INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **28**: 361–370 (2008)
Published online 1 August 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/joc.1533

ROYAL METEOROLOGICAL SOCIETY

Non-parametric short-term forecasts of the Great Salt Lake using atmospheric indices

Young-I1 Moon,[a]* Upmanu Lall[b] and Hyun-Han Kwon[c]

[a] *Professor, Civil Engineering Department, University of Seoul, Seoul, Korea*
[b] *Professor, Earth and Environmental Engineering Department, Columbia University, NY, USA*
[c] *Associate Research Scientist, Earth and Environmental Engineering Department, Columbia University, NY, USA*

ABSTRACT: A multivariate, non-parametric model for approximating the non-linear dynamics of hydroclimatic variables is developed and applied for forecasting the volume of the Great Salt Lake (GSL) of Utah. The monthly volume of the GSL is presumed to depend on recent volumes of the lake, and on three atmospheric circulation indices. The indices considered are the Southern oscillation index (SOI), the pacific/North America (PNA) climatic index, and the central North Pacific (CNP) climatic index. Locally weighted polynomials with automatically and locally chosen parameters are used for developing a non-linear forecasting model. Estimated average mutual information (M.I) is used to select appropriate lags across each time series. Iterated and direct multi-step predictions of lake volumes for up to 2 years in the future with and without the atmospheric indices are compared. The atmospheric circulation information can lead to significant improvements in the predictability of the lake. Copyright © 2007 Royal Meteorological Society

KEY WORDS non-parametric forecasting; non-linear dynamics; atmospheric indices; Great Salt Lake

Received 18 October 2006; Revised 23 February 2007; Accepted 4 March 2007

1. Introduction

Short-term forecasts of stream flow and lake volumes are made routinely by various methods and are used to guide the operation of water resource facilities. Recently (Yakowitz and Karlsson, 1987; Smith, 1991; Kember *et al*., 1993; Lall *et al*., 2006), non-parametric regression methods have been proposed for forecasting hydrologic time series. Lall *et al*. (2006) were able to forecast the volume of the Great Salt Lake (GSL) for up to 4 years in advance during extreme conditions. They formulated a forecasting model using recent techniques (Abarbanel *et al*., 1993) for reconstructing the dynamics of a non-linear system from a single observed state variable. Multivariate Adaptive Regression Splines (MARS) due to Friedman (1991) were used to non-parametrically recover the non-linear forecasting function from the time series of GSL volume. Such methods for time series analysis are computationally intensive, and can also require long high quality records. In this paper, we present (1) forecasts of the GSL volume using time series of the GSL volume and of three atmospheric circulation indices defined over the Pacific Ocean and (2) the application of multivariate, locally weighted polynomial regression with locally chosen parameters for non-parametrically approximating the dynamics of the system at each point of prediction.

Surface hydrologic systems are ultimately forced by atmospheric circulation. Hence, the use of information on large-scale circulation patterns is potentially useful for improving short- term forecasts. One can hope that including such information may reduce the length of data needed for reconstructing the dynamics, and thus broaden the applicability of the methods. However, atmospheric circulation indices may have a time scale of fluctuation that is considerably shorter than that for a large lake, complicating such a reconstruction.

Efficient parameter selection is important for non-parametric function approximation. The strategy provided here is capable of automatically selecting the size of the neighbourhood and the order of the polynomial used at each point of estimate. This allows one to represent linear (e.g. classical AR models) or polynomial dynamics, as well as locally approximating more complex dynamics. A measure of the expected prediction error at each time step is also developed.

Background information on recent analyzes of the GSL and connections of its fluctuations to large scale circulation is first provided. This motivates the forecasting algorithm which is presented next. Forecasts of the GSL volume with and without atmospheric circulation information follow. Actual forecast skill is compared with predicted skill.

2. The great salt lake and climate

The GSL of Utah is a closed lake in the lowest part (elevation 1280 m. above Mean Sea Level) of the Great

* Correspondence to: Professor Young-I1 Moon, Civil Engineering Department, University of Seoul, Seoul, Korea.
E-mail: ymoon@uos.ac.kr

Basin (latitudes 40°20′ and 41°40′N, longitudes 111°52′ and 113°06′W), in the arid western United States. Usually, closed lakes are in arid regions of the world where the long-term average evaporation rate exceeds the average precipitation. They integrate the basin's hydrologic response and represent it over a variety of time scales through their level, salinity, and sediments. The GSL is approximately 113 km long and 48 km wide, with a maximum depth of 13.1 *m* and an average depth of 5.0 *m*. The large surface area and shallow depth make the lake very sensitive to fluctuations in long-term climatic variability. Fluctuations of the GSL's level are of direct concern to mineral industries along the shore, the Salt Lake City airport, the Union Pacific Railroad, and Interstate 80. They are also well correlated with regional water supply conditions. As shown in Figure 1, the lake level has varied considerably over decadal time scales during the last 140 years.

Recent investigations (Sangoyomi, 1993; Mann *et al*., 1995; Lall and Mann, 1995; Abarbanel *et al*., 1996; Moon and Lall, 1996; Sangoyomi *et al*., 1996; Lall *et al*., 2006) provide evidence of quasi-periodic interannual and interdecadal variability in the GSL fluctuations, as well as the connection of these fluctuations to similar patterns in regional precipitation, temperature, streamflow and Northern Hemisphere sea level pressure (SLP) time series. The appropriateness of a non-linear dynamical representation of the GSL time series is also shown by Sangoyomi (1993). This predictability was related to the quasi-periodic low frequency variability in the lake and in hemispheric atmospheric circulation. As stated earlier, our interest here is in seeing if good forecasts of the GSL are possible with a shorter, more coarsely sampled time series of the GSL volume augmented by information on atmospheric circulation.

From the recent work of Klein and Bloom (1987), Kiladis and Diaz (1989), Cayan and Peterson (1989), Leathers *et al*. (1991), and Lins (1993), among numerous others, it is clear that atmospheric and oceanic conditions in the Pacific basin exert considerable influence on the low frequency patterns of North American climatic and hydrologic variability. Hirschboeck (1987) showed that there is linkage between anomalous circulation patterns and severe floods in the western, central, and eastern regions of the United States.

