Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm

Francisco Assis Souza Filho

Fundação Cearense de Meteorologia e Recursos Hídricos (FUNCEME), Fortaleza, Ceara, Brazil

Upmanu Lall

Department of Earth and Environmental Engineering and International Research Institute for Climate Prediction, Columbia University, New York, New York, USA

Received 11 April 2002; revised 5 February 2003; accepted 23 June 2003; published 1 November 2003.

[1] A semiparametric approach for forecasting streamflow at multiple gaging locations on a river network conditional on climate precursors is developed. The strategy considers statistical forecasts of annual or seasonal streamflow totals at each of the sites and their disaggregation to monthly or higher resolution flows using a k nearest neighbor resampling approach that maintains space-time consistency across the sites and subperiods. An application of the approach to forecasting inflows at six reservoirs in the state of Ceara in northeastern Brazil is presented. The climate precursors used are the NINO3 index for the El Niño-Southern Oscillation and an equatorial Atlantic sea surface temperature index. Forecasts of January through December streamflow are made at three lead times: in January of the same year and in October and July of the preceding year. The skill of the ensemble forecasts generated is evaluated on subsets of the historical data not used for model building. Correlations with the equatorial Atlantic index and with NINO3 translate into useful streamflow forecasts for the next 18 months of reservoir operation and water management. *INDEX TERMS:* 1833 Hydrology: Hydroclimatology; 1860 Hydrology: Runoff and streamflow; 3220 Mathematical Geophysics: Nonlinear dynamics; *KEYWORDS:* ENSO, forecast, streamflow, nonparametric, nonlinear, climate

Citation: Souza Filho, F. A., and U. Lall, Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm, *Water Resour. Res.*, *39*(11), 1307, doi:10.1029/2002WR001373, 2003.

1. Introduction

[2] Arid regions, such as northeast Brazil are particularly vulnerable to climate fluctuations and to their impact on water supply. Consequently, there is interest in the development of long range streamflow forecasts that could be used for reservoir operation and water allocation for competing demands. Both dynamical and statistical methods are being used to develop such forecasts. One pathway for the development of such forecasts is the use of general circulation models (GCMs) of the ocean and the atmosphere, followed by "downscaling" using Regional Climate Methods or statistical approaches, followed by lumped or distributed rainfall-runoff models. This is a useful research direction. At this time, issues related to uncertainty propagation along the pathway, process parameterization, final forecast skill and utility in the context of resource management at relevant space and timescales of interest are still being evaluated. An alternative is the direct development of statistical forecasts for water supply and demand using a suitably selected set of climate precursors. A new method for generating probabilistic statistical forecasts of river flows is developed and applied here.

[3] Our goal was to develop a procedure that is consistent with the information needs and analysis methods of a water agency responsible for operating a network of reservoirs on a river system. This translates into estimates of spatially and temporally consistent monthly inflows to a set of reservoirs over a storage cycle (e.g., 3 months to 2 years). Many system operators make water and storage allocation decisions for the upcoming storage cycle by simulating the system using inflow sequences resampled from the historical record (e.g., using the index-sequential record method [*Kendall and Dracup*, 1991]) and projected demands. We provide a capability for conditionally resampling the historical record considering the state of key climate precursors. Selected multivariate regression strategies are first explored to describe the relationship between annual (or seasonal) streamflow at the sites of interest and a set of potential climate predictors. Key issues here are the potential nonlinearity of the relationships, and the nonnormality of regression residuals. This analysis is used to prescribe a transformation of the predictor state-space. A nonparametric approximation to the conditional probability density of the matrix of monthly streamflows at all sites, for the future period of interest, is then employed, and Monte Carlo simulations of the future inflows are generated using the k-nearest neighbor method [*Lall and Sharma*, 1996]. Disaggregations to daily timescales, using related methods demonstrated by *Kumar et al.* [2000] can also be considered.

[4] In an analysis of global forecast skills of the leading ocean-atmosphere general circulation models for seasonal precipitation, *Rajagopalan et al.* [2002] find that northeast

0043-1397/03/2002WR001373

Figure 1. Location of Reservoir Inflow Locations in Ceara, Brazil. 1, Oros; 2, Banabuiu; 3, Pedras Branca; 4, Pacajus; 5, Pacoti Riachao; 6, Gaviao. The major irrigation demand areas are indicated by squares, and the municipal and industrial demand areas served are indicated by circles. Only features of the Juagaribe and Metropolitan basins are filled in. Other basin boundaries are marked.

Brazil (Nordeste) is one of a few regions in the world where there is consistent and statistically significant skill during the primary rainy season (January–May). Drought is a perpetual concern in the state of Ceara in NE Brazil and reservoir systems are often stressed even though they are typically designed for a 2–3 year storage cycle. Given the potential for successful long range forecasts, and the high utility of such information, Ceara provides an important case study for the methods developed here. The forecasting methodology developed is discussed through an application to 6 reservoirs in Ceara. In forthcoming papers, we explore forecast uncertainty using Bayesian methods and show how the forecasts are used for water allocation and reservoir operation.

2. Background

[5] Information on drought in Ceara, attributes of water supply and demand, and the historical variation in reservoir inflows is provided first. There has been considerable interest in hydroclimatic forecasting recently [e.g., *Liu et al.*, 1998; *Cordery and McCall*, 2000; *Piechota et al.*, 2001]. Here only past efforts at statistical forecasts of streamflow or precipitation in the region are reviewed in the context of the identification of a suitable set of climatic predictors for Ceara.

[6] The water system of interest is the Jaguaribe-Metropolitano Hidrossytem (JMH) in Ceara, shown in Figure 1. This is the most important water system in the state. Six major reservoirs (see Table 1: the first three are in Jaguaribe Basin and the rest in the Metropolitan Basin) supply the primary irrigated areas and the largest metropolitan area (Fortaleza). The first four reservoirs are in the semiarid region of the state. Precipitation in the Pacoti-Riachao and Cocó River basins is influenced by orography. Rainfall records for each river basin are available since about 1911. Streamflow records at the different inflow sites vary in their start date from 1912 to 1970. Consequently, calibrated rainfall-runoff models have been used to reconstruct the inflow at each reservoir. The quality of the inflow data is expected to be the best for the Oros reservoir, and weakest for Pacoti-Riachão. Note the frequency of drought implied by the pattern of demand-supply deficit in Figure 2. The annual inflow was near zero in several of the years.

[7] The annual inflow at all sites is highly variable and skewed (Table 1). Ninetyfive percent of the annual reservoir inflow typically occurs during January through June. The seasonal variation of the Oros inflow is shown in Figure 3. The Ceara Water Resources Plan [*Secretaria de Recursos Hídricos do Estado do Ceará* (*SRH*), 1991] and Water Basin Plan [*Companhia de Gestão dos Recursos Hídricos* (*COGERH*), 1999a, 1999b] provide demand projections for JMH. The Jaguaribe Basin water demand is 80% Irrigation and 20% urban. The Metropolitan Basin water demand is predominantly for Urban and Industrial use. Consequently, the demands in the Metropolitan basin are relatively uniformly distributed during the year, while those in the Jaguaribe basin are concentrated in the irrigation season (August through November).

[8] Some recent efforts relating precipitation and streamflow in NE Brazil to climate pre-cursors and describing the attendant climatic mechanisms are discussed by *Uvo et al.* [1998, 2000], *Uvo and Graham* [1998], and *Marengo et al.* [1998]. The rainfall in the region is highly variable in space, within the rainy season and over years [*Kousky*, 1979]. The seasonality of regional rainfall, and hence of streamflow is governed largely by the north/south migration of the inter-tropical convergence zone (ITCZ). *Uvo et al.* [1998]

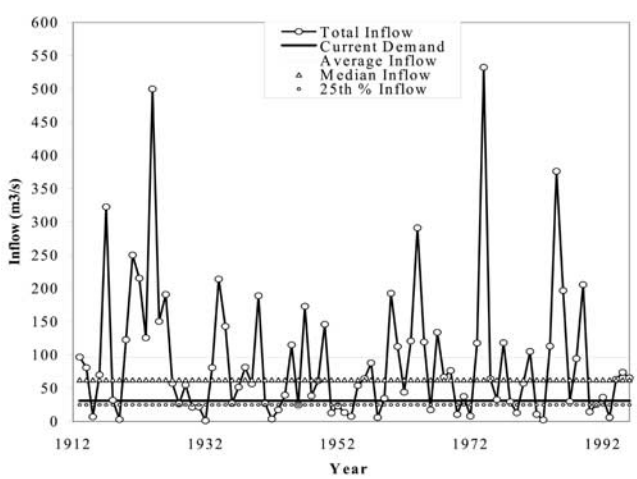


Figure 2. The 1912–1990 time series of total inflow and demand for the Metropolitan River Basin. Note the zero annual inflows in several years, and the possibility of deficit nearly 25% of the time.

