USGS
science for a changing world

Time-Series Model, Statistical Methods, and Software Documentation for R–QWTREND—An R Package for Analyzing Trends in Stream-Water Quality

Time-Series Model, Statistical Methods, and Software Documentation for R–QWTREND—An R Package for Analyzing Trends in Stream-Water Quality

By Aldo V. Vecchia and Rochelle A. Nustad

Open-File Report 2020–1014
Version 1.2, March 2023

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
DAVID BERNHARDT, Secretary

U.S. Geological Survey
James F. Reilly II, Director

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment—visit https://www.usgs.gov or call 1–888–ASK–USGS.

For an overview of USGS information products, including maps, imagery, and publications, visit https://store.usgs.gov/.

Acknowledgments

Contents

Figures

Tables

Conversion Factors

U.S. customary units to International System of Units

Multiply	By	To obtain
Flow rate		
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)
Mass		
pound, avoirdupois (lb)	0.4536	kilogram (kg)
ton, short (2,000 lb)	0.9072	metric ton (t)
ton, long (2,240 lb)	1.016	metric ton (t)

Supplemental Information

Concentrations of chemical constituents in water are given in either milligrams per liter (mg/L) or micrograms per liter (µg/L).

Abbreviations

>	greater than
<	less than
$-2\ln[Lik]$	minus two times the natural logarithm of the likelihood function
BCY	beginning calendar year
ecode	error code from the nonlinear optimization program
ECY	ending calendar year
GLR	generalized likelihood ratio
Lik	likelihood function
LIK	maximum value of the likelihood function
modnum	model number
MPCA	remark code denoting sample was collected by Minnesota Pollution Control Agency
npnN	nitrate plus nitrite as nitrogen parameter name
PARMA	periodic autoregressive moving average
PTot	total phosphorus parameter name
p-value	attained significance level, or probability value
USGS	U.S. Geological Survey
USGSND	remark code denoting sample was collected by U.S. Geological Survey and analyzed by North Dakota Health Department laboratory
USGSNL	remark code denoting sample was collected by U.S. Geological Survey and no outside laboratory is specified

Time-Series Model, Statistical Methods, and Software Documentation for R–QWTREND—An R Package for Analyzing Trends in Stream-Water Quality

By Aldo V. Vecchia and Rochelle A. Nustad

Abstract

As part of a U.S. Geological Survey water-quality study started in 2018, in cooperation with the International Joint Commission, North Dakota Department of Environmental Quality, and Minnesota Pollution Control Agency, a publicly available software package called R–QWTREND was developed for analyzing trends in stream-water quality. The R–QWTREND package is a collection of functions written in R, an open source language and a general environment for statistical computing and graphics. The package uses a parametric time-series model to express logarithmically transformed concentration in terms of flow-related variability, trend, and serially correlated model errors. Flow-related variability captures natural variability in concentration on the basis of concurrent and antecedent streamflow. The trend identifies systematic changes in concentration in terms of potential step trends, piecewise monotonic trends, or user-specified trends. Maximum likelihood estimation is used to estimate model parameters and determine the best-fit trend model. This report describes the time-series model and statistical methodology behind R–QWTREND and provides formal documentation for installing and using the package.

Introduction

A statistical time-series modeling methodology for analyzing trends in stream-water quality, originally developed by Vecchia (2000) and modified as described in Vecchia (2003, 2005), has been used in several subsequent water-quality studies by the U.S. Geological Survey (USGS; Galloway and others, 2012; Risch and others, 2014; Sando and others, 2014a, b, 2015; Giorgino and others, 2018) and other agencies (Jones and Armstrong, 2001; Johnson and others, 2009; Paquette, 2011; Metropolitan Council, 2014). The time-series methodology uses maximum likelihood estimation (Graybill, 1976) to handle complex (nonmonotonic) trends, complex flow-related variability, and seasonal serial correlation structure. The software used in these previous studies for fitting the time-series model to water-quality monitoring data and using the model for parametric statistical inference (hypothesis tests, probability [p]-values, confidence intervals) was not publicly available, and the methodology was referred to informally as "QWTREND."

As part of a USGS water-quality study started in 2018, in cooperation with the International Joint Commission, North Dakota Department of Environmental Quality, and Minnesota Pollution Control Agency, a publicly available software package called R–QWTREND was developed for applying the time-series methodology. The R–QWTREND package is a collection of functions written in R (R Development Core Team, 2019), an open source language and a general environment for statistical computing and graphics. Several enhancements to the original methodology are included in the new package, including the ability to model step trends based on remark codes, improved handling of censored data, expanded graphical output for verifying and interpreting the model results, and the capability to estimate constituent flux (load).

Purpose and Scope

This report describes the time-series model and statistical methodology behind R–QWTREND and provides formal documentation for installing and using the package. This report, along with the accompanying software package, practice datasets, and examples, provides all of the necessary materials and documentation for using R–QWTREND to analyze and interpret trends in stream-water quality based on long-term (10 or more years) datasets on constituent concentration from discrete stream-water samples collected multiple times per year (quarterly or more frequent sampling) and for which the stream-water sampling location is colocated with a streamflow-gaging station from which a complete record of daily mean streamflow is available.

Time-Series Model

The statistical time-series model used for R–QWTREND is a modified version of the model described in Vecchia (2005). Each month is divided into six approximately (~) 5-day time intervals corresponding to days 1–5, 6–10, …, 21–25, and 26–N, where N is the number of days in the given month. Thus, each year consists of 72 (6×12) time intervals. Let t denote the decimal year of the midpoint of each time interval, let $C(t)$ be the observed concentration from, at most, a single water-quality sample collected sometime during the interval, and let $Q(t)$ be the average of the daily mean streamflow values for the interval. A complete record of daily mean streamflow is assumed to be available from a streamflow-gaging station at the water-quality sampling site, and $Q(t)$ is the average of the daily mean streamflow values for the individual days within each time interval. Hereafter, $Q(t)$ is referred to simply as "streamflow." Each time interval is assumed to have, at most, a single observed concentration from a discrete stream-water sample collected sometime during the ~5-day interval. If there are intervals with more than one sample, the sample nearest to the beginning of the interval is selected. For typical monitoring programs, consisting of weekly or less frequent sampling, $C(t)$ may be missing for most of the time intervals, but there can be no missing values for $Q(t)$.

The logarithmically transformed concentration for each time interval is expressed in terms of a constant mean, flow-related variability, trend, and the model error, as follows:

$$\log[C(t)]=MLC+FRVAR(t)+TREND(t)+E(t) \qquad (1)$$

where

log	is the base-10 logarithm;
C	is constituent concentration, in milligrams or micrograms per liter;
t	is the discrete time interval (72, ~5-day time intervals per year), in decimal years;
MLC	is the mean of the logarithmically transformed concentration;
$FRVAR$	is flow-related variability (dimensionless);
$TREND$	is the concentration trend (dimensionless); and
E	is the model error (dimensionless).

Note that in previous reports based on the model described in Vecchia (2005), an ~10-day time interval (3 intervals per month, or 36 per year) was used. For R–QWTREND, the time interval was cut in half. The various model components of equation 1 are defined in more detail in the following paragraphs.

$FRVAR$ is designed to capture as much natural, flow-related variability in logarithmically transformed concentration as possible and is a function of specially crafted variables, called flow anomalies, that depend on concurrent and antecedent streamflow:

$$\log[Q(t)]=MLQ+LTFA(t)+MTFA(t)+STFA(t) \qquad (2)$$

where

Q	is (daily mean) streamflow, in cubic feet per second;
t	is the discrete time interval (72, ~5-day time intervals per year), in decimal years;
MLQ	is the mean of logarithmically transformed streamflow;
$LTFA$	is the long-term flow anomaly (dimensionless);
$MTFA$	is the midterm flow anomaly (dimensionless); and
$STFA$	is the short-term flow anomaly (dimensionless).

$LTFA$ represents long-term (interannual) streamflow variability and is given by a 1-year trailing moving average of the deviation of logarithmically transformed streamflow from its mean,

$$LTFA(t) = \frac{1}{72}\sum_{j=0}^{71}\{log[Q(t-j/72)]-MLQ\} \qquad (3)$$

where

j is the index of summation.

$MTFA$ represents midterm (seasonal) streamflow variability and is given by

$$MTFA(t) = \frac{1}{9}\sum_{j=0}^{8}\{log[Q(t-j/72)]-MLQ \\ -LTFA(t-j/72)\}. \qquad (4)$$

$STFA$ represents short-term (day-to-day) streamflow variability and is given by

$$STFA(t) = log[Q(t)]-MLQ-LTFA(t)-MTFA(t). \qquad (5)$$

$LTFA$ represents variability in geometric mean streamflow for the previous year (72 time intervals) with respect to the long-term geometric mean, $MTFA$ represents variability in geometric mean streamflow for the previous 45 days (9 time intervals) with respect to geometric mean streamflow for the previous year, and $STFA$ represents variability in logarithmically transformed streamflow for the current time interval with respect to geometric mean streamflow for the previous 45 days. Using the flow anomalies as potential explanatory variables to model flow-related variability often explains much more variability in concentration than using only concurrent flow, $log[Q(t)]$. Because a trailing (rather than centered) moving average is used to compute the anomalies, the shorter term (higher frequency) anomalies lead the longer term (lower frequency) anomalies.

The flow anomalies (eqs. 2–5) are used as regression variables in an equation for computing flow-related variability as follows:

$$FRVAR(t) = \begin{cases} \beta_1 LTFA(t) + \beta_2 MTFA(t) + \beta_3 MTFA(t)^2 + \beta_4 STFA(t) + \beta_5 MTFA(t) \times STFA(t) \\ + \beta_6 \cos(2\pi t) + \beta_7 \sin(2\pi t) + \beta_8 \cos(4\pi t) + \beta_9 \cos(4\pi t) \end{cases} \tag{6}$$

where

$\beta_1, \beta_2, ..., \beta_9$　are model coefficients.

The periodic functions (cosine and sine functions) with periods of 1 year and one-half year are included to model seasonal variation in concentration that is not captured by the flow anomaly terms. Other publications may distinguish between flow-related variability and seasonality but for R–QWTREND, "flow-related variability" includes the flow variables and the seasonal terms.

The flow-adjusted concentration is useful for interpreting the model output and is obtained by subtracting $FRVAR$ from the observed concentration as follows:

$$FAC(t)=\log[C(t)]-FRVAR(t)=MLC+TREND(t)+E(t) \tag{7}$$

where

FAC　is flow-adjusted concentration.

The $TREND$ term in equation 7 is used to model potential temporal trends in the mean of FAC and is assumed to consist of a linear combination of various trend functions as follows:

$$TREND(t) = \sum_{j=1}^{J} c_j F_j(t) \tag{8}$$

where

J　is the total number of trend terms,

c_j　is the jth trend coefficient, and

$F_j(t)$　is the jth trend function.

The trend functions are assumed to consist of four specified types as follows:

1. Piecewise monotonic trends

$$F_j(t) = \begin{cases} -1/2, \text{ if } t < B_j \\ -1/2 + (3/2)\left[\dfrac{t-B_j}{E_j-B_j}\right] - (1/2)\left[\dfrac{t-B_j}{E_j-B_j}\right]^3, \text{ if } B_j \leq t \leq E_j \\ +1/2, \text{ if } t > E_j \end{cases} \tag{9}$$

where

B_j　is the beginning time (decimal year) for the monotonic trend and

E_j　is the ending time (decimal year) for the monotonic trend.

2. Step trends based on a specified time interval

$$F_j(t) = \begin{cases} +1/2, \text{ if } B_j \leq t \leq E_j \\ -1/2, \text{ otherwise} \end{cases} \tag{10}$$

where

B_j is the beginning time (decimal year) for the step trend and

E_j is the ending time (decimal year) for the step trend.

3. Step trends based on specified sample attribute

$$F_j(t)= \begin{cases} +1/2, \text{ if } t \in S_j \\ -1/2, \text{ otherwise} \end{cases} \qquad (11)$$

where

S_j is a subset of times corresponding to a specified sample attribute and

$t \in S_j$ indicates that t is in the subset S_j.

4. Ancillary (user-specified) trends

$$F_j(t)=A(t) \qquad (12)$$

where

$A(t)$ is a user-specified function of t.

Piecewise monotonic trends model gradual, multiyear changes in flow-adjusted concentration because of potential anthropogenic causes such as changes in land use/land cover (urbanization, crop acreage or type, and so on). These trends might occur at different times and in different directions (up or down). Interval-based step trends model abrupt changes in flow-adjusted concentration because of anthropogenic causes occurring at known times (for example, a sewage treatment

upgrade or dam removal). Step trends based on sample attributes model the potential bias (systematic tendency for sample concentrations to over or underestimate actual concentrations) associated with a particular collection or preservation method, laboratory analytical method, or another sample attribute. Ancillary trend variables can be any known, user-specified time series computed for the upstream drainage basin, such as total fertilizer use, percentage of the basin in a given land use/land cover category (urban, cropland, forested), or any other variable that might explain water-quality changes in the upstream drainage basin.

Next, consider the model error (E in eq. 1). For a standard parametric regression analysis, the errors are often assumed to be normally distributed with a mean of zero, constant variance, and no serial correlation. However, in practice, water-quality datasets rarely satisfy all these assumptions because the error variance often differs depending on the time of year and the errors often are serially correlated, especially as the sampling frequency increases and the spacing between adjacent samples decreases. Therefore, a special type of time-series model, called a periodic autoregressive moving average (PARMA) model is assumed for the model errors (Salas and others, 1985; Vecchia, 1985). In R–QWTREND, there are three options for specifying the PARMA model number (table 1). For model 1, the error is expressed as the product of a periodic autoregressive coefficient ($\phi_1[t]$) times the error for the previous time point plus a periodic moving average coefficient ($\theta_0[t]$) times a Gaussian white noise process, $Z(t)$. $Z(t)$ is assumed to consist of a time series of independent and identically distributed standard normal random variables. For model 1, the autoregressive and moving average coefficients are expressed in terms of a single pair of cosine and sine functions with a period of 1 year,

Table 1.   Periodic autoregressive moving average model choices for R–QWTREND.

[E, model error; t, discrete time interval (72, approximately 5-day intervals per year, expressed as decimal year); ϕ_1 and ϕ_2, periodic autoregressive coefficients; ϕ_{1j} and ϕ_{2j} ($j=0,1,\ldots,4$), periodic autoregressive parameters; θ_0, periodic moving average coefficient; σ and θ_{0j} ($j=0,1,\ldots,4$), periodic moving average parameters; Z, time series of uncorrelated standard normal random variables]

Model number	Model equation	Model coefficients	Number of model parameters
1	$E(t)=\phi_1(t)\,E(t-1/72)+\theta_0(t)\,Z(t)$	$\phi_1(t)=\phi_{10}+\phi_{11}\cos(2\pi t)+\phi_{12}\sin(2\pi t)$ $\theta_0(t)=\sigma[\,1+\theta_{01}\cos(2\pi t)+\theta_{02}\sin(2\pi t)\,]$	6
2	$E(t)=\begin{cases} \phi_1(t)\,E(t-1/72)+\phi_2(t)\,E(t-6/72) \\ +\theta_0(t)\,Z(t) \end{cases}$	$\phi_1(t)=\phi_{10}+\phi_{11}\cos(2\pi t)+\phi_{12}\sin(2\pi t)$ $\phi_2(t)=\phi_{20}+\phi_{21}\cos(2\pi t)+\phi_{22}\sin(2\pi t)$ $\theta_0(t)=\sigma[\,1+\theta_{01}\cos(2\pi t)+\theta_{02}\sin(2\pi t)\,]$	9
3	$E(t)=\begin{cases} \phi_1(t)\,E(t-1/72)+\phi_2(t)\,E(t-6/72) \\ +\theta_0(t)\,Z(t) \end{cases}$	$\phi_1(t)=\begin{cases} \phi_{10}+\phi_{11}\cos(2\pi t)+\phi_{12}\sin(2\pi t) \\ +\phi_{13}\cos(4\pi t)+\phi_{14}\sin(4\pi t) \end{cases}$ $\phi_2(t)=\begin{cases} \phi_{20}+\phi_{21}\cos(2\pi t)+\phi_{22}\sin(2\pi t) \\ +\phi_{23}\cos(4\pi t)+\phi_{24}\sin(4\pi t) \end{cases}$ $\theta_0(t)=\sigma\begin{cases} 1+\theta_{01}\cos(2\pi t)+\theta_{02}\sin(2\pi t) \\ +\theta_{03}\cos(4\pi t)+\theta_{04}\sin(4\pi t) \end{cases}$	15

and there are six model parameters that need to be estimated (three parameters for the autoregressive coefficient and three for the moving average coefficient). For model 2, there is an additional autoregressive coefficient relating $E(t)$ to itself at a lag of 6 time intervals (1 month) and a total of 9 parameters. Model 3 is the same as model 2 except the autoregressive and moving average coefficients are expressed in terms of sine and cosine functions with periods of 1 year and one-half year, for a total of 15 model parameters. The methods for estimating the model parameters and guidelines for selecting the model number are given later in the "Statistical Methods" section of this report.

