Statistical Nonparametric Model for Natural Salt Estimation

James R. Prairie[1]; Balaji Rajagopalan[2]; Terrance J. Fulp[3]; and Edith A. Zagona, M.ASCE[4]

Abstract: Many rivers in the Western U.S. suffer from high salinity content due to both natural and human-induced causes. Computer simulation models are often used to estimate future salinity levels and identify mitigation needs. To date, estimation of future natural salt loading has utilized linear relationships between natural flow and natural salt. We develop a nonparametric regression technique to fit a functional relationship between natural flow and natural salt. The main advantages of the nonparametric technique are: (1) No prior assumptions have to be made as to the underlying form of the relationship and (2) any arbitrary relationship (linear or nonlinear) can be modeled. In addition, we develop a resampling scheme to provide confidence intervals of the natural salt estimates from the nonparametric model. We apply this model to data from a stream gauge at Glenwood Springs, Colo., on the Colorado River. We show that the new natural salt model reduces the average overprediction of salt mass shown in the existing natural salt model for the period 1941–1995 by approximately 15% (78,000 metric tons).

DOI: 10.1061/(ASCE)0733-9372(2005)131:1(130)

CE Database subject headings: Colorado River; Regression models; Nonlinear systems; Salinity; Simulation.

Introduction

In arid and semiarid river basins with significant irrigation needs, salinity tends to be high due to both natural causes, such as saline springs, and human-induced causes, such as return flows from agricultural use. Locations such as the Brazos River basin in Texas and the Colorado River basin in the western United States are afflicted with salinity problems. In the Colorado River basin, salinity levels in the river must be maintained to meet fixed numeric criteria at several points in the Lower Basin. Minute No. 242 of the International Boundary and Water Commission, United States and Mexico stipulated that water delivered to Mexico have an average flow-weighted salinity of no more than 115 mg/L±30 mg/L above the average annual salinity at Imperial Dam (U.S. Department of the Interior 2001). Although similar standards have not been set in the Brazos River basin, high salinity levels impact the management of reservoirs and the usability of water for irrigation (Wurbs and Karama 1995; Wurbs et al. 1995).

To ensure that future requirements are met, computer simulation models have been used to determine salinity control needs in these basins. Such models require several inputs, including the assumed future hydrologic inflows, the salt loading associated with those inflows, the future projections of development throughout the basin, and the additional salt loading associated with that development. Future hydrologic inflows and the associated salt loading can be estimated from "historic natural flow." The term historic natural flow refers to the flow that would have occurred in the absence of any human development, i.e., no upstream reservoir regulation or upstream depletion [Bureau of Reclamation (BOR) 1987]. Historic measured flows are altered by human development that has varied through time. To remove these variations, the variability of human development from measured flows, natural flows are derived from historic streamflow measurements. In some cases, upstream depletions may be negligible or, at least, invariant with respect to time, so that only a correction for upstream reservoir regulation is warranted. These flows are commonly referred to as "unregulated flows" (Saleh 1993; Wurbs et al. 1995). In either case, the salt loading that would be associated with the inflows, herein referred to as "natural salt" loading, must be estimated. In this way, we can separate the natural and human-induced variability for flows and associated salt loading entering the river. Thus, estimating natural salt is an issue important for the development of data essential to drive a simulation model. Unfortunately, estimating natural salt is not as easy as determining natural flow. For example, estimating crop consumptive use can be readily calculated from measured data. However, the salts returning with irrigation return flows are not easily measured. Indirect methods of estimating natural salt mass must be developed.

A literature search revealed that little has been published regarding natural salt estimation. Most references found focus on the socioeconomic impacts of salinity (Huizenga 1980; Olson 1980; Brown 1984; Lee et al. 1993). However, modeling efforts in the Brazos and Colorado River basins published their salt estimation techniques and we briefly review those efforts

Wurbs et al. (1995) calculated natural salt in the Brazos River basin by adding the salt load associated with reservoir storage and

[1]Hydraulic Engineer, Bureau of Reclamation, Univ. of Colorado, UCB 421, Boulder, CO 80309-0421. E-mail: prairie@colorado.edu

[2]Assistant Professor, Dept. of Civil, Environmental, and Architectural Engineering., Univ. of Colorado, UCB 426, Boulder, CO. 80309-0426. E-mail: rajagopalan.balaji@colorado.edu

[3]Boulder Canyon Operations Office Manager, Bureau of Reclamation, Lower Colorado Region, Boulder City, NV 89006. E-mail: tfulp@lc.usbr.gov

[4]Director, Center For Advanced Decision Support for Water and Environmental Systems (CADSWES), Univ. of Colorado, UCB 421, Boulder, CO 80309-0421. E-mail: zagona@cadswes.colorado.edu

Note. Associate Editor: Mark J. Rood. Discussion open until June 1, 2005. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on August 20, 2002; approved on November 18, 2003. This paper is part of the ***Journal of Environmental Engineering***, Vol. 131, No. 1, January 1, 2005. ©ASCE, ISSN 0733-9372/2005/1-130–138/$25.00.

diversions, and removing the salt load associated with reservoir releases and return flows from the measured historic salt load. Similarly, natural flow was found from measured historic flow minus the effects of reservoir regulation and evaporation. A linear regression was used to develop a relationship between natural salt and natural flow. To estimate the uncertainty in the relationship, the residuals from the regression were first fit to a normal distribution and a value generated from that residual distribution was then added to values generated using the regression.

In the Colorado River basin, Malone et al. (1979) calculated natural salt by removing the estimated agricultural salt loading and measured point source salt loading from the measured historic salt loading. The agricultural salt loading was estimated by two techniques. The first technique developed a relationship for agricultural salt loading dependent upon a base leaching factor and diversion efficiency. The base leaching factor related soils information to the amount of salinity added for a given return flow. The second technique computed the base leaching factor dependent upon the change in historic flow and salt over time, the diverted flow, and the evapotranspiration from agricultural lands. These techniques computed different estimates for agricultural salt loading. Malone incorporated the difference between the two techniques as an error term on the agricultural salt loading. Therefore, the natural salt loading included natural salt loading, unknown diffuse source salt loading, and any measurement error. The natural flow was calculated by removing human development (including reservoir regulation and consumptive use) from the historic flow.