Table 1. Basic Data and Statistics of Annual Reservoir Inflow Based on 1913–1990 Record[a]

	Reservoir					
	Oros	Banabuiu	Pedras Branca	Pacajus	Pacoti Riachao	Gaviao
River	Jaguaribe	Banabuiú	Sitiá	Choró	Pacoti-Riachão	Cocó
Basin Area (km^2)	24563	14931	1787	4060	1108	95
Storage (hm^3)	1956	1800	434	148	420	54
First quartile	6.9	4.5	0.7	2.9	2.8	0.4
Median	18.5	15.1	2.1	17.3	7.1	0.8
Third quartile	37.2	34.3	5.9	32.7	12.0	1.5
Mean	30.0	26.6	5.2	24.6	8.5	1.2
Standard deviation	37.8	31.8	8.0	29.5	7.8	1.1
CV	1.3	1.2	1.5	1.2	0.9	0.9
Skew	2.5	1.9	3.0	2.1	1.2	1.5

[a]Inflow is in units of m^3/s.

synthesize a description of the connection between rainfall and the ITCZ based on past research. They indicate that the principal rainy season is initiated between February and March, as the ITCZ over the tropical Atlantic Ocean reaches it southernmost position. The northward migration of the ITCZ signals the end of the rainy season. However, the timing of this return is highly variable, and significantly affects the seasonal rainfall total. The January–February rainfall is affected by cold fronts or their remnants [*Kousky*, 1979]. Connections between the eastern Pacific and tropical Atlantic ITCZ behavior have been studied by *Nobre and Shukla* [1996], *Saravanan and Chang* [2000], and *Chiang et al.* [2000] with two contrasting hypotheses. Nobre and Shukla explain the connection between a mature ENSO in the boreal winter and the northern part of the tropical Atlantic SST in the winter and the following spring in terms of a northern midlatitude ''atmospheric bridge''. Saravanan and Chang point to the role of an anomalous Walker Circulation as the connection. Chiang et al offer observational support for this mechanism and analyze its interdecadal variations (these relate directly to the changing frequency of El Nino and La Nina events over 21 year moving windows).

[9] The rainfall variability has been related to variations in sea surface temperatures [*Markham and McLain*, 1977; *Moura and Shukla*, 1981; *Hastenrath*, 1984, 1990; *Ward et al.*, 1988]. *Ward and Folland* [1991] found that it is best to use the EOFs of only the tropical Atlantic sea surface temperatures as predictors of the Nordeste rainfall. The Pacific Anomalies associated with ENSO play a weaker role. *Ward et al.* [1993] indicate that the Atlantic EOF spatial patterns are often not robust with respect to the period of analysis and speculate on various reasons for the changes. They demonstrate statistically significant skill in their forecasts of aggregate seasonal Nordeste rainfall with 0 to 2 months lead time using multiple linear regression and linear discriminant analysis. The work of *Uvo et al.* [1998] considers a more detailed multivariate space (105 stations) and time (monthly and seasonal) analysis of the Nordeste precipitation and its relationship to SSTs. Their results indicate that warm SST anomalies in the Southern equatorial Atlantic are associated with an earlier migration of the ITCZ, leading to enhanced rainfall in parts of the Nordeste including Ceara. In agreement with previous studies they find that the position of the ITCZ in April and May and hence the end of the Nordeste rainy season is determined to a great extent by a North-South gradient in the equatorial Atlantic SST. The correlation with an ENSO-Pacific index during this period is also significant. Based on the results of their multivariate analysis, Uvo et al constructed SST indices for the Central Pacific, the equatorial North Atlantic and the equatorial South Atlantic, and for the difference between the North Atlantic and the South Atlantic. The ENSO/Pacific index during the rainy season is highly correlated with the North Atlantic index, but not with the South Atlantic index. They note that the equatorial Atlantic ''dipole'' index is a better predictor of Nordeste precipitation 1 to 3 months in advance, and confirm prior work that recognizes the utility of such an index. They find that the months of April and May were the most important contributors to the interannual variations of Nordeste precipitation, and that their Atlantic dipole index is highly correlated with these fluctuations. The ENSO index plays a smaller but statistically significant role, and is associated with precipitation in January/February, and in April–May.

[10] A context for these observations is provided by the analysis of *Chiang et al.* [2000] who also looked at the dependence of a Ceara rainfall index on ENSO and the cross-equatorial tropical Atlantic SST gradient. They note that as NINO3 increases, the mean and the range of the Ceara rainfall tend to decrease. Their interpretation is that when little convection occurs over the eastern equatorial Pacific (La Nina), the tropical Atlantic ITCZ is influenced

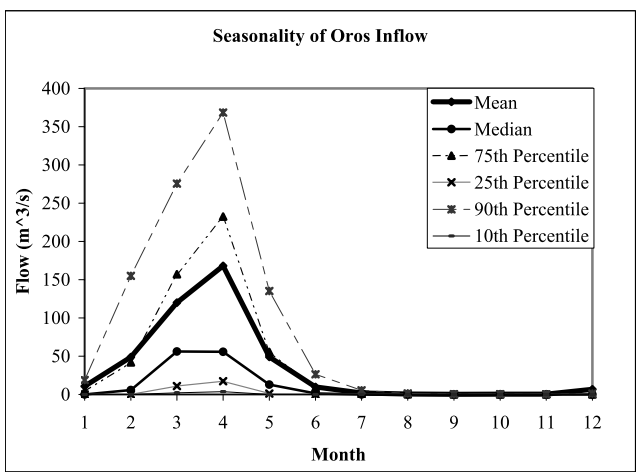


Figure 3. Seasonality of Oros inflow, illustrated through monthly flow quantiles.

Table 2. Correlation of Annual Inflows Across All Sites

	Oros	Banabuiu	Pedras Branca	Pacajus	Pacoti Riachao	Gaviao
Oros	1.00	0.76	0.78	0.78	0.65	0.64
Banabuiu	0.76	1.00	0.83	0.73	0.63	0.56
Pedras Branca	0.78	0.83	1.00	0.83	0.73	0.67
Pacajus	0.78	0.73	0.83	1.00	0.84	0.82
Pacoti Riachao	0.65	0.63	0.73	0.84	1.00	0.94
Gaviao	0.64	0.56	0.67	0.82	0.94	1.00

by other factors, primarily the Atlantic cross-equatorial SST gradient. As convection increases over the eastern equatorial Pacific, anomalous subsidence over the tropical Atlantic reduces the northeast Brazil rainfall and its variation. The nonlinearity in the relation between Pacific SSTs and convection and its influence on the tropical Atlantic SST and Ceara Rainfall is identified as a factor in the apparent change in correlation between NINO3 and its teleconnections to the Atlantic over time.

[11] One and two seasons ahead streamflow forecasts in the adjoining region of Amazonia were considered by *Uvo et al.* [2000] using neural network regression and prior season SSTs in the equatorial Pacific and tropical Atlantic for the 1946–1992 period. They report correlations between observed and forecasted river flows at nine sites that range from 0.35 to 0.76. These were improved over a canonical linear regression model used by *Uvo and Graham* [1998]. The Amazonia region has a rather different climate than the Nordeste, and hence it is not clear if such results can be extrapolated to Ceara. Further, the efficacy of the neural network approach to generate probabilistic scenarios at multiple sites that maintain proper subperiod structure across sites, and can be easily communicated to the reservoir operators, is unclear.

[12] On the basis of the results from prior investigations, the two time series selected as predictors are the NINO3 time series defined as the average sea surface temperature anomaly in the region bounded by the eastern equatorial Pacific 150°W to 90°W and 5°S to 5°N, and a cross-equatorial Atlantic SST gradient (EAD) series defined as the difference in the monthly average of the SST anomaly in the region bounded by North Atlantic (5°–20°N, 60°–30°W) and the monthly average of the region bounded by South Atlantic (0°–20°S, 30°W–10°E). The monthly time series for the indices were derived from the gridded