Statistical Methods

This section describes the statistical methods used in the R–QWTREND software package to estimate the time-series model parameters and complete statistical inference (hypothesis tests, p-values, and so on) useful for water quality trend analysis. The methods for computing model output, such as estimated annual geometric mean concentration, flow-weighted average concentration, flux, and exceedance frequencies, are also described. Finally, the methods used to handle censored concentration data and occasional short gaps in the streamflow record are described and guidelines are given regarding the minimum data requirements recommended for application of the time-series model.

Maximum Likelihood Estimation

Fitting the trend model requires estimation of the intercept (*MLC*, eq. 1), nine coefficients for *FRVAR* (eq. 6), the trend coefficients (eq. 8), and the PARMA model parameters (table 1). These coefficients/parameters are estimated jointly using Gaussian maximum likelihood estimation, as described in Vecchia (2000). The estimation method is complicated because of the PARMA model for the errors (table 1) and the presence of missing values for log[$C(t)$]. A periodic Kalman filter (Jimenez and others, 1989) with missing data is used to recursively filter the data to remove serial correlation. The periodic Kalman filter is a linear filter, which, when applied to the data, filters out serial correlation:

$$PMF\{Y_i-MLC-FRVAR_i-TREND_i\}=R_i \qquad (13)$$

where

PMF{.}	is the PARMA model filter,
$Y_i=\log[C(t_i)]$	is the ith (nonmissing) observation,
t_i	is the observation time of the ith observation,
FRVAR_i	is *FRVAR* (eq. 6) for the ith observation,
TREND_i	is the trend (eq. 8) for the ith observation, and
R_i	is the PARMA model residual for the ith observation.

If the PARMA model assumptions are satisfied (see table 1), the PARMA model residuals should be uncorrelated and have an approximate normal distribution with a mean of zero and a variance that depends on i,

$$R_i=\sigma s_i Z_i \qquad (14)$$

where

σs_i	is the standard deviation of R_i and
Z_i	is the standardized PARMA model residual for the ith observation.

The Gaussian likelihood function is given by

$$-2\ln[Lik]=\sum_{i=1}^{N}\ln[\sigma s_i]+\sum_{i=1}^{N}[R_i/(\sigma s_i)]^2 \qquad (15)$$

where

ln	is the natural (base-e) logarithm,
Lik	is the Gaussian likelihood function, and
N	is the total number of nonmissing concentration values.

The Gaussian maximum likelihood parameter estimates are obtained by minimizing equation 15 with respect to the intercept (*MLC*), the coefficients for *FRVAR* (the β_j values in eq. 6), the trend coefficients (the c_j values in eq. 8), and the PARMA model parameters (table 1). An executable FORTRAN program for computing the maximum likelihood parameter estimates is provided as part of the R–QWTREND software package. A numerical method (modified Gauss-Newton method; Dennis and Schnabel, 1996) is used to minimize equation 15. In general terms, if $-2\ln[Lik]$ has a unique, well-defined minimum with respect to the model parameters (as indicated by a positive definite Hessian matrix; Dennis and Schnabel, 1996), the data are sufficient for fitting the trend model. If the numerical method is unable to converge to a unique, well-defined minimum, the data are not sufficient to determine all the model parameters. Minimizing $-2\ln[Lik]$ is equivalent to maximizing *Lik*. To distinguish between the likelihood function (*Lik* in eq. 15) and its maximum value obtained by minimizing equation 15, *LIK* will be used hereafter to denote the maximized value of the likelihood function.

Trend Coefficients and Probability Values

Trend analysis involves specification of one or more potential trend models (combination of one or more piece-wise monotonic trends, step trends, or user-defined ancillary trends; eq. 8), estimation of the trend coefficients, and evaluation of the attained significance levels, or probability values (p-values) of the specified trend models. The overall p-value of a specified trend model is used to test the null hypothesis that all the trend coefficients equal zero versus the alternative that at least one is nonzero. The overall p-value can be determined using the generalized likelihood ratio (GLR) test (Graybill, 1976). The GLR test statistic is

$$G = -2\ln\{LIK_0/LIK_J\} = (-2\ln[LIK_0]) - (-2\ln[LIK_J]) \quad (16)$$

where

G is the GLR test statistic,

LIK_0 is the maximum of the likelihood function for the NULL (no-trend) model, and

LIK_J is the maximum of the likelihood function for the model with J trend coefficients.

The approximate overall p-value for testing the null hypothesis H_0: $\{c_j=0$ for all $j=1,...,J\}$ versus the alternative hypothesis H_A: $\{c_j\neq 0$ for at least one $j\}$ is

$$P = \text{Prob}[X_J > G] \quad (17)$$

where

X_J is a chi-square random variable with J degrees of freedom.

The approximate p-values of the individual trend coefficients can be determined in a similar manner by comparing the likelihood function for the trend model with all the coefficients and the likelihood function for the trend model excluding each individual coefficient,

$$P_j = \text{Prob}[X_1 > (-2\ln[LIK_J]) - (-2\ln[LIK_{J-1|-j}])] \quad (18)$$

where

P_j is the approximate p-value for the jth trend coefficient;

$LIK_{J-1|-j}$ is the maximum value of the likelihood function for the model with $J-1$ trend coefficients, excluding c_j; and

X_1 is a chi-square random variable with 1 degree of freedom.

Evaluating equation 18 for all the trend coefficients requires fitting $J+1$ trend models (the full model plus a reduced model for each of the J coefficients). An alternative approach for computing approximate individual p-values is based on an asymptotic (large sample size), normal approximation for the probability distribution of the maximum likelihood parameter estimates:

$$c_j^* \sim c_j + s_j^* Z \quad (19)$$

where

c_j^* is the maximum likelihood estimator of the jth trend coefficient,

s_j^* is the approximate standard error of the estimated coefficient, and

Z is a standard normal random variable.

The approximate standard errors can be computed as part of the numerical minimization of the full likelihood function with all the trend coefficients; thus (unlike equation 18), there is no need to fit additional models for each j. Based on equation 19, the approximate p-value for testing H_0: $c_j=0$ versus H_A: $c_j\neq 0$ is given by

$$P_j = 2\text{Prob}\{Z > |c_j^*/s_j^*|\} \quad (20)$$

where

$|.|$ denotes absolute value.

Although the approximate individual p-values using equations 18 and 20 generally should be similar, equation 18 is more robust with respect to potential serial correlation of the model errors (as captured using the PARMA model).

When interpreting the trend coefficients, it is useful to express them in terms of raw (untransformed) concentration rather than logarithmically transformed concentration. The piecewise monotonic trends (eq. 9) are defined in such a way that the percentage change in flow-adjusted concentration during the specified trend interval is given by

$$\Delta_j = 100(10^{\wedge}c_j - 1) \quad (21)$$

where

Δ_j is the percentage change in the geometric mean of FAC, in milligrams or micrograms per liter, from the beginning ($t=B_j$) to the end ($t=E_j$) of the trend,

c_j is the trend coefficient for the jth piecewise monotonic trend, and

$\wedge$ denotes exponentiation.

Selecting the Best Trend Model from Several Alternatives

The generalized likelihood ratio tests previously described (eqs. 16–18) are examples of nested alternatives in which the simpler model (corresponding to LIK_0 or $LIK_{J-1|-j}$) is obtained from the more complex model (corresponding to LIK_J) by setting one or more of the trend coefficients equal to zero. There often are applications in which several alternative, nonnested, trend models may be postulated for a given dataset, all of which seem to be reasonable alternatives, and the investigator wishes to determine the most appropriate, or "best," model from the alternatives. The GLR test statistic (G) can be framed in a more general context as follows. Let

$$G_{1\sim2} = (-2\ln[LIK1_{J-K}]) - (-2\ln[LIK2_J])$$
$$P_{1\sim2} = \text{Prob}[X_K > G_{1\sim2}] \quad (22)$$

where

$G_{1\sim2}$ is the GLR statistic for testing if models 1 and 2 are equivalent,

$LIK1_{J-K}$ is the maximum of the likelihood function for a model 1,

$J-K$ is the number of trend coefficients for model 1,

$LIK2_J$ is the maximum of the likelihood function for a model 2,

J is the number of trend coefficients for model 2,

$P_{1\sim 2}$ is the *p*-value, and

X_K is a chi-square random variable with *K* degrees of freedom.

The smaller the *p*-value, the stronger the indication that the more complex model (model 2) is a better trend model than the less complex model (model 1). To avoid overly complex models, especially when making multiple comparisons among several models, it is recommended that a high significance level (low *p*-value), such as a *p*-value less than ($<$) 0.01, be used. When both models have the same number of parameters ($K=0$), the *p*-value is undefined. In that case, a good rule of thumb is to consider models 1 and 2 equivalent unless $G_{1\sim 2}$ is greater than ($>$) 1; in which case, model 2 is preferred over model 1.

Annual Geometric Mean Concentration

The annual geometric mean concentration is a useful statistic for evaluating overall water-quality conditions at a specified sampling location in relation to applicable aquatic benchmarks or in relation to overall water quality at other sampling locations. The annual geometric mean concentration for a given year can be expressed as follows:

$$\log[GMC_y] = \tfrac{1}{72}\sum_{t\in y} EV\{\log[C(t)]\}$$

$$EV\{\log[C(t)]\} = MLC + SFRVAR(t) + TREND(t)$$
$$+ FRVAR(t) - SFRVAR(t)$$

$$SFRVAR(t = y + i/72) = \frac{1}{N}\sum_{j=BCY}^{ECY} FRVAR(j + i/72)$$

(23)

where

GMC_y is the annual geometric mean concentration for year *y*, in milligrams or micrograms per liter.

y   is a specified calendar year;

$EV\{.\}$ denotes the expected value (mean) of the quantity in braces;

$t\in y$ indicates that the summation is for all time point in year *y*;

SFRVAR   is seasonal flow-related variability;

i   is the seasonal time index ($i=1,\ldots, 72$);

j   is an index of summation;

BCY and *ECY*   are the beginning and ending calendar years of the period of record; and

$N=ECY-BCY+1$ is the number of calendar years in the period of record.

SFRVAR is interpreted as the flow-related variability that would occur under the hypothetical assumption that streamflow conditions were the same year after year. Equation 23 can be expressed as

$$GMC_y = TRGMC_y \times FRVGMC_y$$

$$TRGMC_y = 10^{MLC+FABCF+TREND_y}$$

$$FRVGMC_y = 10^{\frac{1}{72}\sum_{t\in y}[FRVAR(t)-SFRVAR(t)]}$$

(24)

$$FABCF = \frac{1}{72}\sum_{t\in y} SFRVAR(t)$$

$$TREND_y = \frac{1}{72}\sum_{t\in y} TREND(t)$$

where

$TRGMC_y$ is the trend in annual *GMC* for year *y*;

$FRVGMC_y$ is the flow-related variability in annual *GMC* for year *y*;

FABCF   is a flow-averaged bias correction factor; and

$TREND_y$ is the annual average of *TREND(t)* for year *y*.

The estimated value of GMC_y is obtained by substituting the maximum likelihood estimators of the model parameters in place of the true parameters in equation 24. The trend in annual geometric mean concentration defined in equation 24 is often called a "flow-averaged" or "flow-normalized" trend because it is an unbiased estimate of the annual geometric mean concentration that would be observed under the hypothetical assumption that streamflow, and hence flow-related variability, was the same year after year ($FRVAR=SFRVAR$).

Annual Flow-Weighted Mean Concentration

When using the trend model (eq. 1) for estimating quantities such as annual flow-weighted mean concentration or annual mean flux, it is important to ensure that the estimates are approximately unbiased when expressed in terms of raw (untransformed) concentration or flux. Annual flow-weighted mean concentration is defined as

$$FWMC_y = \frac{\sum_{t\in y} EV\{C(t)\}Q_V(t)}{\sum_{t\in y} Q_V(t)}$$

(25)

where

$FWMC_y$ is the annual flow-weighted mean concentration for calendar year *y*, in milligrams or micrograms per liter, and

$Q_V(t)$ is the total flow volume, in cubic meters, for time interval *t*.

Given the trend model (eq. 1) and the PARMA model for the errors (table 1), concentration has a lognormal distribution with expected value

$$EV\{C(t)\} = 10^{\wedge}\{MLC + FRVAR(t) + TREND(t)$$
$$+ (1/2)\ln[10]\times Var[E(t)]\}$$

(26)

where

$Var[E(t)]$ is the error variance.

In equation 26, the error variance depends on the time of year and can be computed directly from the PARMA model (table 1) given the PARMA model parameters. The annual *FWMC* can be expressed in a similar manner to equation 24 is as follows:

$$FWMC_y = TRFWMC_y \times FRVFWMC_y$$
$$TRFWMC_y = 10 \wedge \{MLC + FABCF* + TREND_y\} \quad (27)$$

where

$TRFWMC_y$ is the trend in in annual *FWMC* for year y,

$FRVFWMC_y$ is the flow-related variability in annual *FWMC* for year y, and

$FABCF*$ is a flow-averaged bias correction factor.

The trend ($TRFWMC_y$) defined in equation 27, which can also be referred to as "flow-averaged " trend, is the trend that would have occurred under the hypothetical assumption that flow conditions for the analysis period were the same every year. Because *FWMC* (eq. 25) is expressed in terms of a flow-weighted average of raw (untransformed) concentration, *FABCF** and $FRVFWMC_y$ (eq. 27) are much more complicated compared to *FABCF* and $FRVGMC_y$ (eq. 24), where the latter formulas were easily derived because *GMC* (eq. 23) is a simple average of logarithmically transformed concentration. In R–QWTREND, *FABCF** and $FRVFWMC_y$ are computed internally using a numerical algorithm and explicit formulas are not provided.

The estimated value of $FWMC_y$ is obtained by substituting the maximum likelihood estimators of the model parameters in place of the true parameters in equation 25. By the asymptotic (large sample) properties of maximum likelihood estimation (Graybill, 1976), as the sample size becomes large, the resulting estimated value is approximately unbiased and optimal (minimum variance). However, for small sample sizes, the estimator may have substantial bias (Cohn and others, 1989). As a general rule of thumb, the estimated value of *FWMC* should be approximately unbiased for sample sizes of at least 60 observations and record lengths of at least 10 years, provided the samples are reasonably spread out among seasons (time of year), flow conditions (as represented by the flow anomaly terms), and calendar years. The minimum data requirements recommended for the application of R–QWTREND are described later in this section.

