consideration to time scales in data collection and analysis (Loftis et al. 1991).

Gupta (1973) considered the information content of hydrologic records when reduced from daily values to weekly and monthly records. Gupta (1973) performed a number of analyses considering the information loss which occurs when records were smoothed to means which represent a week or a month.

The present analysis considers this same question when applied to records at a finer time scale; considering the information loss which accompanies the smoothing of records digitized at 15 minute intervals and smoothed to 30 minute, hourly etc. durations. In much of the coastal area of British Columbia, there is considerable detail in records collected at 15 minute intervals (Whitfield and Wade 1992; Wade and Whitfield 1994). The present work considers the information loss that happens when data from a 15 minute sampling interval is smoothed to longer time intervals.

Today, we are increasingly concerned with details, mechanisms, and environmental processes than in the past. Ecosystems operate on a wide range of time and space scales, ranging from fractions of second to thousands of years, and from millimeters to global extent (Steele 1991). Standard methods and standard reporting scales were expedient and efficient for dealing with engineering and economic assessments. However, today we are seeking different types of information from the records, and the availability of hardware and software to process large amounts of data allows us to consider more detailed examination of the records. Specifically we are seeking information about the processes which result in the observed hydrologic events and what the events tell us about the environments and ecosystems.

In the coastal areas of British Columbia rainfall driven streamflow can vary quickly; often very significant changes can occur over a few hours or less. The analysis here relates the information loss to two physical properties of the stations at which the records were collected. First, the examination of information content in relation to drainage area, and second to runoff. It was expected that small watershed with high runoff would have much more complex records, and more information would be lost when records from these stations were smoothed than stations which have large drainage areas and less annual runoff per unit area.

Methods

For each of the cases considered here the primary measurement interval is 15 minutes. These values are the integration of streamflow over 15 minutes rather than discrete samples. Twenty-two hydrometric stations from different hydrologic regions of British Columbia and over a broad range of drainage areas were examined. For each station three separate years of data were assessed. These stations are listed in Table 1 with the drainage area, the long term annual runoff, and the Environment Canada station number. These stations were selected to cover the range of both drainage area and runoff from small to large. A plot of watershed area against runoff in Figure 1 shows the variation amongst the 22 stations studied. These stations were chosen to cover both several orders

TABLE 1. Station number and name for the watersheds with the watershed area and long term annual runoff and the runoff in each of the three years used in this study.

				Mean Annual Discharge in Cubic Meters/second		
	Station	Area (km^2)	Runoff in mms	1991	1992	1993
08HB069	Carnation Creek at 150 m contour	2.53	1576	0.159	0.153	0.143
08HB048	Carnation Creek at the mouth	10.1	2150	0.912	0.740	0.825
08MH098	West Creek near Fort Langley	11.4	704	0.345	0.310	0.261
08NM145	Bulman Creek at the mouth	12.7	176	0.131	0.035	0.089
08MH076	Kanaka Creek near Webster Corners	47.7	1057	2.049	1.640	2.065
08JA016	MacIvor Creek near the mouth	53.4	622	1.744	0.863	1.185
08NK021	Fording River below Clode Creek	104	580	2.955	1.355	2.571
08HC004	Bedwell River above Ursus Creek	114	3258	16.775	12.171	13.425
08NM171	Vaseux Creek above Solco Creek	117	256	1.455	0.607	1.355
08HC002	Ucona River at the mouth	185	2155	15.470	15.552	14.448
08AA009	Giltana Creek near the mouth	194	225	1.772	2.102	1.096
08NE077	Barnes Creek near Needles	201	448	3.571	3.022	3.672
08NG077	St. Mary River below Morris Creek	206	876	6.759	7.734	6.102
08MG001	Chehalis River near Harrison Mills	383	2198	38.435	28.924	28.652
08KA001	Dore River near McBride	404	1365	27.994	21.509	13.393
08JD006	Driftwood River above Kastberg Creek	407	633	11.647	10.807	6.935
08MF068	Coquihalla River above Alexander Creek	720	1113	39.564	25.507	26.377
08KH010	Horsefly River above McKinley Creek	785	772	22.351	21.351	25.415
08NM116	Mission Creek near East Kelowna	811	232	9.096	3.644	8.750
08GA022	Squamish River near Brackendale	2330	2575	256.929	252.589	175.033
08GD004	Homathko River at the mouth	5720	1581	396.308	333.401	301.609
08MF005	Fraser River at Hope	217000	343	3100.473	2931.958	2468.117

of magnitude in drainage area and from low runoff to high runoff watersheds.