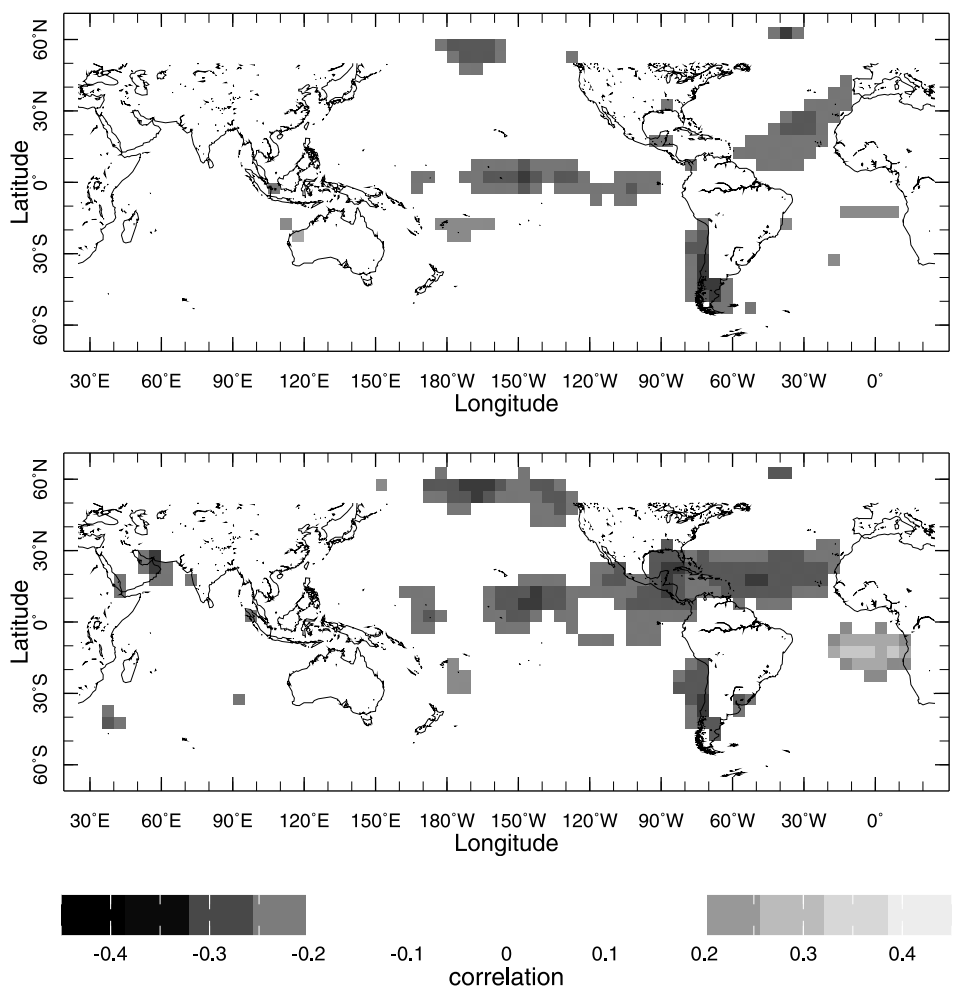


Figure 4. April–May–June and July–August–September SST correlations with Oros annual flow for the next year. The NINO3 index is defined over 5°S to 5°N, 150°W to 90°W; the North Atlantic index is defined over 5° to 20°N and 60° to 30°W, and the South Atlantic index is defined over 0° to 20°S and 30°W to 10°E. The EAD index is the difference between the North and the South Atlantic indices.

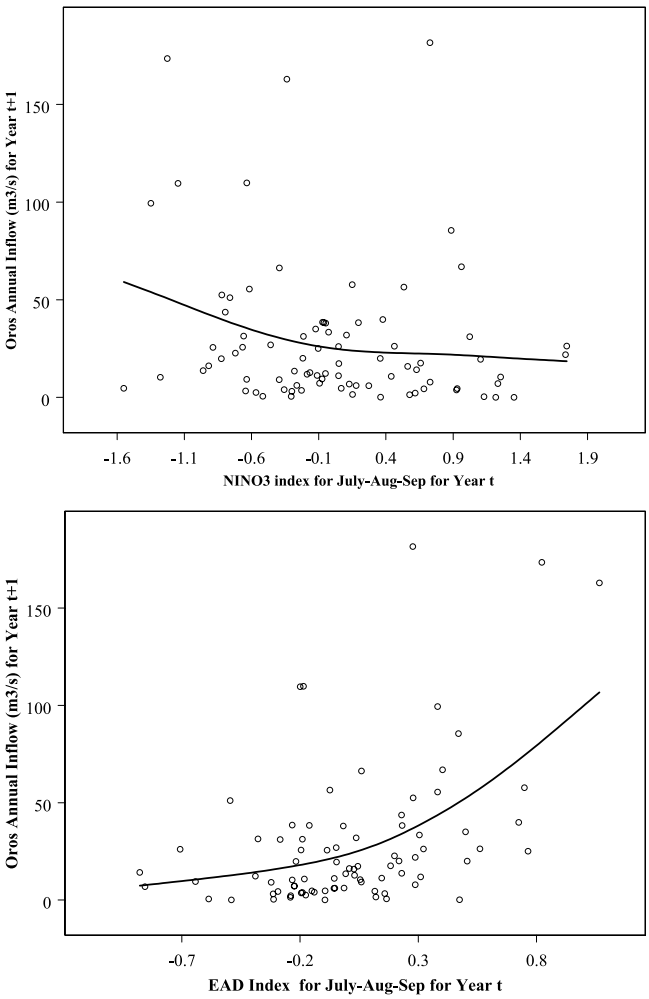


Figure 5. Relations between annual Oros inflow and preceding July–August–September SST indices. Similar relations hold for the other seasons. The line in each plot is a cross-validated smoothing spline fit to the data.

SSTA data sets developed by *Kaplan et al.* [1998], available at http://ingrid.ldeo.columbia.edu/SOURCES/.KAPLAN/.EXTENDED/.

3. Diagnostic Analyses

[13] Basic statistical analyses of the temporal variations and inter-relationships of selected streamflow and climate index time series are presented in this section. We derived seasonally averaged time series for each index, and annually averaged series for each streamflow site. This is in anticipation of the design of a forecasting system that would use the average April–May–June (AMJ), July–August–September (JAS), October–November–December (OND) index to forecast the January–December (ANN) streamflow in the next year, in early July, October and January respectively. Since, the streamflow in July–December is near zero, one could potentially provide up to 18 months (for the ANN forecast from the preceding July) of operational guidance with this design.

[14] The annual streamflow at the six sites had a similar, positively skewed probability distribution. A Principal Components Analysis (using correlation) of the annual streamflow data reveals that the first Principal Component (which represents a weighted average of the flows across sites) accounts for 79% of the variance across the sites. The associated eigenvector has nearly equal, positive weights across all sites, reflecting the mutual correlation (Table 2). Thus the dominant mode of inter-annual evolution of the climate state is seen similarly at all sites. The quality of the reconstructed streamflow data, and hence the signal-to-noise ratio, varies across sites. Consequently, it is useful to analyze the most reliable data (Oros) or with the series for the leading Principal Component. Correlations of the annual Oros inflow with SSTs for two preceding seasons are illustrated in Figure 4. Note the change in the relative importance of the Atlantic and the Pacific predictors as the forecast lead time changes.

[15] Smoothed scatterplots of the Oros flow versus two predictor series are shown in Figure 5. The line in each plot is a cross-validated smoothing spline [*Wahba*, 1990]. The relationships appear to be nonlinear, and heteroskedastic. The corresponding correlations between the Oros inflow and the indices at different lags are provided in Table 3. The Oros correlations are statistically significant at the 5% level, and drop slowly as we increase the forecast lead time. The serial correlation of Oros flow with inflow the previous year is not significant. The EAD index and the NINO3 index do not seem to be correlated.

[16] A wavelet analysis [*Torrance and Compo*, 1998] of the Oros inflow exhibits episodic, multiyear events that are organized on interannual (3–6 year) and interdecadal (12 year) timescales. The NINO3 wavelet spectrum for each season corresponds to the interannual frequency structure and its intermittence, while the EAD spectrum corresponds to the interdecadal mode and its expression. Thus an interdecadal regime in which the EAD persists from year to year in a strongly positive mode may mitigate the impact of an El Nino event (which usually leads to drought). Similarly, a negative EAD regime coupled with an inter-annually persistent El Nino event may be the harbinger of a significant drought in the upcoming year.

4. Forecast Development

[17] The procedure used for developing the forecasts for the six Ceara sites using prior EAD and NINO3 time series is described here. The results from the forecasts are then analyzed. The description of the general algorithm follows.

[18] The main ideas are (1) streamflow at the Ceara sites is highly spatially correlated and is apparently influenced by climate in a similar manner, leading to the possibility of a

Table 3. Correlations Between Oros Annual Inflow and Preceding Season Climate Indices[a]

	EAD OND	Nino3 OND	EAD JAS	Nino3 JAS	EAD AMJ	Nino3 AMJ
EAD OND	1					
Nino3 OND	0.08	11				
EAD JAS	**0.76**	0	11			
Nino3 JAS	0.05	**0.90**	−0.02	11		
EAD AMJ	**0.54**	0.01	**0.83**	0.02	11	
Nino3 AMJ	−0.10	**0.64**	−0.19	**0.74**	**−0.20**	11
OROS	**0.51**	**−0.21**	**0.47**	**−0.20**	**0.33**	**−0.23**

[a]Entries in bold indicate values for which the null hypothesis of zero correlation can be rejected at the 95% significance level.