Lall and Mann (1995) found evidence of quasi-periodic behaviour in the monthly time series of the GSL, local streamflow, precipitation, and temperature, using the midnight mean temperature (MTM) (Thomson, 1982) and single scattered albedos (SSA) (Ghil and Vautard, 1991). Moon and Lall (1996) used SSA, M-SSA, and MTM to identify coherent quasi-periodic patterns in time in the GSL volume and three atmospheric circulation indices. The indices considered are the southern oscillation index

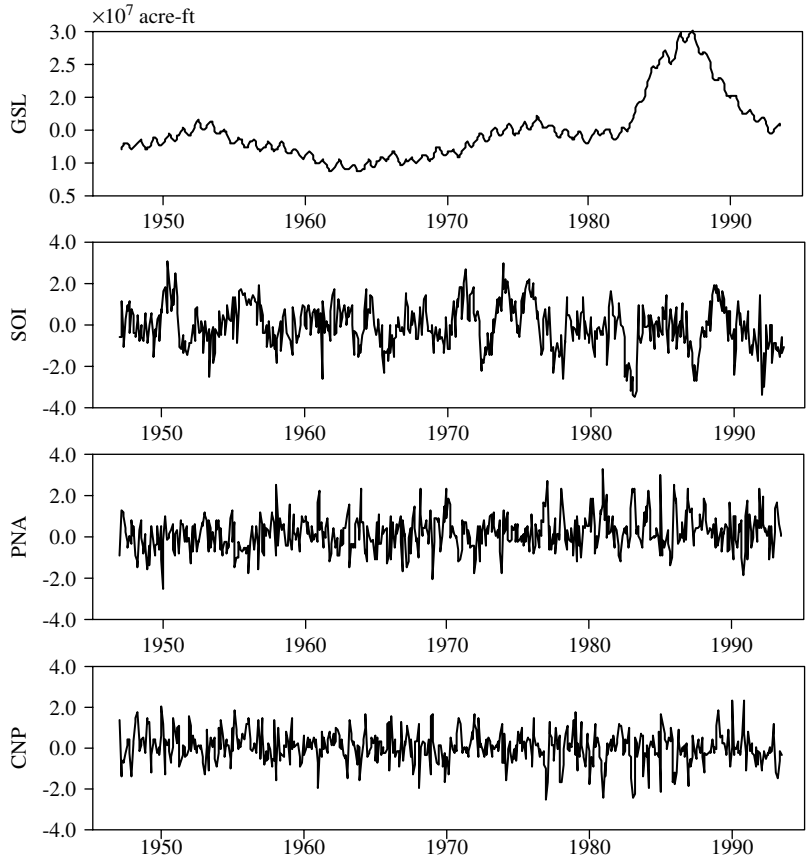

Figure 1. The monthly time series of GSL, SOI, PNA, and CNP. The GSL data were provided by the USGS and the SOI, PNA, CNP data by D. R. Cayan at Scripps Inst. of Oceanography. Note the apparently disparate time scales of GSL, SOI, and PNA/CNP fluctuations.

DOI: 10.1002/joc

(SOI), the pacific/North America (PNA) climatic index, and thecentral North Pacific (CNP) climatic index. The SOI reflects the influence of tropical Pacific ocean-atmospheric variability on climate, while the PNA and CNP reflect the North Pacific extratropical jet stream's influence on western United States climate.

The El Niño Southern Oscillation (ENSO) refers to an event in the tropical Pacific Ocean that is a significant perturbation of general atmospheric circulation. Western United States precipitation and stream flow are enhanced during an El Niño event and drought occurs over the continental United States with a La Nina event (Ropelewski and Halpert, 1987; Keppenne and Ghil, 1992; Kahya and Dracup, 1993).

The PNA was constructed (Horel and Wallace, 1981) as the monthly value of H(170 W,20 N) − H(165 W, 45 N) + H(115 W,58 N) − H(90 W,30 N), where H refers to the height of the 700 mb atmospheric pressure surface at a given location. Cayan and Peterson (1989) note that correlations between PNA and December–August mean streamflow anomalies reflect variations in the strength and position of the mean North Pacific storm track entering North America, as well as shifts in the trade winds over the subtropical North Pacific. Something resembling a reverse PNA pattern was shown to be associated with enhanced precipitation in the GSL region on a decadal time scale by Mann *et al*. (1995).

The CNP was constructed (Cayan and Peterson, 1989) by averaging the SLP (in mb) over the region 35N–55N and 170E–150W, for each month. This index is similar to the PNA index and reflects the circulation far a field over the CNP. SLP has been shown to be an adequate indicator of atmospheric circulation, especially over the extratropical oceans during winter (Emery and Hamilton, 1985) and is reasonably well correlated with precipitation over the west coast (Cayan and Roads, 1984).

The common time period spanned by these indices and the GSL volumes is 1946–1993. These time series are shown in Figure 1. The MTM (Thomson, 1982; Mann and Park, 1993) is applied to estimate the coherence between GSL volume change and atmospheric circulation data at selected frequencies. Figure 2 shows the coherence estimated from the MTM using 3, 2π tapers between the GSL volume change and atmospheric circulation data. There is significant coherence between GSL and SOI at 2.4, 2.6, 3.6, 4.4, 7.8, and 12.9 years. PNA and CNP have significant coherence at 2.3 and 3.9 years with GSL. This analysis lends confidence to the presumption that this atmospheric circulation information may improve GSL volume forecasts.

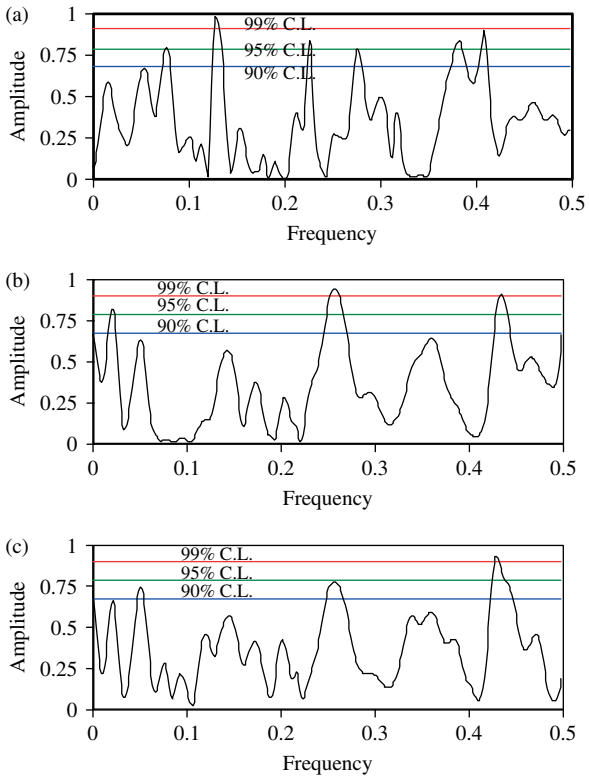

Figure 2. MTM coherence (solid) between GSL monthly volume change and (a) SOI, (b) PNA, and (c) CNP using 3, 2 π tapers, and data from December 1946 to July 1993. The dashed horizontal lines are the 90, 95, and 99% confidence limits (C.L.) for coherence amplitude. This figure is available in colour online at www.interscience.wiley.com/ijoc

