

STATISTICAL AND GEOCHEMICAL WATER QUALITY RELATIONSHIPS IN THE RED RIVER OF THE NORTH (1984 TO 2006)

JOEL L. FISHER (PhD)
HAIMENG ZHANG (PhD)
CHARLES FRITZ

3/15/2007

STATISTICAL AND GEOCHEMICAL WATER QUALITY RELATIONSHIPS IN THE RED RIVER OF THE NORTH (1984 TO 2006)

■ THE STUDIES:

- Dynamics of nutrient sources and loadings
- Chemical behavior of the watershed
- Factors affecting nutrients in the Red River

■ STATISTICAL TOOLS:

- Time series analyses
- Correlation/regression analyses
- Correlation matrices and Principal Component Analysis

STATISTICAL AND GEOCHEMICAL WATER QUALITY RELATIONSHIPS IN THE RED RIVER OF THE NORTH (1984 TO 2006)

THE LATTICE MODEL OF THE RED RIVER

- The river is a one-dimensional grid.
- The nodes are sampling locations.
- The model is adapted from the “Ising” model of statistical thermodynamics.
- Behavior at nodes depends on interactions with nearest neighbor nodes.

STATISTICAL AND GEOCHEMICAL WATER QUALITY RELATIONSHIPS IN THE RED RIVER OF THE NORTH (1984 TO 2006)

THE LATTICE MODEL OF THE RED RIVER

- Interactions between nodes are water quality parameter correlations.
- One moves sequentially from node to node to fix pattern and sequence of tests
- One can also skip or jump over nodes if needed, but the penalty is weaker or no correlations.

STATISTICAL AND GEOCHEMICAL WATER QUALITY RELATIONSHIPS IN THE RED RIVER OF THE NORTH (1984 TO 2006)

OTHER MODEL ATTRIBUTES:

- Use of anchor stations and satellite stations
- Maps water quality gradients
- Studies sites based on grid location
- Exploits a rich mathematical theory
- Correlates small data sets with large data sets
- Can infer dynamic models for limited data sets from the correlations with anchor stations

STATISTICAL AND GEOCHEMICAL WATER QUALITY RELATIONSHIPS IN THE RED RIVER OF THE NORTH (1984 TO 2006)

MODEL LIMITATIONS

- Not easily adapted to systems with backflows or diffusion.
- Limited capabilities in two dimensional systems
- Not easily adapted if interactions between neighboring nodes are non-linear or not subject to linearizing transformations of data.

STATISTICAL AND GEOCHEMICAL WATER QUALITY RELATIONSHIPS IN THE RED RIVER OF THE NORTH (1984 TO 2006)

- Background on Stochastic Version of Ising Model
 - First studied during World War II
 - Existing applications: air quality problems; traffic flow on highways, on bridges and through tunnels; behavior of electrons in alloys; dynamics of xerographic films; behavior of DNA and binding of small ligands to DNA; methods of advancing mathematical theories of Padé determinants, Bessel functions, and solution of integral equations.
 - This Red River study was first major application in hydrology and water chemistry

STATISTICAL AND GEOCHEMICAL WATER QUALITY RELATIONSHIPS IN THE RED RIVER OF THE NORTH (1984 TO 2006)

DATA LIMITATIONS AND PROBLEMS

- Very few comprehensive data sets.
- Data records have gaps and interruptions.
- Nomenclature problems.
- “Censored data.”
- Site designations are inconsistent: name changes, multiple locations with same name
- Poor quality control on data base structure and management

STATISTICAL AND GEOCHEMICAL WATER QUALITY RELATIONSHIPS IN THE RED RIVER OF THE NORTH (1984 TO 2006)

NODE LOCATIONS:

- EMERSON, MB (the international boundary)
- GRAND FORKS, ND
- FARGO-MOORHEAD AREA
- TRIBUTARIES: PEMBINA, HALSTAD AREA

STATISTICAL AND GEOCHEMICAL WATER QUALITY RELATIONSHIPS IN THE RED RIVER OF THE NORTH (1984 TO 2006)

WATER QUALITY CONSTITUENTS STUDIED (only dissolved species):

- Nutrients: total dissolved nitrogen and total dissolved phosphorus
- Major ions: Na, K, Ca, Mg, Cl
- Landscape related substances: Fe, Al, Si