Table 4. Linear Regression Coefficients for Prior Season NINO3 and EAD for Transformed and Standardized Reservoir Inflows

	January		October		July	
Reservoir	Coefficient EAD	Coefficient Nino3	Coefficient EAD	Coefficient Nino3	Coefficient EAD	Coefficient Nino3
Pacajus	1.14	−0.43	0.76	−0.33	0.39	−0.35
Pacoti-Riachão	1.28	−0.38	0.82	−0.26	0.48	−0.25
Gavião	1.46	−0.39	0.98	−0.30	0.59	−0.31
Pedras Branca	1.18	−0.32	0.87	−0.21	0.45	−0.23
Banabuiu	1.03	−0.19	0.92	−0.13	0.61	−0.16
Oros	1.32	−0.37	1.22	−0.37	0.62	−0.40
Pooled	1.23	−0.35	0.93	−0.27	0.52	−0.28

common, underlying model for all sites; (2) the climate indices are autocorrelated, and are approximately normally distributed, while the annual streamflows are neither, suggesting that an appropriate forecasting framework may be to consider conditioning annual/monthly streamflow on a sequence of past climate index values; (3) use of traditional, parametric statistical methods for building a common regression model at the annual scale and then disaggregating to monthly values may be difficult given the highly skewed distributions of annual and monthly flow, the large number of zero flows, and the nonlinear relationship between flow and the climate indices, and between monthly and annual/seasonal flow; and (4) nonparametric methods for regression and density estimation may also have limited success in direct application, given the high dynamic range of the flow data, the dimension of the multivariate state-space, and the limited amount of available data.

[19] Consequently, we develop a semiparametric approach by decomposing the estimation problem into three parts: (1) the transformation of at site annual streamflows, such that individual regressions on climate indices lead to residuals that are near normally distributed with approximately constant variance; (2) a vector regression model (e.g., pooled regression, principal component regression, or canonical regression) with dimension reduction to develop a common projection of the transformed annual streamflows on to the set of climate indices used as predictors at a given forecast lead time; and (3) use this common projection as the conditioning set for nonparametrically resampling ensembles of historical years (and hence a set of monthly/annual flows at all sites), given current values of the climate indices.

[20] The semiparametric ensemble forecasting approach is described in the context of the longest forecast (July) lead time for the 1914–2000 data for the climate indices and the inflows at the six reservoirs. The predictors considered for the July forecast of the annual flows (January–December) for the coming year were the April–May–June (AMJ) values of the NINO3 and the EAD indices. Recall from Table 3 that the indices are essentially uncorrelated. We reserved contiguous blocks of 5 to 10 years at a time for model validation and the balance for model fitting.

[21] Step 1: Power transformations of the annual flow (January to December) at each of the six sites were considered. The cube root transform provided an approximately symmetric probability distribution for the annual flows $\mathbf{a}_s$, at each site, and for the residuals from a linear regression on the two climate indices that had approximately constant variance, but whose distribution had tails fatter than for the Normal distribution. Linear terms in both indices were selected consistently by a stepwise regression procedure at all sites.

[22] Step 2: For the vector regression problem (six transformed annual flow series conditional on two climate indices) for annual flow forecasting, we considered the following candidate models for the transformed and standardized series ($\mathbf{q}_s = (\mathbf{a}_s^{1/3} - \text{mean}(\mathbf{a}_s^{1/3}))/\text{stdev}(\mathbf{a}_s^{1/3})$, where a_s is the annual flow at site s): A. Separate regressions for each series q_s on the AMJ values of NINO3 and EAD; B. A pooled regression across all series; C. Principal Component Regression; D. Canonical Regression.

[23] Approach A is actually inadmissible, since it neither uses the common information at the sites, nor reproduces the spatial structure in the subsequent forecasts. Of the remaining three, pooled regression (i.e., a common regression equation across all sites) is the most parsimonious if it can be justified. The predictand column of length n1*s for pooled regression is the collection of n1 transformed and standardized flows $\mathbf{q}_s$ at each site s, and the predictor matrix is formed by repeating the block of n1 years of EAD and NINO3 values, s times. The pooled regression ($\mathbf{q} = \mathbf{X}\beta + \mathbf{e}$) was not found to be different from the six regressions (see Table 4) for individual sites at the 95% significance level using the Chow test [*Chow*, 1960]. Consequently, only the pooled regression results, rather than those for all four approaches are discussed here.

[24] Step 3: To generate an ensemble forecast for annual flows, we need estimates of the conditional probability distributions $f(\mathbf{a}_{t+f,\,s}|\mathbf{x}_t)$ of a vector of s annual streamflow values, $\mathbf{a}_{t+f,\,s}$. This could be estimated assuming a model (e.g., normal) for the probability distribution of the residuals from the pooled regression, and then transforming back to the original space of the streamflow data. This led to a marginal density function for the $\mathbf{a}_s$ that invariably had significant density for negative flows, and to density functions that did not look like the original data if these density functions were truncated at zero. It was also difficult to preserve the spatial correlation structure in real space across sites after back transformation. For monthly streamflow values at each site, one would also need to estimate the conditional probability distributions $f(\mathbf{m}_{t+f}|\,\mathbf{a}_{t+f},\,\mathbf{x}_t)$, $f(\mathbf{m}_{t+f}|\,\mathbf{a}_{t+f})$ or $f(\mathbf{m}_{t+f}|\,\mathbf{x}_t)$, as appropriate for the data. Disaggregation of the annual streamflows to monthly flows [e.g., *Bras and Rodriguez-Iturbe*, 1993, sect. 3.5] while preserving spatial and temporal summability could be considered. Given the issues with generating the ensemble forecast, and the interest in resampling historical data to match the operational practice, we took a nonparametric approach

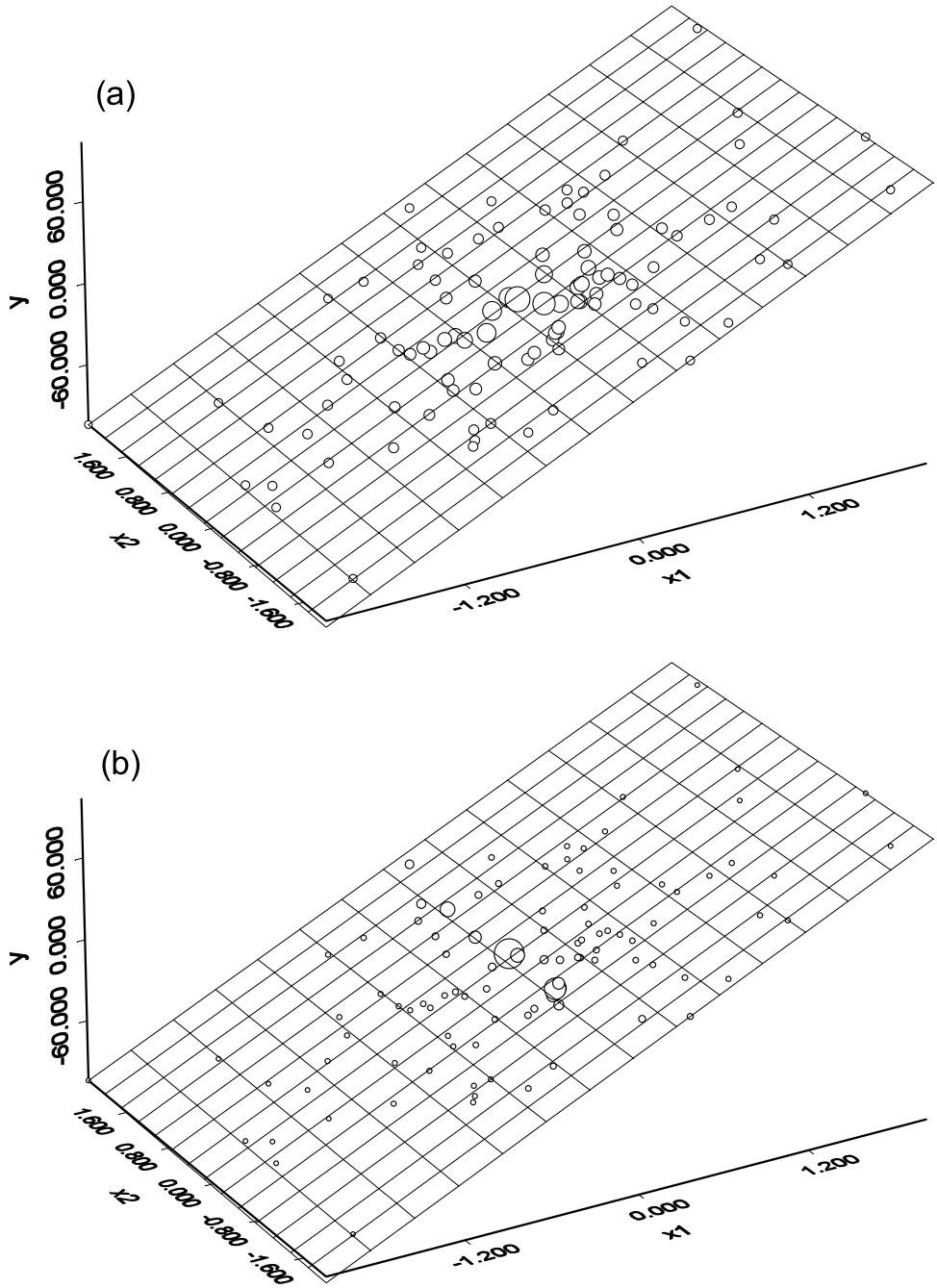


Figure 6. The effect of distance metric on neighbors selected for resampling. The underlying model here is y = 50x1 + x2 + e, and x1 = x2 = 0 is the point about which we seek to resample. The larger the circle, the closer the neighbor using (a) $d_i = (\mathbf{x}^* - \mathbf{x}_i)^T(\mathbf{x}^* - \mathbf{x}_i)$ and (b) a weighted distance $d_i = \{(\mathbf{x}^* - \mathbf{x}_i)\gamma\}^T\{(\mathbf{x}^* - \mathbf{x}_i)\gamma\}$, where the weights $\gamma = [50, 1]$. Figure 6b is equivalent to choosing Euclidean distances with a rescaling that reflects the relative linear importance of each predictor. Note the change in the neighbors identified, favoring the more important predictor. In the examples here the linear model is applied to the parameterically transformed streamflow data, implicitly reflecting a more complex weighting of the coordinates in real space.

from this point. The k-nearest neighbor density estimation approach to time series described by *Lall and Sharma* [1996] and *Karlsson and Yakowitz* [1987] is adapted to the current setting.