Mueller and Osen (1988) proposed a different technique that avoids the need to first estimate the human-induced salt loading. They developed a multiple linear regression to fit historic salt dependent on historic flow and several development variables, including reservoir regulation, consumptive use, exports, and irrigated acres. The development values were then set to zero and the natural flow was substituted for the historic flow, arriving at a relationship between natural salt and natural flow. The natural flow was calculated by removing flows resulting from human development, including reservoir regulation, consumptive use, and exports. This technique was applied to 20 gauges in the Upper Colorado River Basin and is the technique currently used by the BOR to estimate natural salt from natural flow. The technique does not include the information available from the residuals of the regression. Rather, the residuals are equalized and assumed random normal noise.

Recent modeling studies of the Colorado River system have exhibited systematic overprediction of salinity that is likely (possibly) caused by overprediction of natural salt. This paper presents a new natural salt model developed using nonparametric techniques to capture both the observed linear and nonlinear relationships between natural flow and natural salt. The addition of a residual resampling technique incorporates the information available from the residuals of the regression, adding the variance around the regression into the new salt model's results. Unlike the technique used by Wurbs et al., our residual resampling technique does not need to assume a distribution for the residuals. We followed the approach of Malone and determined an estimate of human-induced salt loading. The estimate of human-induced salt loading was used to estimate natural salt loading.

We first provide background information about the importance of modeling salinity in the Colorado River basin and present a description of the salinity sources and remediation methods. Next, we further discuss the existing modeling efforts for estimating natural salt loadings. We then develop the statistical nonparametric model for estimating natural salt and demonstrate its application to data from the stream gauge at Glenwood Springs, Colo., on the Colorado River. We also compare this new salt model to the model proposed by Mueller and Osen and show that the modeled overprediction is reduced by a significant amount (approximately 14%). We conclude with a brief discussion and recommendations for future work.

Background on Salinity in the Colorado River Basin

The salinity of the Colorado River became an important issue when the Mexican government strongly objected to the quality of the water Mexico was receiving in 1962. The average annual salinity of water delivered to Mexico in that year was 1,500 mg/L (Nathanson 1978). Such high salt concentration made the water unsuitable for irrigation, municipal, and industrial water uses.

In response to Mexico's concerns and after years of negotiations, Minute No. 242 of the International Boundary and Water Commission dated August 30, 1973, was signed. Minute No. 242 stipulates that water delivered to Mexico must have an average flow-weighted salinity of no more than 115 mg/L ± 30 mg/L above the average annual salinity at Imperial Dam. Subsequently, the Colorado River Basin Salinity Control Act of 1974 was enacted to ensure that the United States could meet its obligation to Mexico under Minute No. 242.

Minute No. 242 sets a variable salinity standard for the Mexico delivery, but does not set numerical water quality critera at any fixed points in the basin. Numerical criteria resulted from separate U.S. legislation that set policy regarding water quality. The Federal Water Pollution Control Act Amendments of 1972 required the development of fixed point numerical critera for salinity in the Colorado River Basin. The fixed point numeric critera were set in 1975: 723 mg/L below Hoover Dam; 747 mg/L below Parker Dam; and 879 mg/L at Imperial Dam.

These numeric salinity criteria were developed from the 1972 average annual salinity concentrations at each location and are currently unchanged (Lee 1989; U.S. Department of the Interior 2001). To predict these flow-weighted average total dissolved solids concentrations at all locations a computer simulation model is utilized that models the impacts of further human development on total dissolved solids concentration. An important step toward an informative computer simulation model includes understanding the sources of salinity and incorporating the sources in the simulation model.

Salinity Sources

Natural and human-induced salinity results from point and nonpoint sources. Natural point sources that have been identified include seeps and saline springs. Some springs originate from deep geological formations containing brackish water. Natural nonpoint sources of salinity generally originate from the weathering and dissolution of underlying rocks or soils overlaying the rocks.

Human-induced salinity predominantly results from irrigated agriculture. Agriculture increases salinity concentration through two processes: (1) Salt concentration and (2) salt loading. The salt concentration process is a result of evapotranspiration from crops, which consume water but leave salts behind in the soil. Return flows to the river from the diversion typically contain the same salt mass present in the diversion water but with less water,

hence, higher concentration of salt. Additionally, reservoirs concentrate salt by evaporation, when water is lost from the reservoir but salt is conserved.

Salt loading occurs when water transported through soil leaches salts present in the soil and transports them to the river. The water can be introduced into the soil from human-induced sources, such as irrigation practices, or from natural sources, such as precipitation. Irrigation practices increase the flow through soils, which increases the total salt loading from previous natural salt loading levels.

Limited data are available describing agricultural salt loading, termed *salinity pickup*, throughout the Colorado River basin. One extensive study (BOR 1983) explains how salinity pickup was calculated in the Grand Valley using a mass balance of salt averages over 1952 to 1980. The report states that the human-induced salinity pickup for the Grand Valley averages 526,000 metric tons per year±82 metric tons with 95% confidence. The report suggests that variations in salinity pickup cannot be due to changes in irrigation practices because variations in practices could not account for the magnitude of annual variations the data showed. This report indicated that measuring annual variations in salinity pickup from agriculture is extremely difficult. Therefore, for lack of a better assumption, we assume that agricultural salinity pickup is a constant mass for long-term modeling. This assumption is consistent with agricultural consumptive use, which has generally been constant since 1941 in the upper Colorado River above Glenwood Springs, Colo..