STATISTICAL AND GEOCHEMICAL WATER QUALITY RELATIONSHIPS IN THE RED RIVER OF THE NORTH (1984 TO 2006)

WATER QUALITY CONSTITUENTS NOT STUDIED:

- Nutrients: BOD and COD
- Major ions: sulfate
- Trace elements: (e.g., Cu, Hg, Zn, Mn)
- Biological parameters: microbial entities
- Physical/chemical parameters: total cations, total anions, pH, temperature, Eh
- Constituents as suspended solids or particulate matter

EMERSON, MB RESULTS:

- DISSOLVED PHOSPHORUS
 - Seasonal spikes with spring flooding
 - No clear trend since 1995: levels relative to 1984 are elevated
 - Time series based on monthly averaging of data
- TOTAL DISSOLVED NITROGEN
 - Strong seasonal effect and multiple periodicities
 - System appears to accumulate nitrogen – 12-17 month reserve in Red River
 - Nitrogen and nitrogen loading data show possible fractal behavior and may be amenable to methods related to chaos theory
 - Nitrogen and phosphorus are very weakly correlated with each other
 - Time series based on monthly averaging of data

EMERSON, MB RESULTS:

MAJOR IONS AND NUTRIENTS:

- UNCORRELATED AS CONCENTRATIONS;
STRONGLY CORRELATED AS LOADINGS
- HYDROLOGY OVERWHELMS CHEMISTRY

(next slide illustrates effects of hydrology)

EMERSON, MB RESULTS

MAJOR IONS AND NUTRIENTS:

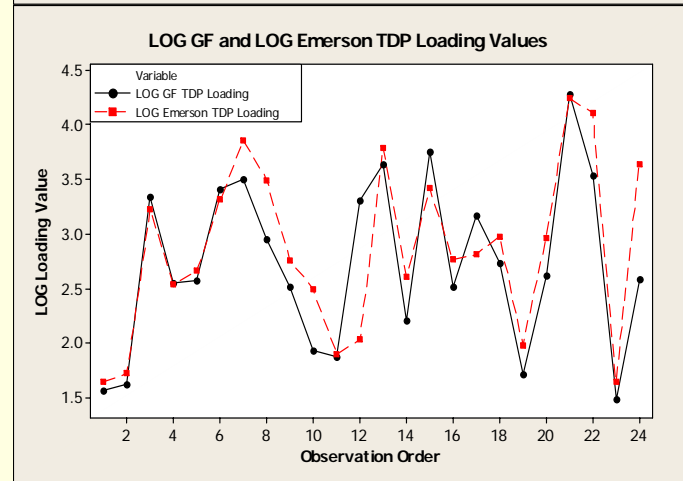
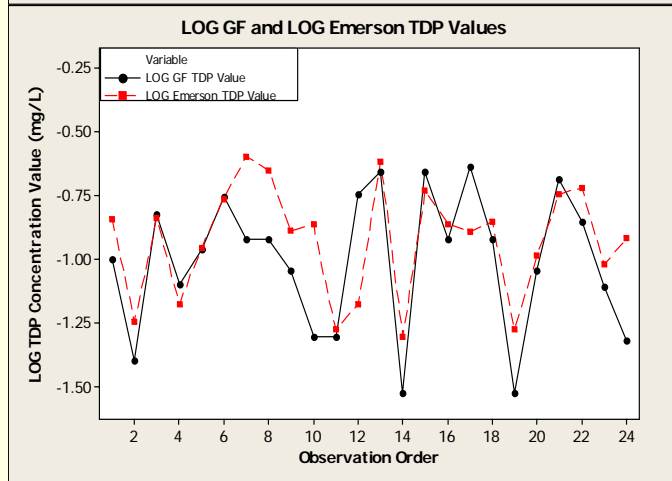
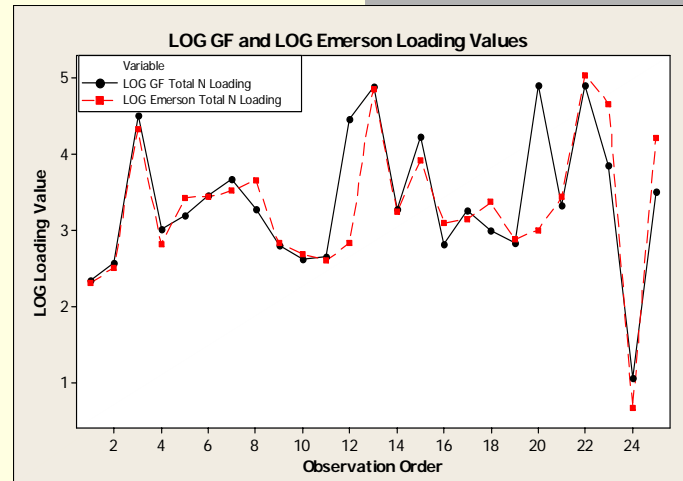
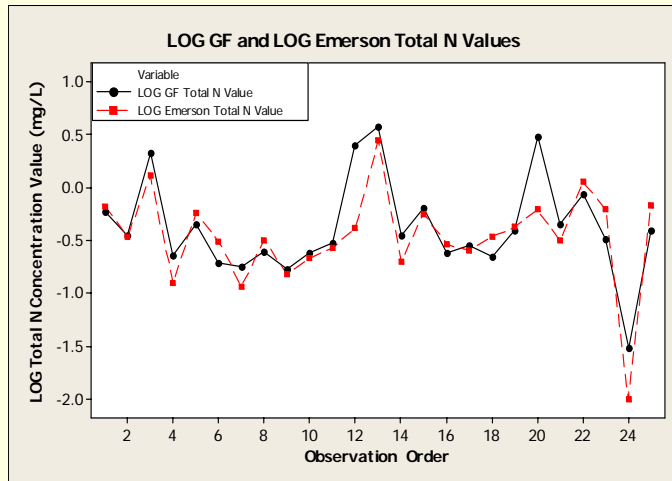
- GEOCHEMICAL CORRELATION MATRIX FOR EMERSON, MB
 - UPPER TRIANGLE IS FOR CONCENTRATIONS, LOWER TRIANGLE IS FOR LOADINGS
 - NUMBERS IN BLACK BOLD FACE ARE SIGNIFICANT AT $P \leq 0.05$
 - NUMBERS IN RED BOLD FACE SIGNIFICANT AT $P \leq 0.01$
 - SILICA NOT INCLUDED IN THIS MATRIX BECAUSE OF INADEQUATE NUMBER OF DATA ENTRIES

