ELSEVIER

Water Research 39 (2005) 2621–2635

WATER RESEARCH

www.elsevier.com/locate/watres

Evaluation of river water quality monitoring stations by principal component analysis

Ying Ouyang*

Department of Water Resources, St. Johns River Water Management District, Reid Street, Palatka, Florida 32178-1429, USA

Received 10 August 2004; received in revised form 18 March 2005; accepted 4 April 2005

Abstract

The development of a surface water monitoring network is a critical element in the assessment, restoration, and protection of stream water quality. This study applied principal component analysis (PCA) and principal factor analysis (PFA) techniques to evaluate the effectiveness of the surface water quality-monitoring network in a river where the evaluated variables are monitoring stations. The objective was to identify monitoring stations that are important in assessing annual variations of river water quality. Twenty-two stations used for monitoring physical, chemical, and biological parameters, located at the main stem of the lower St. Johns River in Florida, USA, were selected for the purpose of this study. Results show that 3 monitoring stations were identified as less important in explaining the annual variance of the data set, and therefore could be the non-principal stations. In addition, the PFA technique was also employed to identify important water quality parameters. Results reveal that total organic carbon, dissolved organic carbon, total nitrogen, dissolved nitrate and nitrite, orthophosphate, alkalinity, salinity, Mg, and Ca were the parameters that are most important in assessing variations of water quality in the river. This study suggests that PCA and PFA techniques are useful tools for identification of important surface water quality monitoring stations and parameters.

Keywords: Monitoring network; Principal component analysis; Surface water quality

1. Introduction

Pollution of surface water with toxic chemicals and excess nutrients, resulting from storm water runoff, vadose zone leaching, and groundwater discharges, has been an issue of worldwide environmental concern. With an increased understanding of the importance of drinking water quality to public health and raw water quality to aquatic life, there is a great need to assess surface water quality. This is true for the lower St. Johns River (LSJR), located in Florida, USA. Pollution of the LSJR with contaminants such as nutrients, hydrocarbons, pesticides, and heavy metals comes from both point and non-point sources. These sources are the results of surface runoff generated from urban, rural, and agricultural lands; discharge from ditches and creeks; groundwater seepage from malfunctioning septic tank systems; aquatic weed control and naturally occurring organic inputs; and atmospheric deposition. The degradation of water quality due to these contaminants has resulted in altered species composition and decreased the overall health of aquatic communities within the river basin (Campbell et al., 1993; Durell et al., 2001; Ouyang et al., 2002).

*Tel.: +1 386 312 2320; fax: +1 386 329 4585.
E-mail address: youyang@sjrwmd.com.

0043-1354/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.watres.2005.04.024

In 1987, the Florida State Legislature enacted the Surface Water Improvement and Management (SWIM) Act, which identified the LSJR as a water body of regional significance and in need of restoration and management. Since then, many efforts have been devoted to restoring and protecting the LSJR. While the control of surface water runoff, ditch discharge, vadose zone leaching, groundwater seepage, and atmospheric deposition is necessary to reduce point and non-point source pollution of the river, development of an effective surface water monitoring network is a critical element in these restoration and protection efforts. Surface water quality monitoring within the LSJR has been conducted by various agencies at varying levels of intensity since 1956. The primary objectives are to identify water quality problems, describe seasonal and spatial trends for developing qualitative and quantitative models of the riverine ecosystem, and determine permit compliance. Since its inception, the monitoring network has become one of the most critical efforts in assessment of surface water pollution in the LSJR and has been a significant resource for others working to prevent pollution of the river (Campbell et al., 1993). However, efforts to determine the effectiveness and efficiency of the monitoring network are still warranted. To this end, the principal component analysis (PCA) and principal factor analysis (PFA) techniques were employed in this study.

PCA and PFA are multivariate statistical techniques used to identify important components or factors that explain most of the variances of a system. They are designed to reduce the number of variables to a small number of indices (i.e., principal components or factors) while attempting to preserve the relationships present in the original data. The problems of indicator parameter or import monitoring station identification, data reduction and interpretation, and characteristic change in water quality parameters can be approached through the use of the PCA and PFA. Details for mastering the arts of PCA and PFA are published elsewhere (Manly, 1986; Davis, 1986; Wackernagel, 1995; Tabachnick and Fidell, 2001).

In recent years, the PCA and PCF techniques have been applied to a variety of environmental issues, including evaluation of ground water monitoring wells, interpretation of groundwater hydrographs, examination of spatial and temporal patterns of heavy metal contamination and identification of herbicide species related to hydrological conditions. Some examples of PCA and PCF applications in environmental practices are described below.

Measurements of water level in wells are a routine part of groundwater studies. Recently, Winter et al. (2000) applied the PCA and PCF techniques to investigate the areal distribution of various types of water level fluctuation patterns within a study area and to determine if fewer wells could be measured while still achieving effective long-term monitoring goals at four small, lake-watershed research sites in the USA. These authors found that the PCA technique was very useful in summarizing information from large data sets to select long-term monitoring wells, which would greatly reduce the cost of monitoring programs.

Gangopadhyay et al. (2001) applied the PCA and PCF techniques to identify monitoring wells important in predicting the dynamic variations in potentiometric head at a location in Bangkok, Thailand. Through the years, the groundwater monitoring networks in the area have expanded tremendously, and many networks today consist of dozens, if not hundreds, of sampling wells. These authors argued that at a certain stage, municipalities have to justify their groundwater monitoring networks and ask questions such as how sampling from a particular well can help explain the dynamic variations of potentiometric head in the aquifer, and which subset of observation wells should be selected to continue monitoring in the near future for a municipality facing budget constraints. To answer these questions, the authors performed PCA on all of the monitoring wells, and developed a ranking scheme based on the frequency of occurrence of a particular well as a principal well. Based on the study results, the decision maker with budget constraints can now opt to monitor only the principal wells and still adequately capture the potentiometric head variations in the aquifer.

Additionally, the PCA technique has been used to estimate spatial and temporal patterns of heavy metal contamination (Shine et al., 1995); to investigate nutrient gradients within a eutrophic reservoir (Perkins and Underwood, 2000); and to identify the major herbicide compositions causing the observed data variations (Tauler et al., 2000). These studies have provided good examples of the effective application of PCA. However, there are few documented examples of the evaluation of the highly dynamic and complex surface water quality monitoring networks in river systems using the PCA or PFA technique.

The aims of this study are to demonstrate the application of these novel data reduction techniques (i.e., the PCA and PFA techniques) to evaluate the potential for reducing the number of ambient water quality monitoring stations located in the main stem of the LSJR for long-term monitoring purposes and to evaluate the importance of various water quality parameters. The specific objectives are to: (1) present detailed procedures on how to interpret PCA and PFA results, (2) identify the non-principal surface water quality monitoring stations, and (3) extract the parameters that are most important in assessing variations in river water quality.

2. Study area

The LSJR basin is located in northeast Florida, USA between 29° and 30° north and between 81.13° and 82.13° west (Fig. 1). It is an area of approximately 2777 square miles. The LSJR is a sixth order, dark-water river estuary, and exhibits characteristics associated with riverine, lacustrine, and estuarine environments. The land uses within the basin largely consist of forested, residential, commercial, industrial, agricultural, wetland, barren land, and water. A number of water quality problems have been identified and addressed since the 1950s including the discharge of point and non-point source pollutants such as nutrients, hydrocarbons, pesticides, and heavy metals (Campbell et al., 1993; Durell et al., 2001).

As part of their commitment to enhance water quality monitoring efforts, the US Environmental Protection Agency (EPA), the Florida State Department of Environmental Protection, the St. Johns River Water Management District (SJRWMD) and other agencies formed a body known as the Integrated Water Resource Monitoring (IWRM) committee in late 1996. The committee was charged with developing strategies and techniques for implementing an integrated monitoring plan that combines surface water, groundwater, and biological monitoring. In 1997, the IWRM committee decided that the most efficient form of monitoring could be achieved by establishing the following three-tiered monitoring network: (1) status monitoring, (2) basin assessment, and (3) regulatory compliance. At the SJRWMD, the LSJR Basin Management Section is responsible for surface water sample collection, habitat assessment, and data analysis in support of this network.