In addition to agricultural salt loading, we need estimates of natural salt loading. Natural salt loading contributes an estimated 47% of the total salinity in the Colorado River basin (U.S. Department of the Interior 2001). Natural flow is calculated by removing the human-induced effects on flow from observed historic flow. Human-induced effects include agricultural consumptive use, exports, and reservoir regulation, all of which are measured or can be estimated. Natural salt can be calculated by removing the human-induced effects on salt from observed historic salt.

Description of Existing Methods for Estimating Natural Salt Loading

As stated in the Introduction, the United States Geological Survey (USGS) developed the technique currently used by the BOR on the Colorado River to estimate natural salt entering the river using historic (observed) flow and salt data from 1941 to 1983 (Mueller and Osen 1988). Fig. 1 shows a typical sequence for the existing simulation model. Natural flow data is input into the model and a regression-based natural salt model provides estimates of natural salt. The historical salt is estimated by adding the salinity picked up by agriculture and subtracting the salt that leaves with water exported from the basin. When applying this model for generating future scenarios, the natural flows have to be generated from a stochastic model (the BOR uses the index sequential method for this purpose) and estimates of salinity from agriculture pickup must be provided. Recently, Prairie (2002) developed a stochastic nonparametric natural salt model in an effort to improve upon the current approach.

The existing simulation model, as described above, when applied to data from the stream gauge 09072500 (Colorado River near Glenwood Springs, Colo.), overpredicts the annual historical salt mass 1941 to 1995, by an average 2,502 kilograms/second (Fig. 2). The overprediction could result from salinity pickup of agriculture being too high, and/or natural salt loading being too high.

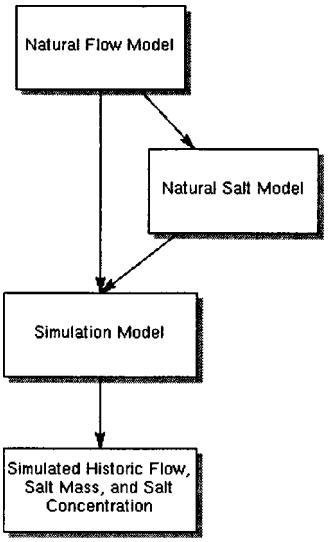

Fig. 1. Flowchart depicting interconnection of existing simulation model

From 1941 to 1995, the historic salt mass in the river passing gauge 09072500 averaged 13,800 kg/s. The relationship proposed by the USGS estimates an average annual natural salt of 16,300 kg/s. For the simulation model to simulate the historic salt mass, the human-induced salinity pickup sources would need to *remove* salt from the river. Current estimates, as reflected in the simulation model, are that human-induced sources contribute 4,000 kg/s from agricultural salinity pickup and exports remove an average 1,300 kg/s. The estimate for salinity pickup by agriculture is developed from an extensive study that quantified estimates of natural and human-induced salt (Iorns et al. 1965). The report estimates that, in 1957, natural sources contributed 15,060 kg/s, and human-induced sources contributed 4,051 kg/s from agricultural salinity pickup and removed 463 kg/s by exports above Glenwood Springs. These values were adjusted for current basin conditions then input in the simulation model.

Using these numbers, if human-induced sources contributed no salt above gauge 09072500, the existing simulation model would still overpredict salt mass. Iorns et al. (1965) indicate that the human-induced sources of salinity are not removing salt, but

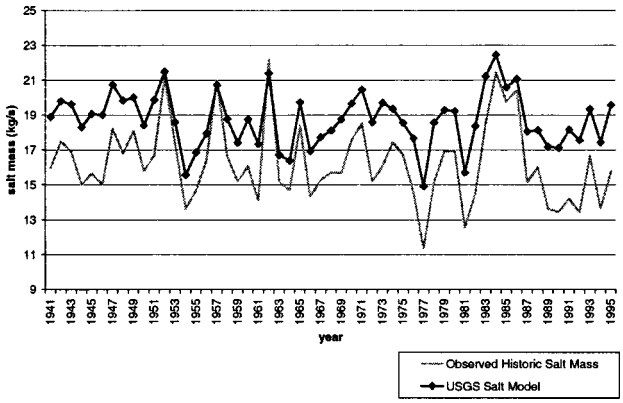

Fig. 2. Observed historical salt and estimates from the United States Geological Survey (USGS) relationship

are adding significant amounts. These findings point to an overestimation of natural salt by the USGS model. Further research intends to extend this analysis to additional stream gauges throughout the Upper Colorado River basin. In this study, we are setting up the analysis framework that will be extended.

Vaill (1999) performed trend analysis throughout the Upper Colorado River basin including gauge 09072500. The report found significant downward trends for flow-adjusted dissolved-solids concentrations and salt load for the period 1986–1993. Butler (1996) also performed a trend analysis on select gauges in the Upper Colorado River basin. This second report found decreasing salinity for a given flow at gauge 09095500: Colorado River at Cameo, Colo. This is the next major gauge downstream of 09072500. The USGS regressions were developed with flow and salinity values from 1941 to 1983; they no longer reflect current trends in the relationship. The recent trend analysis studies support the need to update or replace the USGS regressions to reflect current trends in the flow and salinity relationship.

To refine the USGS model would require a reanalysis of the detailed data on which the regressions were based. Unfortunately, these data are not available. To improve upon the USGS model, we propose a statistical nonparametric model that will relate natural streamflow to natural salt. We also present a technique to provide uncertainty estimates.

Nonparametric Model for Estimating Natural Salt Loading

Nonparametric methods estimate functions locally, in that the estimate of the function at any point is based on a small number of neighboring points (this point will become clear in the following section when we describe the proposed model). As a result, outliers do not exert undue influence on the overall fit, unlike parametric methods (e.g., linear regression or fitting probability density functions). This provides the ability to capture any arbitrary underlying functional form and local features present in the data.