[25] Given a current p*1 vector of predictors $\mathbf{x}^*$, we seek to conditionally resample a vector $\mathbf{a}_{t+f}$ of annual flows and the corresponding vector $\mathbf{m}_{t+f}$ of monthly flows to implicitly reflect the conditional probability distribution $f(\mathbf{M}|\mathbf{x})$. The basic strategy is to select the k-nearest neighbors of $\mathbf{x}^*$ in the historical data set $\mathbf{X}$, estimate appropriate weights or probabilities to assign to each of these neighbors, and then resample the corresponding vector(s) $\mathbf{M}$, given the estimated probabilities. For example, suppose the only predictor was NINO3, and we wish to issue a forecast for

the next year in January using the data for NINO3 for AMJ. Let's say that the latest AMJ value of NINO3 is 2.5. Then one would locate k (e.g., 30) neighbors as the k historical years with the closest values of NINO3. Then, probabilities are assigned to each of these k years based on their "closeness" to a NINO3 value of 2.5. An entire year's (starting with the following January) sequence of monthly flows (**M**) at all sites, are then resampled using these probabilities to select years. This amounts to selectively drawing historical years as a scenario for reservoir operation, rather than drawing them at random (or unconditionally). The key parameters of the algorithm are the number of neighbors, k, to use, the selection of a metric to define "closeness" in predictor space, particularly in the multivariate context, and the probability weight function. In the context defined here this procedure can be stated as follows.

[26] 1. Compute distances d_i between the current predictor vector $\mathbf{x}^*$ and the historical state vectors, $\mathbf{x}_i$, as:

$$di = \{(\mathbf{x}^* - \mathbf{x}_i)\gamma\}^T\{(\mathbf{x}^* - \mathbf{x}_i)\gamma\} \quad (1)$$

where $\mathbf{x}^*$ is a 1*p vector, $\mathbf{x}_i$ is a 1*p vector of predictors for the ith year used in model fitting, and γ is a p*1 vector. For the pooled regression with 2 predictors, $\boldsymbol{\gamma} = [\beta_1\ \beta_2]$, where β_i is the ith regression coefficient of the pooled regression of the standardized and transformed flows $\mathbf{q}$ on the climate indices EAD and NINO3 respectively.

[27] The distances record the similarity of the current predictor condition to each of the past conditions. If one directly uses the Euclidean distance between the current predictor vector and the historical vectors, the relative importance of the components of the predictor vector in determining the future state of the predictand is not used. The parametric variable selection, transformation and regression procedure used in the earlier steps is used here to develop "weights" for each component of the predictor matrix that would yield a good parametric regression of the predictand on to the predictors. Consequently, a weighted Euclidean distance is used to define similarity for selecting the k-nearest neighbors, and to "transfer" the knowledge from the parametric, multivariate regression of the annual flows on to the potential predictors. The difference between conditioning on the original and the rescaled predictor space is illustrated in Figure 6.

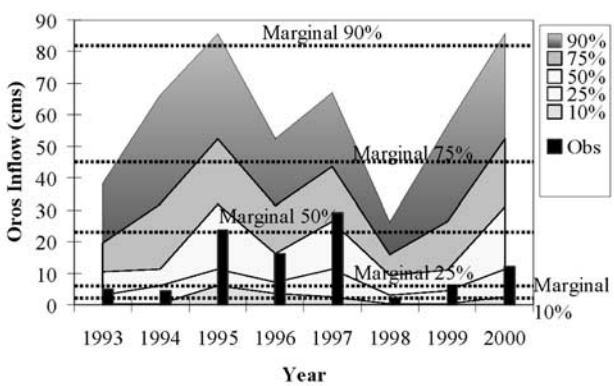


Figure 7. Probabilistic forecasts of 1993–2000 January–December annual inflow into Oros from the preceding July. The 1914–1991 data were used for model fitting. The vertical bars depict the observed values. The shaded areas provide the 10th, 25th, 50th, 75th and 90th percentiles of the k nearest neighbor ensemble forecasts. The dashed lines provide the marginal distribution percentiles. The correlation between the median forecast and the observed values is 0.91.

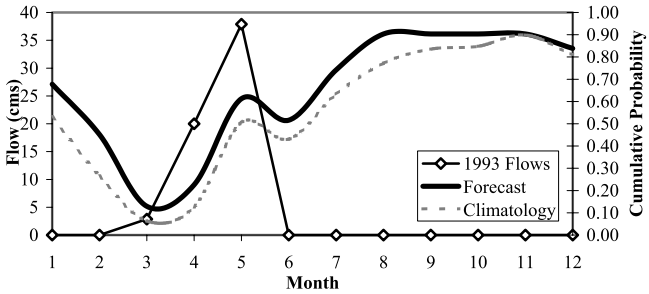


Figure 8. Cumulative distribution function (cdf) of monthly ensemble forecasts for 1993 from July 1992. The observed monthly flows are shown with diamonds, and the cdfs of the monthly k nearest neighbor ensemble and climatology are shown as solid and dashed lines, respectively.

[28] 2. Using the distance vector **d** computed in the previous step, identify the ordered set of nearest neighbor indices **J**. The jth element of this set records the year t associated with the jth closest $\mathbf{x}_i$ to $\mathbf{x}^*$. If the data is highly quantized, it is possible that a number of observations may be the same distance from the conditioning point. The resampling kernel defined in step 3 is based on the order of elements in **J**. Where a number of observations are the same distance away, the original ordering of the data can impact the ordering in **J**. To avoid such artifacts, we copy the time indices t into a temporary array that is randomly permuted prior to distance calculations and creation of the list **J**.

[29] 3. Now, select the number of neighbors to use (k) and the resampling kernel or weight function K(j). Choices for the kernel include

$$\textit{LallandSharma} \qquad K(j) = \frac{1/j}{\sum_{i=1}^{k} 1/i}$$

$$\textit{Uniform} \qquad K(j) = 1/k$$

$$\textit{Power} \qquad K(j) = \eta\left(d_j + \delta\right)^{-\alpha}$$

Different combinations of k and K(j) can give similar results, recognizing the trade-off between bandwidth choice and kernel properties discussed by *Hardle* [1991].

[30] 4. The forecast flow matrix is then resampled using the kernel K(j). If the jth element is drawn from the kernel, the corresponding year is identified from **J**, and the forecast is the set of annual and monthly flows at all s sites for that year. This process is repeated to generate the desired number of ensemble forecasts of $\mathbf{a}_{t+f}$ and $\mathbf{m}_{t+f}$.