Annual Mean Flux

Annual mean flux (often referred to as "load") is an important statistic for determining which subbasins of a large watershed contribute the most constituent mass (load). Annual mean flux is defined as

$$FLUX_y = \tfrac{10^6}{365} FWMC_y \times AFV_y \quad (28)$$

where

$FLUX_y$ is annual mean flux for calendar year y, in metric tons (1,000 kilograms [kg]) per day (if concentration is in milligrams per liter) or kilograms per day (if concentration is in micrograms per liter);

$FWMC_y$ is annual flow-weighted mean concentration (eq. 25), in milligrams or micrograms per liter; and

AFV_y is annual flow volume, in cubic meters.

In a similar manner to equation 27, the annual mean flux can be expressed as

$$FLUX_y = TRFLUX_y \times FRVFLUX_y$$
$$TRFLUX_y = 10 \wedge \{MLC + FABCF** + TREND_y\} \quad (29)$$

where

$TRFLUX_y$ is the trend in in annual flux for year y,

$FRVFLUX_y$ is the flow-related variability in annual flux for year y, and

$FABCF**$ is a flow-averaged bias correction factor.

As described previously for *FWMC*, the estimated value of $FLUX_y$ (obtained by substituting maximum likelihood parameter estimates in place of true parameter values in eq. 28) should be approximately unbiased for a large sample size (at least 60 observations) but may have substantial bias for a smaller sample size.

Flow-Averaged Exceedance Probability

Flow-averaged exceedance probabilities are useful for determining how the likelihood of exceeding a specified concentration threshold during a fixed time of year changes through time in response to potential concentration trends. Given the time-series model (eq. 1) for logarithmically transformed concentration, the probability of exceeding a specified concentration is given by

$$P_e(t) = \text{Prob}\{C(t) > C_e\} =$$
$$\text{Prob}\left\{ Z > \frac{\log C_e - MLC - FRVAR(t) - TREND(t)}{SD\{E(t)\}} \right\} \quad (30)$$

where

P_e is the exceedance probability,

C_e is a specified exceedance level,

$SD\{E(t)\}$ is the standard deviation of the error, $E(t)$, and

Z is a standard normal random variable.

Year-to-year differences in *FRVAR* generally cause substantial interannual variability in the exceedance probability. Therefore, alternative probabilities, called flow-averaged

exceedance probabilities, are computed under the hypothetical assumption that flow-related variability during each year is equal to the seasonal average flow-related variability:

$$FRVAR(t)=SFRVAR(t) \qquad (31)$$

where

$SFRVAR$ is seasonal flow-related variability (eq. 23).

Combining equations 30 and 31 yields

$$FAP_e(t) =$$
$$\text{Prob}\left\{ Z > \frac{\log C_e - MLC - SFRVAR(t) - TREND(t)}{SD\{E(t)\}} \right\} \quad (32)$$

where

FAP_e is the flow-averaged exceedance probability.

For any given decimal time, $t=y+i/72$, $FAP_e(t)$ is interpreted as the probability that $C(t)$ will be greater than the specified exceedance level given "typical" (annually averaged) flow conditions for that time of year.

Annual Flow-Averaged Exceedance Frequency

The flow-averaged exceedance probability (eq. 32) indicates the relative chance of exceeding a specified level of concern during any fixed 5-day time interval. For example, if $FAP_e(t)=0.9$, there is a 90-percent chance that concentration at time t exceeds the level of concern, whereas $FAP_e(t)=0.1$ indicates only a 10-percent chance of exceeding the level of concern. Another important consideration for some applications is to determine the amount of time during a given year that concentration is expected to exceed the specified level of concern. Given the flow-averaged exceedance probabilities (eq. 32), annual flow-averaged exceedance frequency is defined as follows:

$$FAF_e(y) = \frac{1}{72} \sum_{i=1}^{72} FAP_e(y + i / 72) \qquad (33)$$

where

$FAF_e(y)$ is the annual flow-averaged exceedance frequency for year y.

The annual flow-averaged exceedance frequency is interpreted at the expected fraction of time during year y that concentration exceeds the specified level of concern, assuming "typical" (flow-averaged) conditions. For example, if $FAF_e(y)=0.25$, we would expect to see about 3 decimal months (18, 5-day time intervals) during year y with concentrations greater than the specified exceedance level. If $FAF_e(y)=0.001$, which is small in relation to the model time step ($1/72=0.015$), we would expect virtually none of the 5-day intervals to exceed the specified level.

Censored Concentration Data

The maximum likelihood parameter estimation methods previously described in this report assume that the concentration data used to fit the model are known (uncensored) values. Exact maximum likelihood parameter estimation with censored data and PARMA errors (table 1) is not tractable; however, an alternative, approximate maximum likelihood estimation methodology can be used when there is a moderate amount (less than about 25 percent) of censored data. The ith observed concentration is censored if it is known to be less than a specified detection limit,

$$C(t_i)<DL(t_i) \qquad (34)$$

where

t_i is the time index of a censored observation,

$C(t_i)$ is the true (but unknown) concentration for the censored observation, and

$DL(t_i)$ is a specified detection limit.

Concentration values for censored observations are imputed using a preliminary trend model with uncorrelated errors:

$$\log[C(t)]=MLC+FRVAR(t)+TREND^*(t)+\sigma Z^*(t) \qquad (35)$$

where

$TREND^*$ is a preliminary trend model defined below and

$Z^*(t)$ is a time series of uncorrelated standard normal random variables.

The preliminary trend model consists of a quadratic spline with number and placement of interior knots that depend on the record length,

$$TREND^*(t) = \sum_{j=1}^{D+1} c_j^* F_j^*(t) \qquad (36)$$

where

$D=\text{int}[RL/10]$ is the number of full decades in the period of record;

RL is the record length, in calendar years;

$\text{int}[.]$ is the greatest integer less than or equal to the quantity in brackets; and

$F_j^*(t)$ is the jth basis function ($j=1,2,…,D+1$) for a quadratic spline with D interior knots where the knots are spaced 10 years apart and are symmetrically distributed around the midpoint of the record.

For example, if the beginning and ending calendar years of the record are $BCY=1990$ and $ECY=2004$, then $RL=15$, $D=1$, and there is a single interior knot at decimal year 1997.5. If $BCY=1990$ and $ECY=2010$, then $RL=21$, $D=2$, and there are two interior knots at decimal years 1995.5 and 2005.5.

Given the preliminary trend model (eq. 35), the expected value for a censored observation (eq. 34) is computed using the following formula:

$$\log[C*(t_i)] = MLC + FRVAR(t_i) + TREND*(t_i)$$
$$+ \sigma E[Z \mid Z < Z_i^*]$$ (37)

where

$C*(t_i)$ is the expected value for the censored concentration (eq. 34);

$Z_i^* = \frac{\log[DL(t_i)] - MLC - FRVAR(t_i) - TREND*(t_i)}{\sigma}$ is the model error assuming $C(t_i) = DL(t_i)$; and

$E[Z|Z<Z_i^*]$ is the conditional expected value of a standard normal random variable (Z) given Z is less than Z_i^*.

A two-step procedure is used to estimate values for censored observations:

1. The survreg function in the Survival Analysis package (Therneau, 2019) is used to obtain maximum likelihood estimates of the model parameters for the preliminary trend model with censored observations (eq. 35).

2. For censored observations, the parameter estimates from step 1 are used in equation 37 to compute the estimated concentration $C*(t_i)$. For subsequent trend analysis, the resulting estimated value(s) are treated the same as if they were known (uncensored) observations.

This approach is preferred to the naive method of substituting arbitrary values (such as one-half of the detection limit) for censored observations, while still being computationally straightforward. Because the assumed preliminary trend model does not depend on the actual (and a priori unknown) trend, nor on the actual (and a priori unknown) serial correlation structure, the estimated values remain the same no matter what the fitted trend model might look like. Allowing the estimated values to depend on the actual (rather than preliminary) trend model is too computationally intensive to be feasible. However, extra care should be taken when verifying and interpreting the fitted trend models using this approach, particularly with respect to p-values, especially for moderate to high (greater than 25 percent) censoring rates.

Flagging Potential Outliers

The R–QWTREND methodology is based on a parametric time-series model (eq. 1), in which log-transformed concentration is modeled using a multiple linear regression model with Gaussian (normally distributed), serially correlated errors. As is the case with any parametric model, outliers must be carefully examined and dealt with to ensure that the fitted model is not unduly influenced by a very small fraction (or perhaps even one) of the observations. For a general discussion of outliers in a multiple linear regression setting, see Helsel and Hirsch (1992).

Outliers can generally be grouped into two cases: outliers for which the reported concentration value is erroneous (for example, data transcription error, improper sample collection or preservation methods, laboratory equipment malfunction) or outliers for which the reported concentration is accurate but for which the environmental conditions at the time of the sample were extreme compared to the vast majority of samples (for example, dam failure, record flood or drought). In either case, outliers should be either corrected (if feasible) or removed before analyzing trends. For purposes of trend analysis, provided a small fraction (say, less than about 2 percent) of the observations are removed, there should be little in the way of adverse consequences of removing outliers. Conversely, if the outliers are not removed, there can be serious adverse consequences such as misidentification of trends or highly biased model output, depending on how influential the outliers are in determining the maximum likelihood parameter estimates. In any case, the investigator needs to carefully examine potential outliers and, if not removed, at least verify that they are not adversely affecting the model results.

Much time and effort can be saved by identifying and removing potential outliers before performing a rigorous trend analysis. The preliminary trend model for filling in censored concentrations (eq. 35) provides a good tool for identifying potential outliers before performing a rigorous analysis. For R–QWTREND, an observation is flagged as a potential outlier if the absolute value of the standardized residual from the preliminary trend model (eq. 35), $Z*(t)=\{\log[C(t)]-MLC-FRVAR(t)-TREND*(t)\}/\sigma$, is larger than 3.5 in absolute value. The chance that the absolute value of a standard normal random variable exceeds 3.5 is very small (less than 1 in 2,000), therefore, this criterion is conservative in that it is likely to flag only true outliers. In general, if more than a small fraction (2 percent or less) of observations are flagged as potential outliers using this criterion, the data may not be appropriate for analysis using R–QWTREND.

Missing Streamflow Data

The computations previously described require that there be no missing values for the streamflow variable, $Q(t)$. A small amount of missing streamflow data can be filled in using the streamflow interpolation method described in this section. This interpolation method should be used only if there are occasional short gaps (less than 6 months in duration) in the record. In particular, the following minimum streamflow data requirements are recommended:

- At least 80 percent (four out of five) of the calendar years in the period of analysis have a compete streamflow record, including the first and last year of the period of analysis.

- There are no streamflow gaps longer than 6 months in duration.

If these recommendations hold, the following steps may be used to fill in missing streamflow record. Let the missing streamflow values for a generic gap in the record be given by

$$QMIS = \{Q(t), t_b < t < t_e\} \tag{38}$$

where

$QMIS$	is the set of the missing streamflow values for a generic gap in the record,
t_b, t_e	are the beginning and ending times of the gap, and
$Q(t_b), Q(t_e)$	are known (nonmissing) streamflow values bracketing the gap.

The first step for filling in the streamflow gap is to compute a time series of seasonal average streamflow:

$$SAQ(t = k + j / 72) = Ave\{Q \in NM_j\}$$
$$j = 1, 2, ..., 72; k = BCY, ..., ECY \tag{39}$$

where

SAQ	is seasonal average streamflow,
$t=k+j/72$	is the decimal year,
k	is the integer year,
j	is the integer season,
NM_j	is the set of all nonmissing streamflow values for season j, and
$Ave\{.\}$	denotes the average of the values in the braces.

Next, to remove seasonality in the streamflow time series, streamflow is divided by SAQ to obtain flow ratios,

$$FR(t) = Q(t) / SAQ(t) \tag{40}$$

where

FR	is the flow ratio.

Note that the missing values for FR correspond to the missing values for Q. Interpolated streamflow values for the generic gap (eq. 38) are computed by linearly interpolating the streamflow ratios across the gap and multiplying the interpolated ratios by the seasonal average streamflow as follows:

$$Q^{\sim}(t) = SAQ(t)\{FR(t_b) + \frac{t - t_b}{t_e - t_b}[FR(t_e) - FR(t_b)]\}$$
$$t_b < t < t_e \tag{41}$$

where

$Q^{\sim}$	is interpolated streamflow for the generic gap $QMIS$ (eq. 38).

This streamflow interpolation method relies on the assumption that seasonal streamflow conditions (as indicated by the flow ratios) across an occasional short (less than 6 month) gap are similar to conditions immediately before and after the gap. The method may not be appropriate when flow

conditions during the gap are thought to be considerably more extreme (higher or lower) compared to conditions before or after the gap, or for flashy streams where flow conditions can vary rapidly during the gap. In such cases, more rigorous streamflow interpolation methods such as rainfall-runoff modeling or routing/reverse routing of known streamflows from nearby upstream/downstream locations should be considered.

Recommended Minimum Data Requirements

The statistical methods previously described are based upon two general assumptions: (1) that the data can be modeled using the general framework described in the "Time-Series Model" section of this report and (2) that the data available to fit the model are sufficient to allow the asymptotic (large-sample) properties of Gaussian maximum likelihood estimation to be applied. Whether or not a particular dataset is sufficient to obtain reliable estimates of trend coefficients and p-values, annual flow-weighted average concentration, annual flux, and other quantities using R–QWTREND depend on a host of factors, including the number of observations, record length, sampling design (distribution of samples among different years, seasons, and flow conditions), and the complexity of the trend model. Although there are no minimum data requirements that are guaranteed to provide reliable results for every possible water-quality constituent and sampling location, a few general recommendations are provided to lead to reliable results for most applications:

1. At least 10 separate calendar years with 1 or more observations (water-quality samples) in each of the following 3-month windows: January–March, February–April, March–May, April–June, May–July, June–August, July–September, August–October, September–November, and October–December.

2. A total of at least 60 observations.

3. At most 25-percent censored data.

4. Minus 2 times the logarithm of the likelihood function ($-2\ln[Lik]$; eq. 15) has a well-defined minimum (positive definite Hessian matrix) with respect to the model parameters.

5. The model assumptions are reasonable, judging by examination of diagnostic model output (see example applications later in this report for suggested diagnostic output and model verification).

These requirements ensure that observations are reasonably spread out among multiple (10+) years and among sliding 3-month seasons within each year, starting with January–March and ending with October–December. For example, 10 years of bimonthly sampling (sampling every 2 months, 6 samples per year, 60 observations) would satisfy the minimum requirements, as would 15 years of quarterly sampling (sampling every 3 months, 4 samples per year,

60 observations). However, 10 years of monthly sampling during April–September (60 observations) would not satisfy the requirements, nor would monthly sampling for 5 years (60 observations).

R–QWTREND Software Documentation

The R–QWTREND package is a collection of functions written in R (R Development Core Team, 2019), an open source language and a general environment for statistical computing and graphics. Although R can be run on a variety of operating systems, including Linux, Mac OS, UNIX, and Windows, R–QWTREND can only be run using the following requirements:

- Windows 10 operating system

- R (version 3.4 or later; 64 bit recommended)

- RStudio (version 1.1.456 or later).

A flowchart of the R–QWTREND functions, inputs, and outputs is shown in figure 1. In this section, instructions for installing the software are provided, and the components of figure 1 are described in detail. Examples are provided to illustrate the R–QWTREND commands and provide hands-on practice using the software to analyze trends for practice datasets provided with the package.