Nonparametric (i.e., local) methods are more computationally intensive than their parametric counterparts. However, with increasing computational power readily available, nonparametric techniques provide an attractive alternative. Kernel-based nonparametric techniques have been successfully applied to a variety of hydrologic problems—rainfall modeling (Lall et al. 1996); flood frequency (Lall et al. l993, Moon and Lall 1994); streamflow simulation (Sharma et al. 1997; Tarboton et al. 1998); groundwater applications (Adamoski and Feluch 1991), and streamflow forecasting (Smith 1991). More recently, *K*-nearest-neighbor methods have been developed to improve upon the kernel-based techniques and have been applied for streamflow simulation (Lall and Sharma 1996; Prairie 2002); and daily weather generation (Rajagopalan and Lall 1999; Yates et al. 2003). The reader is referred to Lall (1995) for an overview of nonparametric techniques methods and their hydrologic applications.

Here, we use the nonparametric regression based on local polynomials to model natural salt from natural flows. A *K*-nearest neighbor (*K*-NN) bootstrap technique is developed to provide uncertainty estimates. These are described below.

Local Polynomials

Functional fitting problems, such as the one in this case (i.e., natural streamflow and natural salt) involve recovering the underlying relationship between a dependent variable (y) and a set of independent variables (x). In reality, the dependent variable that is observed has noise (or error) in it, which makes the function estimation more challenging. The problem reduces to estimating the function, f, in the model below

$$y = f(x)\beta + e_t \tag{1}$$

where, e_t = error term; β = vector of model parameters.

Typically, β is estimated as the minimizer of the least-squares function over all the data points

$$\min_{\beta}(y_i - X\beta)^T(y - X\beta) \tag{2}$$

Furthermore, parametric techniques fit an equation (linear or nonlinear for f) for the entire data, which restricts the ability to capture non-linearity in the data, as will be seen below. In addition, hypothesis testing (e.g., testing the goodness of fit of the model, the parameters, etc.) requires a Gaussian assumption of the error term and consequently, the data, which further restrict the model (Helsel and Hirsch 1992). If the fitted model (f) does not pass the hypotheses tests, then a different model is assumed and the fitting process [Eq. (2), above] repeated. As can be seen, the model complexity is limited by the sample size, thereby restricting the capability to capture nonlinear features in small samples.

Nonparametric methods, on the other hand, fit the function f locally and make no prior assumption about the functional form, i.e., linear, quadratic. Thereby, providing the capability to capture any arbitrary relationship. Several nonparametric methods exist: Kernel based, splines, and local polynomials. For a detailed description of these methods and comparisons, see Owosina (1992) and references therein. We adopted a local polynomial scheme that has been shown to be easy to implement and effective (Rajagopalan and Lall 1998; Loader 1999). The method and the algorithm are described through the following example (see Fig. 3). We generated a synthetic dataset from a sine wave function with noise added (the noise is normally distributed with mean 0 and variance of 0.1) to it. Traditional linear regression and quadratic fits, as can be seen from Fig. 3, are unable to capture the true underlying sinusoidal function. A very high-order polynomial (as sine function is a higher-order polynomial) will capture the underlying function, but given the small sample size a higher-order fit is not feasible. In the nonparametric approach, the underlying function is evaluated "locally" in that the estimate at any point is obtained by fitting a polynomial to a small number of its neigh-

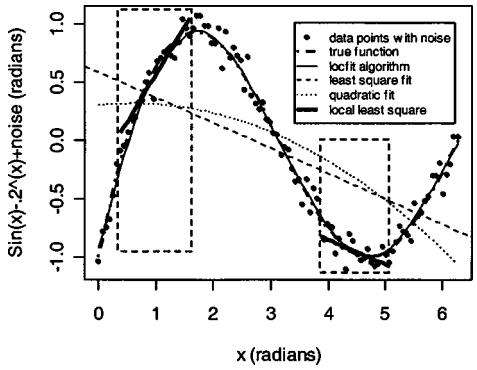

Fig. 3. Several data fitting techniques for data generated from a sine function with noise. The local linear polynomials had an alpha of 0.2. The data is generated via $y_t = \sin(x_t) - 0.2^{x_t} + e_t$ with e_t being the noise from a normal distribution with mean 0 and variance 0.2.

bors. The main parameters then are the order of the polynomial and the size of the neighborhood. Estimation of these parameters is described in the algorithm below. The local polynomial fit to the sine wave data (with a neighborhood size of 20 data points and local linear polynomials) is show in Fig. 3 as the solid line—and it can be seen that this captures the true underlying function very well. One obvious benefit is that outliers or extreme values do not influence the overall fit, as they do in a parametric approach. Furthermore, the local fitting provides the capability to capture any arbitrary features that might be present. The local polynomial algorithm is presented, with reference to Fig. 3, as follows:

1. Let us assume that we want to estimate the function at x_t.
2. A neighborhood is defined around x_t. The size of the neighborhood is $(K=\alpha\times n)$, where α is a smoothing parameter between 0 and 1. Bigger α indicates more smoothing. For example, for $\alpha=1$ and a local linear fit; it is the same as the parametric linear regression. n is the sample size.
3. The neighbors are weighted as per their distance to x_t—in that the nearest neighbor gets the highest weight and the farthest neighbor gets the least weight. The weights can be obtained in many different ways (e.g., using the inverse distance with a smoothing function). The weights form the elements of the diagonal matrix W.
4. For the neighbors captured in the neighborhood (shown in the dashed rectangles), a regression of order p is fit. Typically, a linear fit works very well (shown as the heavy solid line within the neighborhood).
5. The regression is fit using a weighted least squares, i.e., $\min_\beta(y-X\beta)^T W(y-X\beta)$, over all the K-nearest neighbors.
6. The fitted regression is then used to estimate at x_t.
7. This is repeated at all points where we need the estimate.

As mentioned before, the "local fitting" of the regressions provides a great flexibility in modeling any structure that might be present in the data (linear and nonlinear). The neighborhood size (α) provides the amount of smoothing and hence, the flexibility. When the neighborhood size is the same as the number of data points and the fit is linear, we reproduce the traditional linear regression. For a nonlinear function, we would expect a smaller neighborhood size.