3. The non-parametric forecasting model

We refer the reader to Abarbanel *et al*. (1993), Abarbanel *et al*. (1996), and Lall *et al*. (2006) for background information on the reconstruction of the state space of a dynamical system from a scalar time series of one of the state variables. Here, we shall consider directly the problem at hand, i.e. the short-term forecast of the future volume of the GSL given the history of the GSL and of the three atmospheric circulation indices. Generically, we consider that the m step ahead forecast G_{t+m} of the GSL volume can be obtained in terms of a set of current coordinates $\overline{V}_t$ in state space as

$$G_{t+m} = f(\overline{V}_t) + e_t \tag{1}$$

where $f(.)$ is an unknown linear or non-linear function, and e_t represents an 'error' term that may be a stochastic process.

There are two possibilities for implementing the forecasting strategy. The first strategy considers a direct m step ahead forecast from the current point. We will term this a *direct* forecast. The second considers m sequential 1-step-ahead forecasts, each using the prior forecasted values to generate an m-step-ahead forecast. We shall call this the *iterated* forecast. Both strategies are explored.

The forecasting model is specified through a selection of the elements of the coordinates of $\overline{V}_t$, assumptions regarding the behaviour of the error term e_t, and a procedure for recovering the forecasting function $f(.)$. These are discussed in this order in this section.

The state space coordinates $\overline{V}_t$ are defined as

$$\overline{V}_t = \{G_t, \cdots, G_{t-\tau(m_1-1)}, S_{t-\tau_1}, \cdots, S_{t-\tau_1-\tau(m_2-1)}, P_{t-\tau_2-\tau(m_3-1)}, C_{t-\tau_3}, \cdots, C_{t-\tau_3-\tau(m_4-1)}\} \tag{2}$$

DOI: 10.1002/joc

where τ is a time delay applied to the same series; τ_1 is a lag between GSL and SOI; τ_2 is a lag between GSL and PNA; τ_3 is a lag between GSL and CNP; S_t refers to an SOI value, P_t to a PNA value, and C_t to a CNP value; and m_1, m_2, m_3 and m_4 are embedding dimensions (i.e. the number of coordinates needed) for GSL and SOI, PNA, and CNP, respectively. Each series is 'standardized' by subtracting its mean and dividing by its standard deviation.

The lags, τ_1, τ_2 and τ_3, recognize that there may be some phase lag between the circulation indices and the GSL volume. Presuming that these indices represent causal factors, they are considered to 'lead' the GSL volume. This assumption is supported by cross spectral analyzes. The values used for these lags are determined as the ones that maximize the average mutual information (M.I) (Fraser and Swinney, 1986) between the GSL series and the corresponding index.

Mutual information attempts to measure the dependence (linear or non-linear) between two coordinates x_t and $y_{t+\tau}$ as a function of the delay τ. The M.I, which can be thought of as a *non-linear* analog of the *sample correlation function*. We have used M.I. only to locate the leading coordinate for each atmospheric index, pairwise, with respect to the current GSL state. This is a prescriptive choice. The M.I. is estimated using kernel density estimators as described in Moon *et al.* (1995).

For the data sets analysed here, the first maximum of M.I. occurs at a lag of 1 month between GSL and SOI, at 6 months between GSL and PNA, and at 3 months between GSL and CNP. A Monte Carlo resampling analysis showed that the associated M.I. values were significantly different from 0, at the 1% level. The resulting data matrix for $\overline{V}_t$ is then composed of GSL, SOI (1 month lagged), PNA (6 months lagged), and CNP (3 months lagged) coordinates.

At this stage one needs to specify a procedure for recovering the function $f(\overline{V}_t)$ for a given vector $\overline{V}_t$, and for selecting the embedding dimension for each predictor. The algorithm for estimating $f(\overline{V}_t)$ is described next. Procedures for simultaneously selecting the parameters of that algorithm and the embedding dimensions are then summarized.

3.1. Local weighted polynomial regression for estimating $f(\overline{V}_t)$

Let us presume for now that the composition of the vector $\overline{V}_t$ is specified, i.e. a choice has been made for the number of coordinates to use for each variable and the associated lags. Then, if we desire a prediction from a certain time T, we can use the time series values prior to time T to form a data matrix $(\mathbf{x}_i, y_i)$, $i = 1 \ldots n$, where the y_i represent past values of G_{t+m}, and the $\mathbf{x}_i$ the corresponding values of the state vector $\overline{V}_t$. The data matrix $\mathbf{x}$ has dimension $n \times d$ where the column dimension is the sum of the embedding dimensions, i.e., $d = m_1 + m_2 + m_3 + m_4$.

Then the forecast $f(\mathbf{x}_n)$ at time T is obtained through the solution to a general regression model given as

$$y_i = f(\mathbf{x}_i) + e_i \quad i = 1, \ldots, n \tag{3}$$

where the function $f(.)$ can be thought of as a regression function.

A non-parametric regression problem results if we consider a solution of this problem such that (1) no prior assumption is made about the explicit functional form of $f(.)$, (2) the interest is in approximating $f(.)$ at each desired location, presuming that it belongs to a fairly rich class of functions (e.g. differentiable functions), and (3) the estimate is 'local', i.e. the influence of distant points on the regression at a given point diminishes with distance. The target function $f(.)$ may be approximated at a point $\mathbf{x}_n$ by retaining the leading terms in its Taylor's series expansion.

This is equivalent to a low order polynomial approximation of the function at that point using k neighbouring data points. The idea is illustrated for the univariate case in Figure 3. In the multivariate case one uses k neighbours $\mathbf{x}_j$, $j = 1 \ldots k$, of $\mathbf{x}_n$ in a vector space of dimension d, to evaluate a low order polynomial regression using the corresponding y_j. The k neighbours are found as the state vectors that are closest in distance to the vector $\mathbf{x}_n$. Thus, in the time series context we locate the k data patterns that are most similar to the state vector $\overline{V}_t$ we seek to forecast from, identify their m step ahead successors, and evaluate a low-order polynomial regression with these data as an approximation to $f(\overline{V}_t)$.