Initializing R–QWTREND

The R–QWTREND software package is provided as an appendix to this report. The package consists of a folder called QWTREND2018V4, which contains the files described in table 2. This folder can be downloaded as described in the appendix. For the practice problems in this section, the files in this folder are assumed to be at the following pathname:
QWModPath = "C:\\RQWTREND\\QWTREND2018V4\\"
To get started, use RStudio to open the QWTrendV4_practice.RData workspace and the script file StartQWTrendV4.R (fig. 2). To initialize R–QWTREND, run the commands in lines 1–17 of the script file. These commands load the required R packages (truncnorm, survival, and splines); specify the path of the QWTREND2018V4 folder (QWModPath); and create the R–QWTREND functions plotQWtrend, prepQWdata, and runQWmodel (R Development Core Team, 2019; Therneau, 2019; fig. 3). If the pathname of the QWTREND2018V4 folder differs from the one shown above, the pathname needs to be changed to the correct pathname before running line 11. Four dataframes provided in the practice workspace are used for all practice problems in this section.

Input Dataframes

Two dataframes are required as inputs for R–QWTREND: a water-quality dataframe (generic name QWdata) and a daily discharge (streamflow) dataframe (generic name DDdata). These dataframes must be in the format described later in this section. Users are required to import water-quality and streamflow data into R and may need to transform the data into the proper format before running R–QWTREND. The simplest file format for importing data into R is a comma delimited Excel (".csv") file, which can be imported using the read.csv() function. Users with limited R programming skills may wish to transform their data into the proper format using Excel before importing the data into R. Alternatively, the data can be imported first and then transformed into the proper format using R commands. The examples provided in this section use four dataframes that have already been imported into R and transformed into the proper format: RRFargo_Prac_DDdata, RRFargo_Prac_QWdata, RRHalstad_Prac_DDdata, and RRHalstad_Prac_QWdata. These dataframes contain nutrient concentration data and daily streamflow data for the Red River of the North at Fargo, North Dakota (USGS streamflow-gaging station 05054000), and the Red River at Halstad, Minnesota (USGS streamflow-gaging station 05064500), for 1970–2017. The streamflow data used in the examples were downloaded from the USGS National Water Information System database (U.S. Geological Survey, 2019) and the nutrient concentration data were downloaded from the National Water Quality Monitoring Council database (National Water Quality Monitoring Council, 2019). The dataframes must be configured as follows:

- Water-quality input data (for example, RRFargo_Prac_QWdata, see fig. 4):

  - First column must have type character and name "date" and consists of the date of the water-quality sample, in "YYYY-MM-DD" format. This column should be in ascending order, but there can be dates with multiple rows (multiple samples on the same day).

  - Concentration data must be in a column of type numeric with name "P_XXX," where the first two characters of the name must be "P_," "XXX" is the specified parameter name, and there are no spaces. For this example, "P_npnN" designates nitrate plus nitrite as nitrogen (npnN) concentration (hereafter referred to as nitrate plus nitrite concentration) and "P_PTot" designates total phosphorus (PTot) concentration.

  - Immediately following each parameter column ("P_XXX") must be a remark column ("R_XXX"). This column must be of type character and have the format "YYY=" or "YYY<," where "YYY" is a remark code. There can be no blanks, and the last character must be either "=" (if the value is not censored) or
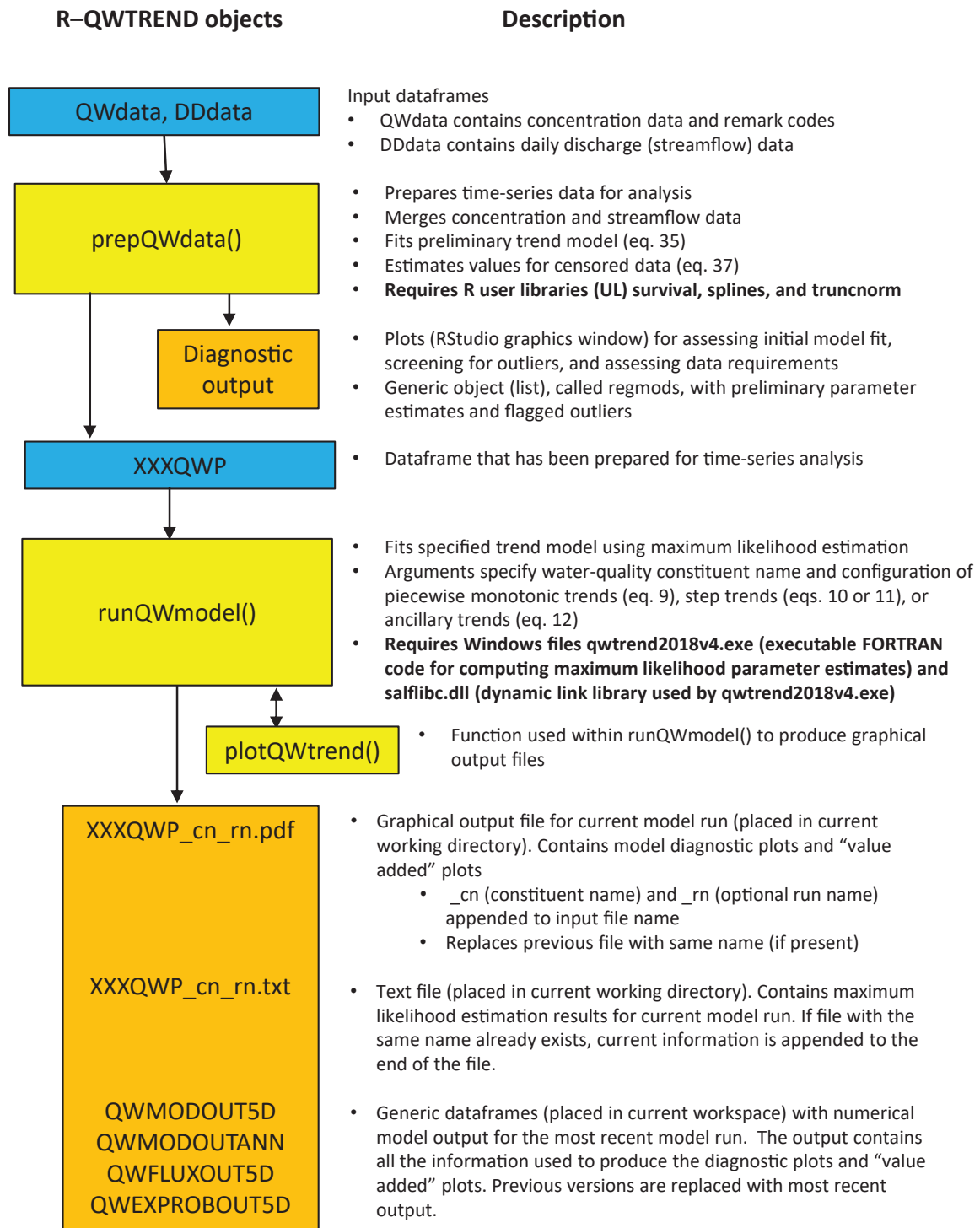
R–QWTREND objects   **Description**

QWdata, DDdata

Input dataframes
- QWdata contains concentration data and remark codes
- DDdata contains daily discharge (streamflow) data

prepQWdata()

- Prepares time-series data for analysis
- Merges concentration and streamflow data
- Fits preliminary trend model (eq. 35)
- Estimates values for censored data (eq. 37)
- **Requires R user libraries (UL) survival, splines, and truncnorm**

Diagnostic output

- Plots (RStudio graphics window) for assessing initial model fit, screening for outliers, and assessing data requirements
- Generic object (list), called regmods, with preliminary parameter estimates and flagged outliers

XXXQWP

- Dataframe that has been prepared for time-series analysis

runQWmodel()

- Fits specified trend model using maximum likelihood estimation
- Arguments specify water-quality constituent name and configuration of piecewise monotonic trends (eq. 9), step trends (eqs. 10 or 11), or ancillary trends (eq. 12)
- **Requires Windows files qwtrend2018v4.exe (executable FORTRAN code for computing maximum likelihood parameter estimates) and salflibc.dll (dynamic link library used by qwtrend2018v4.exe)**

plotQWtrend()

- Function used within runQWmodel() to produce graphical output files

XXXQWP_cn_rn.pdf

- Graphical output file for current model run (placed in current working directory). Contains model diagnostic plots and "value added" plots
  - _cn (constituent name) and _rn (optional run name) appended to input file name
  - Replaces previous file with same name (if present)

XXXQWP_cn_rn.txt

- Text file (placed in current working directory). Contains maximum likelihood estimation results for current model run. If file with the same name already exists, current information is appended to the end of the file.

QWMODOUT5D
QWMODOUTANN
QWFLUXOUT5D
QWEXPROBOUT5D

- Generic dataframes (placed in current workspace) with numerical model output for the most recent model run. The output contains all the information used to produce the diagnostic plots and "value added" plots. Previous versions are replaced with most recent output.

Figure 1.   Flowchart of R–QWTREND functions, inputs, and outputs.

Table 2.   R–QWTREND software files contained in windows folder QWTREND2018V4.

File name	Description
prepQWdataV4.txt	Text file containing R code for creating function prepQWdata().
runQWmodelV4.txt	Text file containing R code for creating function runQWmodel().
plotQWtrendV4.txt	Text file containing R code for creating function plotQWtrend().
qwtrend2018v4.exe	Windows executable file used for computing maximum likelihood parameter estimates.
salflibc.dll	Dynamic link library required by qwtrend2018v4.exe.
QWTrendV4_practice.RData	R workspace with example datasets.
StartQWTrendV4.R	R script for installing and running R–QWTREND.


```
1
2   # Prepare your workspace for running R-QWTREND
3   #
4   # Load required libraries (truncnorm, survival, and splines)
5   #
6       library("truncnorm")
7       library("survival")
8       library("splines")
9   #
10  # Set the pathname for the QWTrend2018V4 folder (use your correct pathname)
11      QwModPath <- "c:\\RQWTREND\\QWTrend2018V4\\"
12  #
13  #  Create the functions prepQWdata, runQWmodel, and plotQWdata
14  #
15      source(paste(QwModPath,"prepQWdataV4.txt",sep=""))
16      source(paste(QwModPath,"runQWmodelV4.txt",sep=""))
17      source(paste(QwModPath,"plotQWtrendV4.txt",sep=""))
18
19  #  Examine practice datasets (RRFargo_Prac_DDdata, RRFargo_Prac_QWdata,
20  #    RRHalstad_Prac_DDdata, RRHalstad_Prac_QWdata)
21
22  # Prepare Fargo data for time series analysis
23      RRFargoQWP <- prepQWdata(RRFargo_Prac_QWdata, RRFargo_Prac_DDdata, 1970, 2017)
24      reqmods
25
26  # Prepare Halstad data for time series analysis
27      RRHalstadQWP <- prepQWdata(RRHalstad_Prac_QWdata, RRHalstad_Prac_DDdata, 1970, 2017)
28      reqmods
29  # Replace 2 outliers for npnN and 6 outliers for PTot and repeat
30      RRHalstad_Prac_QWdata[c(278,480),"P_npnN"] <- NA
31      RRHalstad_Prac_QWdata[c(12,44,45,74,110,217),"P_PTot"] <- NA
32      RRHalstadQWP <- prepQWdata(RRHalstad_Prac_QWdata, RRHalstad_Prac_DDdata, 1970, 2017)
33
34  # Example 1 (npnN for RRFargo)
35      runQWmodel(RRFargoQWP,"npnN",fullout=F)
36      runQWmodel(RRFargoQWP,"npnN",monxx="1975x2015",fullout=F)
37      runQWmodel(RRFargoQWP,"npnN",monxx=c("1985x2000","2000x2015"),fullout=F)
38      # examine ML parameter estimates (see RRFargoQWPnpnN.txt)
39      # rerun the last model with full output and examine graphical output file RRFargoQWPnpnN.pdf
40      runQWmodel(RRFargoQWP,"npnN",monxx=c("1985x2000","2000x2015"),fullout=T)
41
42  #  Example 2 (PTot for RRFargo)
43      runQWmodel(RRFargoQWP,"PTot",runname="_null")
44      runQWmodel(RRFargoQWP,"PTot",monxx=c("1975x1985","1985x2000","2000x2015"),runname="_M1")
45      runQWmodel(RRFargoQWP,"PTot",monxx=c("1975x1985","1985x2000","2000x2015"),remxx=c("MPCA"),runname="_M2")
46  #  Examine parameter estimates (RRFargoQWPPTot_XX.txt) and graphical output (RRFargoQWPPTot_XX.pdf)
47  #  Exceedance probabilities
48      runQWmodel(RRFargoQWP,"PTot",monxx=c("1975x1985","1985x2000","2000x2015"),remxx=c("MPCA"),exlev=c(0.1,0.2,0.3),runname="_ep")
49
50  # Example 3 (PTot for RRHalstad)
51      runQWmodel(RRHalstadQWP,"PTot",monxx=c("1975x1985","1985x2000","2000x2015"),runname="_M1")
52      runQWmodel(RRHalstadQWP,"PTot",monxx=c("1975x1985","1985x2000","2000x2015"),remxx="MPCA",exlev=c(0.1,0.2,0.3),runname="_M2")
53  # Examine graphical output and generic dataframes with numerical output (QWMODOUT5D, QWMODOUTANN, QWFLUXOUT5D, and QWEXPROBOUT5D)
54  # Create your own dataframe of trend variables
55      dyr <- RRHalstadQWP[,"yr"]+(RRHalstadQWP[,"mo"]-1)/12 + RRHalstadQWP[,"da"]/360
56      ancxx <- data.frame(dyr,bs(dyr,knots=seq(1980,2010,10)))
57      runQWmodel(RRHalstadQWP,"PTot",remxx="MPCA",userxx=list(7,ancxx,names(ancxx)[-1]),fullout=T,runname="_M3")
58  # Example of mis-specified trend model which produces error in nonlinear optimization program
59      runQWmodel(RRHalstadQWP,"PTot",remxx="MPCA",monxx=c("1975x1985","1975x1985","2000x2015"),fullout=F)
60  # Close error window before continuing
61
```

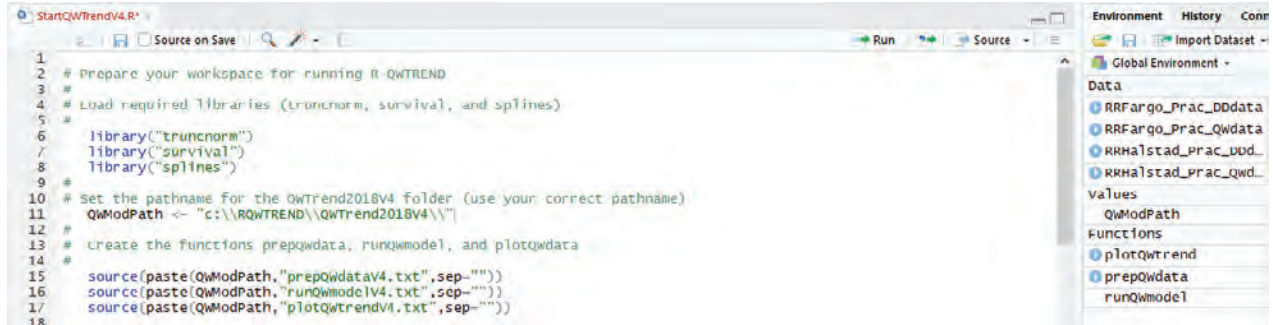**Figure 2.**   Script file (StartQWTrendV4.R) for running R–QWTREND practice problems.

Figure 3.   Initial workspace for running R–QWTREND practice problems.

Figure 4.   Beginning and ending rows of the water-quality dataframe, RRFargo_Prac_QWdata, and the daily discharge dataframe, RRFargo_Prac_DDdata, for input to the prepQWdata() function.

"<" (if the value is left censored). Every nonmissing concentration must have a nonmissing remark code. If there is no remark code, a placeholder such as "_=" or "_<" can be used.