The model parameters (i.e., α and p) can be estimated by minimizing an objective function such as the cross-validation (CV) function:

$$\mathrm{CV}(\alpha,p)=\frac{1}{n}\sum_{i=1}^{n}(Y_i-y_{-i})^2 \tag{3}$$

where y_{-i}=estimate at x_i by dropping x_i, y_i from the fit. The CV function above is computed for several choices of α and p. The choice that minimizes this function is selected. When a dataset is small, Loader (1999) recommends using CV; therefore, we used this technique to find α and p in our applications here.

The generalized CV (GCV) function, on the other hand, is a good estimate of the predictive capability of the model when the dataset is larger (Craven and Whaba 1979). Furthermore, it obviates the need for dropping an observation as in the case of CV, thereby, saving computation time. The GCV function is given as

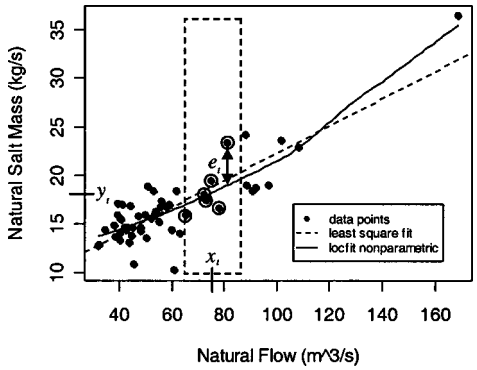

Fig. 4. Scatterplot of natural salt and natural flow for the month of April, along with the linear regression fit and the local polynomial (locfit) fit

$$\mathrm{GCV}(\alpha,\rho)=n\frac{\sum_{i=1}^{n}(Y_i-y_i)^2}{\left(1-\sum_{i-1}^{n}h_{ii}\right)^2} \tag{4}$$

where n=sample size, Y_i-y_i is the residual; and h_{ii}=diagonal terms of the hat matrix H.

The diagonal of hat matrix, termed the influence matrix, explains the weight of a data point on the estimate at that point. The hat matrix is found in matrix algebra as $X(X'X)^{-1}X'$ (Eubank 1999). In the GCV function above, the numerator represents the mean-square error (MSE) while the denominator represents a penalty term that penalizes for increasing the model complexity (i.e., model parameters which depend upon the order of the polynomial, p).

Quantifying Uncertainty

It is important to quantify the uncertainty (i.e., confidence intervals) of the estimates from the nonparametric salt model. In the case of linear regression the uncertainty estimates are obtained by assuming the errors to be normally distributed (Helsel and Hirsch 1992). Here, we developed a K-NN residual resampling technique to quantify the uncertainty of the estimates from the local regression method. In this method, we resample (or bootstrap) residuals within a neighborhood of the point of estimate and add them to the mean estimate from the local regression. This is described with the help of Fig. 4. Let us suppose that the natural flow is x_t. We find the corresponding mean natural salt mass estimate, y_t from the local polynomial regression. Next, we find K-nearest neighbors (within the dashed rectangle box) to x_t shown as a circle around a data point. We resample one of the residuals using a weight function that gives larger weight to the nearest neighbor and smallest weight to the farthest neighbor. Let us say that we picked e_t. This is added to the mean estimate y_t to get a simulated value $y_t^*=y_t+e_t$. We repeat this process several times to obtain an ensemble of natural salt estimates at x_t. The 5th percentile of the ensemble provides the 5% confidence interval and so on. The key point here is that, by resampling residuals locally, non-Gaussian features that might be present in the data can be captured by way of asymmetric confidence intervals, unlike traditional methods, that provide only symmetric intervals assuming Gaussian distribution.

Application of Model

We applied the nonparametric model to natural streamflow and natural salt data from the gauge at Glenwood Springs, Colo. Natural salt mass is "backcalculated" from the observed historic salt mass and salt load data from the simulation model as

natural salt = observed historic salt
+ salt with water exported out of the basin
− salinity pickup from agriculture
(values based on simulation model).

The salt removed by exports and the salt added by agriculture for the period 1941 to 1995 were taken from the data used to drive the simulation model. In the simulation model, agriculture annually adds 124,000 metric tons of salt above gauge 09072500. As stated previously, a constant salinity pickup is a fair assumption, because agricultural consumptive use was basically constant 1941 to 1995. To determine the monthly salt added by agriculture the annual tons were distributed to monthly values as a function of each month's percent of annual return flow. For example, if June 1943 generated 86% of the annual return flow in water year 1943, then in June 1943 the monthly salt added from agriculture would be 124,000 metric tons times 86% or 106,640 metric tons. The exports remove a constant concentration of 100 mg/L. The tons removed by exports vary with flow, according to the relationship between flow and salt mass. The natural flows are the observed historic flows minus the total human-induced consumptive use.

Local regressions were developed separately for each month, just like the existing USGS model. We applied the residual resampling technique to obtain the 5 and 95% confidence levels.

Evaluation Criteria

We first compared the performance of the local polynomial method and the traditional linear regression on a synthetic data set. We then compared results from the nonparametric model to those from the USGS model for the years 1941 to 1995. We evaluated each model's performance on a monthly and annual time scale. The annual time series of flow and salt are obtained by summing the water year months, October through September. We also, performed a blind forecast of annual salt for the last 5 years.

First, we compared the regressions developed from the two models. Second, we used both models to estimate the natural salt for the natural flows during 1941 to 1995 and compared their performance. We also calculated the estimated historic salt from the estimated natural salt mass obtained from the models and compared them against the observed historical salt mass. The estimated historic salt mass is obtained as

estimated historic salt = estimated natural salt − salt from exports
+ salt from agricultural salinity pickup

In addition to visual comparisons, we also provided quantitative estimates such as standardized RMSE (RMSE)—both fitting and cross-validated. The fitting RMSE is computed from the model residuals with respect to the true function (in the case of the synthetic data set) and the natural salt value (for the real data). For the cross-validated case, a point is dropped from the data and the remaining points are used to fit to the model, which is then used to predict the value at the dropped point and consequently, the cross-validated residuals and RMSE. The cross-validated RMSE provides a measure of the predictive capability of the model.