	TDP	Calcium	Chloride	Aluminum	Iron	Magnesium	Potassium	Silica	Sodium	Total N
TDP		-0.184	-0.099	0.340	0.096	-0.188	0.059	-----	-0.120	0.286
Calcium	0.374		0.486	-0.157	-0.341	0.937	0.316	-----	0.597	-0.123
Chloride	0.255	0.930		0.234	-0.111	0.506	0.452	-----	0.964	-0.147
Aluminum	0.848	0.637	0.513		0.482	-0.158	-0.013	-----	-0.150	0.037
Iron	0.343	0.704	0.614	0.734		-0.356	-0.027	-----	-0.170	0.269
Magnesium	0.350	0.996	0.940	0.616	0.685		0.365	-----	0.622	-0.134
Potassium	0.394	0.982	0.910	0.655	0.731	0.973		-----	0.487	0.217
Silica	-----	-----	-----	-----	-----	-----	-----		-----	-----
Sodium	0.314	0.956	0.963	0.540	0.600	0.959	0.920	-----		-0.203
Total N	0.551	0.802	0.757	0.671	0.647	0.790	0.838	-----	0.753	

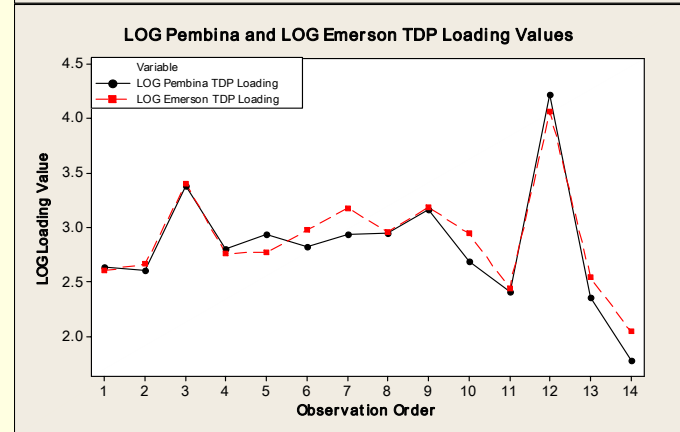
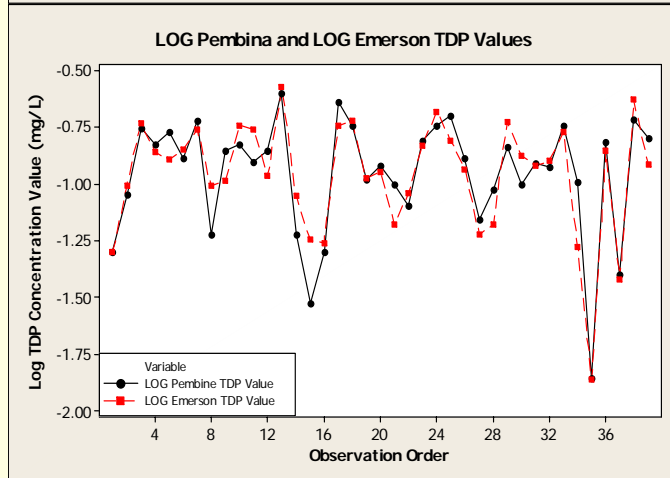
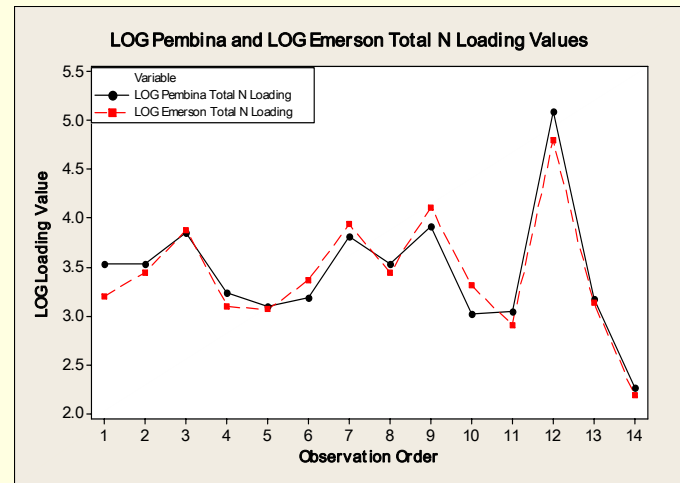
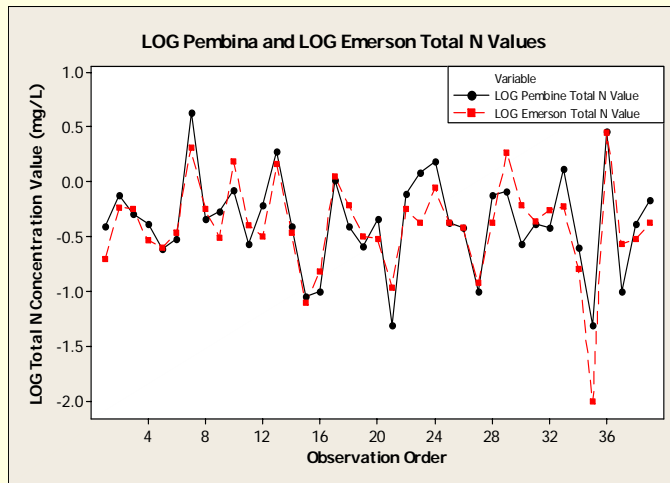
GRAND FORKS, ND RESULTS

- Patterns of nutrients (N and P) similar to Emerson
- System stores nitrogen
- Node behavior inferred from nearest neighbor analyses
- Correlation coefficients of both N and P allow data imputations by regression equations