Two strategies were considered for forming the k nearest neighbourhood of the forecast vector $\mathbf{x}_n$. First

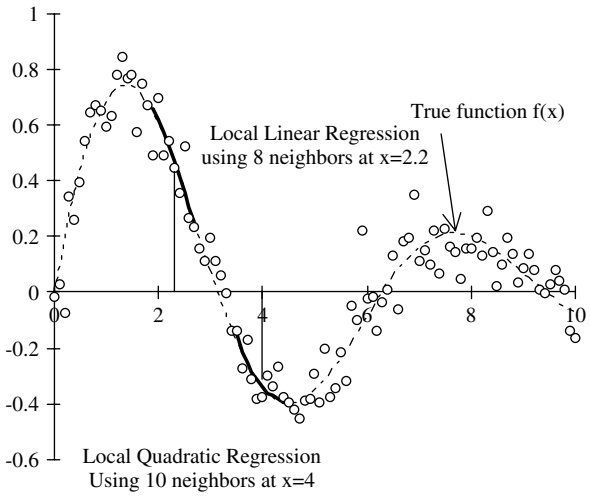

Figure 3. Local linear and local quadratic approximation of $f(x) = \sin(x)e^{-0.2x}$ at two points. This 'damped' oscillation is representative of the quasi-periodic oscillations seen in the GSL data upon bandpassing it at frequencies that have high spectral power. The data (circles) were generated using Equation (3), with $e_i \sim N(0, 0.1)$. The function $f(x)$ is shown as the dashed line, and the local regressions are shown as heavy solid lines. For forecasting, x would be a d dimensional vector in state space, the neighbours would be the closest points in IR^d, and a multivariate local regression will be needed.

DOI: 10.1002/joc

(KM = 1), we considered an unweighted L_1 distance taken across all components of $\mathbf{x}_n$. The distance between two data vectors $\mathbf{x}_i$ and $\mathbf{x}_n$ is then evaluated as

$$d_{\mathbf{x}_i,\mathbf{x}_n} = \sum_{j=1}^{d} |x_{ij} - x_{nj}| \tag{4}$$

The k^{th} nearest neighbour of $\mathbf{x}_n$ is then found by sorting the distances $d_{\mathbf{x}_1,\mathbf{x}_n}$ in ascending order, and locating the k^{th} such distance and its associated data vector.

The second strategy (KM = 2) also tries to account for seasonality factors in determining similarity in data patterns. Consequently, the calendar month, χ_i, associated with the data vector $\mathbf{x}_i$ is also included as a descriptor of the nearest neighbourhood. By considering the calendar month of each observation for defining the nearest neighbourhood, we attempt to provide a treatment of the annual cycle in the hydroclimatic processes. We could have pursued a seasonal model, i.e. fit our model to data only from pre-specified seasons as is often done in time series modeling. However, recognition that the seasonality of a number of such variables may indeed be changing over the last century [see Rajagopalan and Lall, 1995] motivated us to consider an approach that uses the time of the year only as a coordinate in the identification of similarity of data patterns. Months closer to the month of forecast are given more weight, but the actual volume or pressure patterns are also considered. A weighted L_1 distance across the components and the time index χ_i is now considered as

$$d_{\mathbf{x}_i,\mathbf{x}_n} = \sum_{j=1}^{d} |x_{ij} - x_{nj}|\omega_j + \omega_\chi d_{\chi_{i,n}} \tag{5}$$

where ω_j and ω_χ represent weights given by the user to the j^{th} component of x and to the calendar month χ_i respectively, and

$$d_{\chi_{i,n}} = \min\{|\chi_i - \chi_n|, |12 + \chi_i - \chi_n|\} \tag{6}$$

The distance $d_{\chi_{i,n}}$ measures the distance between the calendar months χ_i and χ_n associated with the data vectors $\mathbf{x}_i$ and $\mathbf{x}_n$. In typical applications we varied the weights accorded to the different components in Equation (5). We started with all weights equal to each other, and then explored increasing the relative weight accorded to χ and to the coordinates associated with GSL volume.

A detailed exposition of weighted local regression may be found in Cleveland (1979), Cleveland and Devlin (1988), Cleveland *et al.* (1988), Lall and Bosworth (1995), and Lall *et al.* (2006). Localization of the regression is achieved by using only k neighbours of the prediction point, and also by weighting the data with a monotonic weight function, with weights decreasing as a function of distance of the neighbour from the prediction point.

We consider locally linear ($p = 1$), quadratic ($p = 2$), and quadratic with cross products ($p = 2'$) approximations. Say we denote by $\mathbf{Z}$ a data matrix formed by augmenting the matrix $\{\mathbf{x}\}_{k,n}$ of k nearest neighbours of $\mathbf{x}_n$ to complete a polynomial basis of order p. The order p weighted local regression using k nearest neighbours is then defined through the model

$$\mathbf{y} = \mathbf{Z}\beta + \mathbf{e} \tag{7}$$

where $\mathbf{y}$ is a $k \times 1$ vector, $\mathbf{Z}$ is a $k \times d'$ matrix, β is a $d' \times 1$ vector of regression coefficients and $\mathbf{e}$ is a $k \times 1$ vector of residuals that are assumed to be independent and locally homogeneous.

The quality of such a low-order weighted polynomial approximation depends on the size of the neighbourhood and the order of the polynomial. For a given order, as the size of the neighbourhood increases, the variance of estimate decreases while the bias of estimate may increase. Likewise, increasing the order of the polynomial may reduce the bias or approximation error, while increasing the variance of estimate if the number of points in the neighbourhood is kept the same. This bias-variance trade-off suggests the possibility of searching for an optimal model for local estimation by varying the order of the local polynomial, and the size of the neighbourhood. Lall *et al.* (2006) developed a localized measure of predictive risk that recognizes this trade-off in a cross-validatory or predictive context. This measure, termed local Generalized Cross Validation with leverage (LGCVLEV), is used here for assessing the point wise predictive mean square error as well as for the selection of model parameters (i.e. the order p of the local polynomial and the number of neighbours k, and the embedding dimensions $m_1 - m_4$). See the paper (Lall *et al.*, 2006) for more detail on this algorithm.

$$LGCVLEV(\mathbf{z}_n) = \frac{\mathbf{e}^T\mathbf{W}\mathbf{e}}{((k-p)/k)^2} h_n \tag{8}$$

where

$$h_n = \{\mathbf{z}_n(\mathbf{Z}^T\mathbf{W}\mathbf{Z})^{-1}\mathbf{z}_n{}^T/k\} \tag{9}$$

where $\mathbf{W}$ is a diagonal weight matrix with the diagonal elements.