One of the objectives for this network is to characterize the status of water resources and to determine if the resources are changing with time. This network is largely designed based on known pollution sources in the streams. Data collected are used to support a number of scientific and management investigations, including (1) identifying water quality problem areas or violations of Florida State water

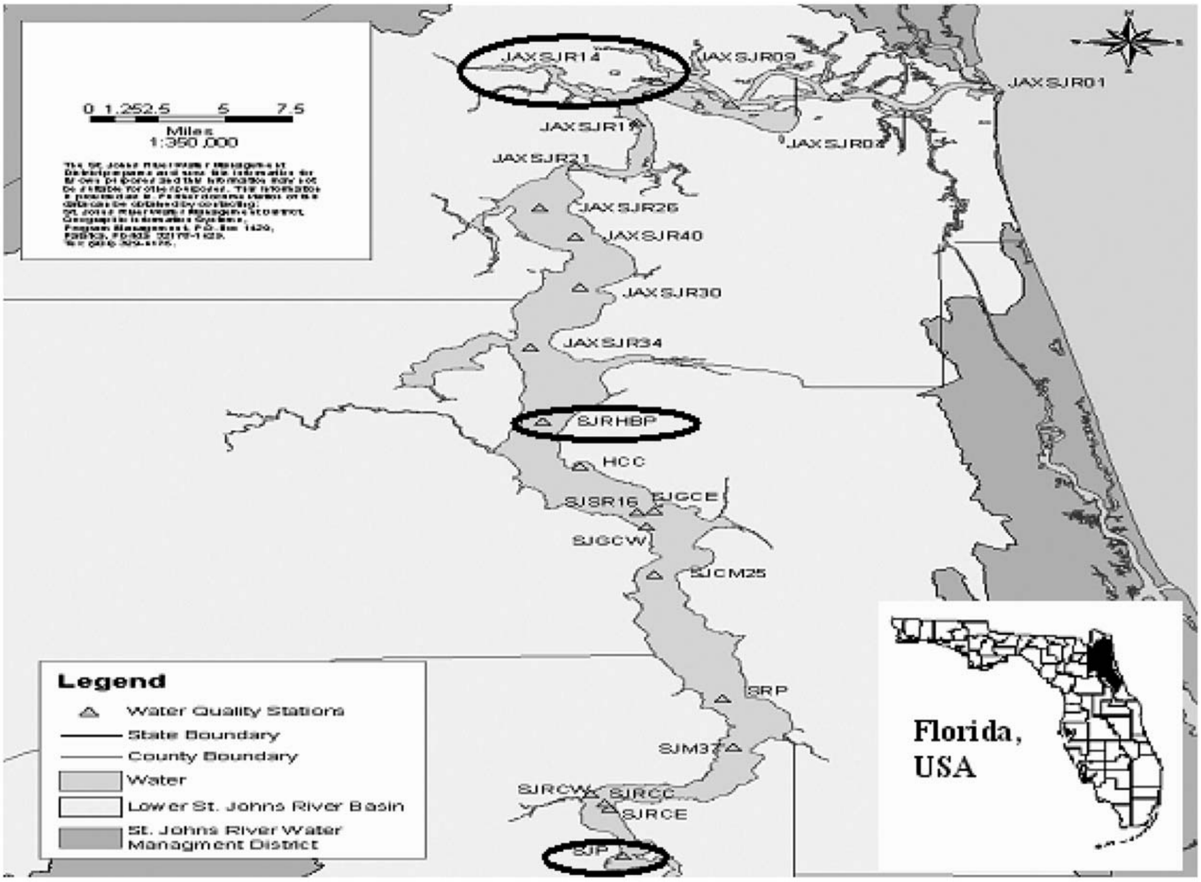

Fig. 1. Location of the Lower St. Johns River, Florida, USA. The symbol "△" represents surface water monitoring stations, whereas the circled stations are the three non-principal stations (i.e., JAXSJR14, SJRHBP, and SJP) obtained by PFA.

Table 1
Water quality data from the monitoring stations located at the main stem of the lower St. Johns River used for this study[a]

Water quality parameter	Storet number	Surface water monitoring stations									
		SJR01	SJR04	SJR09	SJR14	SJR17	SJR21	SJR26	SJR30	HCC	SJR34
Temperature, water (°C)	10	30.18	30.60	16.99	30.83	18.61	31.08	31.18	30.69	22.45	30.39
Temperature, air (°C)	20	18.45	19.90	19.10	19.95	19.20	33.50	34.50	34.50	25.00	33.00
Weather (WMO code 4501)	41	20.00	20.00	20.00	20.00	17.50	80.00	30.00	10.00	50.00	15.00
Transparency, secchi disc (m)	78	1.98	1.20	0.72	1.58	1.18	1.10	0.92	1.20	0.75	1.10
Color (platinum–cobalt units)	80	15.00	25.00	45.00	70.00	70.00	100.00	100.00	150.00	80.00	110.00
Specific conductance, field (umhos/cm at 25°C)	94	46843.50	39940.00	20845.00	28479.50	12457.50	22969.00	15629.00	9756.00	1438.00	8448.00
Sampling station location vertical (m)	98	2.50	3.80	2.50	3.45	7.00	2.50	2.40	2.40	0.50	2.55
Oxygen, dissolved, analysis by probe (mg/L)	299	4.98	4.36	8.32	4.01	8.01	4.28	6.87	6.48	9.07	5.75
BOD, 5 day, 20°C (mg/L)	310	0.90	0.95	1.10	1.10	1.20	1.25	1.40	0.95	1.20	1.45
pH (standard units)	400	7.93	7.79	8.03	7.60	7.67	8.39	8.73	8.27	8.02	7.82
Alkalinity, total (mg/L as $CaCo_3$)	410	110.96	90.82	99.41	80.38	75.62	69.46	63.94	60.52	76.53	59.38
Salinity–parts per thousand	480	30.52	25.55	12.61	17.50	7.30	13.83	9.10	5.50	1.83	4.73
Residue, total nonfiltrable (mg/L)	530	35.50	25.50	34.00	25.00	32.00	30.00	19.50	21.00	10.00	15.00
Nitrogen, Ammonia, dissolved (mg/L as N)	608	-0.04	0.00	0.02	0.03	0.02	0.03	0.01	0.01	0.00	0.01
Nitrogen, Ammonia, total (mg/L as N)	610	0.01	0.03	0.03	0.04	0.02	0.03	0.02	0.02	0.01	0.02
Nitrogen, Kjeldahl, dissolved (mg/L as N)	623	0.14	0.32	0.45	0.56	0.95	0.65	0.67	0.70	1.00	0.77
Nitrogen, Kjeldahl, total (mg/L as N)	625	0.39	0.68	0.65	0.88	1.15	1.09	1.21	1.17	1.26	1.20
Nitrite plus nitrate, diss. 1 det. (mg/L as N)	631	0.04	0.11	0.21	0.26	0.25	0.29	0.28	0.32	0.00	0.14
Phosphorous, dissolved (mg/L as P)	666	0.05	0.07	0.09	0.11	0.11	0.13	0.12	0.10	0.06	0.09
Phosphorous, total (mg/L as P)	665	0.10	0.09	0.11	0.15	0.08	0.15	0.14	0.12	0.03	0.10
Phosphorous, dissolved orthophosphate (mg/L as P)	671	0.00	0.04	0.05	0.06	0.06	0.07	0.07	0.05	0.02	0.05
Carbon, total organic (mg/L as C)	680	2.79	6.23	8.16	11.03	11.53	15.14	14.54	17.73	16.37	15.58
Carbon, dissolved organic (mg/L as C)	681	1.77	5.11	5.86	7.23	11.09	9.94	10.39	12.11	15.44	14.36
Calcium, total (mg/L as Ca)	916	302.45	257.40	222.70	190.50	167.80	95.80	95.90	45.20	48.52	38.40
Magnesium, total (mg/L as Mg)	927	1193.00	1021.50	501.00	507.50	430.50	213.70	159.75	32.14	15.74	14.73
Cadmium, total (μg/L as Cd)	1027	0.85	0.98	0.89	0.86	0.48	0.63	0.45	0.40	0.00	0.54
Chromium, total (μg/L as Cr)	1034	9.82	4.36	4.81	3.60	2.62	3.16	6.53	4.01	0.59	1.20
Copper, total (μg/L as Cu)	1042	7.62	7.37	5.85	5.52	5.51	4.66	3.85	2.12	0.53	2.17
Iron, total (μg/L as Fe)	1045	194.45	189.75	350.60	307.50	280.70	459.90	410.85	389.45	146.25	429.45
Lead, total (μg/L as Pb)	1051	1.88	1.94	1.29	1.52	1.10	1.18	1.27	0.65	0.24	0.68
Manganese, total (μg/L as Mn)	1055	10.51	5.83	6.92	8.86	8.90	12.66	12.25	11.25	10.66	10.86
Nickel, total (μg/L as Ni)	1067	49.25	29.30	39.80	32.45	20.85	14.75	18.05	12.10	0.67	6.07
Zinc, total (μg/L as Zn)	1092	3.31	1.77	3.57	2.70	4.99	5.72	3.93	3.87	1.57	1.96
Chlorophyll-*a* μg/L trichromatic uncorrected	32210	3.83	5.09	4.69	5.08	4.15	7.55	12.79	7.83	20.68	13.73
Chlorophyll-*a* μg/L spectrophotometric acid. meth.	32211	3.33	4.09	3.91	4.61	2.03	4.48	10.53	5.98	16.42	11.20
Chlorophyll-*c* μg/L trichromatic uncorrected	32214	0.43	0.52	0.40	0.48	0.61	0.46	0.95	0.45	2.07	1.12
Pheophytin-*a* μg/L spectrophotometric acid. meth.	32218	1.19	1.45	1.49	1.98	3.97	4.93	3.60	2.53	5.92	3.39
Pheophytin ratio (OD 663) spectro, before/after acid	32219	1.50	1.51	1.49	1.49	1.18	1.38	1.44	1.51	1.51	1.53
Hardness, Ca Mg calculated (mg/L as $CaCo_3$)	46570	5675.00	4865.00	2795.00	2800.00	2325.00	1235.00	935.00	245.00	187.00	156.50
Residue, total filtrable (dried at 180° C), mg/L	70300	37500.00	29050.00	24650.00	15350.00	7285.00	4425.00	3115.00	955.00	502.00	727.00
Turbidity, lab nephelometric turbidity units, NTU	82079	3.07	4.72	6.23	6.23	10.46	8.97	9.59	8.20	3.93	5.73
Depth of bottom of water body at sample site meters	82903	5.30	11.20	11.50	9.20	10.00	9.30	3.10	4.20	1.90	5.90