- Daily discharge input data (for example, RRFargo_Prac_DDdata, see fig. 4):

  - There must be a column named "date," of type character, with the date in "YYYY-MM-DD" format.

  - There must be a column named "flow," of type numeric, with daily mean streamflow, in cubic feet per second.

  - Unlike the concentration data, the daily streamflow data must consist of complete calendar years, and there should be only a limited number of days with missing streamflow values (see the earlier "Missing Streamflow Data" section of this report for guidelines regarding missing streamflow). In this case, the flow data consist of a complete record for calendar years 1970–2017.

Preparing Data for Time-Series Analysis Using PrepQWdata

The function prepQWdata prepares the water-quality and daily streamflow data for time-series analysis. This function requires as inputs the water-quality and daily discharge dataframes described previously. The command for running the function has the following form:

$$\text{XXXQWP} \leftarrow \text{prepQWdata(QWdata, DDdata,} \\ \text{yrbeg, yrend)}$$

where
QWdata	is a water-quality input dataframe,
DDdata	is a daily discharge (streamflow) input dataframe,
yrbeg	is the beginning calendar year (integer),
yrend	is the ending calendar year (integer), and
XXXQWP	is an output dataframe (or dataset) that has been prepared for analysis.

The user can assign any name for the output, but it is recommended to use "QWP" as the last three characters to indicate that the data have been prepared for time-series analysis; for example, executing line 23 of the practice script (fig. 2) produces a dataset called RRFargoQWP. Note that a shorter period of record can be used, if desired, by specifying either a beginning year later than 1970 or ending year earlier than 2017, or both. The format of RRFargoQWP is shown in figure 5, which shows the rows corresponding to June–August 1971. Each month consists of six ~5-day time intervals. The first nonmissing concentration value for this example was for parameter npnN from a sample collected on June 3,

1971 (during the first interval of the month, days 1–5). A day of 3 (the midpoint of the interval) is assigned to this observation. The original concentration was left censored (<0.05), and the concentration value for that observation (0.023) is the estimated value from the preliminary trend model (eq. 37). The streamflow, or discharge, value (column name "dis") for the same time interval (324.8) is daily mean streamflow, in cubic feet per second, for June 1–5, 1971. There are no missing values for streamflow. The next nonmissing concentration value for this example was for the third time interval (days 11–15) of July 1971 and consists of an npnN concentration of 0.430. Although the original concentration was from July 14, 1971, day 13 (the midpoint of the interval) is used for the day. Columns with the remark codes are at the end of this dataframe (remark codes are defined below).

Several pages of rough plots are produced by prepQWdata (fig. 6). These plots are shown in the default RStudio Plots window. Page 1 of the plots shows the flow anomalies (eqs. 3–5, fig. 6A). The horizonal black line in all three plots equals MLQ. The top plot shows log[Q] (black line) along with $MLQ+LTFA$ (coral line). For this site there were 12 days with zero flow (all during 1976). To allow logarithmic transformation, flow values less than 0.1 cubic foot per second are replaced by 0.1 when running prepQWdata. The middle plot shows log[Q]–$LTFA$ (black line) along with $MLQ+MTFA$ (coral line). $MTFA$ captures seasonal flow variability remaining after removing $LTFA$. The bottom plot shows $MLQ+STFA$ (coral line), which captures short-term (day-to-day) flow variability after removing $LTFA$ and $MTFA$.

Two additional pages of plots are produced for each water-quality parameter (fig. 6B–C). Page 2 of the plots shows logarithmically transformed npnN concentration versus decimal year and decimal season or decimal month (fig. 6B). Colors indicate distinct remark codes detected in the input data. In this case, there were three distinct remark codes: MPCA, sample collected and analyzed by Minnesota Pollution Control Agency; USGSND, sample collected by USGS and analyzed by the North Dakota Health Department laboratory; and USGSNL, sample collected by USGS and no outside laboratory specified (generally indicating the sample was analyzed by the USGS National Water-Quality Laboratory). Open circles indicate that the original concentration value was left censored and was replaced by the estimated value using equation 37.

Page 3 of the plots (fig. 6C) is the second page of plots for npnN and can be used to detect and correct potential outliers and spot potential concerns related to different remark codes, gaps in the water-quality record, or other irregularities that may affect the ability to analyze trends. These plots are based on the preliminary trend model (eq. 35). The top plot shows flow-adjusted concentration (log[C]–$FRVAR$) along with the preliminary trend ($MLC+TREND*$, eq. 35). In this case the water-quality record spans four complete decades, so the preliminary trend consists of a quadratic spline with five basis functions (eq. 36). The second plot shows flow-adjusted

	yr	mo	da	dis	P_npnN	P_PTot	R_npnN	R_PTot
103	1971	6	3	324.8	0.023	NA	USGSNL<	NA
104	1971	6	8	345.4	NA	NA	NA	NA
105	1971	6	13	297.2	NA	NA	NA	NA
106	1971	6	18	446.4	NA	NA	NA	NA
107	1971	6	23	383.2	NA	NA	NA	NA
108	1971	6	28	415.4	NA	NA	NA	NA
109	1971	7	3	651.0	NA	NA	NA	NA
110	1971	7	8	1443.4	NA	NA	NA	NA
111	1971	7	13	800.2	0.430	NA	USGSNL=	NA
112	1971	7	18	683.2	NA	NA	NA	NA
113	1971	7	23	389.6	NA	NA	NA	NA
114	1971	7	28	343.5	NA	0.69	NA	MPCA=
115	1971	8	3	211.4	NA	NA	NA	NA
116	1971	8	8	185.4	NA	NA	NA	NA
117	1971	8	13	178.8	NA	NA	NA	NA
118	1971	8	18	110.8	NA	NA	NA	NA
119	1971	8	23	96.6	NA	0.17	NA	MPCA=
120	1971	8	28	76.0	0.120	NA	USGSNL=	NA

Figure 5.   Rows of practice dataset produced by prepQWdata, RRFargoQWP, corresponding to June–August 1971.

and detrended concentration (log[*C*]–*FRVAR*–*TREND**, eq. 35). The solid horizontal line corresponds to *MLC*. In this case, there were no apparent outliers or other abnormalities.

Page 5 of the output from prepQWdata (fig. 6*D*) is the second page of plots for total phosphorus (PTot). The dashed lines in the bottom plot correspond to plus and minus 3.5 times the estimated error standard deviation (these lines were outside of the plot limits for npnN, fig. 6*C*). Points outside of these bounds may indicate potential outliers. In this case, like npnN, there were no apparent outliers. However, after about 2003, there seems to be a higher frequency of negative errors for the USGSND and USGSNL remark codes and positive errors for the MPCA remark code (differences related to remark codes are explored in more detail later in this section).

Next, run line 27 of the script file (fig. 2) to create the practice dataset RRHalstadQWP and examine page 5 of the plot output (fig. 7). In this case, there were five observations for PTot for this site for which the flow-adjusted and detrended concentrations were well outside of the dashed lines, indicating potential outliers. A generic list, called regmods, is produced in the current workspace when running prepQWdata. This list contains information regarding the preliminary regression models, including flagged outliers. The element of this list corresponding to the outliers for PTot is shown

below the plots. These observations correspond to rows 12, 44, 45, 74, and 110 of the original water-quality dataframe (RRHalstad_Prac_QWdata). Another observation during 1985 (row 217) also was determined to be an outlier. Outliers should be carefully examined and either removed or corrected in the original water-quality dataframe before completing a formal trend analysis. For the practice problems in this section, run lines 30–32 of the script file (fig. 2) to replace 6 outliers for PTot and 2 outliers for npnN with missing values and rerun prepQWdata to prepare RRHalstadQWP with outliers removed. After removing outliers, the results for PTot for this site (fig. 8) were similar to the previous results for RRFargoQWP (fig. 6*D*). Although there were no remaining outliers, there tended to be a higher frequency of negative flow-adjusted and detrended concentrations for the USGSND and USGSNL remark codes compared to the MPCA remark code.

Analyzing Trends Using RunQWmodel

After preparing the data using prepQWdata, the next step is to use maximum likelihood estimation to fit alternative trend models, select the best model from among the alternatives, and verify the model assumptions for the selected trend model.
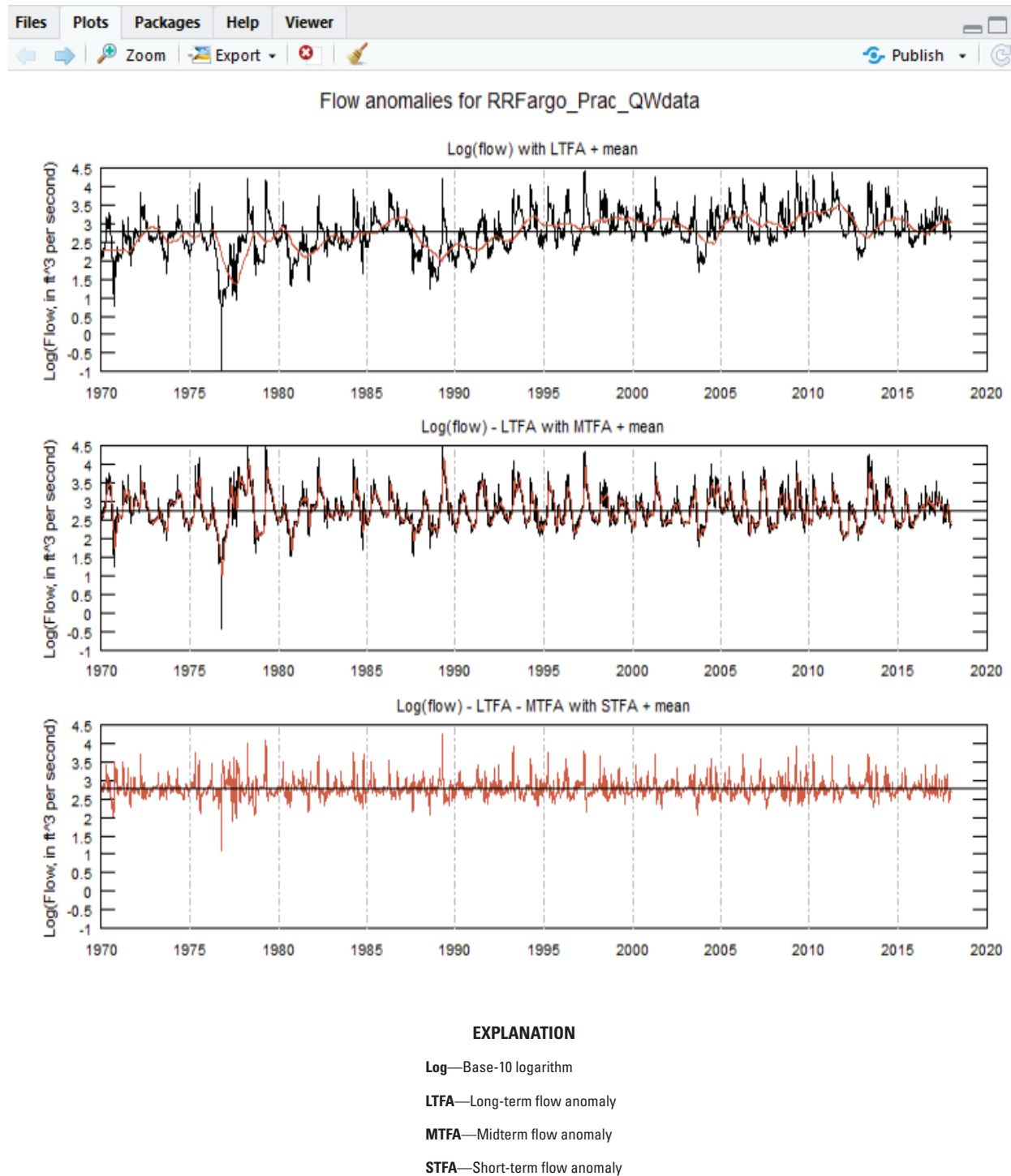
A

EXPLANATION

Log—Base-10 logarithm

LTFA—Long-term flow anomaly

MTFA—Midterm flow anomaly

STFA—Short-term flow anomaly

Figure 6.   Plots produced by prepQWdata for practice dataset, RRFargoQWP. *A*, page 1, flow anomalies; *B*, page 2, nitrate plus nitrite (npnN) concentration data versus decimal year and decimal month; *C*, page 3, nitrate plus nitrite (npnN) flow-adjusted and flow-adjusted and detrended concentration data versus decimal year; *D*, page 5, total phosphorus (PTot) flow-adjusted and flow-adjusted and detrended concentration data versus decimal year.

B

Figure 6. Plots produced by prepQWdata for practice dataset, RRFargoQWP. *A*, page 1, flow anomalies; *B*, page 2, nitrate plus nitrite (npnN) concentration data versus decimal year and decimal month; *C*, page 3, nitrate plus nitrite (npnN) flow-adjusted and flow-adjusted and detrended concentration data versus decimal year; *D*, page 5, total phosphorus (PTot) flow-adjusted and flow-adjusted and detrended concentration data versus decimal year.—Continued
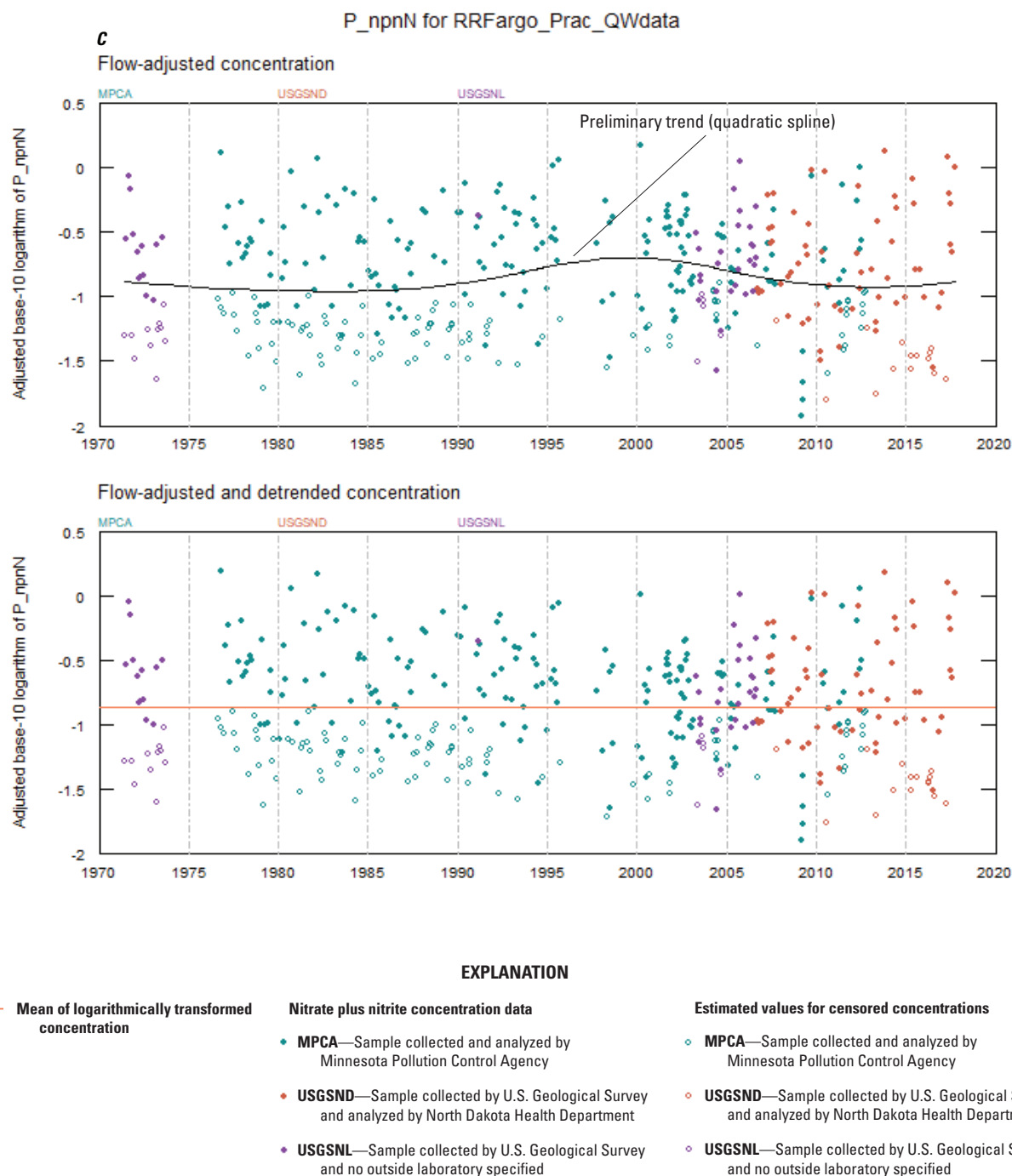
Figure 6.   Plots produced by prepQWdata for practice dataset, RRFargoQWP. *A*, page 1, flow anomalies; *B*, page 2, nitrate plus nitrite (npnN) concentration data versus decimal year and decimal month; *C*, page 3, nitrate plus nitrite (npnN) flow-adjusted and flow-adjusted and detrended concentration data versus decimal year; *D*, page 5, total phosphorus (PTot) flow-adjusted and flow-adjusted and detrended concentration data versus decimal year.—Continued