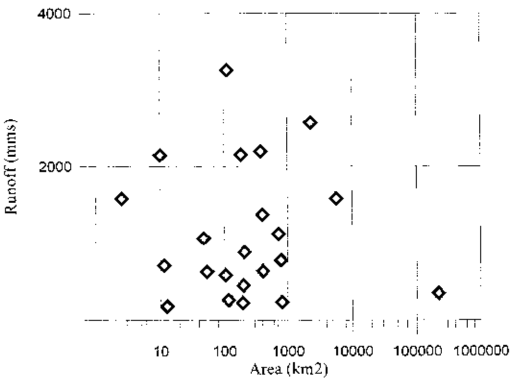

Figure 1. Runoff and drainage area relationships for the 22 hydrologic stations described in Table 1.

Each series was smoothed from 15 minute observations to 30 minute averages, 1 hour averages, 2 hour, 4 hour, 8 hour, 1 day and 2 day averages. Each of these smoothed series has the same mean as the initial series. The autocorrelation function of each of these series was calculated. One problem with the interpretation of the lag one autocorrelation is the differences in the degrees of freedom between the 15 minute data and the smoothed versions. When smoothed to one day there is approximately a loss of 100 degrees of freedom. While the total degrees of freedom for all cases is large, it is not equal between the various curves.

The assessments of information loss was based on the assessment of two properties of the original and smoothed records. The first of these is the autocorrelation of the values in the time series. The autocorrelation is a measure of the relationship between each observation and the next one in the series and is calculated using:

$$r_k = \frac{\sum(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum(x_t - \bar{x})^2}$$

where x_t is an observation at time *t*, and $\bar{x}$ is the mean of the series. When values in the sequence are similar the autocorrelation value will approach a value of one. If the series is random, the value will approach zero.

The second property is a measure of the amount of information contained in the variance of the values in the series compared to the total variance of the entire data set. While several estimators of the information content have been proposed von Neumann's variance ratio was chosen (von Neumann et al.; 1941; von Neumann 1941). The Von-Neumann's variance ratio was calculated using:

$$\frac{\sigma^2}{S^2} = \frac{\sum(x_t - x_{t-1})^2 / (T - 1)}{\sum(x_t - \bar{x})^2 / T}$$

Where x_t is an observation at time *t*, $\bar{x}$ is the mean of the series and *T* is the total number of observations. Von-Neumann's ratio is essentially the ratio of the variance of the points in the series, one to the next, against the statistical variance of the entire series. This provides a measure of the information content of the series of observations, and as it approaches unity indicates that the point to point variance approaches the statistical variance. An important feature of Von-Neumann's variance ratio is that the effect of sample sizes associated with the smoothing of the raw data is removed (von Neumann 1941).

Results

The difference between 15 minutes accumulations and daily mean discharges are shown in Figure 2. A sixteen day interval is shown for a small watershed (Figure 2a) and for the same period in a large watershed (Figure 2b). These are typical examples of the impact of smoothing. For Carnation Creek, a small high discharge stream, the impact of the smoothing is quite obvious (Figure 2a). Here the observed streamflow is only grossly represented by the daily mean discharge. For the Fraser River at Hope, a large river with a large drainage area, the daily mean streamflow more closely represents the original observations (Figure 2b). In particular, the smoothing which takes place for a large river results in the expected integration, however, for the smaller coastal stream there is a complete obliteration of the signal. This then demonstrates that the records of the daily mean discharges significantly alters the information content of the original measurements.

Sampling Rate

Estimates of the lag 1 autocorrelation coefficient and the Von-Neumann's variance ratio were obtained for each of the 22 rivers and streams, for three separate years (1991, 1992, and 1993). The results for each of the three years were similar, and the average lag 1 autocorrelation for the three years for each stream and each level of smoothing is presented in Figure 3. For all streams, the autocorrelation coefficient increases with the