4.1. Results

[31] We explored values of k ranging from 10 to 30, and the first two kernels indicated above. The differences across kernels are minor, and the median forecast is very similar

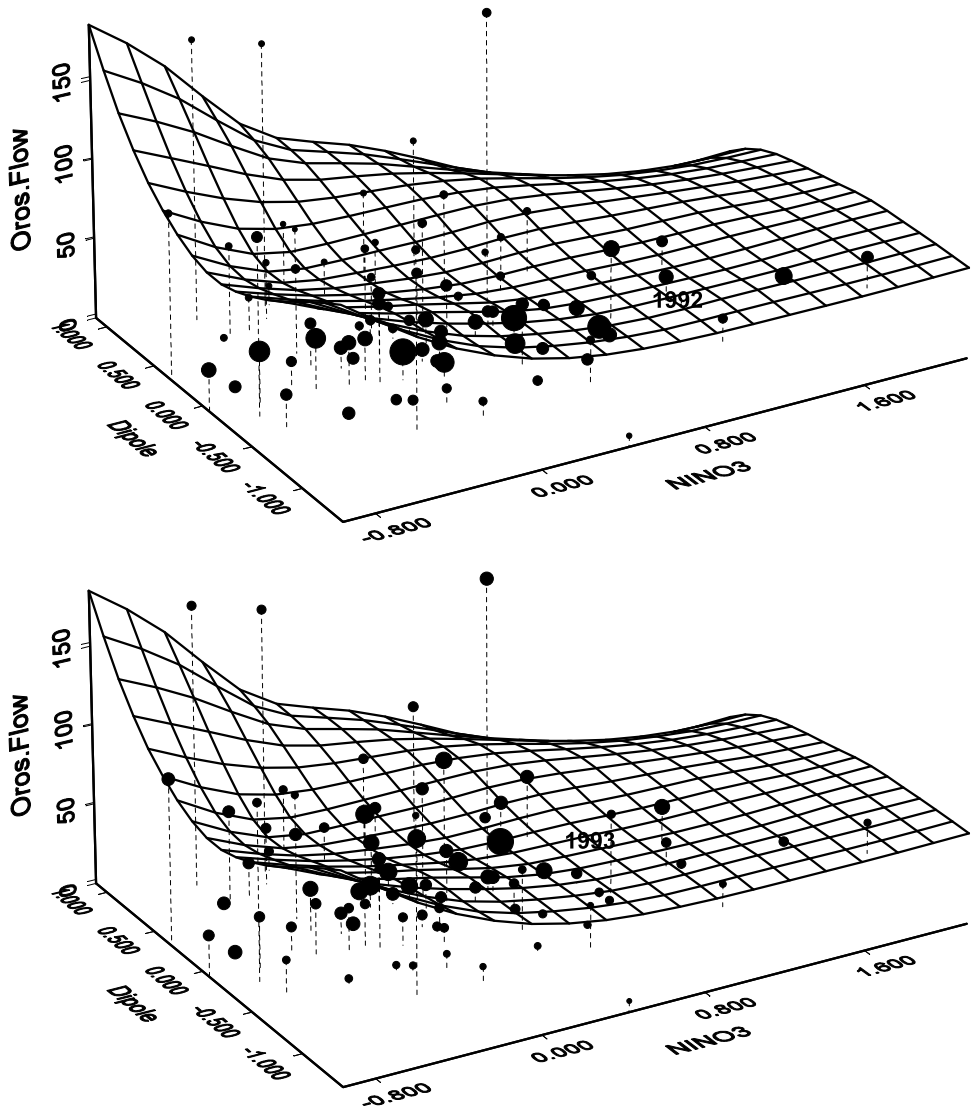


Figure 9. Neighbors selected for k nearest neighbor ensembles following pooled regression of transformed and standardized annual inflows on EAD and NINO3 for (a) July 1992 forecast for 1993 and (b) for July 1993 forecast for 1994. The location of the observed values for July 1992 and 1993 is marked in each figure. The size of the ball indicates the similarity of conditions. Note the more diffuse situation in July 1993. These differences translate into the more diffuse conditional probability distribution for 1994 seen in Figure 7.

with 20 or 30 neighbors. Results for the July forecasts for Oros inflows, using 30 nearest neighbors, and the uniform kernel are presented in Figures 7 and 8 for the recent part of the data reserved for model validation. Quantiles for annual and the monthly flow for each year at Oros are computed from the k-nearest neighbors and are used to compare the forecast with the observations for the years 1993–2000 (not used in model fitting). The "climatology" quantiles or the quantiles of the marginal distribution of the corresponding flows are also computed. From Figure 7 we see that the forecast median is usually closer to the observation than climatology, and the forecast quantile spread is generally smaller than that for climatology. The correlation of the median forecast with the observations is 0.91 over these 8 years.

[32] The 1993 and 1998 drought years are especially well marked, while the forecast for the 1994 and 2000 years is more diffuse. The difference between the situation in 1993 and 1994 is explored in Figure 9. For AMJ 1992, the values of the EAD and NINO3 are −0.2 and 1.32 respectively, indicating moderate El Nino conditions in the Pacific, and that the North Atlantic area close to Ceara is colder than the South Atlantic area close to Africa. Neighbors of these conditions in the historical years using the weighted Euclidean metric indicated earlier are shown in Figure 9 (top). The size of the symbol used to plot the Oros inflow indicates the similarity of the July conditions to those in 1992. The nearest neighbors indicate dry conditions, leading to the relatively tight 1993 ensemble forecast for dry conditions.

[33] In July 1993, the EAD and NINO3 values were 0.31 and 1.13 respectively, indicating that the moderate El Nino conditions persisted over the year, but the EAD has changed its sign. From the regression equation and from Figure 9

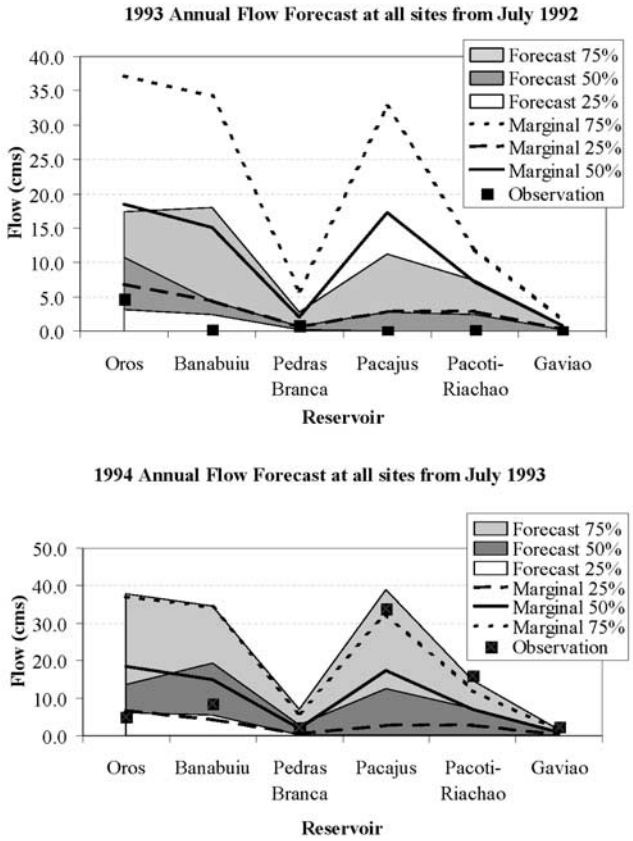


Figure 10. Annual flow forecasts at all sites for (a) 1993 and (b) 1994. All sites were extremely dry in 1993, and 75% of the forecast is consistently at or below the median for climatology. There is considerable variability across sites in 1994, and spatially the forecast is not too different from a climatology forecast, even though it seems to provide slightly better coverage.

(bottom), we can see that a positive EAD for a fixed NINO3 condition suggests wetter conditions. The change in the structure of the nearest neighbors of the July climate conditions translates into a change in the possibilities for 1994. For example, 1924, which was a very wet year, now shows up as a near neighbor. Thus, while 1994 was nearly as dry as 1993, the forecast reflects the possibility that it could have been much wetter. Indeed, 1994 was an anomalous year in that while Oros was dry, inflows to the other reservoirs were at or above the median. Oros is at the southern extremity of the region, and the associated river basin was just beyond the influence of the major storms in 1994.

[34] The k-nearest neighbor conditioned ensembles for monthly flow at all sites represent the full year of monthly flow for each ensemble member. A reservoir operator would use each such ensemble directly as a supply scenario. Potential changes in seasonality of inflows, and the spatial structure of inflows would thus be directly accounted for. The 1993 forecast for Oros inflows relative to the subsequently observed monthly inflow sequence is presented in Figure 8. Note that 1993 inflows into Oros were nonzero only for March through May. Recall from Figure 3 that the median inflow is usually the highest for the months of March and April and for above median inflow conditions, May is on the recession limb of the hydrograph. The 1993 hydrograph peak in May is consequently anomalous and representative of dry conditions in which there is a failure of the rainfall systems that bring moisture in the early part of the wet season. For each of the months, we compute the empirical probability that the inflow would be less than or equal to that observed from the historical data, and from the forecast ensemble. For 1993, the observed annual flow corresponds to about the 20th percentile for climatology and the 30th percentile for the forecast (Figure 9, top). The historical data for January, February and for June through December contains a lot of zeros. Hence the higher cumulative probability for a zero inflow in the forecast for these months, relative to climatology, indicates confidence in drier then average conditions.

[35] The spatial expression of the forecasts across the six reservoir inflow sites is presented in Figure 10. The 1993 Oros forecasts are drier than climatology for all sites, and the observations are consistently dry across sites, ranging from below the 25% to 50% of the forecast, and at or below the 25% of climatology. The inter-quartile range (75%−25%) of the forecast is consistently smaller than that of climatology for the 1993 forecast. Thus a decision to operate as if in a drought across the region would be indicated by the forecast and would then be borne out by the subsequent experience.