Figure 6. Plots produced by prepQWdata for practice dataset, RRFargoQWP. *A*, page 1, flow anomalies; *B*, page 2, nitrate plus nitrite (npnN) concentration data versus decimal year and decimal month; *C*, page 3, nitrate plus nitrite (npnN) flow-adjusted and flow-adjusted and detrended concentration data versus decimal year; *D*, page 5, total phosphorus (PTot) flow-adjusted and flow-adjusted and detrended concentration data versus decimal year.—Continued
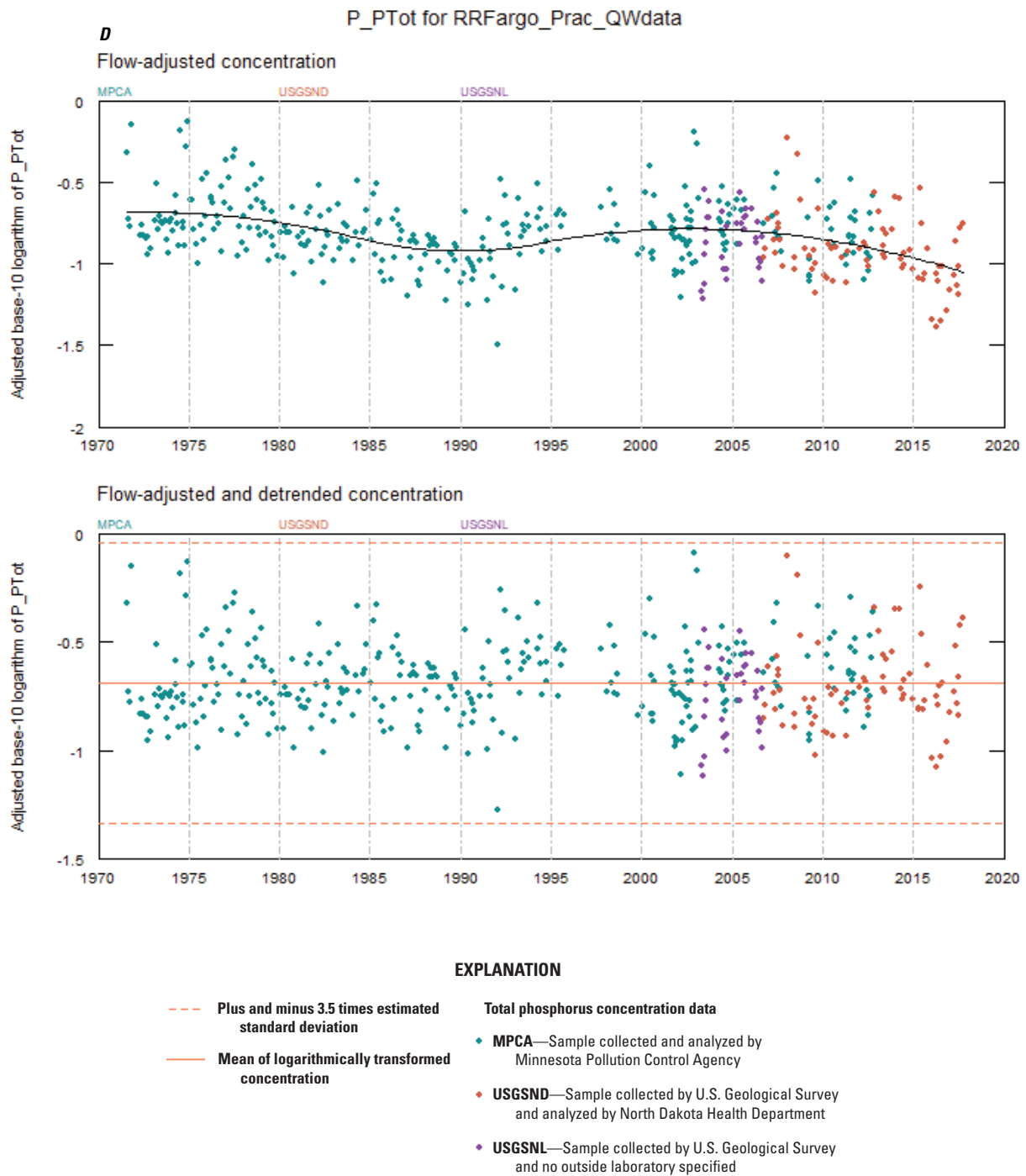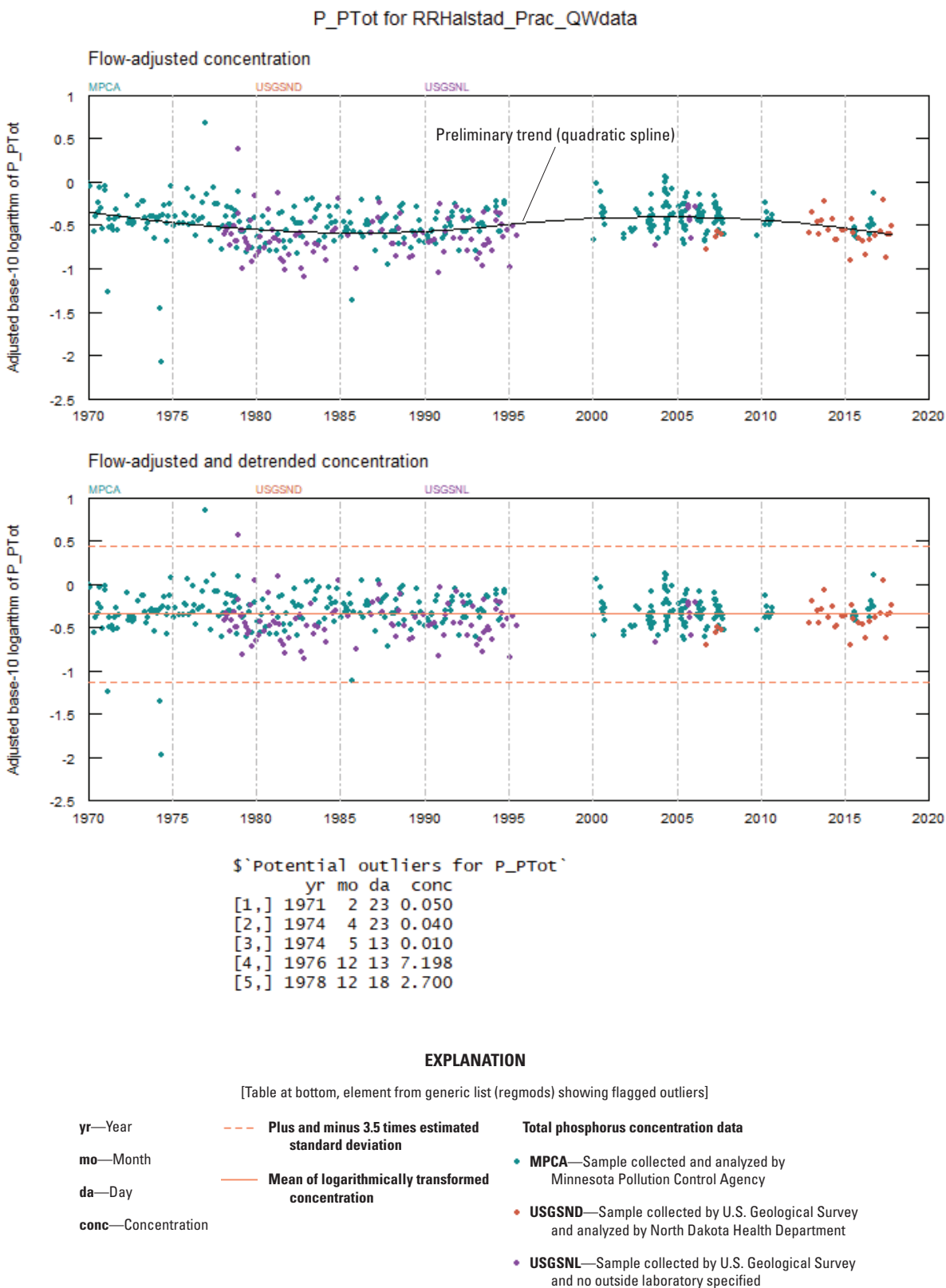
Figure 7.   Page 5 of plots produced by prepQWdata for practice dataset, RRHalstadQWP, showing total phosphorus (PTot) flow-adjusted and flow-adjusted and detrended concentration data versus decimal year.
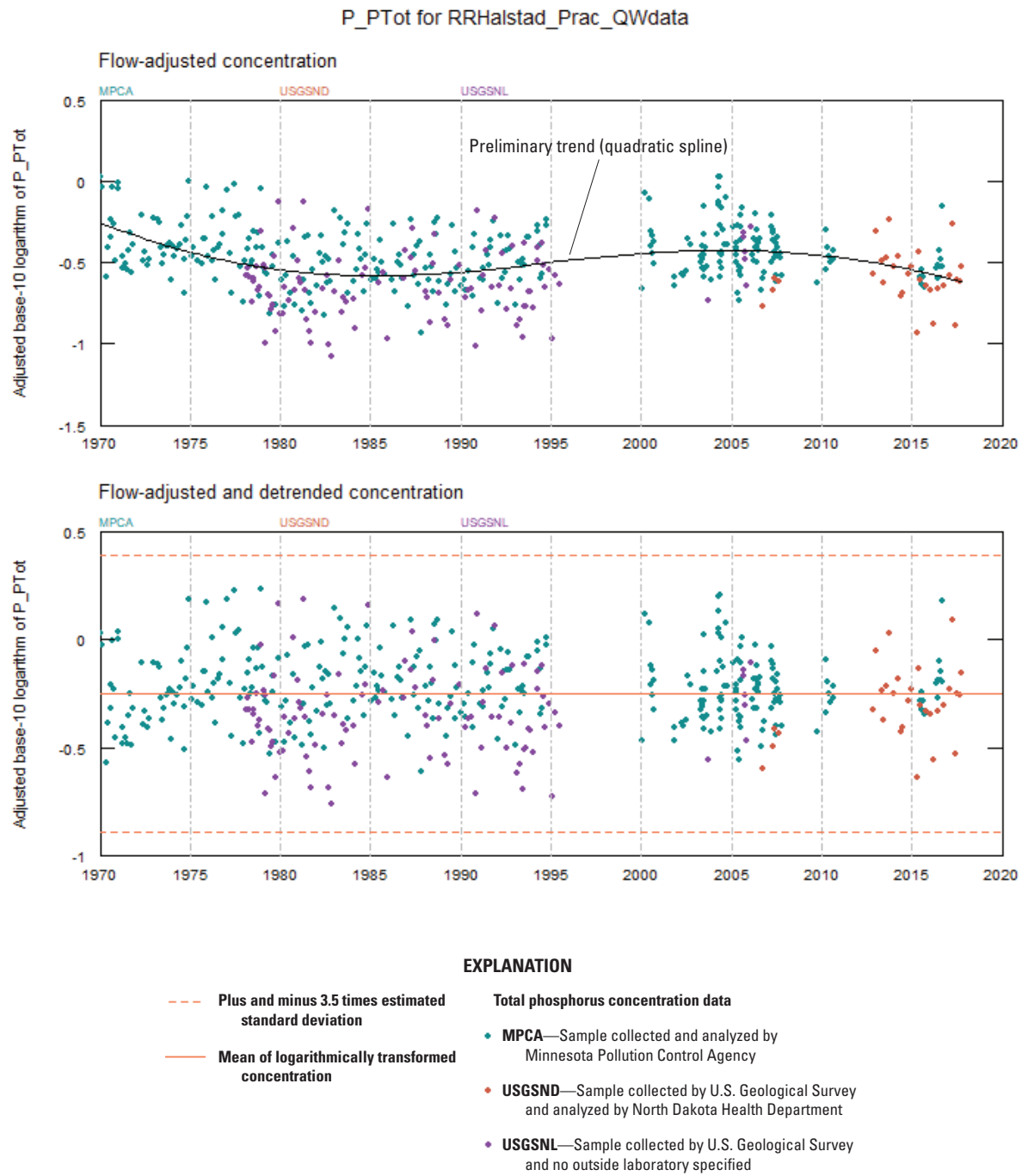
Figure 8. Page 5 of plots produced by prepQWdata for practice dataset, RRHalstadQWP, with outliers removed, showing total phosphorus (PTot) flow-adjusted and flow-adjusted and detrended concentration data versus decimal year.

The function runQWmodel is used to compute maximum likelihood parameter estimates and specify alternative trend models. The function is executed using the following command:

    runQWmodel(XXXQWP, pname, monxx=NULL,
    stepxx=NULL, remxx=NULL, userxx=NULL, fullout=T,
    units=1, runname="", modnum=3, exlev=NULL)

where
XXXQWP	is a time-series dataframe (dataset) produced by prepQWdata;
pname	is a character variable specifying the parameter name for analysis;
monxx	specifies piecewise monotonic trends (eq. 9);
stepxx	specifies step trends based on time interval (eq. 10);
remxx	specifies step trends based on sample attributes, or remark codes (eq. 11);
userxx	specifies trends based on ancillary, or user-specified, variables (eq. 12);
fullout	specifies whether to provide full output (fullout=T) or reduced output (fullout=F);
units	specifies whether concentration is in milligrams per liter (units=1) or micrograms per liter (units=2);
runname	is an optional character string to append to output files;
modnum	specifies the PARMA model number (table 1); and
exlev	is a vector of exceedance levels for computing exceedance probabilities.

The first two arguments (XXXQWP and pname) are mandatory, and the remaining arguments are optional. Default values of NULL for the specified trend components indicate there are no trend variables of that type in the model. The methods for specifying trends is described later in this section.