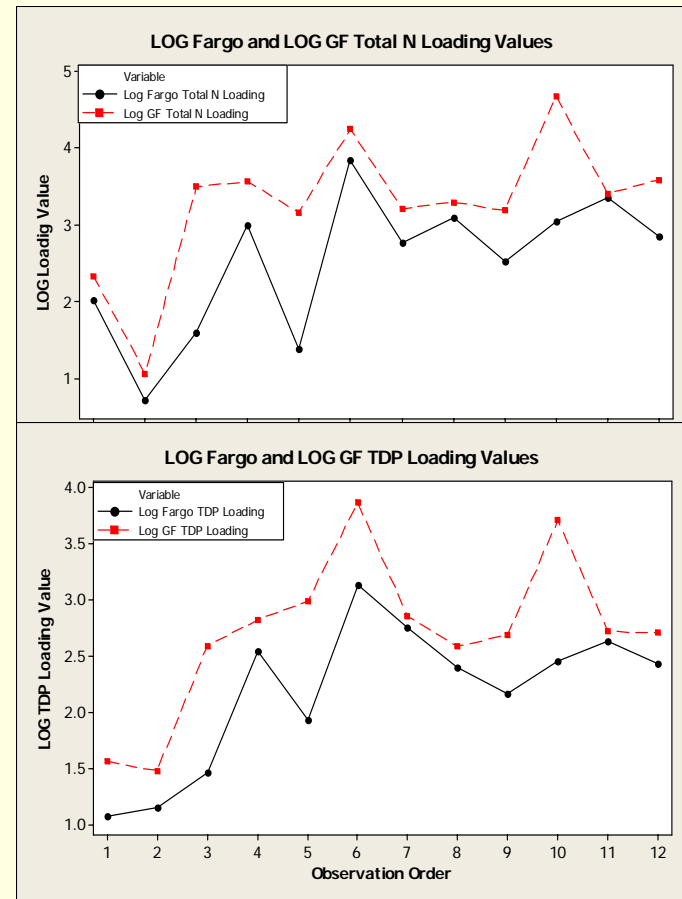
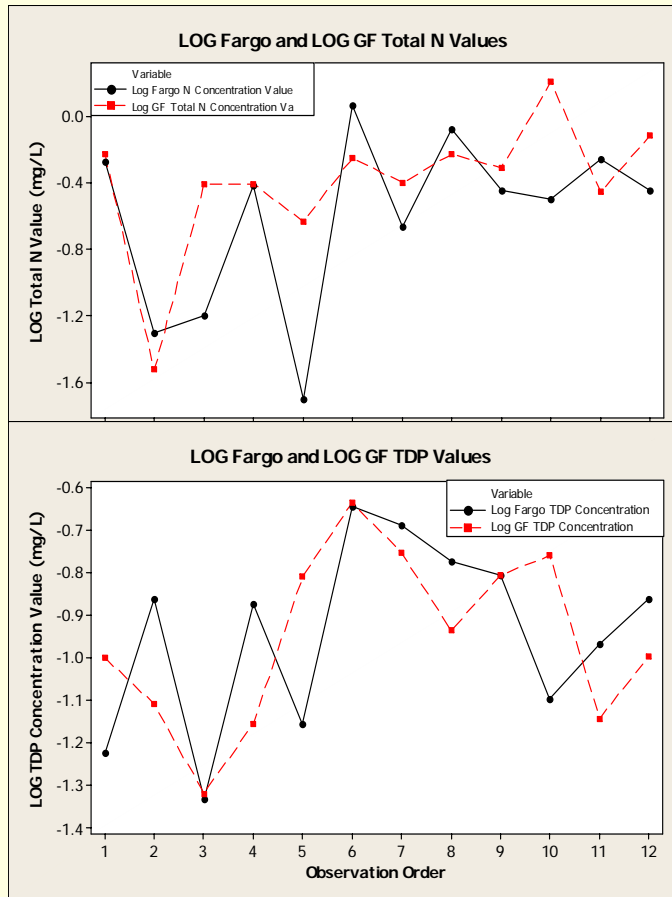
EMERSON-GRAND FORKS CORRELATIONS FOR NUTRIENTS AND THEIR LOADINGS USING NEAREST NEIGHBOR MODEL