[36] The situation in 1994 is different. As we noted earlier, Oros was drier than normal, but with a larger spread for the inflow than in 1993. The spatial forecast scenarios, suggest that the forecast may not be too different from a climatology forecast, with perhaps a slight chance for being wetter in the north and drier in the south. The reservoirs are ordered approximately from south to north in the Figure 10. The interquartile range of the forecast is generally comparable to or larger than the interquartile range of climatology.

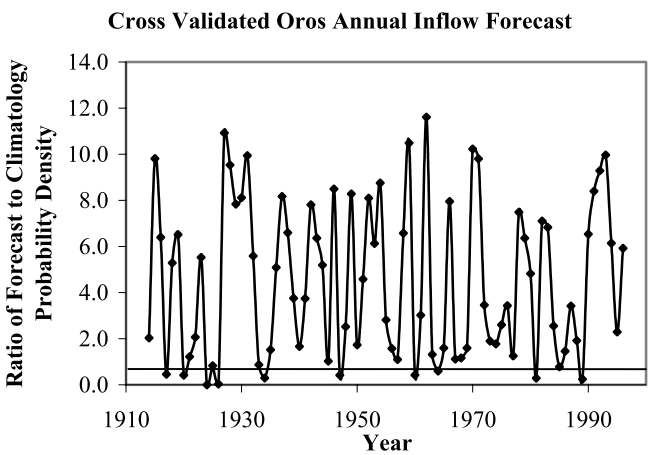


Figure 11. The ratio of the probability density of the leave one out forecast evaluated at the observation that was left out, to the probability density of the marginal distribution of flows (“climatology”) for the 1914 to 1996 period. The probability densities were estimated using a kernel density estimator (kde) with a biweight kernel. The kde was applied to the raw data and to data simulated from the forecast density for each year. The normalized likelihood ratio for the entire period is 2.89.

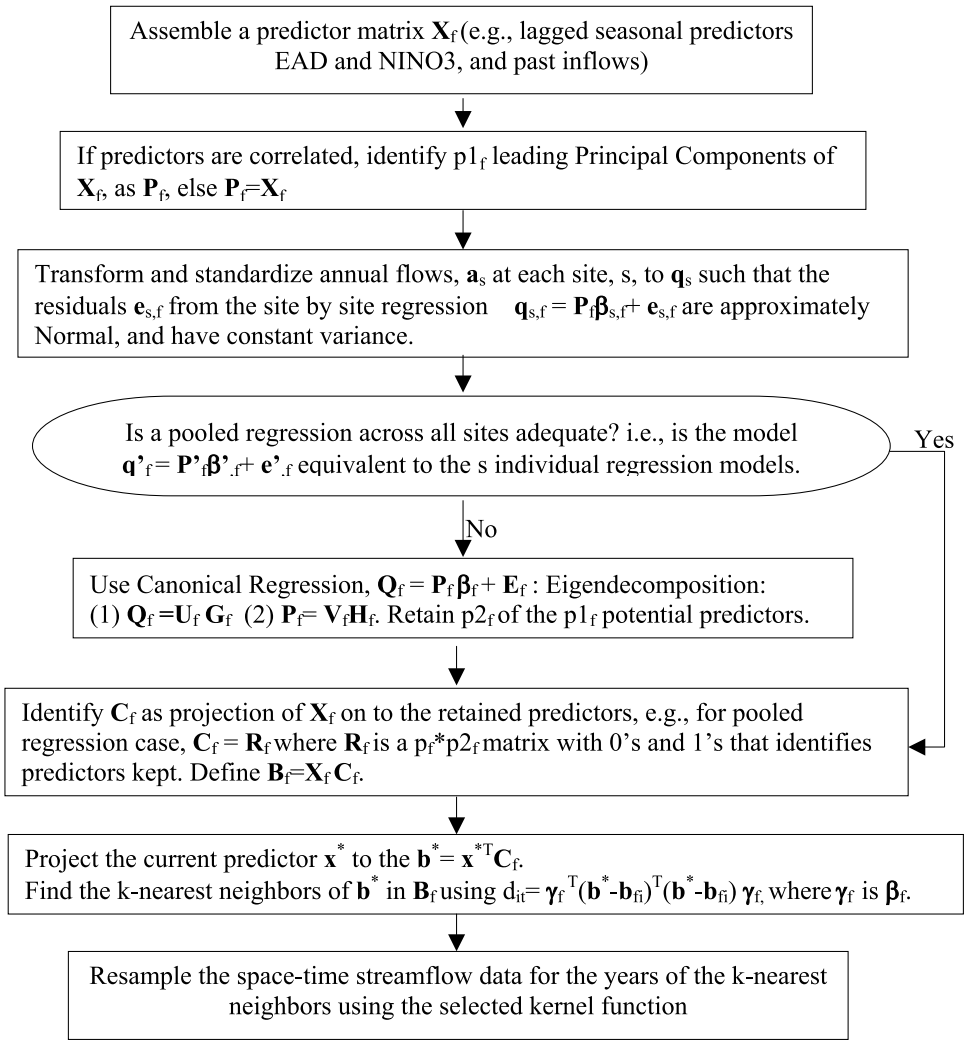


Figure 12. Flowchart for a general version of the semiparametric, multivariate forecasting algorithm for annual and monthly flow ensemble generation given a set of climate predictors.

The observations for 1994 are generally consistent with this interpretation of the forecast. Given the broader range and the spatial spread, a water system operator would likely have hedged any bets on the forecast. Recall that long term inflows at all sites are positively correlated with each other, and an year in which there is spatial opposition in the departure from the median is consequently unusual.

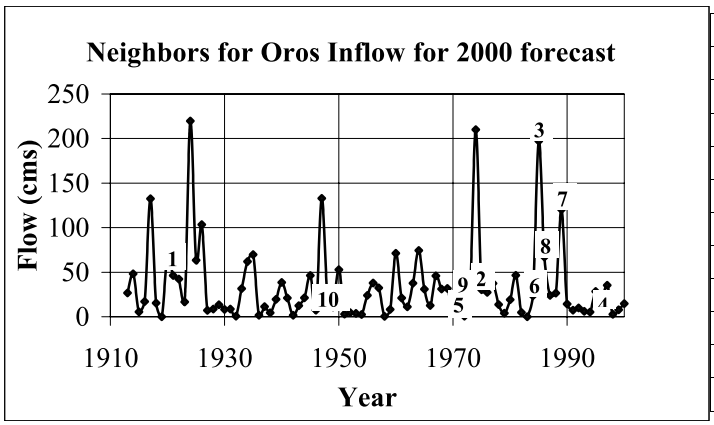
[37] An assessment of the algorithm applied in a leave one out cross-validation mode, relative to a climatology forecast (i.e., using the marginal distribution of flows) is provided in Figure 11. Kernel density estimates using the biweight kernel were used to estimate the marginal probability distribution of annual flow at Oros, and of the cross-validated ensemble forecast for each year. In the latter case, a sample was simulated from the ensemble and the kernel density estimate developed from the sample to simulate the manner in which the information may be used. The normalized likelihood ratio (LR) of the two methods is defined below. It represents the average ratio of the likelihood of the forecast being superior to climatology in a given year. Figure 11 illustrates the variation in this ratio from year to year. Note that the ratio for 1994 is lower than that for 1993, as we may expect in light of the preceding discussion.

$$LR = \left(\frac{\prod_{i=1}^{n} f_F(x_i)}{\prod_{i=1}^{n} f_C(x_i)} \right)^{1/n} \tag{2}$$

where $f_F(x_i) = \frac{\sum_{j=1}^{n} K(u_{ij})}{nh}$, $f_C(x_i) = \frac{\sum_{j=1}^{ns} K(v_{ij})}{n_s h}$, $K(w) = \frac{15}{16}(1-w^2)^2$, $u_{ij} = \frac{x_i - f_j}{h}$, $v_{ij} = \frac{x_i - x_j}{h}$, f_j is the jth forecast ensemble member, and n and n_s are the sample sizes for the historical record, and for the forecast ensemble, respectively.

4.2. General Forecasting Procedure

[38] A procedure for semiparametric forecasting of annual and monthly streamflow at multiple sites that generalizes the presentation for the July forecast using two predictors is presented in Figure 12. The predictor matrix can include polynomial terms in the primary predictors. A Principal Component Analysis is indicated in the second step if the



Index	Year	Weight
1	1921	14%
2	1975	9%
3	1985	6%
4	1996	6%
5	1971	5%
6	1984	5%
7	1989	5%
8	1986	3%
9	1973	3%
10	1949	3%
Sum		59%

Figure 13. Example of forecast presentation: 2000 annual inflow forecast from July 1999. The top 10 of 30 neighbors are identified on a plot of the historical time series and the weight associated with each based on the *Lall and Sharma* [1996] kernel is indicated in the accompanying table.

predictors are mutually correlated. The individual and pooled regressions are then applied to the leading orthogonal vectors or PCs of the eigenvalue decomposition of the predictor matrix. In case the pooled regression does not adequately represent the regressions at the individual sites, a canonical regression of the multisite transformed and standardized annual flows on the predictor matrix is indicated. We retain a subset of potential predictors and record the projections of the original data matrix that were used in the final multivariate, parametric regression model. The current predictor vector is then mapped to this reduced space projection and distances are found to historical reduced space predictor vectors, with scaling by the inverse of the corresponding regression coefficients as presented earlier.