Example 1—Parameter NpnN for RRFargoQWP

The command on line 35 of the script file (fig. 2) fits the NULL (no trend) model for parameter npnN for the RRFargoQWP dataset created previously, with reduced output (fullout=F). When running this command, a message ("fitting the trend model …") should appear in the commands window indicating that the maximum likelihood estimates are being computed. Depending on the speed of your computer, it typically takes a few seconds to no more than 1 minute to complete the estimation; at which time the message "Program terminated normally" will appear in the command window. For reduced output, a single page of plots will be produced in the RStudio Plots window (fig. 9). The plots are similar to the output from prepQWdata (see fig. 6C). The top plot in figure 9 shows the flow-adjusted data along with the fitted trend (in this case, a flat line for the null model). The second plot shows the flow-adjusted, detrended, and PARMA filtered data (after removing serial correlation) along with a quadratic spline

to help spot possible lack of fit of the specified trend model. The second plot in figure 9 is similar to figure 6C, except that the effects of serial correlation have been removed in figure 9, whereas for the preliminary model (fig. 6C), there was assumed to be no serial correlation.

In general, the no-trend (null) model usually is fitted first and the results are examined before specifying more complicated trend models. The quadratic spline fitted to the flow-adjusted, detrended, and PARMA filtered observations for the null model (fig. 9) can be used to help postulate potential piecewise monotonic trends for the data. For this example, two trend models were postulated, one with a single monotonic trend from 1975 to 2015 (see line 36 of the script, fig. 2) and the other with two piecewise monotonic trends from 1985 to 2000 and from 2000 to 2015 (see line 37 of the script, fig. 2). Monotonic trends are specified using a character vector for the monxx argument. For example, the model with two piecewise monotonic trends (line 37) uses monxx=c("1985x2000", "2000x2015") to specify the two trends. There can be no spaces in the character names of the trends and the single character "x" is used to separate the beginning and ending times of each trend. Note that the beginning and ending times are in decimal years; for example, the first trend starts at the beginning of 1985 (decimal year 1985.0) and ends at the beginning of 2000 (decimal year 2000.0). Character name "1985.5x2001", for example, would specify a trend starting in the middle of 1985 (July 1, 1985) and ending at the beginning of 2001 (January 1, 2001). Run lines 36 and 37 of the script and examine the graphical output results (figs. 10 and 11A).

Next, examine the maximum likelihood estimation results for the three fitted trend models. These results are in the text file called RRFargoQWPnpnN.txt which is in the current working directory. The elements of each table are defined as follows (see text file RRFargoQWPnpnN.txt, fig. 11B):

- −2ln*Lik* is minus two times the natural logarithm of the maximized likelihood function (the minimum value of eq. 15).

- AIC is a penalized likelihood value (not used for this report).

- ecode is the error code from the nonlinear optimization program.

  - Values ecode=1 (absolute convergence) or ecode=2 (iterates within tolerance) signify that the optimization was successful.

  - Values ecode=3 (function too nonlinear to obtain convergence) or ecode=4 (iteration limit exceeded) indicate that the results are not reliable. If ecode=3 or ecode=4, the PARMA model number (modnum) should be reduced from the default (modnum=3) to a lower value (modnum=2 or modnum=1) until convergence is achieved. If convergence can-
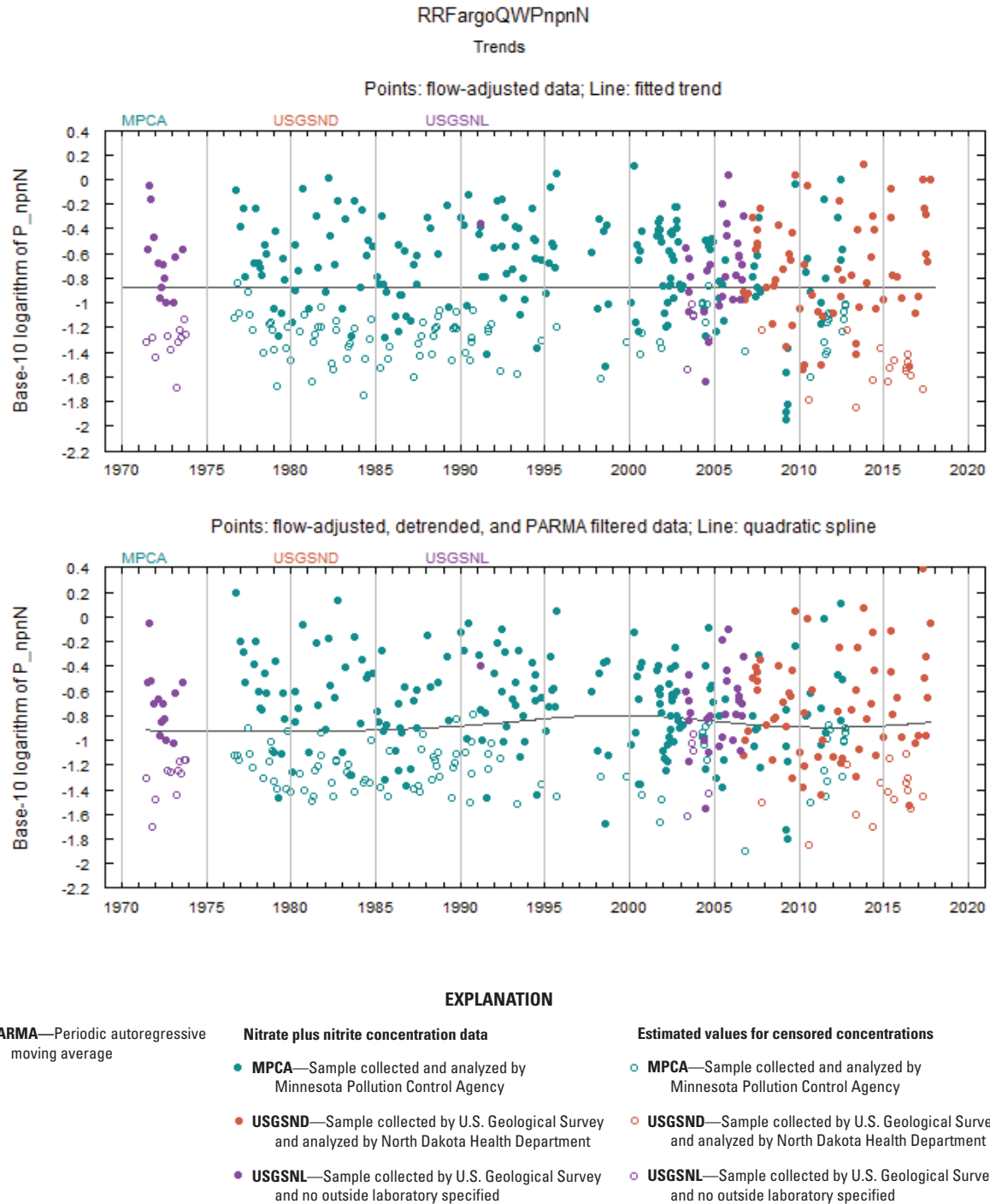
Figure 9. Reduced graphical output (fullout=F) from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the null (no trend) model.
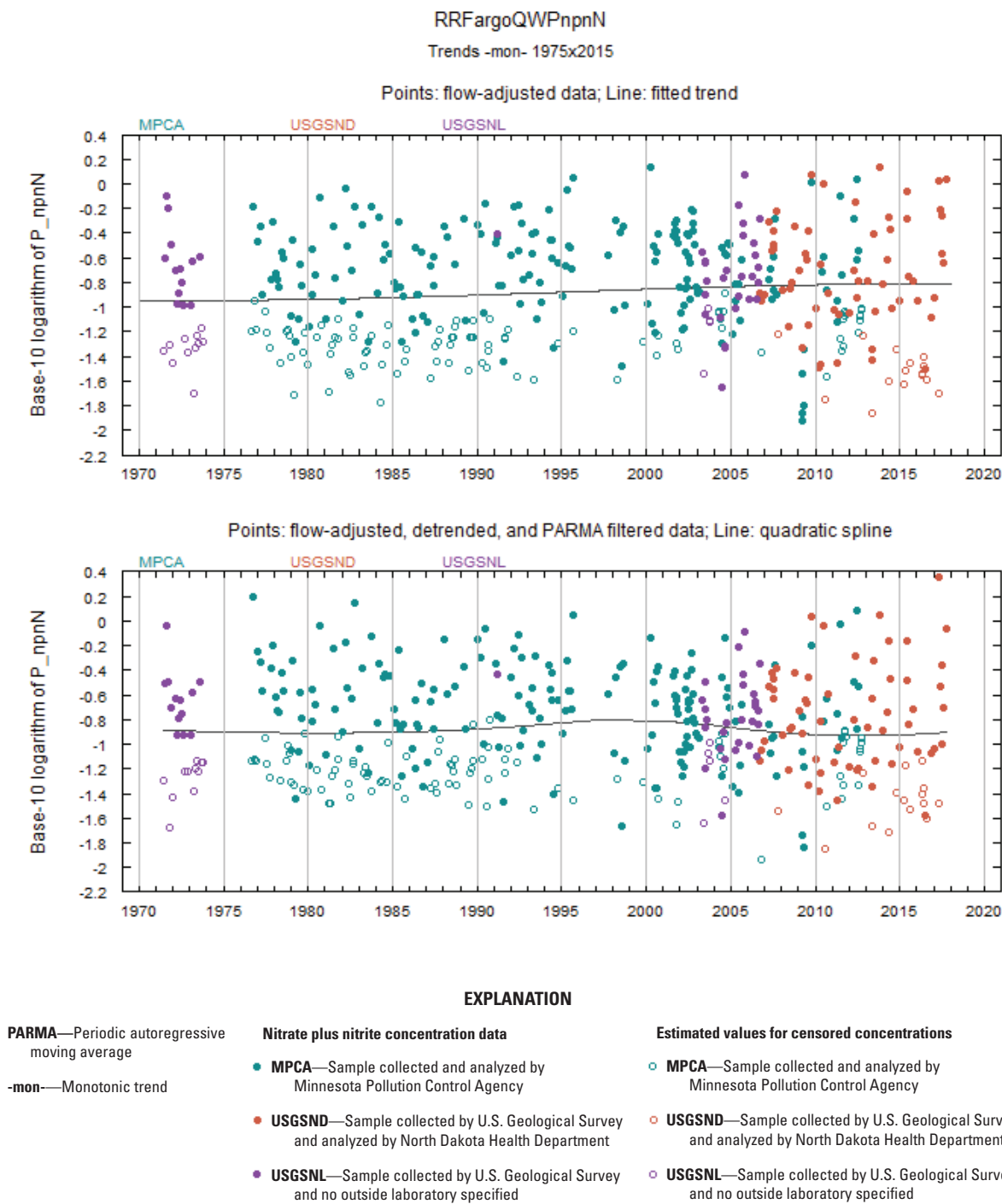
Figure 10.   Reduced graphical output (fullout=F) from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the model with a single monotonic trend.
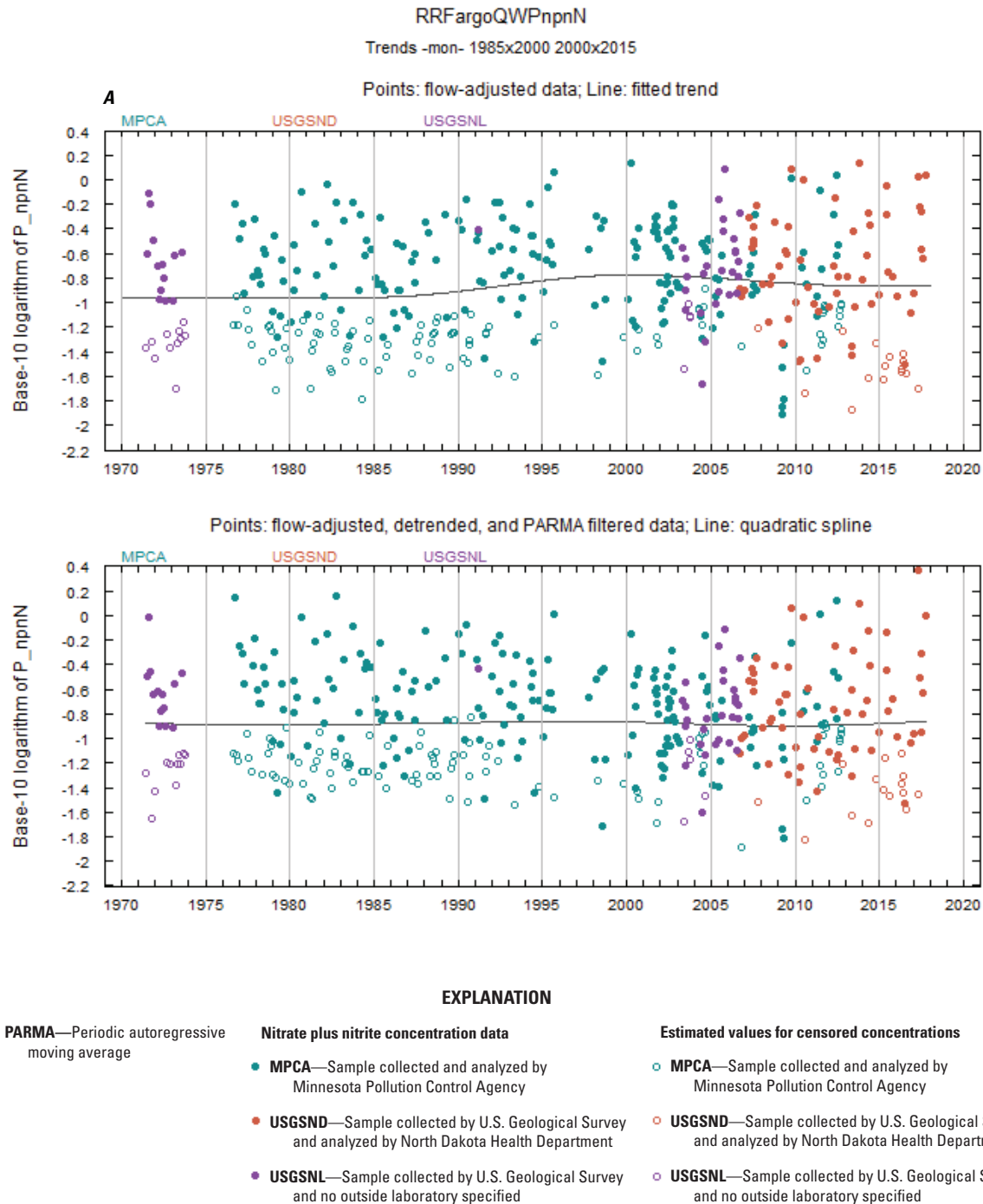
Figure 11. Reduced output (fullout=F) from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the model with two piecewise monotonic trends. *A*, graphical output; *B*, maximum likelihood estimation results from text file RRFargoQWPnpnN.txt.