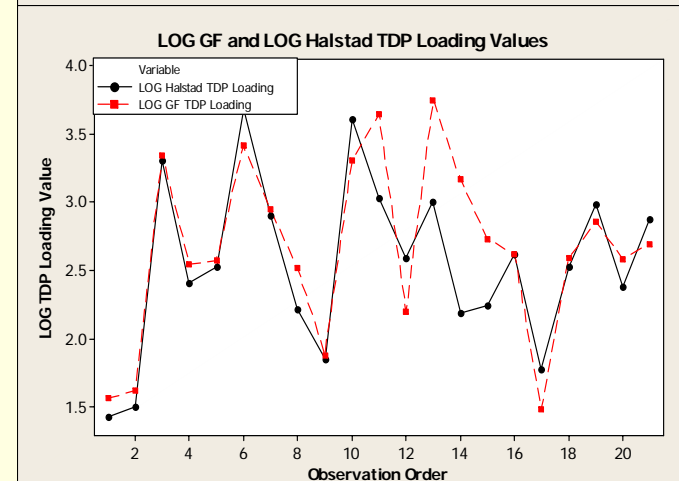
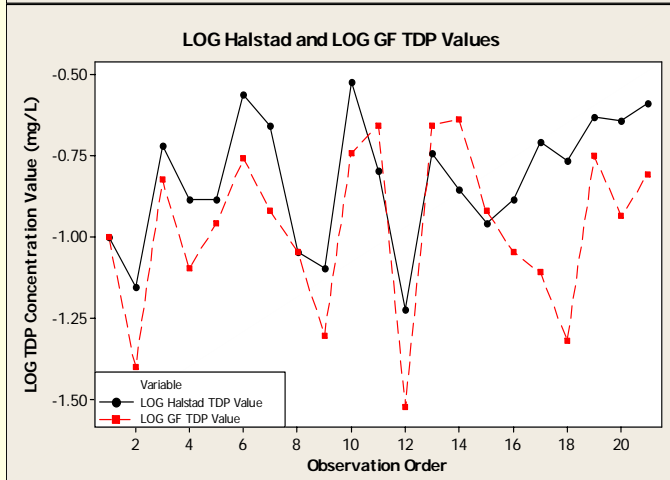
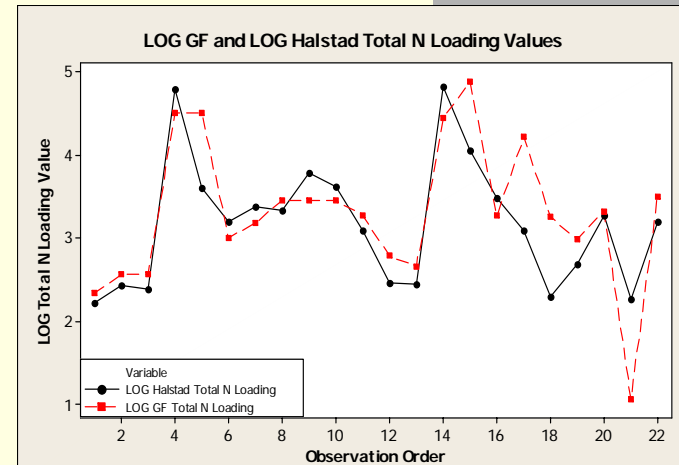
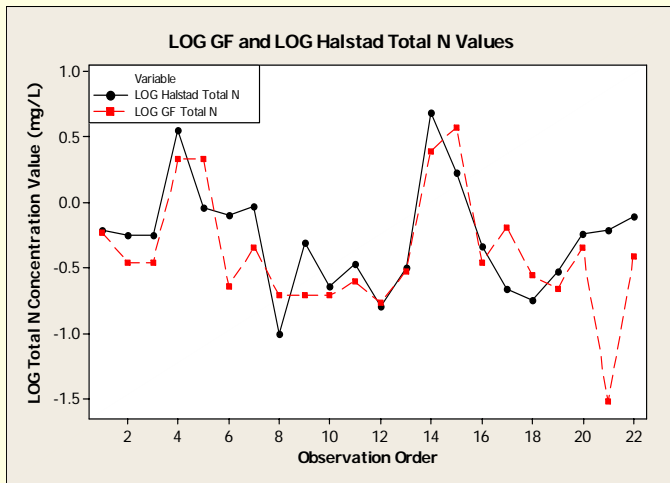
EMERSON-PEMBINA CORRELATIONS USING NEAREST NEIGHBOR CORRELATIONS



GRAND FORKS-FARGO CORRELATIONS USING NEAREST NEIGHBOR CORRELATIONS



GRAND FORKS-HALSTAD USING NEAREST NEIGHBOR CORRELATIONS



GRADIENT OF CORRELATIONS FROM BETWEEN STATION STUDIES (NEAREST NEIGHBORS) FOR NUTRIENTS

Location	TDP	TN
Emerson – Emerson	$r = 1.000$	$r = 1.000$
Emerson – Pembina	$r = 0.907$	$r = 0.847$
Emerson - Grand Forks	$r = 0.704$	$r = 0.857$
Halstad – Grand Forks	$r = 0.695$	$r = 0.648$
Grand Forks – Fargo	$r = 0.489$	$r = 0.582$

r = Pearson correlation coefficient (logarithmic)