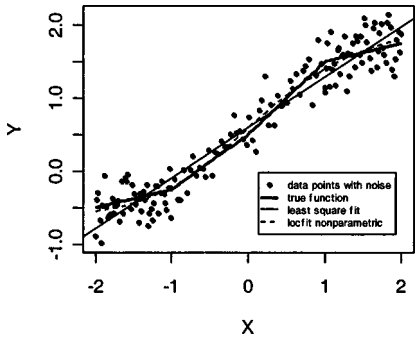

Fig. 5. Scatterplot of synthetic data set with mild nonlinearity, along with local polynomial and United States Geological Survey regression fits

Results

We discuss results from synthetic data first, followed by the results from the real application.

Synthetic Data

The first synthetic data set is the sine wave data that was described in the earlier section. As can be seen from that figure (Fig. 3) the local polynomial captures the underlying function very well and it is practically indistinguishable from the true function. Furthermore, the cross-validated RMSE with respect to the true value of the function from the local polynomial is 0.033, which is significantly lower than the parametric alternatives.

We then generated a synthetic data set with mild nonlinearity at the extremes buried in noise (Fig. 5). The generated data are shown as dots in the figure. The linear regression fits the data quite well and is statistically significant (with a p value of close to 0 on the F test). However, the linear regression is unable to capture the mild nonlinearity at the ends. On the other hand, the local polynomial captures the true function very well. The fitting RMSE with respect to the true function is 0.047 and 0.129, respectively, for the local polynomial and linear regression. The cross-validated RMSE with respect to the true value of the function is 0.049 and 0.130, respectively, for the local polynomial and linear regression.

Owosina (1992) and Loader (1999) compared nonparametric regression methods in general, and local polynomials in particular, to a wide range of synthetic data sets on a variety of measures (RMSE, bias, etc.) and they find that the nonparametric approaches perform extremely well.

Data from Glenwood Springs, Colorado

Table 1 shows the fitting and cross-validated RMSE of the local polynomial method and linear regression for the monthly and the annual regressions. It can be seen that the two methods exhibit comparable RMSE suggestive of a linear relationship for the most part, with local polynomial providing a lower fitting RMSE. In all

Table 1. Fitting and Cross-Validated Root-Mean-Square Error (RMSE)

	Fitting RMSE		X-val RMSE	
	LOCFIT	LSFIT	LOCFIT	LSFIT
Jan	0.792	0.794	0.847	0.827
Feb	0.733	0.739	0.784	0.771
Mar	0.754	0.756	0.784	0.778
Apr	0.698	0.700	0.737	0.729
May	0.601	0.609	0.640	0.636
Jun	0.582	0.588	0.622	0.617
Jul	0.421	0.461	0.500	0.518
Aug	0.630	0.635	0.670	0.665
Sep	0.571	0.571	0.600	0.592
Oct	0.586	0.588	0.616	0.609
Nov	0.757	0.762	0.802	0.790
Dec	0.720	0.723	0.791	0.764
Annual	0.401	0.411	0.430	0.432

Note: X-val=cross validated; LOCFIT=local polynomial method; and LSFIT=least square linear regression method.

months the alpha (i.e., the size of the neighborhood) was between 0.9 and 1 further indication, that the relationships are generally linear.

Fig. 6 shows the local polynomial fit and the existing USGS salt model fit for April and June—the circles show the data points. It can be seen that the local polynomial fit does a better job of capturing the relationship indicated by the scatter in the data. It can also be seen that the USGS relationship underestimates the salt in April and overestimates the salt in June. Furthermore, the wider scatter of the data points indicates there is significant variability around the relationship. The USGS relationship provides no estimate of the variability i.e., error estimates of the relationship.

We applied these two models to natural flow data from 1941 to 1995 and estimated the natural salt. Fig. 7 (upper graph) shows the estimated natural salt from these two models. The estimated natural salts were generated at a monthly time step (using the monthly relationships) and then summed to obtain the annual values. The estimates from the USGS model are higher than those from the nonparametric model. The lower graph replaces the nonparametric salt model line with the annual natural salt. The annual natural salt is captured between the 5 and 95% confidence. The USGS model, as expected overestimates the salt mass by 15%, or 78,000 metric tons, greater than the annual average observed historic salt mass. Meanwhile, the nonparametric salt model reduced the difference between the annual average observed historic salt mass and the estimated historic salt mass to 0.8%, or 3,600 metric tons.

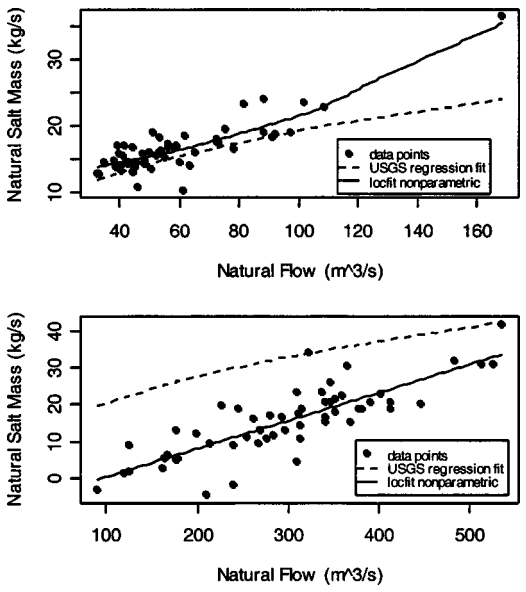

Fig. 6. Same as Fig. 4 along with the fit from the United States Geological Survey (USGS) salt model