	SJB40	SJCM25	SJGCE	SJGCW	SJM37	SJP	SJBCC	SJBCW	SJBHRP	SJSB16	SJBCE	SBP
Temperature, water (°C)	20.64	22.70	22.15	22.60	22.80	23.50	20.50	20.50	16.60	21.00	20.70	20.50
Temperature, air (°C)	23.50	23.00	25.50	24.50	22.50	25.00	25.00	25.00	20.50	25.00	24.50	22.00
Weather (WMO code 4501)	40.00	65.00	45.00	52.50	70.00	70.00	60.00	65.00	10.00	47.50	60.00	65.00
Transparency, secchi disc (m)	1.15	0.69	0.70	0.73	0.53	0.57	0.57	0.48	0.82	0.79	0.59	0.54
Color (platinum–cobalt units)	60.00	75.00	75.00	75.00	75.00	80.00	80.00	85.00	80.00	75.00	75.00	75.00
Specific conductance, field (umhos/cm at 25°C)	5480.00	925.50	1045.50	964.00	1002.00	1065.50	1015.00	1018.50	864.50	973.50	1021.00	990.50
Sampling station location vertical (m)	2.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Oxygen, dissolved, analysis by probe (mg/L)	8.29	8.78	8.73	8.94	9.35	8.63	9.14	8.76	8.38	9.13	8.56	8.70
BOD, 5 day, 20°C (mg/L)	1.65	1.90	1.50	1.45	2.65	1.75	2.60	2.70	0.80	1.70	2.40	3.15
pH (standard units)	7.81	8.18	8.10	7.95	8.32	8.33	8.35	8.26	7.62	8.07	8.26	8.45
Alkalinity, total (mg/L as $CaCo_3$)	72.14	80.97	77.42	78.38	90.07	83.89	89.69	89.73	72.99	77.48	88.79	89.55
Salinity–parts per thousand	2.60	0.54	0.59	0.55	0.58	0.53	0.57	0.57	3.10	0.54	0.58	0.56
Residue, total nonfiltrable (mg/L)	15.00	10.00	10.00	13.00	14.00	11.50	13.00	11.50	10.50	8.00	12.50	13.50
Nitrogen, Ammonia, dissolved (mg/L as N)	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Nitrogen, Ammonia, total (mg/L as N)	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.01
Nitrogen, Kjeldahl, dissolved (mg/L as N)	0.73	1.03	1.02	0.87	1.01	0.93	0.91	1.03	1.00	1.02	0.94	0.98
Nitrogen, Kjeldahl, total (mg/L as N)	1.06	1.43	1.30	1.40	1.65	1.30	1.37	1.56	1.28	1.29	1.38	1.74
Nitrite plus nitrate, diss. 1 det. (mg/L as N)	0.08	0.01	0.00	0.00	0.00	0.06	0.02	0.03	0.08	0.01	0.01	0.01
Phosphorous, dissolved (mg/L as P)	0.09	0.07	0.06	0.08	0.08	0.07	0.07	0.08	0.07	0.06	0.08	0.09
Phosphorous, total (mg/L as P)	0.07	0.02	0.02	0.03	0.01	0.02	0.02	0.02	0.04	0.03	0.01	0.02
Phosphorous, dissolved orthophosphate (mg/L as P)	0.04	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.02	0.01	0.00	0.01
Carbon, total organic (mg/L as C)	12.47	16.14	16.61	15.83	16.20	14.75	15.41	16.92	15.42	16.24	15.73	16.37
Carbon, dissolved organic (mg/L as C)	12.28	15.64	15.41	14.67	15.02	14.70	14.60	16.38	15.00	14.18	15.03	16.03
Calcium, total (mg/L as Ca)	57.35	52.75	50.16	51.50	58.80	51.90	56.20	55.45	46.28	52.15	55.55	58.25
Magnesium, total (mg/L as Mg)	88.80	16.77	18.92	17.78	18.26	15.66	17.80	17.56	16.04	17.74	17.95	18.76
Cadmium, total (μg/L as Cd)	0.42	0.02	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.00	0.00
Chromium, total (μg/L as Cr)	2.74	0.43	0.51	0.69	0.41	0.42	0.40	0.50	0.40	0.46	0.39	0.95
Copper, total (μg/L as Cu)	2.82	0.96	0.58	0.94	0.80	0.62	0.57	0.87	0.82	0.74	0.54	0.97
Iron, total (μg/L as Fe)	391.65	120.85	130.80	202.40	117.45	159.95	103.90	133.80	208.90	124.85	118.40	112.55
Lead, total (μg/L as Pb)	0.97	0.05	0.24	0.42	0.04	0.15	0.04	0.17	0.21	0.00	0.06	0.20
Manganese, total (μg/L as Mn)	13.10	13.08	15.48	17.10	15.06	8.47	14.49	18.63	14.11	15.44	14.92	16.67
Nickel, total (μg/L as Ni)	12.40	0.31	0.50	0.48	0.47	0.41	0.34	0.93	0.23	0.62	0.62	1.94
Zinc, total (μg/L as Zn)	4.56	4.56	5.49	4.71	4.50	2.48	-0.49	7.22	4.07	2.20	0.04	0.00
Chlorophyll-*a* μg/L trichromatic uncorrected	8.17	29.13	20.67	28.06	50.55	23.65	39.37	40.55	10.16	23.32	3.28	4.26
Chlorophyll-*a* μg/L spectrophotometric acid. meth.	6.48	23.66	15.65	21.36	43.41	18.37	35.12	35.98	8.64	18.81	40.80	54.21
Chlorophyll-*c* μg/L trichromatic uncorrected	1.03	2.67	1.98	2.51	4.06	2.51	3.11	2.70	1.33	2.36	35.58	48.33
Pheophytin-*a* μg/L spectrophotometric acid. meth.	2.95	7.98	6.41	8.36	6.74	6.50	6.01	5.14	4.27	5.99	3.12	3.55
Pheophytin ratio (OD 663) spectro, before/after acid	1.50	1.52	1.53	1.51	1.57	1.54	1.62	1.61	1.44	1.55	5.92	6.17
Hardness, Ca Mg calculated (mg/L as $CaCo_3$)	509.00	200.00	202.50	200.00	220.50	194.00	213.50	211.00	180.50	202.50	1.59	1.60
Residue, total filtrable (dried at 180° C), mg/L	5855.00	534.50	605.00	566.50	573.50	499.50	580.00	577.50	486.50	562.00	212.50	218.50
Turbidity, lab nephelometric turbidity units, NTU	4.69	5.53	5.09	4.08	6.06	6.07	5.78	4.20	5.07	5.61	5.10	3.50
Depth of bottom of water body at sample site meters	5.00	9.50	3.11	2.85	1.50	1.60	7.60	3.40	4.30	2.75	8.55	4.95

[a] These data are three-year annual median values from 1999 to 2001.

quality standards; (2) determining the amount, or mass load, of pollutants entering rivers from tributaries and point sources; (3) estimating daily, seasonal, and long-term water quality trends; (4) assessing the ability of best management practices to improve non-point source pollution; (5) examining how water quality affects the plants and animals living within the river; and (6) investigating how water quality varies within different reaches of the river. In this study, 22 water quality monitoring stations located in the main stem of the LSJR were selected for analysis (Fig. 1). The physical, chemical, and biological parameters collected from those 22 monitoring stations and used in this study are given in Table 1.