GRADIENT OF CORRELATIONS FROM BETWEEN STATION STUDIES (NEAREST NEIGHBORS) FOR NUTRIENT LOADINGS

Location	TDP Flux	TN Flux
Emerson – Emerson	$r = 1.00$	$r = 1.00$
Emerson - Pembina	$r = 0.973$	$r = 0.961$
Emerson - Grand Forks	$r = 0.832$	$r = 0.890$
Grand Forks – Halstad	$r = 0.870$	$r = 0.800$
Grand Forks - Fargo	$r = 0.814$	$r = 0.749$

r = Pearson correlation coefficient (logarithmic)

COMMENTS ON CORRELATION GRADIENTS

- Loading correlations suitable for regression analyses to impute time series data
- Correlations based on logarithmically transformed data and show that water quality constituent data and loading data both have log-normal distributions
- Data suggests using log-normal probability distribution functions to impute missing data from regressions
- Pattern of correlation coefficients suggests that tributaries have limited effects on water quality constituents in Main Stem of Red River from Fargo to Emerson, even during periods of high flow

Use of Probability Distributions to Impute Data

- Establish frequency distribution of loading data time series and obtain cumulative distribution as a function of value of a particular loading. Obtain mean and variance assuming that loadings follow a log-normal distribution.
- Perform same calculations for flow data and nutrient concentration data
- Choose a value of flow. Find the position of this value of flow on the cumulative distribution curve for flow. Match that position on the cumulative distribution function for loading. Estimate concentration from value of loading divided by value of flow, making suitable adjustment for dimensional constants. (To be used when hydrological data exist, but not concentration data.)
- Choose a value of a nutrient from its cumulative distribution function. Match its position on the loading cumulative distribution function, and estimate loading directly (to be used when concentration data exist, but not hydrological data.)

CRITERIA FOR USE OF REGRESSION EQUATIONS TO IMPUTE MISSING DATA

- At least 8 degrees of freedom for regression line
- A correlation coefficient of 0.83 regardless of numbers of degrees of freedom
- Assures standard error of regression line is less than 30% based on using coefficient of variation. Typical S.E. is less than 12% when this was used

GEOCHEMICAL CORRELATIONS

BASED ON PRINCIPAL COMPONENT
ANALYSIS COMBINED WITH
CORRELATION MATRICES

Grand Forks Major Ion Correlation Matrix

	TDP	Calcium	Chloride	Aluminum	Iron	Magnesium	Potassium	Silica	Sodium	Total N
TDP		-0.189	-0.145	-----	0.102	-0.228	0.526	0.528	-0.07	0.378
Calcium			0.478	-----	-0.127	0.95	0.248	-0.206	0.645	-0.218
Chloride				-----	-0.114	0.48	0.284	-0.236	0.816	0.061
Aluminum					-----	-----	-----	-----	-----	-----
Iron						-0.144	0.006	-0.066	-0.168	0.039
Magnesium							0.183	-0.296	0.685	-0.24
Potassium								0.142	0.305	0.356
Silica									-0.186	0.028
Sodium										-0.085
Total N										

- Note: March 1995 outlier removed. 27 Si values (1991 – 2005). 29 TDP and TN monthly values (1992 – 2005). All other variables have 63 monthly values (1985 – 2005). Only 5 Al measurements - Al was not included in the analysis. Upper triangle = concentration correlation coefficients. Lower triangle = flux correlation coefficients. Black Bold = 0.01 < P value < 0.05, Red Bold = P value < 0.01.

HALSTAD MAJOR ION CORRELATION MATRIX

	TDP	Calcium	Chloride	Aluminum	Iron	Magnesium	Potassium	Silica	Sodium	Total N
TDP		-0.309	0.264	-0.012	-0.121	-0.117	0.168	-0.59	0.312	0.192
Calcium			0.314	-0.354	-0.227	0.85	0.57	0.014	0.454	0.068
Chloride				-0.376	0.062	0.459	0.497	-0.159	0.817	0.267
Aluminum					0.495	-0.345	-0.077	0.239	-0.345	0.406
Iron						-0.378	0.47	0.093	-0.196	0.332
Magnesium							0.486	-0.073	0.596	-0.007
Potassium								-0.042	-0.634	0.431
Silica									-0.254	0.251
Sodium										0.343
Total N										

•Note: 31 total data values. All values are from 1985 –. Black Bold = 0.01<P 1994 due to lack of TN values after 1995.
 Upper right triangle = concentration correlation coefficients value<0.05, Red Bold = P value <0.01.