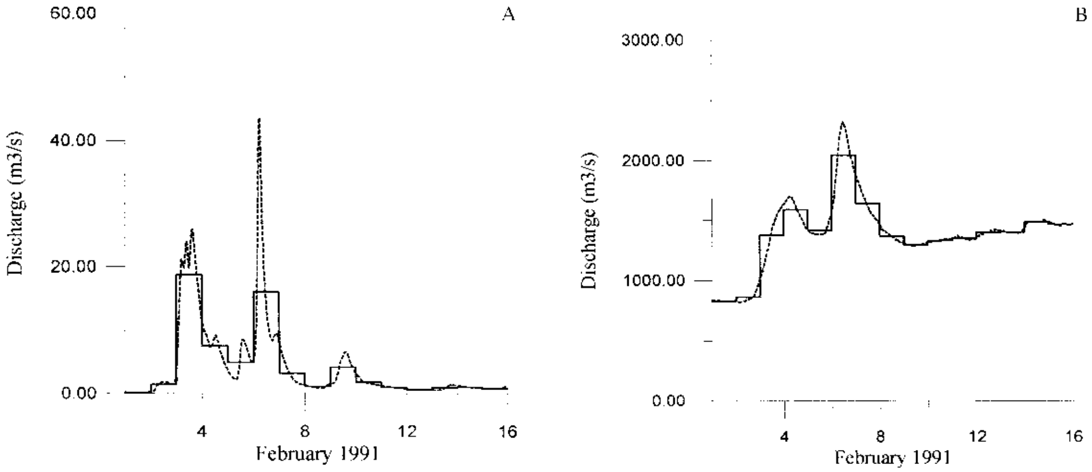

Figure 2. Streamflow records as measured and as reported for a six day period for a high runoff small watershed A and for a large drainage basin B. The dashed line is the series of 15 minute measurements and the solid line is the reported daily value.

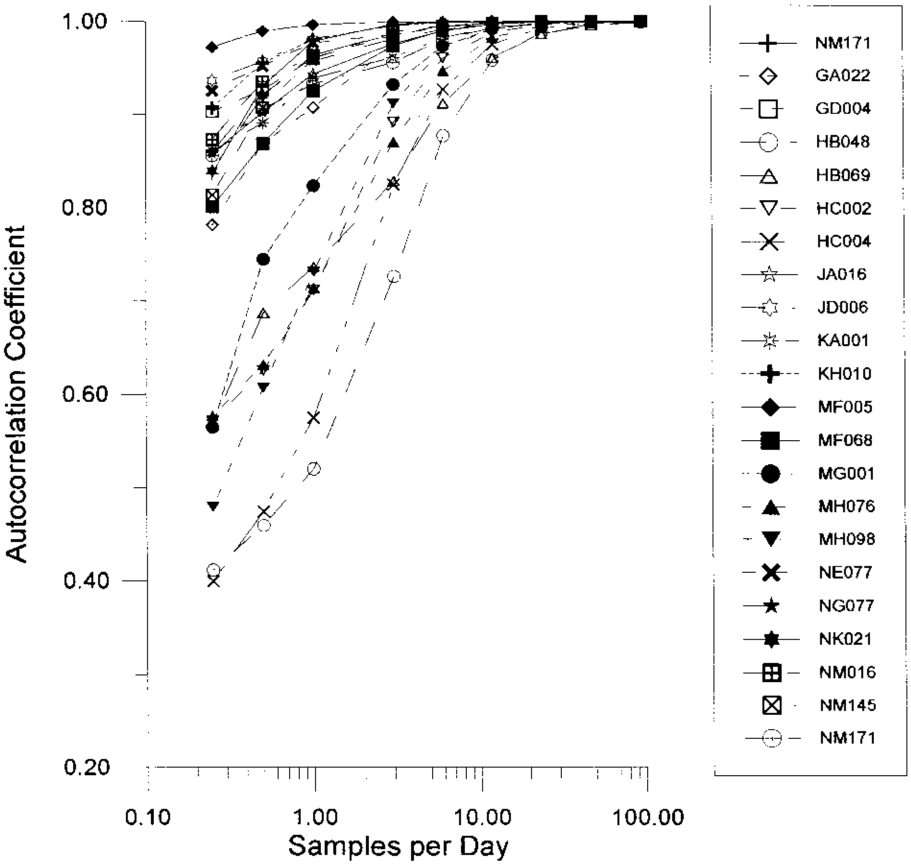

Figure 3. Mean lag 1 autocorrelation coefficient as a function of reporting interval for the 22 hydrologic stations.

number of samples reported per day, indicating that as the density of samples increases the time series which are generated contain less independence between measurements. In larger streams, such as the Fraser River at Hope [MF005] the lag 1 autocorrelation remains near unity until the reporting interval is multiple days. Small coastal streams such as Carnation Creek at the Mouth [HB048] and Bedwell River [HC004] have very low lag 1 autocorrelations when the number of samples reported per day is reduced. In particular, consider the range of observed autocorrelations at the reporting interval of one per day. The lag 1 autocorrelation coefficient varies from 0.5 to nearly 1. This result alone suggests that the information content of such records is different, and the amount of redundant information contained in those records is also different.