3. Analysis procedures

PCA was first performed in this study to identify the potential for reducing the number of monitoring stations. This analysis investigated the annual variations in water quality parameters measured from the ambient water quality monitoring stations of the LSJR over a 3-year time period (1999–2001). Forty-two water quality parameters from the 22 stations were examined in this study (Table 1). The procedures used for PCA are described below.

3.1. Selection of water quality data

The ambient water quality monitoring databases from the main stem of the LSJR were used in this study. These databases contain the agency monitoring stations, STORET numbers (measured parameters), latitude and longitude information and dates of historic and current data. Station locations are where conditions are the most representative and homogeneous, away from transitional areas such as point source mixing zones and near-shore regions. Some stations are sampled daily or monthly and a couple of stations are sampled seasonally due to budget constraints. Timing of sample collection is routine and not intended to capture any specific flow or rainfall events.

Data from these stations were collected at different times of day and/or different days of the year for each parameter. Plots of all of the data in Excel spreadsheets show that they are not normally distributed and are positively skewed. For the purpose of this analysis, the annual median values for each parameter were used. The choice of the median values rather the mean values was based upon the fact that the measured parameter values are very skewed. In general, when this is the case, the median is a better measurement than the mean (Anderson and Sclove, 1986).

In this study, a 3-year time period was selected based on the following reasons: (1) no complete data set is available to include all of the water quality parameters used in this study beyond the 3-year period. In other words, although some parameters have been measured for a period of 20 years, others have only recently been added; and (2) the PCA requires no missing values in a data set.

3.2. Selection of computation method

The Statistical Analysis System (SAS) package (version 8), developed by SAS Institute Inc. (SAS, 1999), was employed to perform principal component and factor analyses. This software has the PRINCOMP and FACTOR modules that can perform the analyses.

Mathematically, PCA and PFA normally involve the following five major steps: (1) start by coding the variables x_1, $x_2, \ldots, x_\mathrm{p}$ to have zero means and unit variance, i.e., standardization of the measurements to ensure that they all have equal weights in the analysis; (2) calculate the covariance matrix $\mathbf{C}$; (3) find the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_\mathrm{p}$ and the corresponding eigenvectors $\boldsymbol{a_1}$, $\boldsymbol{a_2}, \ldots, \boldsymbol{a_\mathrm{p}}$; (4) discard any components that only account for a small proportion of the variation in data sets (Manly, 1986); and (5) develop the factor loading matrix and perform a varimax rotation on the factor loading matrix to infer the principal stations. The very last step is primarily used in PFA. For this study, a factor correlation coefficient that is greater than 0.75 (or 75%) was considered significant. This conservative criterion was selected because the study area was large and the river system was highly non-linear and dynamic. Stations that do not have any factors with correlation coefficients greater than this value were considered as non-principle stations.

4. Results and discussion

4.1. Principal component analysis

In a PCA, the number of components is equal to the number of variables. However, a component is not only comprised of a single variable but rather all of the variables used in a study. For example, there are 22 variables (stations) used in this study, which produce 22 components. In each component, there are 22 variables (or stations) as shown in Eq. (1) below. The PCA results showed that of the 22 components, the first component accounted for about 94.6% and the second component accounted for about 4.5% of the total variance in the data set. These two components together accounted for about 99.1% of the total variance and the rest of the 20 components only accounted for about 0.9%. Therefore, our discussions should focus only on the first two components.

From the eigenvectors obtained in the PCA, the first component, Z_1, can be given as

$$\begin{aligned} \mathrm{Z}_1 = {} & 0.21x_1 + 0.21x_2 + 0.20x_3 + 0.22x_4 + 0.22x_5 \\ & + 0.21x_6 + 0.21x_7 + 0.20x_8 + 0.22x_9 + 0.20x_{10} \\ & + 0.20x_{11} + 0.22x_{12} + 0.22x_{13} + 0.22x_{14} + 0.22x_{15} \\ & + 0.22x_{16} + 0.22x_{17} + 0.22x_{18} + 0.22x_{19} \\ & + 0.22x_{20} + 0.21x_{21} + 0.21x_{22}, \end{aligned} \tag{1}$$

where x is the monitoring station, the subscripts denote the station numbers, and the coefficients are the eigenvectors. This component had almost equal loadings (i.e., similar coefficient values in Eq. (1)) on all variables and therefore is a measure of overall performance of the stations.

It is apparent that Z_1 has an extremely high correlation with the measured data as it accounts for 94.6% of the data variance, which would indicate that only one major source of data variation is present. This finding is somewhat different from other studies where many more components are needed to explain the same amount of variance (Bengraine and Marhaba, 2003). A possible explanation of the discrepancy is that in this study, the variables are the monitoring stations rather than the water quality parameters. It is expected that the monitoring stations (which are primarily controlled by hydrogeological conditions) would have higher correlations than the water quality parameters (which are controlled by hydrogeological, chemical, and biological conditions). If the variables used in this study were water quality parameters, one would expect to have more components to explain the same amount of the variance in such a highly dynamic and complex river system. Further discussion of this issue is given in the water quality parameter identification section below.

Similarly, the second component can be expressed as

$$\begin{aligned} \mathrm{Z}_2 = {} & -0.23x_1 - 0.181x_2 - 0.41x_3 - 0.04x_4 - 0.09x_5 \\ & + 0.27x_6 + 0.26x_7 + 0.36x_8 + 0.12x_9 + 0.37x_{10} \\ & - 0.36x_{11} - 0.08x_{12} - 0.08x_{13} - 0.08x_{14} - 0.07x_{15} \\ & + 0.01x_{16} - 0.07x_{17} - 0.07x_{18} - 0.06x_{19} \\ & - 0.08x_{20} + 0.27x_{21} + 0.26x_{22}. \end{aligned} \tag{2}$$

This equation shows that the second component, Z_2, will be high if x_6 to x_{10}, x_{16}, x_{21} and x_{22} are high but x_1 to x_5, x_{11} to x_{15}, and x_{17} to x_{20} are low. Hence, Z_2 represents a difference among the stations. The low coefficients of x variables such as those associated with x_4, x_{16}, and x_{19} mean that the values of these variables have little effect on Z_2.

A graphical representation of the first two component loadings is given in Fig. 2. This diagram was constructed using the eigenvectors from the first two components. It becomes clear that the first component had similar loadings (eigenvectors) for all of the monitoring stations and therefore this component represents the overall performance of all the monitoring stations (Fig. 2A), while the second component measured the difference among the stations (Fig. 2B). Furthermore, the second component shows that three monitoring stations, namely SJR14, SJP, and SJRHBP, had the lowest absolute loading (eigenvector) values, which could indicate that these stations were less important in monitoring water quality variations. However, any conclusion based upon Z_2 would be inappropriate since Z_2 only accounted for 4.5% of the total variance. It should be pointed out that a loading reflects only the relative importance of a variable within a component, and does not reflect the importance of the component itself (Davis, 1986).

4.2. Extraction of important monitoring stations

Although the results of PCA have identified two important components that account for 99.1% of the annual variance in the data set, they did not provide any information on which monitoring stations explain most of the variance. To further identify the monitoring stations that are important in revealing surface water quality variations, a PFA was employed. Similar to PCA, the number of factors is equal to the number of variables in this analysis as well. For instance, if there are 22 variables (stations), each variable will have 22 factors. In this study, the criterion (i.e., eigenvalue) used to retain the principal factors was $>10^{-6}$ (a default value assigned by SAS), which resulted in the retention of 14 factors (Table 2). The results of PFA showed that the eigenvalue of factor 1 accounted for 94.7% and of factor 2 accounted for 4.5% of the total variance in the data set. These two factors together accounted for 99.1% of the total variance, which was the same as those obtained from PCA.

Table 2 shows the rotated factor correlation coefficients for all of the 22 stations. In this study, the factor correlation coefficient considered significant is one that is greater than 0.75 (or $>75\%$). This conservative criterion was selected because the study area was large and the river system was highly non-linear and dynamic. The stations with correlation coefficients less than this value were not considered principal stations. From Table 2, stations SJR14, SJP, and SJRHBP have coefficient values less than 0.75 for all of the factors. These stations (shown in Fig. 1) are considered less important in explaining the annual variance of the data set, and thereby could be the non-principal stations. This finding is similar to the result from the second component (Z_2) in the PCA.