3.2. The forecasting algorithm

An individual forecast proceeds as described below. If a direct m-step-ahead forecast is desired, this process is undergone m times changing the vector $\mathbf{y}$ each time to bring in the next set of successor values of GSL volumes, for the same $\mathbf{x}_i$ values, and the forecast is made from $\mathbf{x}_n$. If a m step iterated forecast is desired, then this process is repeated m times, with the latest forecast used to augment the $\mathbf{x}$ matrix.

(1) Select a candidate set of values for the embedding dimensions m_1, m_2, m_3 and m_4, delay τ, and time

DOI: 10.1002/joc

lags τ_1, τ_2, and τ_3. The lags are selected using the cross-mutual information function.
(2) At the current prediction point $\overline{\boldsymbol{V}}_t$, form the data set $(\mathbf{x}_i, y_i)(i = 1, \ldots, n)$.
(3) Specify the number of neighbours (k) of $\overline{\boldsymbol{V}}_t$ to use and the order (p) of the local polynomial.
(4) Specify the weights ω_j and ω_χ, as well as the method (KM) to be used to define the k nearest neighbourhood.
(5) Form a basis matrix $\mathbf{Z}$ for this neighbourhood.
(6) Fit the local regression.
(7) Evaluate LGCVLEV.
(8) Repeat for other candidate choices of $k, p, m_1, m_2, m_3, m_4, \tau, \omega_j, \omega_\chi$ and KM.
(9) Pick the combination of parameters that minimizes LGCVLEV.

4. Application

The investigations of the applicability of the forecasting model to the GSL focused on the following questions:

(1) Is it better to use direct or iterated m step ahead forecasts, if we are reconstructing the dynamics from a single scalar time series (GSL in this case)?
(2) Do the forecasts using the GSL and the atmospheric indices improve on the forecasts using just the GSL from the same point?
(3) Can one successfully identify situations where predictability is high or low using LGCVLEV; i.e. do the actual forecast errors fall within the indicated confidence intervals, and do high/low values of LGCVLEV correctly indicate poor/good forecasting ability?
(4) Is the dynamics linear or non-linear; i.e., do the forecasts made by the method presented here improve on those from the best fit linear autoregressive (AR) model?
(5) How does the predictability of the system, as measured by the time rate of growth of forecast errors, vary by situation?

Forecasts of the GSL volume for up to 2 years ahead were considered from various points in time in the record. The time lags, τ_1, τ_2, and τ_3 were pre-selected using MI as described, and KM, ω_j and ω_χ were specified after some initial testing from various forecast points. Based on initial tests, KM = 2, i.e. treatment of annual cycle using a calendar month coordinate, with all coordinate weights taken to be equal (i.e. $\omega_j = \omega_\chi = 0.2$; $j = 1.4$), was selected. The remaining parameters (i.e. k, p, m_1, m_2, m_3, m_4 and τ) were selected locally for each forecast as the values that minimize LGCVLEV for that forecast. The number of nearest neighbours, k, was varied from 50 to 160, depending on the values of p, d, τ and the available sample size n. Locally linear, quadratic, and quadratic with cross product terms models ($p = 1, 2, 2'$) were considered in each case. The embedding dimensions m_1, m_2, m_3 and m_4 were varied between 1 and 10, again depending on how much data was available at the forecast point. The delay τ was varied between 1 and 12. We found that in most cases a locally linear model with $k = 120$–140 was selected, with $\tau = 5$ or 6, and m_1, m_2, m_3 and m_4 were near 5. Since the number of coordinates that can be used is limited by the data available at the prediction point, we allowed m_1, m_2, m_3 and m_4 to vary by prediction point, and tried fixing the other parameters in the range indicated above. The motivation for this prescription was to achieve parsimony and to reduce the possibly increased variance in the forecasts due to repeated parameter selection.

We also tried to use a t-test for the coefficients of a linear regression to choose only the coefficients of the local regression that were important. Typically this test chose most of the coefficients in the model. An F-test to check all the terms belonging to SOI, PNA, and CNP were needed in a local fit. We found that while some of the local fits considered only one of the PNA or CNP sets, overall better results were obtained by retaining both indices.

We considered direct m-step forecasts of the GSL based on past GSL, SOI, CNP, and PNA data; direct m-step GSL forecasts based only on GSL data; iterated m-step GSL forecasts based only on GSL data; and a univariate AR forecast (using only the GSL) with the order of the AR model chosen using the Akaike information criteria. We wanted to limit the computational effort to forecasting only the GSL. This precludes an iterated m-step-ahead forecast using all the variables. The performance of this approach is compared with the iterated m-step-ahead forecast using only the GSL to see if there is an improvement from using the atmospheric variables, and with the AR model. In all cases, it was assumed that the training set for each model is given by the time series available up to the forecast date, and that subsequent data are not available.

Three such forecasts that are representative of different conditions in state space as well as rather different record lengths are now compared. The first set of forecasts is from March 1961 to February 1963 using the segment of the time series from December 1946 to February 1961 as the training data. It was assumed that records beyond February 1961 were not available for either fitting the model or using as input to forecasting. The results from this analysis are presented in Figure 4.