There is a relationship between the reporting frequency and the von Neumann's variance ratio for all streams. As the number of samples reported per day increase there is a reduction of the ratio (Figure 4). This reduction in the ratio is in total several orders of magnitude. For all rivers there is a substantial difference between almost every sampling interval. In effect, as the reporting frequency increases there is more information about the signal captured. Considering the reporting interval of one day, there is more than a two order of magnitude variation [1.0 to less that .01] across the stations considered. When the ratio is less than unity the point to point variance in

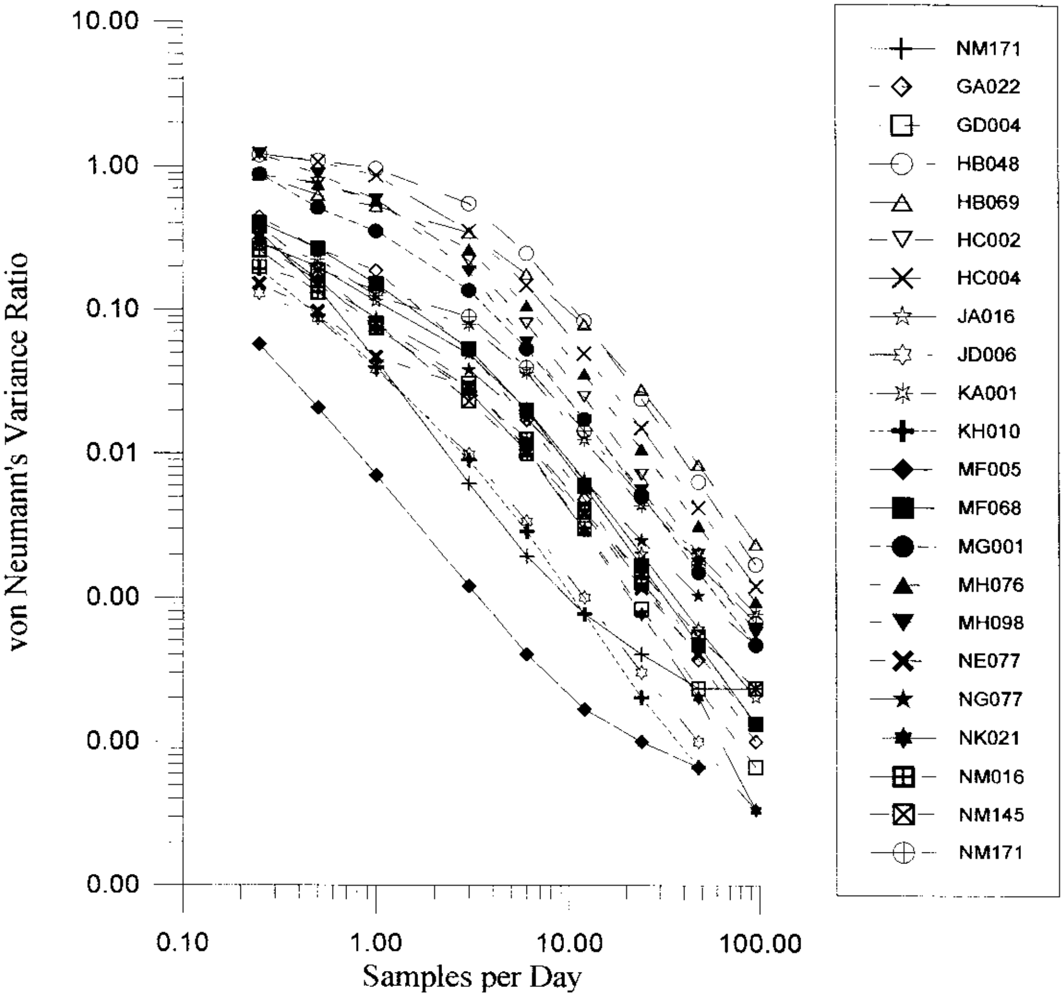

Figure 4. Mean von Neumann's variance ratio as a function of reporting interval for the 22 hydrologic stations.

the series is less than the total variance, indicating that there is significant serial correlation between successive points. As the statistic increases towards unity there is less serial correlation in the data. Thus, for one reporting frequency (i.e. one per day) there is a substantial difference between this statistic across the 22 streams considered.

On the other hand, consider that it might be desirable to hold this statistic constant for all streams in a network, in a case where the same level of information about each station was needed. Consider that a Von-Neumann's variance ratio of 0.01 was important for providing adequate hydrologic data. In such as case, some river could be reported once per day (Fraser River at Hope [MF005]) while others would require more frequent reporting (e.g. Carnation Creek [HB069] ~50 times per day).