4.3. Validation of PCA results

Before applying the above finding, its scientific reliability must be validated using other independent

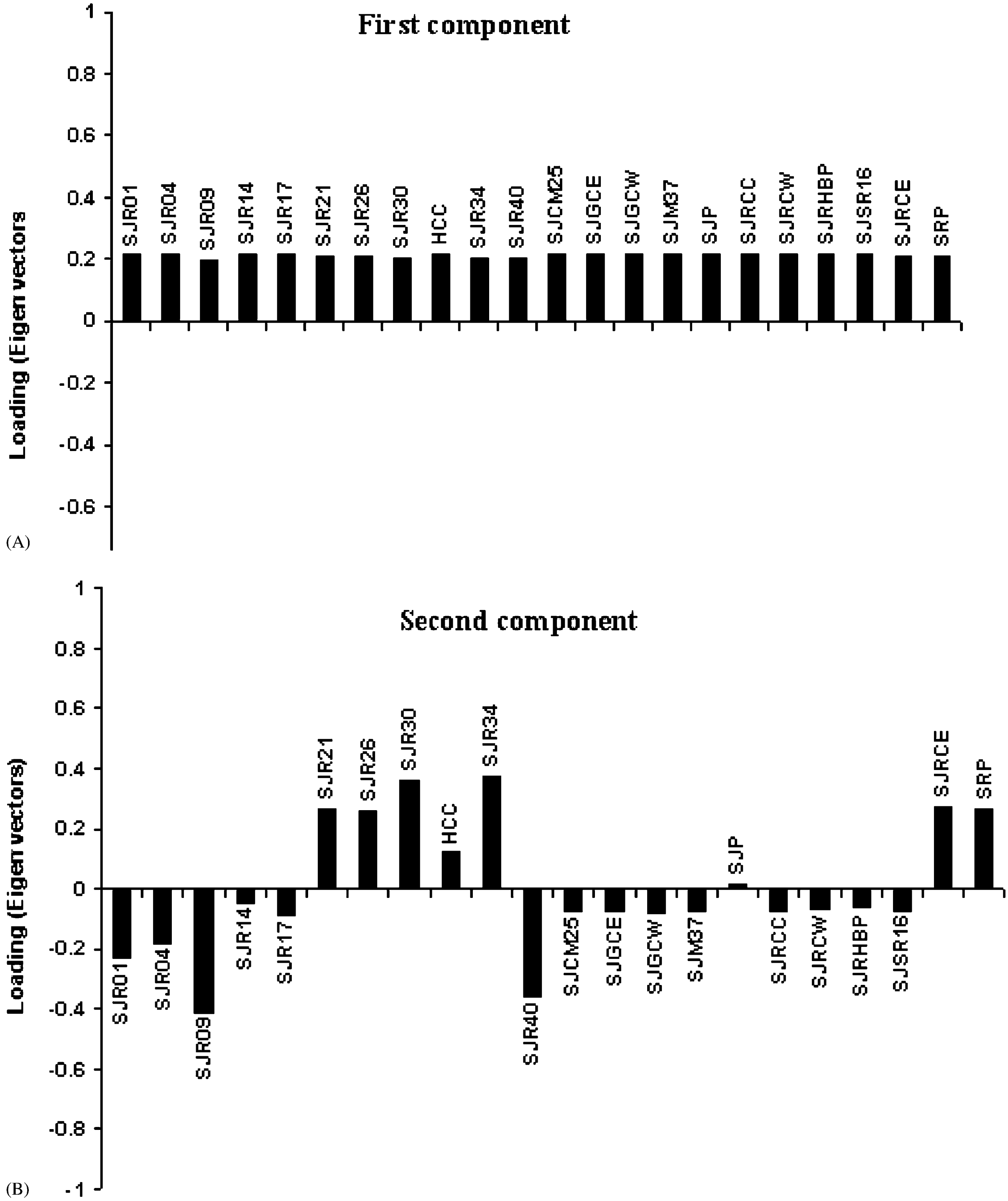

Fig. 2. Component loadings for the first component (A) and the second component (B).

methods. One way to achieve this goal is to compare the water quality data with and without the three non-principal stations. In this study, two cases were developed for comparisons. In the first case, data from the principal stations were used to formulate the following four relationships by regression: (1) dissolved organic carbon (DOC) versus water color; (2) chlorophyll *a* versus total phosphorous (TP); (3) biochemical

Table 2
Rotated factor correlation coefficients for the 22 ambient water quality monitoring stations[a]

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9	Factor10	Factor11	Factor12	Factor13	Factor14
SJR01	0.85753	0.51379	0.01152	−0.01567	0.01197	0.01103	−0.0028	−0.00007	0.00056	−0.00046	0.00093	0.00062	0.00105	0.00037
SJR04	0.83265	0.55257	0.01784	−0.02406	0.01529	0.01458	−0.00274	−0.00001	0.00068	−0.00052	0.00102	0.00067	0.0012	0.00041
SJR09	0.9409	0.33822	−0.00119	0.01649	−0.00477	−0.00285	−0.00155	−0.00016	−0.00028	0.00007	−0.00097	−0.00078	−0.0005	−0.00049
SJR14	0.74855	0.66144	0.02592	−0.03254	0.01616	0.01306	−0.003	−0.0003	0.00075	0.00015	−0.00045	−0.00073	−0.00025	0.00059
SJR17	0.77409	0.62261	0.07419	−0.07099	0.03007	0.0412	−0.00082	0.00036	0.00003	0.00004	−0.00006	0.00002	−0.00004	0.00002
SJR21	0.51005	0.85758	0.03924	−0.04906	0.018	0.01102	−0.00049	0.00112	0.00094	0.00045	0.00024	−0.00042	−0.00089	0.00135
SJR26	0.51575	0.85379	0.04662	−0.04995	0.0153	0.01198	−0.00047	−0.00043	0.00106	−0.00012	0.0001	−0.0008	−0.00089	0.00027
SJR30	0.42394	0.90374	0.04676	−0.0363	0.0021	−0.00076	−0.00289	−0.00136	−0.0005	−0.00008	−0.00085	0.00157	0.00087	−0.00079
HCC	0.62507	0.76652	0.14566	−0.01369	0.01531	0.00365	−0.00712	0.00561	0.00166	−0.00055	−0.00085	0.0001	0.00152	−0.00037
SJR34	0.4121	0.90916	0.05039	−0.03152	−0.00547	−0.00537	0.00252	−0.00159	−0.00029	−0.00007	−0.0001	−0.00023	0.00037	−0.00001
SJR40	0.91939	0.38949	0.01534	0.03772	−0.0303	−0.01961	0.00724	0.00157	−0.00108	−0.0002	0.00018	0.00001	−0.00055	0.0001
SJCM25	0.75668	0.62304	0.19344	0.0011	0.04029	0.00715	−0.00509	0.00953	0.00092	0.00567	0.00111	−0.00045	0	−0.00006
SJGCE	0.75994	0.62671	0.17081	−0.00342	0.0175	0.00764	−0.01123	0.0034	0.00479	0.00119	0.00073	0.00038	0.00032	0.00138
SJGCW	0.75512	0.61543	0.22417	0.01169	−0.01726	−0.00686	0.01689	0.0031	0.00133	0.00031	0	−0.00017	−0.00005	0.00026
SJM37	0.75456	0.62532	0.18967	−0.00133	0.05943	0.00599	0.00354	−0.0055	−0.00065	−0.00346	−0.00095	−0.00157	0.00024	−0.00048
SJP	0.69667	0.68876	0.19907	0.00355	0.01747	0.0009	0.00216	0.01726	−0.00036	−0.00058	0.0001	−0.00023	−0.00003	0.00002
SJRCC	0.75584	0.62805	0.17583	−0.0027	0.05543	0.00962	−0.01152	−0.00439	−0.00182	0.00136	0.00116	0.00015	−0.00026	0.00081
SJRCW	0.75118	0.63029	0.19167	0.00536	0.04053	0.00243	−0.00274	−0.0028	−0.00004	−0.0003	−0.00099	0.00463	0.00005	0.00001
SJRHBP	0.73872	0.62576	0.24319	0.00498	−0.05804	−0.00059	−0.00369	−0.01085	−0.00275	−0.00141	0.00004	−0.0007	−0.00006	−0.00005
SJSR16	0.75835	0.62469	0.18359	−0.00471	0.02565	0.01012	−0.01133	0.00283	0.00512	0.00049	−0.00024	−0.00042	−0.00021	−0.00096
SJRCE	0.49649	0.85553	0.10911	0.09787	−0.0022	−0.00269	−0.00368	0.00298	0.00017	0.00134	0.00535	−0.00057	−0.00005	0.00007
SRP	0.50482	0.84932	0.10733	0.11032	0.00633	0.00555	0.00304	−0.00188	−0.00017	−0.00086	−0.0036	0.00032	−0.00014	−0.00001

[a] Fourteen factors were retained by the proportion criterion ($\geqslant 10^{-6}$) which was set up by SAS.

oxygen demand (BOD) versus total organic carbon (TOC); and (4) chlorophyll *a* versus total dissolved nitrogen (TDN). In the second case, data from all of the stations (i.e., principal and non-principal stations) were used to formulate the aforementioned four relationships by regression. These two cases were then compared to determine if the addition of data from the three non-principal stations improved the regression relationships.