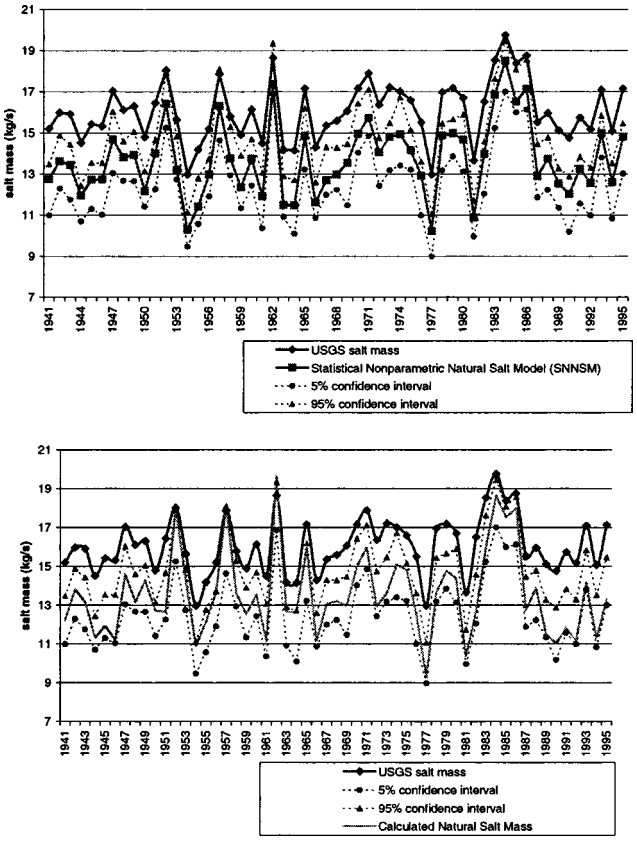

Fig. 7. The upper graph shows the natural salt estimates from the United States Geological Survey (USGS) relationship and the nonparametric regression. The lower graph shows the annual natural salt and the corresponding estimates from the USGS relationship. The confidence intervals are obtained from the *K*-NN residual resampling technique.

We computed the confidence levels of the estimates from the nonparametric model, using the residual resampling technique. The 5 and 95% confidence levels were computed as described earlier and are plotted along with the estimates from the nonparametric model and the USGS model. This demonstrates that the estimates from the USGS model fall outside the 95% confidence levels of the nonparametric model suggesting that the estimates from these two models are significantly different. Furthermore, the confidence intervals are asymmetric, unlike the confidence intervals one would obtain from parametric models; this suggests that the assumption of normal distribution of the errors is not quite valid.

Finally, we performed blind forecast of salt for the last 5 years using data prior to 1991. The RMSE values are 0.56 and 0.58, respectively, for local polynomial and linear regression methods.

Here too, the local polynomial and linear regression show similar performance with the nonparametric method showing a slightly lower RMSE.

Summary and Conclusions

We outlined a technique to calculate the natural salt based on the observed historic salt mass. We then developed a nonparametric regression method using local polynomials to obtain relationships between natural salt and natural flow. Further, we incorporated a residual resampling technique in the nonparametric model to enable the quantification of uncertainty in the estimates. We showed that this approach can generate realistic ensembles of salinity and also seems to improve upon the USGS salt model.

It is evident that some of the variability captured with the residual resampling technique could be attributed to data uncertainty and not natural variability. To allow the development of the regression framework presented, we used the "best" available data for a single stream gauge. An advantage of the regression framework presented in this paper is that as data uncertainty is addressed the revised data can easily be applied to develop updated regressions; this was not an option with the USGS salt model. As with all regressions, the writers recognize regression relationships are only as good as the underlying data, therefore, information interpolated from the regressions should be viewed accordingly.

Data uncertainty is an important issue with this work because natural flow and salinity are not directly measured. Current efforts to address this issue include improving methods to compute natural flow and recomputing natural flow with the improved methods. Further, new research intends to improve modeling the salt load attributed to agriculture. This is an extremely difficult value to estimate as shown by the study measuring salt loading in the Grand Valley (BOR 1983). The new research intends to develop better methods to model agricultural salt loading in order to reduce the uncertainty of this value.

Nonparametric models, like parametric methods, are not without drawbacks. Short and poor quality datasets can make the estimation of the neighborhood size (i.e., alpha) in the local polynomial method difficult and also increase the variance of the estimates. Often times, the GCV function might not provide a clear minimum (especially in short data sets) and, in such cases, the alpha is chosen subjectively by looking at the estimated fit. Extrapolating values too far out from the data set can result in large variance.

The flexibility of the nonparametric approach allows it to be portable across various sites. This is a very useful feature for agencies such as BOR that like to prescribe a uniform method across sites without having to worry about model fitting, parameter estimation, and hypothesis estimation issues. We intend to extend the nonparametric natural salt model framework to the remaining twenty streamflow gauges throughout the Upper Colorado River basin. Preliminary results from this effort are encouraging and corroborate the findings reported in this paper. Additional work includes exmaining the salt mass as a function of natural flow relationship for different time periods.

Acknowledgments

The BOR funded this work. The writers would like to thank Dave Trueman for his strong support and encouragement. Valuable comments from three anonymous reviewers in improving the manuscript are thankfully acknowledged.

Notation

The following symbols are used in this paper:

- e_t = error term at time t;
- H = hat matrix;
- h_{ii} = diagonal terms of the hat matrix;
- h_{ij} = any individual term of the hat matrix;
- i = index term;
- j = index term;
- K = number of neighbors;
- n = sample size;
- p = order of the polynomial;
- t = time index;
- W = weight function;
- X = matrix of the independent variable x;
- x = independent variable;
- y = dependent variable;
- y^* = dependent variable plus an error term;
- α = smoothing parameter;
- β = vector of model parameters; and
- $\hat{\mu}$ = estimate of the mean.

References

Adamoski, K., and Feluch, W. (1991). "Application of nonparametric regression to groundwater level prediction." *Can. J. Civ. Eng.*, 18, 600–606.

Brown, R. D. (1984). "Economic evaluation of salinity control." *Salinity in watercourses and reservoirs*, R. H. French, ed., Butterworth, Boston 125–134.