Drainage Area

In the preceding results it was suggested that smaller watershed have a more complex hydrologic record, and that more frequent reporting provides more information. A plot of drainage area against lag 1 autocorrelation coefficient (Figure 5) shows that there is a tendency for low lag 1 autocorrelations to be associated with small

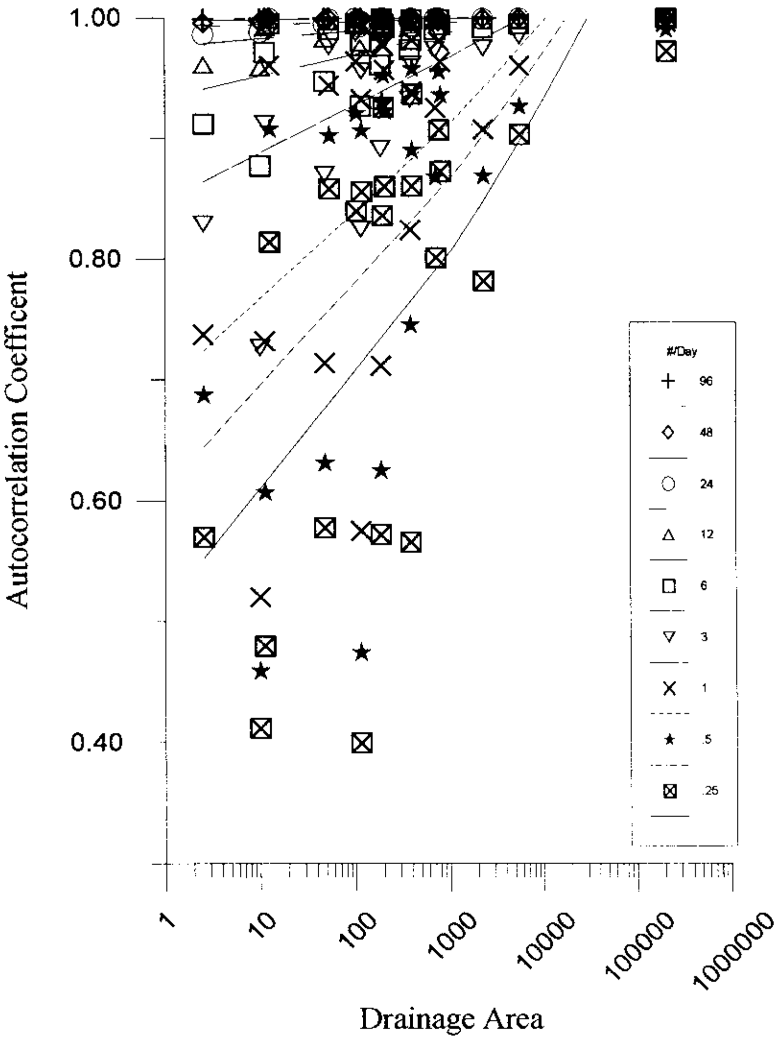

Figure 5. Mean lag 1 autocorrelation coefficient as a function of drainage area for the nine reporting intervals for the 22 hydrologic stations.

drainage basins. This is not a consistent trend, as there are basins of similar area which have very different coefficients for a sampling frequency, yet have similar areas. This is best illustrated by considering one of the lower reporting frequencies. At one sample per day, for example, there the coefficient is high [>0.9] at the largest drainage areas [10000 km^2] but much more variable in smaller drainage basins, ranging from 0.5 to >0.9.

A much clearer pattern exists for the relationship of Von-Neumann's variance ratio to drainage area. There is a consistent decrease in the ratio across the various reporting intervals for the range of drainage areas considered here (Figure 6). This decrease is about one order of magnitude for each of the reporting frequencies from the smallest watershed to the largest.

Runoff

The other hydrologic variable which was considered to be important was runoff. The lag 1 autocorrelation decreased with increasing runoff (Figure 7) particularly for highly smoothed data (0.25). This suggests that there is less correlation