Comparison of the relationship between DOC and water color obtained using data from all of the stations with that obtained using data from the principal stations (Fig. 3) showed that the addition of the three

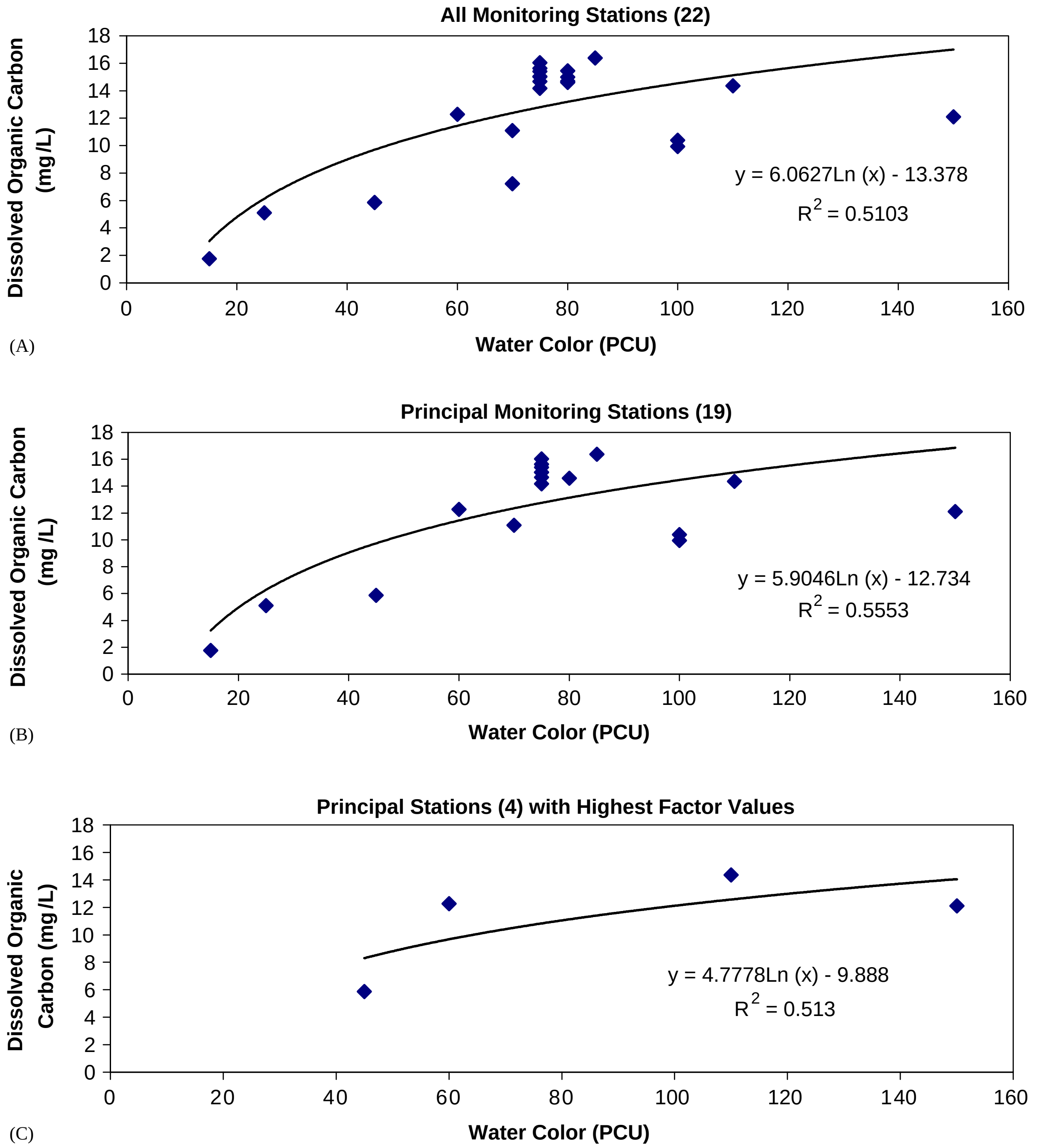

Fig. 3. Relationship between dissolved organic carbon (DOC) and water color for data from the principal stations (A), all stations (B), and four principal stations with the highest factor values (C).

non-principal stations did not improve the curve-fitting between DOC (y) and water color (x), as indicated by correlation coefficients (i.e., R^2 values). The R^2 value for the regression equation ($y = 6.0627\mathrm{Ln}(x) - 13.378$) for data from all of the 22 stations was 0.5103, whereas the R^2 value for the regression equation ($y = 5.9046\mathrm{Ln}(x) - 12.734$) for data from the 19 principal stations was 0.5553. The latter is slightly better than the former. Similar results were obtained for the relationships between chlorophyll *a* versus TP (Fig. 4), BOD versus TOC (Fig. 5), and chlorophyll *a* versus TDN (Fig. 6). That is, the R^2 values obtained for data from the 19 principal stations were slightly better than those from all of the 22 stations. A published literature search reveals that few efforts have been devoted to investigating the correlations among the variables used in this study in an estuarine environment like the LSJR although similarly close R^2 values for

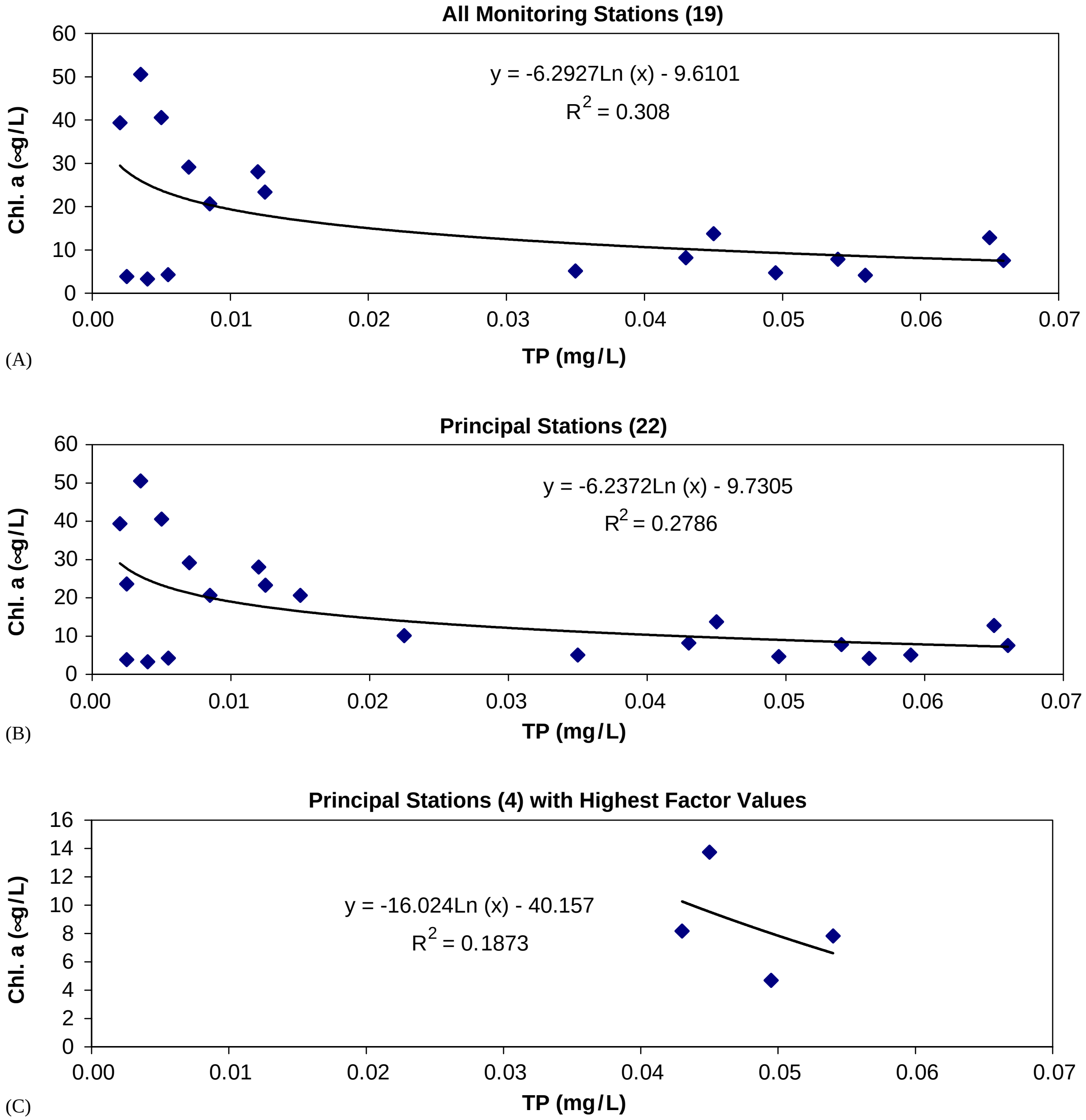

Fig. 4. Relationship between chlorophyll *a* and total phosphorous (TP) for data from the principal stations (A), all stations (B), and four principal stations with the highest factor values (C).