This is a rather short record length, even of fitting a simple AR model for time series prediction. From Figure 4(a) we see that the forecasts based on the methods introduced in this chapter are all markedly superior to those from the AR(13) model fitted to the data. In this case, the m-step-ahead forecasts using the direct or the iterated method using only the GSL data are very similar and are comparable to those based on the GSL and the atmospheric data for the first year of the forecast. The latter appears to be better during the second year. The limited ability to fit a multivariate model $(GSL + Atm.)$ with this short data set may account

DOI: 10.1002/joc

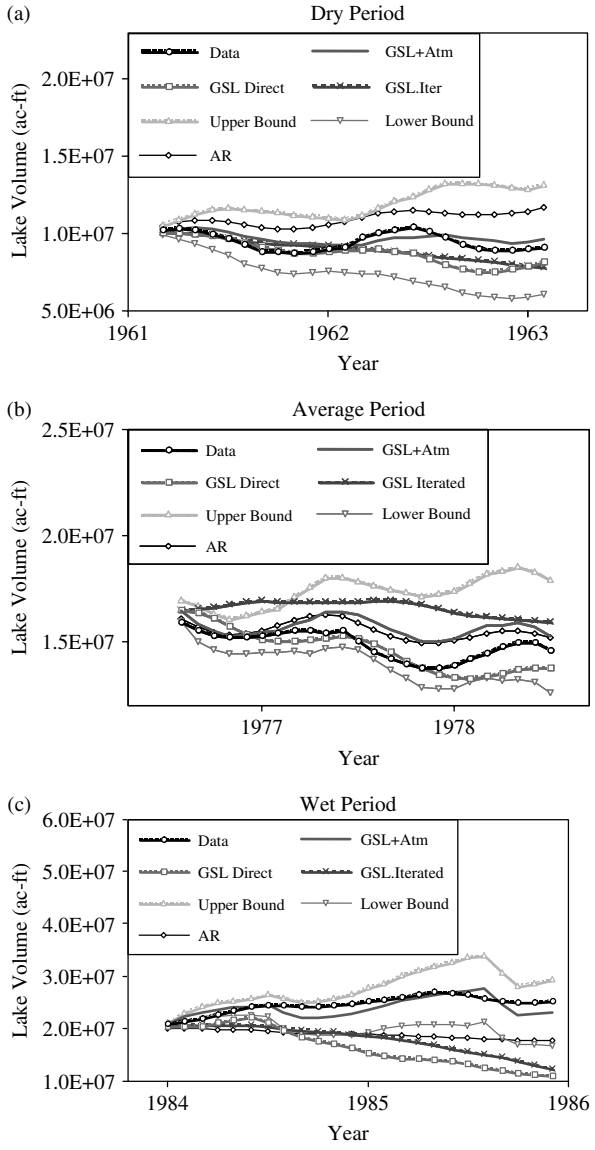

Figure 4. Comparative forecasts of $GSL + Atm$ in a) the Dry period, b) the Average Period and c) the Wet Period: Direct prediction using GSL and atmospheric indices ($k = 130$, $p = 1, m_1 = 5, m_2 = m_3 = m_4 = 3$); GSL Direct: Direct prediction using GSL ($k = 130$, $p = 1, m_1 = 5$); GSL Iterated: Iterated prediction using GSL ($k = 130$, $p = 1, m_1 = 5$); and AR : AR (13) predictions using GSL only. The confidence intervals shown are for the direct forecast based on $GSL + Atm$. Indices.

for some of this observation. Nevertheless, the extended range (second year) forecast clearly seems to benefit from the atmospheric information. It is interesting to see that the future GSL values are contained within the confidence intervals associated with the $GSL + Atm$ forecasts. The AR forecast does not lie within these confidence intervals, and the forecasts that do not use the atmospheric data also do not lie within these confidence intervals during the second year.

The 'predicted errors' (standard errors) of each forecast as measured by half the width of the confidence interval are estimated. Absolute forecast errors from the $GSL + Atm$ direct forecast are also evaluated. We observe that the predicted errors are consistent with the relative quality of the forecasts in Figure 4(a), and that using the prediction intervals (equivalently LGCVLEV) as a criterion for selecting across the models, one would make good choices. The general trend of the predicted errors appears to be quite consistent with the trend of the actual forecast errors.

The next situation corresponds to a 2 year forecast of the GSL volume for conditions near the average volume of the GSL, using 10 more years of data (from December 1946 to July 1977). Results corresponding to the previous forecast are shown in Figure 4(a). From Figure 4(b), we observe that the $GSL + Atm$. forecast is comparable to that from an AR(13) model, is better than the forecast from an iterated forecast using only the GSL data, but worse than a direct forecast using only the GSL data. The future 24 month values of the GSL still lie within the prediction intervals from the $GSL + Atm$ model. However, the confidence intervals this time are quite wide, and include virtually all the forecasts from the different methods.

A third 2 year forecast during a period of high GSL volume is based on data from December 1946 to December 1983. Results are presented in Figure 4(c). From the time series plot in Figure 1, we see that this is an unprecedented period in the history of the GSL that is used for forecasting in this paper. From Figure 2, prospects of useful predictions of the GSL during this period would appear bleak. We observe from Figure 4(c) that the $GSL + Atm$ forecasts are actually quite credible and soon become markedly superior to those from any of the other methods. The confidence intervals are indeed quite wide reflecting the extrapolation of the historical dynamics. While the future observations fall well within the confidence intervals of the $GSL + Atm$ forecasts, the forecasts from the other methods fall out of the confidence intervals after the first year.

A sequence of 1- and 2-year-ahead forecasts for the 1976–1987 period using the $GSL + Atm$ method is presented in Figures 5 and 6. It is interesting to see that the predictability of the system appears to be the least near or at the turning points of the GSL time series. Some trajectories (e.g. 1976, 1977,1981,1983) that appear to be departing from the observed time series at the end of 1 year actually do quite well in the second year. Other trajectories (e.g. 1979, 1982, 1983) that are near turning points diverge substantially in the second year. The crucial trajectories near the top in 1987–1988 are marked by increased variability (as would be expected given the degree of extrapolation) as well. Of course, in a multivariate setting as the one considered here, it is not too valuable to talk about the behaviour in terms of the 'turning' points of one of the time series.

Given the short record length, and the extrapolatory nature of the forecasts, and the sparsity of the observed phase space, it is clearly quite difficult to recover the moderate to high dimensional dynamics of the underlying dynamical system. Considering this, it is remarkable that the local polynomial forecasts are as successful as they are.

DOI: 10.1002/joc

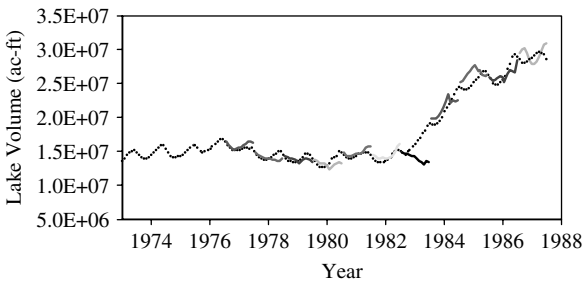

Figure 5. A sequence of 1 year blind forecasts of the GSL using the direct m step method and the GSL and the atmospheric index time series (SOI, CNP, and PNA) from July 1976 to June 1987. The dots represent the observed GSL time series. The solid lines represent 12 forecasts, one for each month of the next year. Only data available up to the beginning of the forecast period are used for fitting and forecasting.