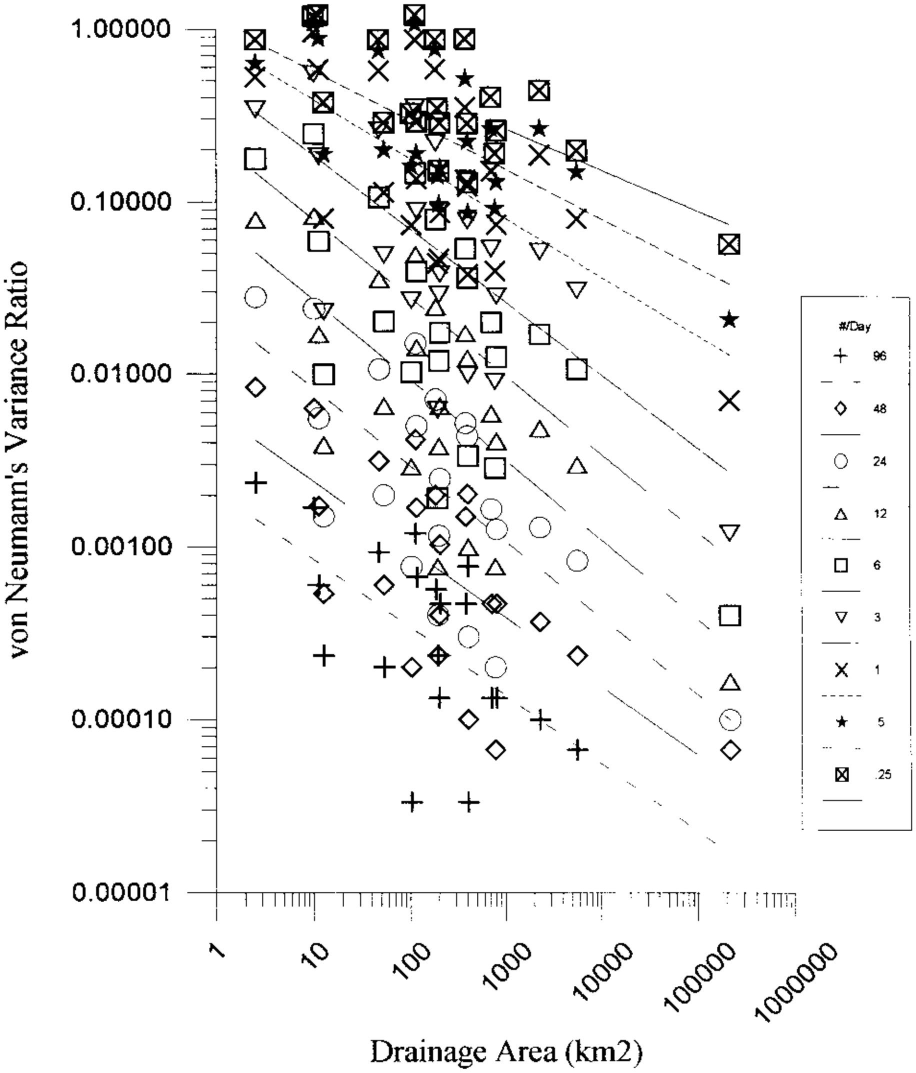

Figure 6. Mean von Neumann's variance ratio as a function of drainage area for the nine reporting intervals for the 22 hydrologic stations.