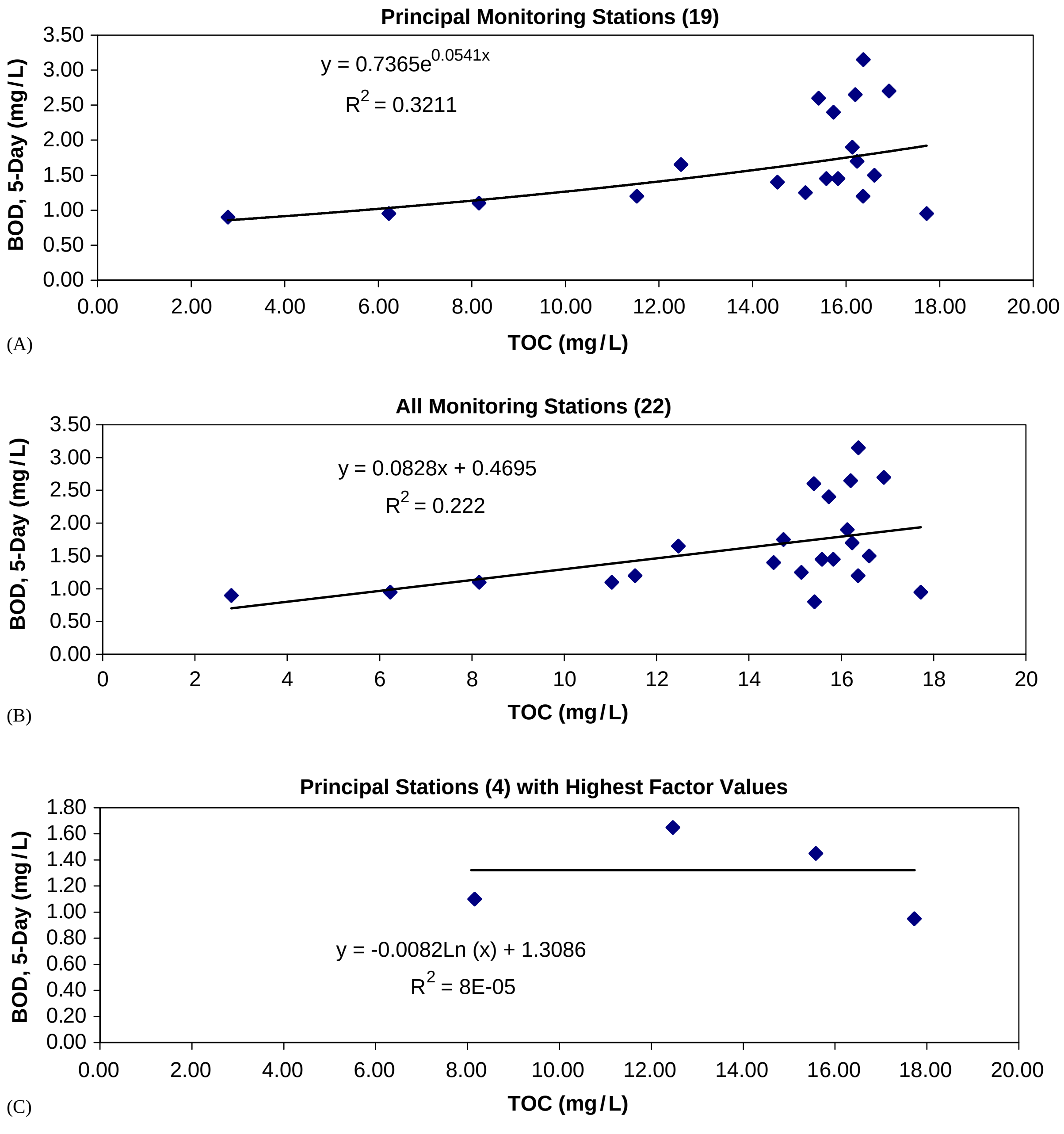

Fig. 5. Relationship between BOD and TOC for data from the principal stations (A), all stations (B), and four principal stations with the highest factor values (C).

some of the variables were obtained in fresh water lakes in Florida, USA (Canfield et al., 1984; Haven, 2003).

To test the statistical significance of the R^2 values between all of the 22 stations and the 19 principal stations, a t-test analysis was performed with a 5% level of significance (i.e., $\alpha = 0.05$). With the t value (2.724) greater than the t critical region (2.353), it rejected the hypothesis of equal means, indicating that the R^2 values from the 19 principal stations were statistically better than those from all of the 22 stations. Therefore, the three monitoring stations are considered to be less important stations since the addition of data from these three stations did not improve the curve-fittings.

Further validation was also performed by comparing those 19 principal stations with the four principal stations (i.e., SJR09, SJR30, SJR34, and SJR40) that have highest factor correlation coefficients ($<90\%$). The R^2 values from the four principal stations were much

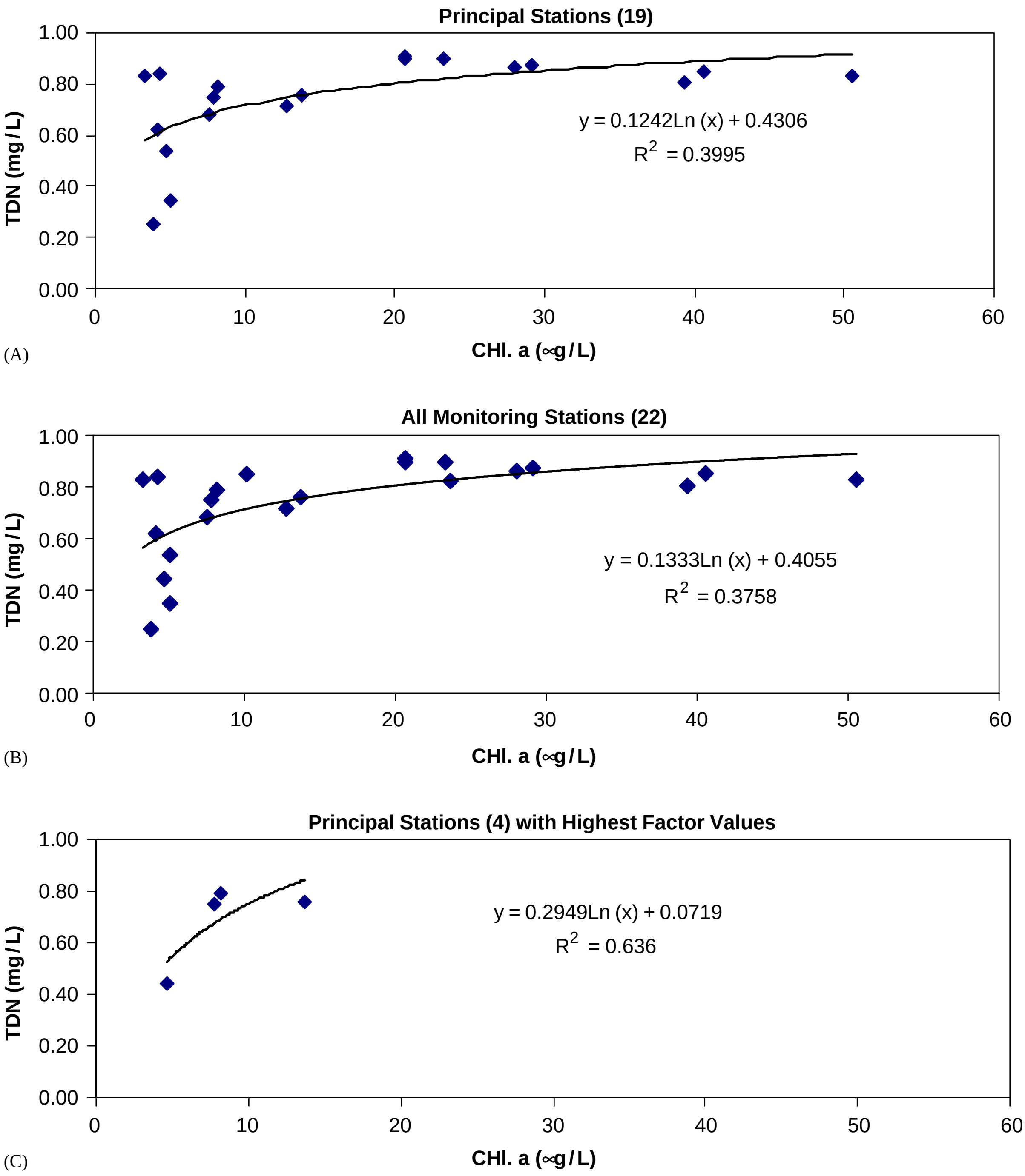

Fig. 6. Relationship between total dissolved nitrogen (TDN) and chlorophyll *a* for data from the principal stations (A), all stations (B), and four principal stations with the highest factor values (C).

lower than those from the 19 principal stations (Figs. 3C–5C) except for TND versus chlorophyll *a* (Fig. 6C). Results suggest that the selection of only four principal stations with highest PFA correlation values was insufficient to monitor water quality variations in the LSJR, and thereby confirming that those 19 principal stations are needed for monitoring the water quality variations in the main stem of the LSJR.