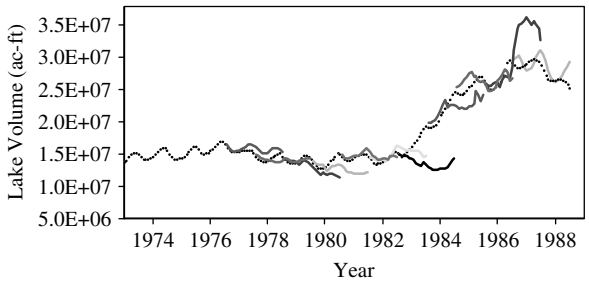

Figure 6. A sequence of 2 year blind forecasts of the GSL using the direct m step method and the GSL and the atmospheric index (SOI, CNP, and PNA) time series from July 1976 to June 1988. The dots represent the observed GSL time series. The solid lines represent 24 forecasts, one for each month of the next year. Only data available up to the beginning of the forecast period is used for fitting and forecasting.

Another way of assessing the predictability of certain states is to examine the rate of growth of errors with time. For a dynamical system, the rate of growth of errors is related to the Lyapunov exponents of the system [see Abarbanel *et al*., 1993]. The Lyapunov exponents for a d dimensional system tell us how small perturbations in the trajectory of a d dimensional dynamical system grow or decay as a function of time. The largest Lyapunov exponent, λ_1, determines the horizon of predictability of the system, through the rate of divergence of nearby trajectories of the system. If two points in state space are separated by a distance ε at time t, they will be separated by a distance $\varepsilon e^{\lambda_1 \Delta t}$ at time $(t + \Delta t)$. The exponent λ_1 can vary dramatically over the state space of the system. Often, only a global average λ_1 is computed. Abarbanel *et al*. (1996) computed the average global λ_1 as approximately 1/100 days from the biweekly 1847–1991 time series of the GSL. The short, noisy, time series available here precludes a direct estimation of local Lyapunov exponents. Indeed, since the forecasts are made here through what can be considered a non-linear time series model, predictability or forecast error growth with time is more appropriately discussed in that context.

For a linear time series model, the error variance of the forecast is independent of the initial state from which the forecast is initiated. Thus, differences in the rate of growth of errors with forecast time would also be indicative of non-linearity in the system. In our case, we have approximation error involved in the estimation of the map $f(\overline{V}_t)$, in addition to the possible divergence of trajectories in the system if the system parameters were known precisely. We have available to us, at different times, the actual forecast error, as well as a measure of forecast error as provided by LGCVLEV$^{0.5}$. From the discussion of Lyapunov exponents, we expect an exponential rate of growth of errors with forecast time, which may vary with location in state space. Both the initial error (ε above) and its rate of growth (λ_1) are of interest. These are examined in Figure 7 for the average situations covered by Figure 4(b), by plotting ln(LGCVLEV$^{0.5}$) *versus* forecast time in each case. The intercept of these plots estimates ε, and the slope estimates λ_1.

Considerable variation in the rate of growth of predicted errors with state is observed. From Figure 7, we see some dependence of predictability on the annual cycle, suggesting that either the treatment of the annual cycle in the forecasting model through the weights used to select the k nearest neighbourhood is inadequate, or predictability varies seasonally. Variation of the weights for the forecasts in these situations was not investigated systematically due to computational constraints. Both the dry and the wet situations were explored, but the detailed results are not included in here because of the length limitation of the paper. The results of this analysis are consistent with those presented in Figure 7. The sensitivity to the seasonal cycle appears most pronounced for the dry period. The rate of growth of errors, as well as the intercept of the regressions of ln(LGCVLEV$^{0.5}$) with prediction time, seems to vary with state as one compares across the three cases, with a tendency for both parameters to increase as the volume of the lake increases. The

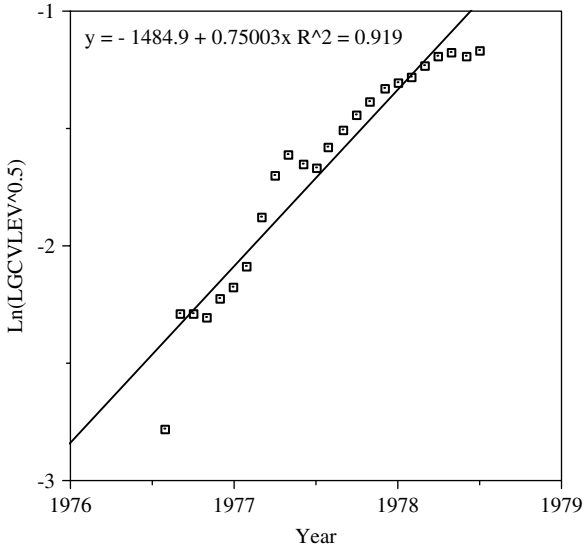

Figure 7. Ln(LGCVLEV$^{0.5}$) *versus* forecast time for August 1976–1978 (average period) forecasts of GSL volume.

DOI: 10.1002/joc

situation in the wet period, where the rate of loss of predictability over the first 6 months is considerably higher than for the remaining 17 months, is also interesting. The detailed results are available on request from the authors.

5. Conclusions

The application of a new methodology for time series forecasting to the prediction of the GSL was presented. Prospects for the improvement of the forecasts by including information on selected atmospheric circulation indices were discussed and demonstrated. Variations in the predictability of the GSL dynamics as a function of state and at different times were demonstrated. The LGCVLEV criterion was shown to be useful for decisions on parameter estimation, model specification, for judging the quality of the forecasts, and for developing prediction intervals.

Further refinements of the actual forecasts by varying the composition of the predictor variables (use of other indices, local rainfall and temperature, variable delays by coordinate) are also possible. The primary contribution of the work presented here is to demonstrate the feasibility of such forecasts, their utility, and the innovation of a local predictive risk measure to guide the process.