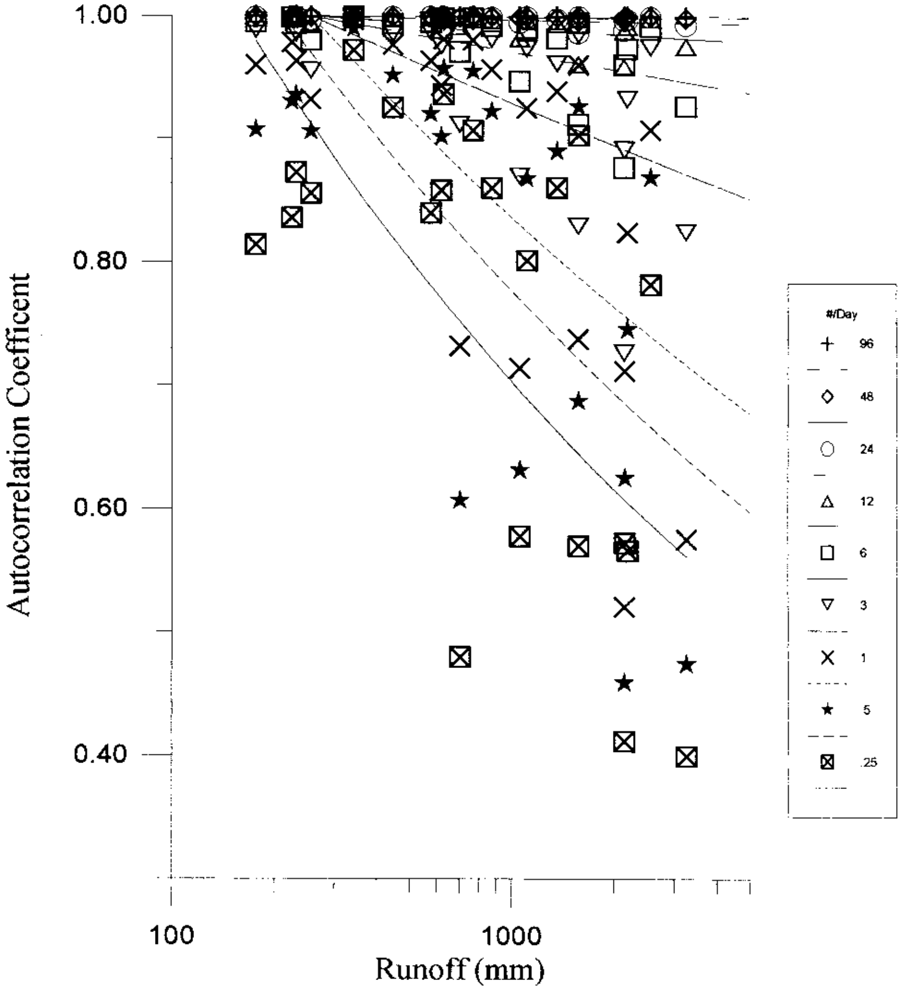

Figure 7. Mean lag 1 autocorrelation coefficient as a function of runoff for the nine reporting intervals for the 22 hydrologic stations.

amongst values in a series from a high runoff stream which has been highly smoothed. There is significant scatter between coefficients for stations of similar area despite the general trend. For the highest reporting frequencies, the autocorrelation for lag one are near unity.

Von-Neumann's variance ratio results have a relationship to runoff. Von-Neumann's ratio increases with increasing runoff for any reporting interval (Figure 8). Generally, there is about one order of magnitude increase in the ratio between the lowest runoff watershed and the highest; however there is considerable scatter about the relationship. This result reflects the higher complexity of the signal in higher runoff streams.

Discussion

The choice of sampling interval is an important decision in the design of data acquisition systems (Gupta 1973; Loftis et al. 1991). The final decision regarding the interval is a compromise between maximizing the information content of the signal and minimizing the collection and storage of redundant information. One possible approach is a sampling interval which minimizes variance and thus maximizes the information content (Gupta 1973). To achieve this Gupta suggests setting a threshold based on information content which balances the information content of the signal against the collection of redundant information. In his study he used a threshold of 0.80 for two