4.4. Identification of important water quality parameter

Characterization of changes in surface water quality is an important aspect for evaluating the potential impact of natural or anthropogenic point and non-point sources of pollution on ecosystem health. In this study, 20 surface water quality parameters commonly used for determining surface water quality in Florida watersheds (Table 3) were selected for analysis. A PFA technique was employed to extract the parameters that are most important in assessing variations in LSJR water quality. It should also be noted that 20 rather than 42 surface water quality parameters were selected for analysis because of the limitation of the SAS software. That is, when the variables (e.g., water quality parameters) are greater than the observations (e.g., stations), one will obtain singularity solutions (unstable solutions) in estimating covariance and correlation matrices in PCA or PFA using SAS. Some studies reported that in a PCA, the number of observations must be greater than the number of variables, which is the minimum requirement to provide a stable solution (Yu et al., 1998). Other studies, however, showed that PCA could be applied to any type of data matrix (i.e., regardless of the number of variables and observations) (Golub and van Loan, 1989). This discrepancy could be due to the differences in solution algorithms used in these studies.

Table 3
Rotated factor correlation coefficients for the selected 20 water quality parameters

Parameters	Factor1	Factor2	Factor3
Temperature	0.10532721	−0.1935633	0.83747929
Secchi depth	−0.9853376	0.06851073	0.16881688
Water color	0.89060774	−0.1252071	0.16033381
Dissolved O_2	0.78006238	−0.2795847	0.0582299
Biochemical O_2 demand	0.42880924	−0.7562247	−0.2468532
pH	−0.2115872	−0.7463909	−0.4166407
Alkalinity	−0.9756057	−0.5342094	−0.1230259
Salinity	−0.9883456	0.01093491	−0.0150069
Total nitrogen	0.94955744	−0.1464032	−0.0484042
Dissolved nitrate & nitrite	0.10361241	0.97906551	−0.2339517
Total phosphorus	−0.1644928	0.84794959	−0.2766875
PO_4^{-3}	0.15962185	0.96848453	−0.2869372
Total organic carbon	1.00223167	0.05381193	0.04769762
Dissolved organic carbon	0.97941929	0.01521998	0.08475127
Ca	−1.0046182	−0.0254255	0.00127302
Mg	−0.9938031	0.00446019	−0.0060257
Fe	−0.1374931	0.89303095	0.17938807
Mn	0.69849057	−0.010781	0.45585082
Chlorophyll *a*	0.54653002	−0.6851524	−0.1671335
Turbidity	0.116587	0.8491219	−0.1896772

In PFA, eigenvalues are normally used to determine the number of principal components or factors to be retained for further study. Results show that the first three principal factors have eigenvalues greater than unity (a criterion used to determine the number of factors retained (Yu et al., 1998)) and explain 61.9%, 23.7%, and 6.9% of the total variances in the original data set. Therefore, the first three factors were used for further analysis. Table 3 shows the rotated correlation coefficients for the first three factors for each parameter. In this study, any water quality parameter with an absolute factor correlation coefficient value greater than 0.95 (or >95%) was considered to be a most important parameter in contribution to variations of LSJR water quality. Based on this selection criterion, the most important water quality parameters that can be used to evaluate variations of the LSJR water quality are the organic-related parameters (i.e., total organic carbon and dissolved organic carbon), the mineral-related parameters (i.e., alkalinity, salinity, Ca and Mg), the nutrient parameters (i.e., dissolved nitrate and nitrite and orthophosphate), and a physical parameter (i.e., Secchi depth). The mineral-related and physical parameters were negatively involved in water quality variations. They may be interpreted as representing influences from natural inputs. In contrast, the organic-related and nutrient parameters positively participated in water quality variations. They may be interpreted as representing influences from both natural and anthropogenic inputs.

5. Summary and conclusions

This study was undertaken to evaluate the ambient water quality monitoring stations located in the main stem of the LSJR and, if necessary, to refine the stations based on the scientific findings. The outcome showed that there was a potential for improving the efficiency and economy of the monitoring network in the main stem of the LSJR by reducing the number of monitoring stations from 22 to 19. This reduction may result in significant cost savings without sacrificing important surface water quality data. However, it should be noted that only the 3-year annual median values of water quality parameters were used in this study. Prior to making any critical decision in eliminating water quality monitoring stations in the LSJR, the PCA and PFA with a longer time scale (i.e., more than 3 years) should be performed, provided sufficient data are available. Furthermore, temporal variations (e.g., seasonal data) at the monitoring stations should also be analyzed. Results from this analysis could prove valuable if budget constraints require the refining of the current monitoring network. Further study is also warranted to identify the principal physical, chemical, and biological parameters that are important in predicting seasonal variations in surface water quality for the entire LSJR monitoring network. This effort could evaluate the potential for reducing the suite of water quality parameters measured.

Acknowledgement

The author thanks the St. Johns River Water Management District for providing the field measured data.

References

Anderson, T.W., Sclove, S.L., 1986. The Statistical Analysis of Data, second ed. The Scientific Press.

Bengraine, K., Marhaba, T.F., 2003. Using principal component analysis to monitor spatial and temporal changes in water quality. J. Hazard. Mater. 100, 179–195.

Campbell, D., Bergman, M., Brody, R., Keller, A., Livingston-Way, P., Morris, F., Watkins, B., 1993. SWIM Plan for the Lower St. Johns River Basin. St. River Water Management District, Palatka, Florida.

Canfield Jr., D.E., Linda, S.B., Hodgson, L.M., 1984. Relations between color and some limnological characteristics of Florida lakes. Water Resour. Bull. 20, 323–329.

Davis, J.C., 1986. Statistical and data analysis in geology, second ed. John Wiley & Sons, New York.

Durell, G.S., Seavey, J.A., Higman, J., 2001. Sediment Quality in the Lower St. Johns River and Cedar-Ortega River Basin: Chemical Contaminant Characteristics. Battelle, 397 Washington Street, Duxbury, MA.

Gangopadhyay, S., Gupta, A.D., Nachabe, M.H., 2001. Evaluation of ground water monitoring network by principal component analysis. Ground Water 39, 181–191.

Golub, G.H., van Loan, C.F., 1989. Matrix Computations, 2nd Ed. The John Hopkins University Press, Baltimore.

Haven, K.E., 2003. Phosphorus-Algal bloom relationships in large lakes of south Florida: Implications for establishing nutrient criteria. Lake Reservoir Manage. 9, 222–228.

Manly, B.F.J., 1986. Multivariate Statistical Methods: A Primer. Chapman & Hall, London.

Ouyang, Y., Higman, J., Thompson, J., O'Toole, T., Campbell, D., 2002. Characterization and spatial distribution of heavy metals in sediment from cedar and ortega rivers Basin. J. Contam. Hydrol. 54, 19–35.

Perkins, R.G., Underwood, G.J.C., 2000. Gradients of chlorophyll a and water chemistry along an eutrophic reservoir with determination of the limiting nutrient by in situ nutrient addition. Water Res. 34, 713–724.

SAS Institute Inc., 1999. SAS Proprietary Software Release 8.2. SAS Institute, Cary, NC.

Shine, J.P., Ika, R.V., Ford, T.E., 1995. Multivariate statistical examination of spatial and temporal patterns of heavy metal contamination in New Bedford Harbor marine sediments. Environ. Sci. Technol. 29, 1781–1788.

Tabachnick, B.G., Fidell, L.S., 2001. Using Multivariate Statistics. Allyn and Bacon, Boston, London.

Tauler, R., Barcelo, D., Thurman, E.M., 2000. Multivariate correlation between concentrations of selected herbicides and derivatives in outflows from selected US Midwestern reservoirs. Environ. Sci. Technol. 34, 3307–3314.

Wackernagel, H., 1995. Multivariate Geostatistics. An Introduction With Applications. Springer, New York and London.

Winter, T.C., Mallory, S.E., Allen, T.R., Rosenberry, D.O., 2000. The use of principal component analysis for interpreting ground water hydrographs. Ground Water 38, 234–246.

Yu, J.C., Quinn, J.T., Dufournaud, C.M., Harrington, J.J., Roger, P.P., Lohani, B.N., 1998. Effective dimensionality of environmental indicators: a principal component analysis with bootstrap confidence intervals. J. Environ. Manage. 53, 101–119.