The work presented here should be considered as a beginning and not an end product. As far as predictions of the GSL are concerned, forecasts based on the 1847–1994 biweekly record presented in Lall *et al.* (2006) are quite good. However, most locations will not have as much high resolution data. Our purpose here was in exploring the utility of the local polynomial regression approach for time series prediction and of the possible utility of atmospheric information for stabilizing and improving the predictions from shorter, coarser time series. On both counts the results are encouraging. The complexity of natural systems and the difficulty of modeling them statistically with automatically chosen parameters were also demonstrated to us during the work. It is presumptuous to believe that one can recover nonlinear dynamics of moderate to high dimensional systems from only a finite time series of selected variables. Nevertheless, the degree of success we have met with in this and related work is encouraging. One must look forward to the development of modeling strategies that further improve error estimates of forecasts, include physically based descriptions where possible, and allow intelligent combinations of forecasts from different methods.

References

Abarbanel HDI, Brown R, Sidorowich JJ, Tsimring LS. 1993. The analysis of observed chaotic data in physical systems. *Review of Modern Physics* **65**(N4): 1331–1392.

Abarbanel HDI, Lall U, Moon Y-I, Mann M, Sangoyomi T. 1996. Non-Linear dynamics of the Great Salt Lake: A predictable indicator of regional climate. *Energy* **21**(7/8): 655–665.

Cayan DR, Roads JO. 1984. Local relationships between United States west coast precipitation and monthly mean circulation parameters. *Monthly Weather Review* **112**: 1276–1282.

Cayan DR, Peterson DH. 1989. The influence of North Pacific atmospheric circulation on streamflow in the west. In *Geophysical Monograph*, Vol. 55. American Geophysical Union: Washington, DC.

Cleveland WS. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**(368): 829–836.

Cleveland WS, Devlin SJ. 1988. Locally weighted regression: An approach to regression. Analysis by local fitting. *Journal of the American Statistical Association* **83**(403): 596–610.

Cleveland WS, Devlin SJ, Grosse E. 1988. Regression by local fitting. *Journal of Econometrics* **37**: 87–114.

Emery WJ, Hamilton K. 1985. Atmospheric forcing of interannual variability in the northeast Pacific Ocean: Connections with El Nino. *Journal of Geophysical Research* **90**: 857–868.

Fraser AM, Swinney HL. 1986. Independent coordinates for strange attractors from mutual information. *Physical Review A* **33**(2): 1134–1140.

Friedman JH. 1991. Multivariate adaptive regression splines. *Annals of Statistics* **19**(1): 1–141.

Ghil M, Vautard R. 1991. Interdecadal oscillations and the warming trend in global temperature time series. *Nature* **350**: 324–327.

Hirschboeck KK. 1987. Hydroclimatically-defined mixed distributions in partial duration flood series. *Hydrologic Frequency Modeling*, Singh VP. (ed) D. Reidel: Norwell, Mass., 107–128. inedited by.

Horel JD, Wallace JM. 1981. Planetary scale atmospheric phenomena associated with the Southern Oscillation. *Monthly Weather Review* **109**: 813–829.

Kahya E, Dracup JA. 1993. U.S. streamflow patterns in relation to the El Nino/Southern Oscillation. *Water Resources Research* **29**(8): 2491–2503.

Kember G, Flower AC, Holubeshen J. 1993. Forecasting river flow using non-linear dynamics. *Stochastic Hydrology and Hydraulics* **7**: 205–212.

Keppenne CL, Ghil M. 1992. Adaptive filtering and prediction of the southern oscillation index. *Journal of Geophysical Research* **97**: 20449–20454.

Kiladis GN, Diaz HF. 1989. Global climatic anomalies associated with extremes in the Southern Oscillation. *Journal of Climate* **2**: 1069–1090.

Klein WH, Bloom HJ. 1987. Specification of monthly precipitation over the United States from the surrounding 700 mb height field. *Monthly Weather Review* **115**: 2118–2132.

Lall U, Bosworth K. 1995. Non-parametric statistical inference and function estimation for space and time hydrologic data. *Trends in Hydrology*, Menon J (ed). CSRI: Trivandrum, India.

Lall U, Mann ME. 1995. The Great Salt Lake: A barometer of low frequency climatic variability. *Water Resources Research* **31**: 2497–2502.

Lall U, Moon Y-I, Kwon H-H, Bosworth K. 2006. Locally weighted polynomial regression: Parameter choice and application to forecasts of the Great Salt Lake. *Water Resources Research* **42**: W05422, DOI: 10.1029/2004WR003782.

Leathers DJ, Yarnal B, Palecki M. 1991. The Pacific/North American teleconnection pattern and United States climate. Part I: Regional temperature and precipitation associations. *Journal of Climate and Applied Meteorology* **24**: 463–471.

Lins HF. 1993. Seasonal hydrologic variability and relations with climate, Ph.D. dissertation, University of Virginia, Reston; 201.

Mann ME, Park J. 1993. Spatial correlations of interdecadal variation in global surface temperatures. *Geophysical Research Letters* **20**: 1055–1058.

Mann M, Lall U, Saltzman B. 1995. Decadal to century scale climatic variability: Understanding the rise and fall of the Great Salt Lake. *Geophysical Research Letters* **22**: 937–940.

Moon Y-I, Lall U. 1996. Atmospheric flow indices and interannual Great Salt Lake variability. *ASCE Journal of Hydrologic Engineering* **1**(2): 55–62.

Moon Y-I, Rajagopalan B, Lall U. 1995. Estimation of mutual information using kernel density estimators. *Physical Review E* **52**(2): 2318–2321.

Rajagopalan B, Lall U. 1995. Seasonality of precipitation along a meridian in the western U.S. *Geophysical Research Letters* **22**(8): 1081–1084.

Ropelewski CF, Halpert MS. 1987. Global and regional scale precipitation patterns associated with the El Nino/Southern Oscillation. *Monthly Weather Review* **115**: 1606–1626.

DOI: 10.1002/joc

Sangoyomi TB. 1993. Climatic Variability and Dynamics of Great Salt Lake Hydrology, PhD. dissertation, Utah State University, Logan; 247.

Sangoyomi T, Lall U, Abarbanel HDI. 1996. Non-linear dynamics of The Great Salt Lake: Dimension Estimation. *Water Resources Research* **32**: 149–160.

Smith JA. 1991. Long-range streamflow forecasting using non-parametric regression. *Water Resources Bulletin* **27**(1): 39–46.

Thomson DJ. 1982. Spectrum estimation and harmonic analysis. *IEEE Proceedings* **70**: 1055–1096.

Yakowitz S, Karlsson M. 1987. Nearest neighbour methods with application to rainfall/runoff prediction. In *Stochastic Hydrology*, Macneil JB, Humphries GJ (eds). D. Reidel: Hingham, MA; 149–160.

DOI: 10.1002/joc