[39] This general strategy was tested with the Ceara data sets for forecasts from the three lead times indicated earlier. Comparative results are available from the authors. The pooled regression approach was typically adequate, leading to the simpler formulation presented earlier.

5. Summary and Discussion

[40] Climatic factors associated with variations in water supply in Ceara, Brazil were reviewed, and some diagnostic analyses were pursued to assess the nature of teleconnections between the Atlantic and the Pacific and streamflow in the region. A multivariate, semiparametric algorithm for resampling historical annual and monthly streamflows conditional on climatic predictors was introduced, and results for the applications of a subset of the algorithm were presented. The feasibility of up to 18 month ahead forecasts of streamflow at a collection of six sites was demonstrated. Examples of forecast performance in different situations were reviewed using default parameters as to the number of nearest neighbors, without an attempt to tune them to each forecast situation. We chose to present results from the most parsimonious form of the model described to focus on the communication of the basic ideas, rather than the complexity of choice offered by the approach Results for the verification period presented here (1993–2000) are representative of other blocks of similar length that were reserved for testing. The correlation of the median forecast with the observed annual flows is consistently high (0.9) for the period reported here. The disaggregated monthly and reservoir forecasts are also informative. Like other nonparametric methods, the k-nearest neighbor approach leads to some biases in the estimation of complex underlying relationships from finite data sets. These biases are evident in the asymmetry of coverage of the highly skewed flow data, particularly in the extremely dry years. Details of methods and other applications are available from the authors.

[41] The semiparametric method used here allows one to tailor forecasts to different user groups. One can present the cumulative distribution function plots as in Figures 7 and 8, illustrate the relationship with predictors and neighbors as in Figure 9, provide traces of monthly streamflow at each site as done in the ESP (ensemble streamflow prediction) process, or present a diagram (Figure 13) showing the neighbor years and their relative weights. We've found each of these methods effective individually and in combination.

[42] Improvements of the method presented here to consider combinations of forecasts from different methods, and parameter uncertainty are underway. Reservoir optimization models have been linked to the forecast methodology and the use of the two tools for operational decision making has been demonstrated to the Ceara Water agencies. Refinements to make the tools directly relevant to a stakeholder driven decision process are underway.

[43] **Acknowledgments.** We gratefully acknowledge discussions with Antonio Divino Moura, Yochanan Kushnir, Steve Zebiak, Andrew Gelman and others who educated us on various aspects of the hydrology and climate of Ceara and the attributes of statistical approaches to the problem.

References

Bras, R., and I. Rodriguez-Iturbe, *Random Functions and Hydrology*, 559 pp., Dover, Mineola, N.Y., 1993.

Chiang, J. C. H., Y. Kushnir, and S. E. Zebiak, Interdecadal changes in eastern Pacific ITCZ variability and its influence on the Atlantic ITCZ, *Geophys. Res. Lett.*, *27*(22), 3687–3690, 2000.

Chow, G., Tests of equality between sets of coefficients in two linear regressions, *Econometrica*, *28*, 591–605, 1960.

Companhia de Gestão dos Recursos Hídricos (COGERH), Plano de Gerenciamento da Bacia do Jaguaribe, Fortaleza, Ceará, Brazil, 1999a.

Companhia de Gestão dos Recursos Hídricos (COGERH), Plano de Gerenciamento das Bacias Metropolitanas, Fortaleza, Ceará, Brazil, 1999b.

Cordery, I., and M. A. McCall, A model for forecasting drought from teleconnections, *Water Resour. Res.*, *36*(3), 763–768, 2000.

Hastenrath, S., Predictability of northeast Brazil droughts, *Nature*, *307*, 531–533, 1984.

Hastenrath, S., Prediction of northeast rainfall anomalies, *J. Atmos. Sci.*, *35*, 2222–2231, 1990.

Kaplan, A., M. Cane, Y. Kushnir, A. Clement, M. Blumenthal, and B. Rajagopalan, Analyses of global sea surface temperature 1856–1991, *J. Geophys. Res.*, *103*, 18,567–18,589, 1998.

Karlsson, M., and S. Yakowitz, Rainfall-runoff forecasting methods, old and new, *Stochastic Hydrol. Hydraul.*, *1*, 303–318, 1987.

Kendall, D. R., and J. A. Dracup, A comparison of index-sequential and AR(1) generated hydrological sequences, *J. Hydrol.*, *122*, 335–352, 1991.

Kousky, V. E., Frontal influences on northeast Brazil, *Mon. Weather Rev.*, *107*, 1140–1153, 1979.

Kumar, D. N., U. Lall, and M. Peterson, Multi-site disaggregation of monthly to daily streamflow, *Water Res. Res.*, *36*(7), 1823–1834, 2000.

Lall, U., and A. Sharma, A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resour. Res.*, *32*(3), 679–693, 1996.

Liu, Z., J. B. Valdes, and D. Entekhabi, Merging and error analysis of regional hydrometeorologic anomaly forecasts conditioned on climate precursors, *Water Resour. Res.*, *34*(8), 1959–1969, 1998.

Marengo, J. A., J. Tomasella, and C. R. Uvo, Trends in streamflow and rainfall in tropical South America: Amazonia, eastern Brazil, and northwestern Peru, *J. Geophys. Res.*, *103*, 1775–1783, 1998.

Markham, C. G., and D. R. McLain, Sea surface temperatures related to rain in Ceara, northeast Brazil, *Nature*, *265*, 320–323, 1977.

Moura, A. D., and J. Shukla, On the dynamics of droughts in northeast Brazil: Observation, theory and numerical experiments with a general circulation model, *J. Atmos. Sci.*, *38*, 2653–2675, 1981.

Nobre, P., and J. Shukla, Variations of sea surface temperature, wind stress, and rainfall over the tropical Atlantic and South America, *J. Clim.*, *9*(10), 2464–2479, 1996.

Piechota, T. C., F. H. S. Chiew, J. A. Dracup, and T. A. McMahon, Development of an exceedance probability streamflow forecast, *J. Hydrol. Eng.*, *6*(1), 20–28, 2001.

Rajagopalan, B., U. Lall, and S. E. Zebiak, Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles, *Mon. Weather Rev.*, *130*(7), 1792–1811, 2002.

Saravanan, R., and P. Chang, Interaction between tropical Atlantic variability and El Nino-Southern Oscillation, *J. Clim.*, *13*(13), 2177–2194, 2000.

Secretaria de Recursos Hídricos do Estado do Ceará (SRH), Plano Estadual de Recursos Hídricos, Fortaleza, Ceará, Brazil, 1991.

Torrance, C., and G. P. Compo, A practical guide to wavelet analysis, *Bull. Am. Meteorol. Soc.*, *79*, 61–78, 1998.

Uvo, C. B., and N. E. Graham, Seasonal runoff forecast for northern South America: A statistical model, *Water Resour. Res.*, *34*(12), 3515–3524, 1998.

Uvo, C. B., C. A. Repelli, S. E. Zebiak, and Y. Kushnir, The relationship between tropical Pacific and Atlantic SST and northeast Brazil monthly precipitation, *J. Clim.*, *11*(4), 551–562, 1998.

Uvo, C. B., U. Tolle, and R. Berndtsson, Forecasting discharge in Amazonia using artificial neural networks, *Int. J. Climatol.*, *20*, 1495–1507, 2000.

Wahba, G., *Spline Methods for Observational Data*, 169 pp., Soc. for Indust. and Appl. Math., Philadelphia, Pa., 1990.

Ward, M. N., and C. K. Folland, Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperatures, *Int. J. Climatol.*, *11*, 711–743, 1991.

Ward, M. N., S. Brooks, and C. K. Folland, Predictability of the Seasonal rainfall in the northern Nordeste Region of Brazil, in *Recent Climate Change*, edited by S. Gregory, pp. 237–251, Belhaven, London, 1988.

Ward, M. N., C. K. Folland, K. Maskell, A. W. Colman, D. P. Rowell, and K. B. Lane, Experimental seasonal forecasting of tropical rainfall at the Meteorological Office, U.K., in *Prediction of Interannual Climate Variations*, edited by J. Shukla, pp.192–216, Springer-Verlag, New York, 1993.

F. A. Souza Filho, Fundação Cearense de Meteorologia e Recursos Hídricos (FUNCEME), 60115-221 Fortaleza, Ceara, Brazil.

U. Lall, Department of Earth and Environmental Engineering and International Research Institute for Climate Prediction, Columbia University, Mail Code 4711, 500 West 120th Street, New York, NY 10027, USA. (ula2@columbia.edu)