Bureau of Reclamation (BOR). (1983). "Draft report—Grand Valley salt pickup calculations." United States Department of the Interior, Grand Junction, Colo.

Bureau of Reclamation (BOR). (1987). "Colorado River simulation system, system overview." United States Department of the Interior, Denver.

Butler, D. L. (1996). "Trend analysis of selected water-quality data associated with salinity-control projects in the Grand Valley, in the Lower Gunnison River basin, and at Meeker Dome, western Colorado." United States Geological Survey, Water Resources Investigation *Rep. No. 95-4274*, Denver.

Craven, P., and Wahba, G. (1979). "Smoothing noisy data with spline functions." *Numer. Math.*, 31, 377–403.

Eubank, R. (1999). *Spline smoothing and nonparametric regression*. Marcel Dekker, New York.

Helsel, D. R., and Hirsch, R. M. (1992). *Statistical methods in water resources*. Elsevier, Reston, Va.

Huizenga, L. J. (1980). "The River Rhine as an example of international environmental policy in Europe." *Water Services*, 84, 1013–1018.

Irons, W. V., Hembree, C. H., and Oakland, G. L. (1965). "Water resources of the upper Colorado River basin—technical report." United States Geological Survey, Professional Paper, United States Government Printing Office, Washington, D.C., 441.

Lall, U. (1995). "Recent advances in nonparametric function estimation: hydraulic applications." *U.S. Natl. Rep. Int. Union Geod. Geophys. 1991–1994, Rev. Geophys.*, 33, 1093–1102.

Lall, U., Rajagopalan, B., and Tarboton, D. G. (1996). "A nonparametric wet/dry spell model for resampling daily precipitation." *Water Resour. Res.*, 32(9), 2803–2823.

Lall, U., and Sharma, A. (1996). "A nearest neighbor bootstrap for resa-

mpling hydrologic time series." *Water Resour. Res.*, 32(3), 679–693.
Lall, U., Moon, Y.-I., and Bosworth, K. (1993). "Kernel flood frequency estimators: Bandwidth selection and kernel choice." *Water Resour. Res.* 29(4), 1003–1015.
Lee, D. (1989). "Salinity in the Colorado River basin: A dynamic modeling approach to policy analysis." PhD thesis, Univ. of California at Davis, Davis, Calif.
Lee, D., Howitt, R., and Marino, M. (1993). "A stochastic model of river water quality: application to salinity in the Colorado River." *Water Resour. Res.*, 29(12), 3917–3923.
Loader, C. (1999). *Local regression and likelihood*. Springer, New York.
Malone, R. F., Bowles, D. S., Grenney, W. J., and Windham, M. P. (1979). "Stochastic analysis of water quality." Utah Water Research Laboratory, Utah State Univ., Logan, Utah.
Moon, Y.-I., and Lall, U. (1994). "Kernel function estimator for flood frequency analysis." *Water Resour. Res.*, 30(11), 3095–3103.
Mueller, D. K., and Osen, L. L. (1988). "Estimation of natural dissolved-solids for the upper Colorado River basin." *Water Resources Investigation Rep. No. 87–4069*, United States Geological Survey, Denver.
Nathanson, M. N. (1978). "Updating the Hoover Dam documents." Bureau of Reclamation, United States Department of the Interior, Denver.
Olson, K. W. (1980). "Economics of controlling natural salt sources in the Arkansas River Basin." *Water Resour. Bull.*, 16(2), 295–299.
Owosina, A. (1992). "Methods for assessing the space and time variability of ground water data." MS thesis, Utah State Univ., Logan, Utah.
Prairie, J. R. (2002). "Long-term salinity prediction with uncertainty analysis: Application for Colorado River above Glenwood Springs, CO." MS thesis, Univ. of Colorado, Boulder, Colo.
Rajagopalan, B., and Lall, U. (1998). "Locally weighted polynomial estimation of spatial precipitation." *J. Geograph. Inf. Decision Anal.*, 2(3), 48–57.
Rajagopalan, B., and Lall, U. (1999). "A *K*-nearest-neighbor simulator for daily precipitation and other weather variables." *Water Resour. Res.*, 35(10), 3089–3101.
Saleh, I. (1993). "Synthesis of streamflow and salt loads." MS thesis, Texas A&M University, College Station, Tex.
Sharma, A., Tarboton, D. G., and Lall, U. (1997). "Streamflow simulation: A nonparametric approach." *Water Resour. Res.*, 33(2), 291–308.
Smith, J. A. (1991). "Long-range streamflow forecasting using nonparametric regression." *Water Resour. Res.*, 27(1), 39–46.
Tarboton, D. G., Sharma, A., and Lall, U. (1998). "Disaggregation procedures for stochastic hydrology based on nonparametric density estimation." *Water Resour. Res.*, 34(1), 107–119.
U.S. Department of the Interior. (2001). "Quality of water Colorado River basin: Progress report 20." ⟨http://www.uc.usbr.gov/progact/salinity/index.html⟩ (October 15, 2001).
Vaill, J. E., and Butler, D. L. (1999). "Streamflow and dissolved-solids trends, through 1996, in the Colorado River basin upstream from Lake Powell—Colorado, Utah, and Wyoming." U.S. Geological Survey, Water Resources Investigation, *Rep. No. 99-4097*, Denver.
Wurbs, R. A., and Karma, A. S. (1995). "Salinity and water-supply reliability." *J. Water Resour. Plan. Manage.*, 121(5), 352–358.
Wurbs, R. A., Saleh, I., and Karma, A. S. (1995). "Reservoir system reliability constrained by salinity." *Water Resour. Dev.*, 11(3), 273–287.
Yates, D., Gangopadhyay, S., Rajagopalan, B., and Strzepek, K. (2003). "A technique for generating regional climate scenarios using a nearest neighbor bootstrap." *Water Resour. Res.*, 39(7), 1199.