FARGO MAJOR ION CORRELATION MATRIX

	TDP	Calcium	Chloride	Aluminum	Iron	Magnesium	Potassium	Silica	Sodium	Total N
TDP		-0.071	-0.622	-----	0.584	-0.171	0.241	0.758	0.029	0.511
Calcium			-0.085	-----	-0.529	0.963	0.026	-0.106	0.768	-0.041
Chloride				-----	-0.65	0.119	0.426	-0.147	0.155	-0.244
Aluminum					-----	-----	-----	-----	-----	-----
Iron						-0.705	-0.241	0.137	-0.568	0.264
Magnesium							0.144	-0.11	0.839	-0.141
Potassium								0.44	0.576	0.101
Silica									0.019	0.558
Sodium										-0.054
Total N										

•Note: Only 9 AI values – AI was removed from the analysis. All other variables = 11 measurements. Upper right triangle = concentration correlation coefficients. Black Bold = 0.01<P value<0.05, Red Bold = P value <0.01.

SOURCE ANALYSIS – (dissolved constituents only)

- 1ST PCA Cluster: Major Ions – Cl, Na, K, Ca, Mg
 - Constituents originate naturally in watershed
 - Largest geochemical cluster at all locations
 - Most stable cluster in Red River
 - Accounts For 36-41% of the data variance
 - **No** Correlation as a cluster with nutrients (P, N)

SOURCE ANALYSIS: 2nd PCA CLUSTER: AGRICULTURAL RUN-OFF

- P, N, silica, potassium and iron (a chemical composition of some commercial fertilizers)
- Often related to Si
- Only cluster with potassium outside of major ion cluster with a nutrient; no other major ions
- Second largest cluster at all sites except Fargo
- Cluster weakens between Halstad and Fargo. Correlations with Fe and Si become marginal
- Accounts for 17-28% of the data variance

SOURCE ANALYSIS 3rd PCA CLUSTER: OFTEN RELATED TO IRON

- Fe behaves anomalously in Red River
- May have a mineralogical and/or biological origin
- Accounts for 8-12% of the data variance when it appears outside of agricultural run-off cluster

SOURCE ANALYSIS OBSERVATIONS - FARGO

- Fargo has unusual second PCA cluster which contains chlorine as only positive variable
- Accounts for 25% of variance
- Not related to 1st PCA cluster, suggesting a second chlorine source besides natural watershed and soil sources
- Analysis is NOT capable of identifying this second source

OTHER GEOCHEMICAL RELATIONSHIPS

- NO CORRELATIONS BETWEEN Mg AND N
 - Suggests that vegetation is NOT a major source of nutrients (Mg-N correlation is indicative of chlorophyll molecule in some systems)
- NO CORRELATIONS BETWEEN Al and Si
 - Suggest that erosion of landscape does NOT contribute to water quality except under flooding conditions (Al-Si correlations suggest clay minerals in some systems)
- Si LEVELS ARE QUITE LOW, MAIN SOURCES IS FROM AGRICULTURAL RUNN-OFF
 - Suggests that levels may limit diatom development in Red River system.
- Fe-Cl, N-Cl, Al-Cl CORRELATIONS SOMETIMES APPEAR IN ONE OF THE MINOR CLUSTERS (5-9+)
 - Suggests small constituent sources related to use of chemicals in water and wastewater treatment at selected locations

PRINCIPAL COMPONENT ANALYSES

- The first 4 principal component (clusters) accounted for between 78 and 93% of data variability at all sites
- Trace of determinant of correlations equals the statistical thermodynamic partition function of lattice at location studied
- Partition function expresses, in probability terms, the thermodynamic functions of the lattice and enables use of entropy-based statistical analyses from information theory and physics

NEXT STEPS

- Application of deterministic chaos theory to Nitrogen concentration and Nitrogen flux time series – determination of Lyapunov exponents, scaling relationships, stability parameters (fractal studies)
- Application of entropy-based statistical methods – determination of similarity-difference relationships of sites relative to each other
- Extension of model to new locations having data to obtain more complete descriptions of the River, tributaries and watershed
- Application of intervention analysis and extreme value statistics to loading data for flood conditions for the River
- Study of additional parameters: sulfate, BOD/COD