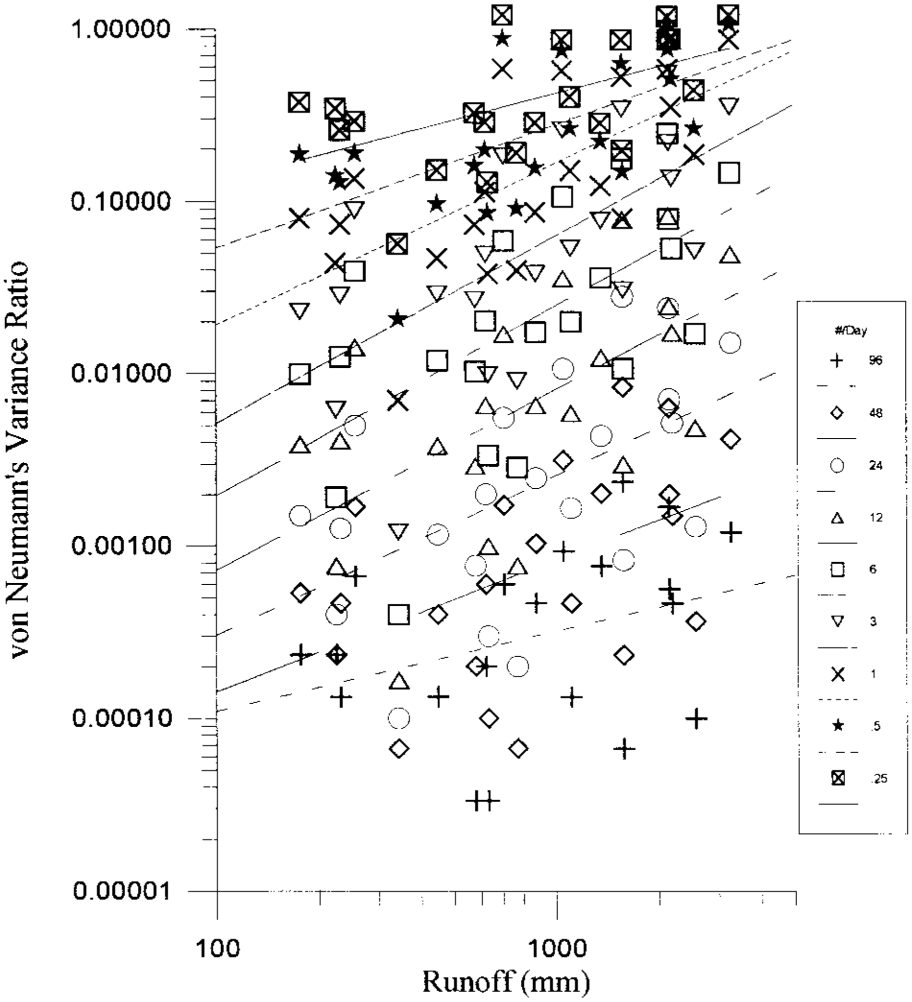

Figure 8. Mean von Neumann's variance ratio as a function of runoff for the nine reporting intervals for the 22 hydrologic stations.

water years from three sites in watersheds with areas of about 100 km^2. In addition Gupta actually resampled the record at discrete intervals—a record which had been created by accumulation and averaging to daily mean flows.

The results presented here show that selecting one size watershed from one geographic region could be misleading. The information content of streamflow records is related to both the runoff and the area of the watershed. The correlation function is a measure of the reservoir mechanism in the watershed (Glottschalk 1977). Caselton and Husain (1980) suggest that there should be a scientific basis for the design and operation of hydrologic networks. One of the important decisions which must be made when developing a network is how the results will be reported. The results here suggest that reporting on one time interval results in a significant loss of detail contained in the original signal. While this information loss may be small in terms of central tendency measures, these details are important in other types of analysis. High frequency water quality monitoring has revealed information about processes taking place within the ecosystem (Whitfield and Wade 1992; Wade and Whitfield 1994). As the interest in ecosystems increases the focus is shifting from means and variances to detailed analysis of processes.

From this analysis it appears that both watershed area and runoff have a strong influence on the information content of hydrologic records. These difference indicated that the response characteristics of watersheds vary with both size and with runoff. This may be considered as a property related to the residence time of water within the watershed. Small watersheds with high runoff have highly variable hydrographs. The information content of records is reduced and the signal obliterated when reported on a time interval greater than the scale at which events take place in that watershed. The information content of such records is not equal across a broad range of drainage areas, nor across a range of annual runoffs.

Conclusions

The information content of records is a useful perspective to consider when designing the collection of a hydrologic record (Gupta 1973). Caselton and Husain (1980) suggest that information content is an important criteria to consider when developing the scientific basis for network operation. The present results support this observation and demonstrate the loss of information which results from operating in a single reporting scale.

Acknowledgments

The author would like to express his appreciation to Lynn Campo and the Water Survey of Canada staff who produced the data on which this study is based, and to Dr. Rory Leith for the contribution of his comments and suggestions to this work.

Literature Cited

Caselton, W. F. and T. Husain. 1980. Hydrologic Networks: Information Content. J. Water Resources Planning and Management Division ASCE 106:503-520.

Glottschalk, L. 1977. Correlation structure and time scale of simple hydrologic systems. Nordic Hydrology 8:129-140.

Gupta. V. L. 1973. Information Content of Time-variant data. J. Hydraulics Div. ASCE 99:383-394.

Loftis, J. C., G. B. McBride, and J.C. Ellis. 1991. Considerations of scale in water quality monitoring and data analysis. Water Resources Bulletin 27:255-264.

Matalas, N. C., and W. B. Langbein. 1962. Information content of the mean. J. Geophysical Research 67:3441-3448.

Quimpo, R. G. and J. Y. Yang. 1970. Sampling considerations in stream discharge and temperature. Water Resources Research 6:1771-1774.

Steele, J. H., 1991. Marine ecosystem dynamics: comparison of scales. Ecological Research 6:175-183.

von Neumann, J. 1941. Distribution of the ratio of mean square successive difference to the variance. Annals of Mathematical Statistics 12:367-395.

von Neumann, J., R. H. Kent, H. R. Bellenson, and B. I. Hart. 1941. The mean square successive difference. Annals of Mathematical Statistics 12:153-162.

Wade, N. L., and P. H. Whitfield. 1994. Observing transient water quality events using electronic sensors. In: National Symposium on Water Quality. American Water Resources Association. pp 105-112.

Whitfield, P. H.. 1983. Evaluation of water quality sampling locations on the Yukon River. Water Resources Bulletin 19:115-121.

Whitfield, P. H.. and N. L. Wade. 1992. Monitoring transient water quality events electronically. Water Resources Bulletin 28:703-711.

Received 30 July 1997
Accepted for publication 13 November 1997