B

```
                I
                ****************
                ** -2lnLik = -478.35
                **     AIC = -436.35
                **   ecode = 1
                **   modnum = 3

                **  Parameter estimates

                    Variable  CV(log)  SE(log)  Pvalue   CV(pct)
                         int  -0.8760  0.0307  0.00000     NA
                       xcos1   0.1515  0.0379  0.00254   41.74
                       xsin1   0.1789  0.0403  0.00126   50.97
                       xcos2   0.1212  0.0300  0.00236   32.19
                       xsin2   0.1588  0.0355  0.00120   44.15
                        falt   0.5903  0.0655  0.00000   64.55
                        famt   0.4853  0.0733  0.00006   57.32
                      famtsq   0.1604  0.1040  0.15406   29.80
                        fast   1.0302  0.0971  0.00000   76.35
                     fastxmt   0.5576  0.2167  0.02772   43.98
```

```
****************
** -2lnLik = -480.5
**     AIC = -436.5
**   ecode = 2
**   modnum = 3

**  Parameter estimates

    Variable  CV(log)  SE(log)  Pvalue   CV(pct)
         int  -0.8818  0.0303  0.00000     NA
       xcos1   0.1497  0.0377  0.00221   41.16
       xsin1   0.1839  0.0400  0.00076   52.72
       xcos2   0.1158  0.0300  0.00268   30.56
       xsin2   0.1600  0.0352  0.00084   44.54
        falt   0.5077  0.0852  0.00010   53.47
        famt   0.4701  0.0730  0.00005   55.10
      famtsq   0.1767  0.1021  0.11147   33.29
        fast   1.0203  0.0978  0.00000   75.39
     fastxmt   0.5670  0.2165  0.02389   44.86
   m1975x2015  0.1300  0.0845  0.15228   34.90
```

```
****************
** -2lnLik = -483.33
**     AIC = -437.33
**   ecode = 1
**   modnum = 3

**  Parameter estimates

    Variable  CV(log)  SE(log)  Pvalue   CV(pct)
         int  -0.8800  0.0299  0.00000     NA
       xcos1   0.1498  0.0376  0.00181   41.19
       xsin1   0.1875  0.0394  0.00047   53.99
       xcos2   0.1162  0.0297  0.00208   30.68
       xsin2   0.1550  0.0355  0.00091   42.89
        falt   0.4943  0.0833  0.00007   51.74
        famt   0.4719  0.0722  0.00003   55.36
      famtsq   0.1756  0.1021  0.11105   33.05
        fast   1.0075  0.0972  0.00000   74.16
     fastxmt   0.5684  0.2156  0.02171   45.00
   m1985x2000  0.1814  0.0782  0.03867   51.84
   m2000x2015 -0.0938  0.0855  0.29454  -19.43
```

EXPLANATION

[Top, null model; bottom left, model with a single monotonic trend; bottom right, model with two piecewise monotonic trends]

–2lnLik—Minus two times the natural logarithm of the maximized likelihood function

AIC—Penalized likelihood value (not used)

ecode—Error code from nonlinear optimization program

modnum—Model number (table1)

CV(log)—Estimated coefficient value

SE(log)—Approximate standard error of estimated coefficient

Pvalue—Approximate probability value of coefficient

CV(pct)—Estimated coefficient value expressed as a percentage

int—Intercept

xcos1—Cosine with period 1 year

xsin1—Sine with period 1 year

xcos2—Cosine with period 6 months

xsin2—Sine with period 6 months

falt—Long-term flow anomaly

famt—Midterm flow anomaly

famtsq—Square of midterm flow anomaly

fast—Short-term flow anomaly

fastxmt—Product of short-term and midterm flow anomalies

m1975x2015—Monotonic trend from 1975 to 2015

m1985x2000—Monotonic trend from 1985 to 2000

m2000x2015—Monotonic trend from 2000 to 2015

Figure 11.   Reduced output (fullout=F) from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the model with two piecewise monotonic trends. *A*, graphical output; *B*, maximum likelihood estimation results from text file RRFargoQWPnpnN.txt.—Continued

not be achieved with a lower model number, the data are not sufficient for trend analysis using R–QWTREND.

- Column labeled "variable" specifies the variable corresponding to the intercept (int or *MLC*, eq. 1), the coefficients of the model for *FRVAR* (eq. 6, xcos1 or $cos(2\pi t)$, xsin1 or $sin(2\pi t)$, xcos2 or $cos(4\pi t)$, xsin2 or $sin(4\pi t)$, falt or *LTFA*, famt or *MTFA*, famtsq or *MTFA²*, fast or *STFA*, and fastxmt or *STFA* times *MTFA*), and coefficients for the specified trend model (for the NULL model, there are no trend coefficients).

- Column labeled "CV(log)" is the estimated coefficient value from the maximum likelihood estimation in base-10 logarithmic (log) units.

- SE(log) is the approximate standard error of the estimated coefficient (eq. 19).

- Pvalue is the approximate individual *p*-value for the coefficient (eq. 20).

- CV(pct) is the estimated coefficient value, expressed as a percentage. For the flow anomaly terms (X= falt, famt, famtsq, fast, or fastxmt), CV(pct)= $100(10^{SD(X)CV(\log)}-1)$, where SD(X) is the standard deviation of X. For the sine and cosine variables and all trend variables, CV(pct)=$100(10^{CV(\log)}-1)$. In this case, all of the variables except famtsq were significant (Pvalue<0.05) and all of the coefficient values for the flow anomalies were positive, indicating that high flow anomalies tend to produce high npnN concentrations for all time scales (annual, seasonal, and daily). For example, for fast and the null model, CV(pct)=76.35, which means that concentrations tend to be about 76 percent higher when fast is high (equal to plus one SD[fast]) compared to concentrations when fast is low (equal to negative one SD[fast]).

For the model with a single monotonic trend (figs. 10, 11*B*, bottom left table), there was an estimated uptrend of about 35 percent from 1975 to 2015 (m1975x2015 in variable column), with an approximate *p*-value using the normal approximation (eq. 19) of about 0.152. The value of −2ln*Lik* for this model was −480.5 with ecode=2 compared to −478.35 with ecode=1 for the null model. Note that an ecode of 3 or 4 would indicate that the model is probably too complex to be determined from the observed data and should not be considered as a good alternative. In this case, both models are acceptable alternatives. The GLR test statistic for comparing the null and single trend models (eq. 16) was (−478.35)−(−480.5)=2.15, and the overall *p*-value using the GLR test statistic (eq. 17) was calculated to be

$$P=1–pchisq(2.15, df=1)=0.142$$

where

$pchisq()$ is the R function for computing probabilities for the chi-square distribution.

Note that this *p*-value is similar to the approximate *p*-value for the trend coefficient (0.152, fig. 11*B*) because there is a single trend coefficient.

For the second trend model (fig. 11*A–B*), there was an estimated uptrend of about 52 percent from 1985 to 2000 (approximate *p*-value 0.039) followed by an estimated downtrend of about 19 percent from 2000 to 2015 (approximate *p*-value 0.295). The value of −2ln*Lik* was −483.33 compared to –478.35 for the null model, for a difference of −478.35–(–483.34)=4.98 and the overall *p*-value of the two-trend model using the GLR test statistic was calculated to be P=1−pchisq(4.98, df=2)=0.083. Because the overall *p*-value of the two-trend model (0.083) was smaller compared to the one-trend model (0.142), the two-trend model appears to be a viable alternative to the one-trend model. To determine if the two-trend model is preferred, the GLR test for comparing the two models (eq. 22) can be used. The GLR test statistic is $G_{1\sim2}$=(–480.5)–(–483.33)=2.83 and the *p*-value is P=1−pchisq(2.83, df=1)=0.092.

Based on the GLR statistics, the two-trend model is better (*p*<0.1) than the one-trend model. Furthermore, the flow-adjusted, detrended, and PARMA filtered observations for the two-trend model (fig. 11*A*, bottom plot) seem not to have any obvious trends remaining. Therefore, in this example the two-trend model seems to be a better alternative than the one-trend model.

After selecting the "best," or at least a "good," trend model, the full diagnostic output should be examined to verify the model assumptions before using the "value-added" output from the model (such as estimated annual flow-weighted average concentration or annual flux). The full diagnostic output can be obtained by using the default (fullout=T) option (see line 40 of the script). After running line 40, 10 pages of plots are produced in a graphical output file (RRFargoQWPnpnN.pdf) in the current working directory. The first six plots show detailed model diagnostic information, and the remaining plots show the "value-added" model output (see fig. 12*A–I*):

- Page 1: Observed data along with flow-related variability (eq. 6).

- Page 2: Flow-adjusted data (eq. 7) along with the fitted trend (*MLC+TREND*).

- Page 3: Flow-adjusted and PARMA filtered data (*MLC+TREND+R*, where *R* is defined in eq. 13) along with the fitted trend.

- Page 4: Flow-adjusted, detrended, and PARMA filtered data (*MLC+R*) along with a quadratic spline to indicate potential residual trends.
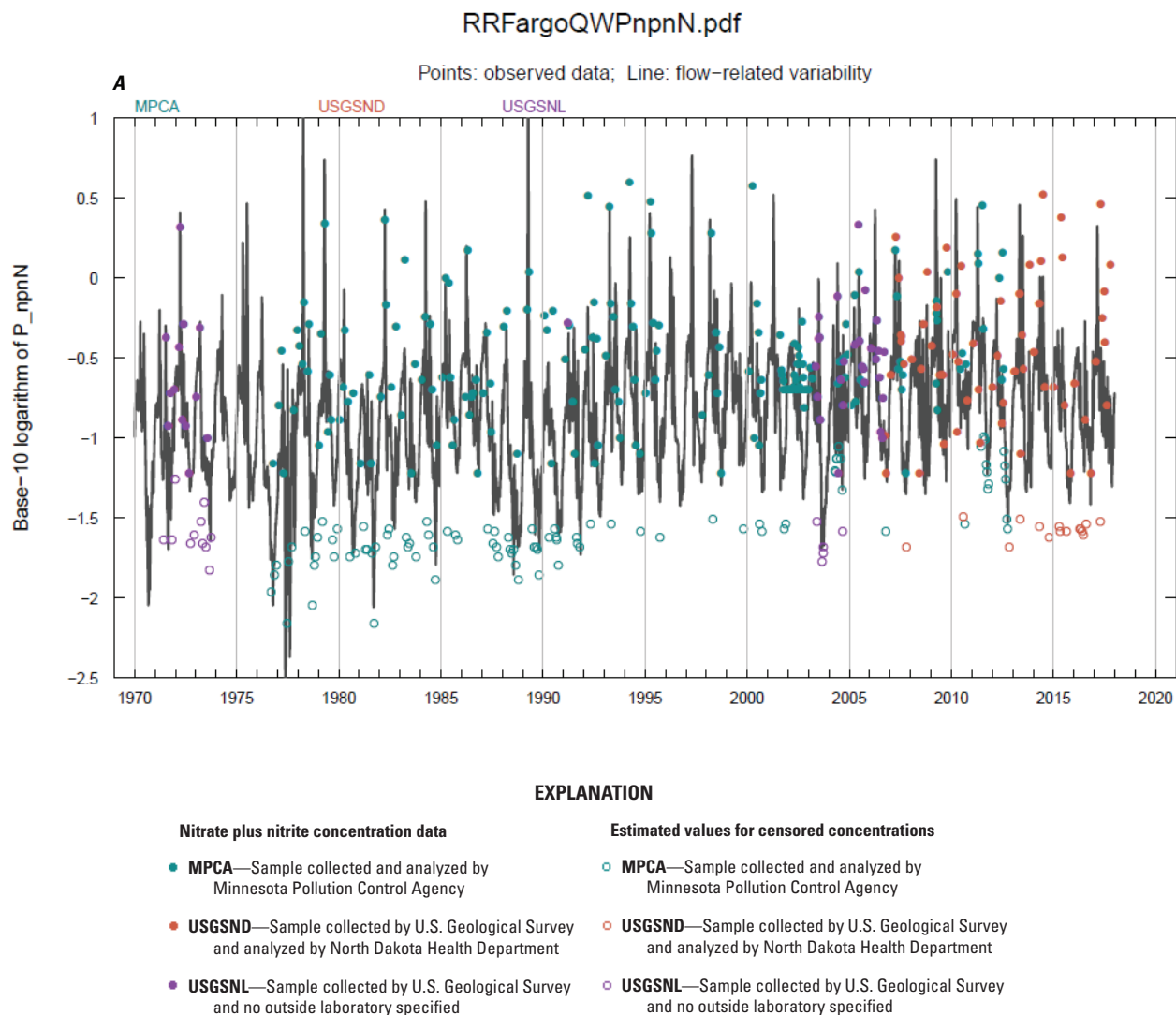
Figure 12.   Full graphical output from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the model with two piecewise monotonic trends. *A*, page 1; *B*, page 2; *C*, page 3; *D*, page 4; *E*, page 5; *F*, page 6; *G*, page 7; *H*, page 8; *I*, page 9; *J*, page 10.
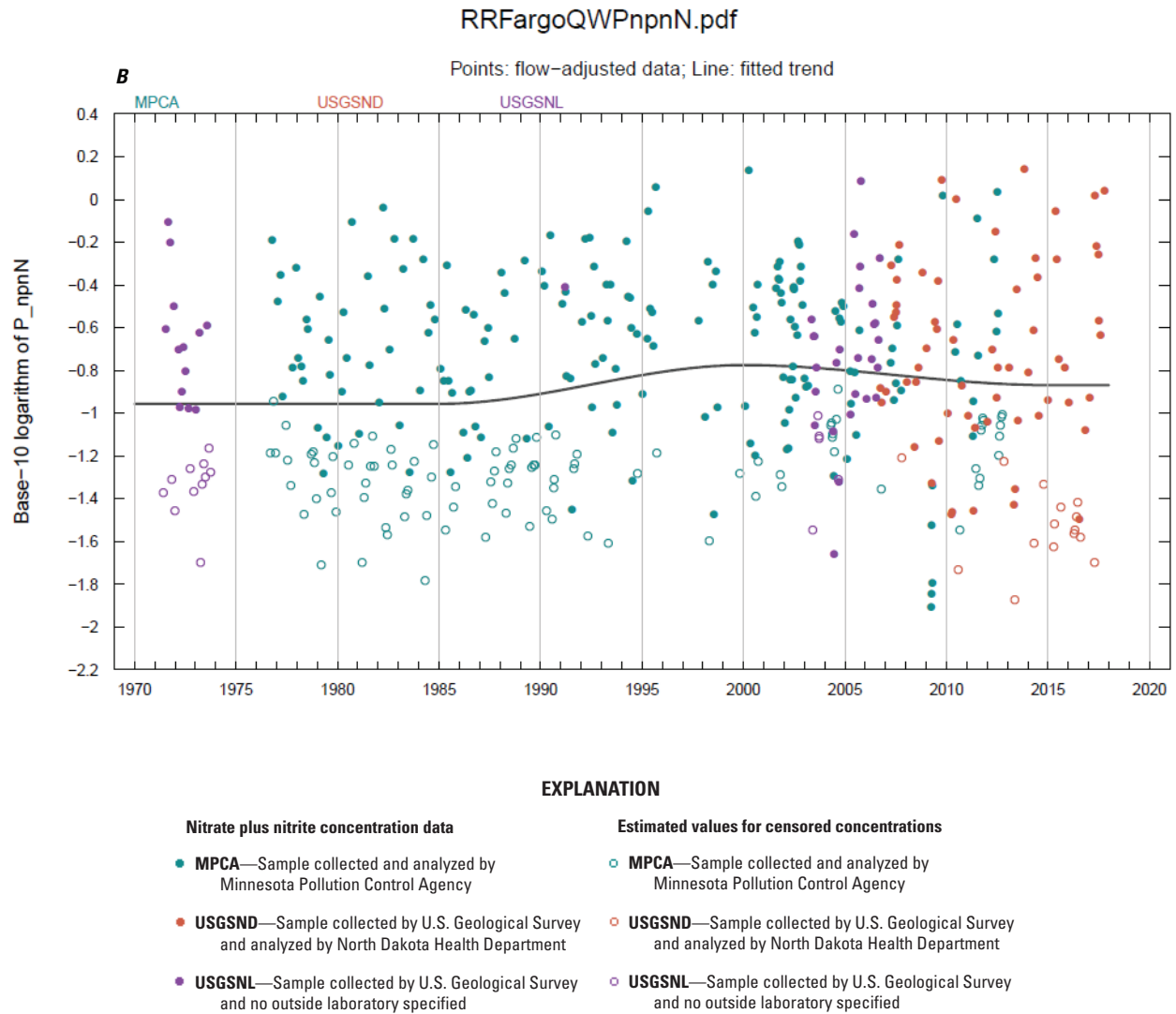
Figure 12. Full graphical output from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the model with two piecewise monotonic trends. *A*, page 1; *B*, page 2; *C*, page 3; *D*, page 4; *E*, page 5; *F*, page 6; *G*, page 7; *H*, page 8; *I*, page 9; *J*, page 10.—Continued
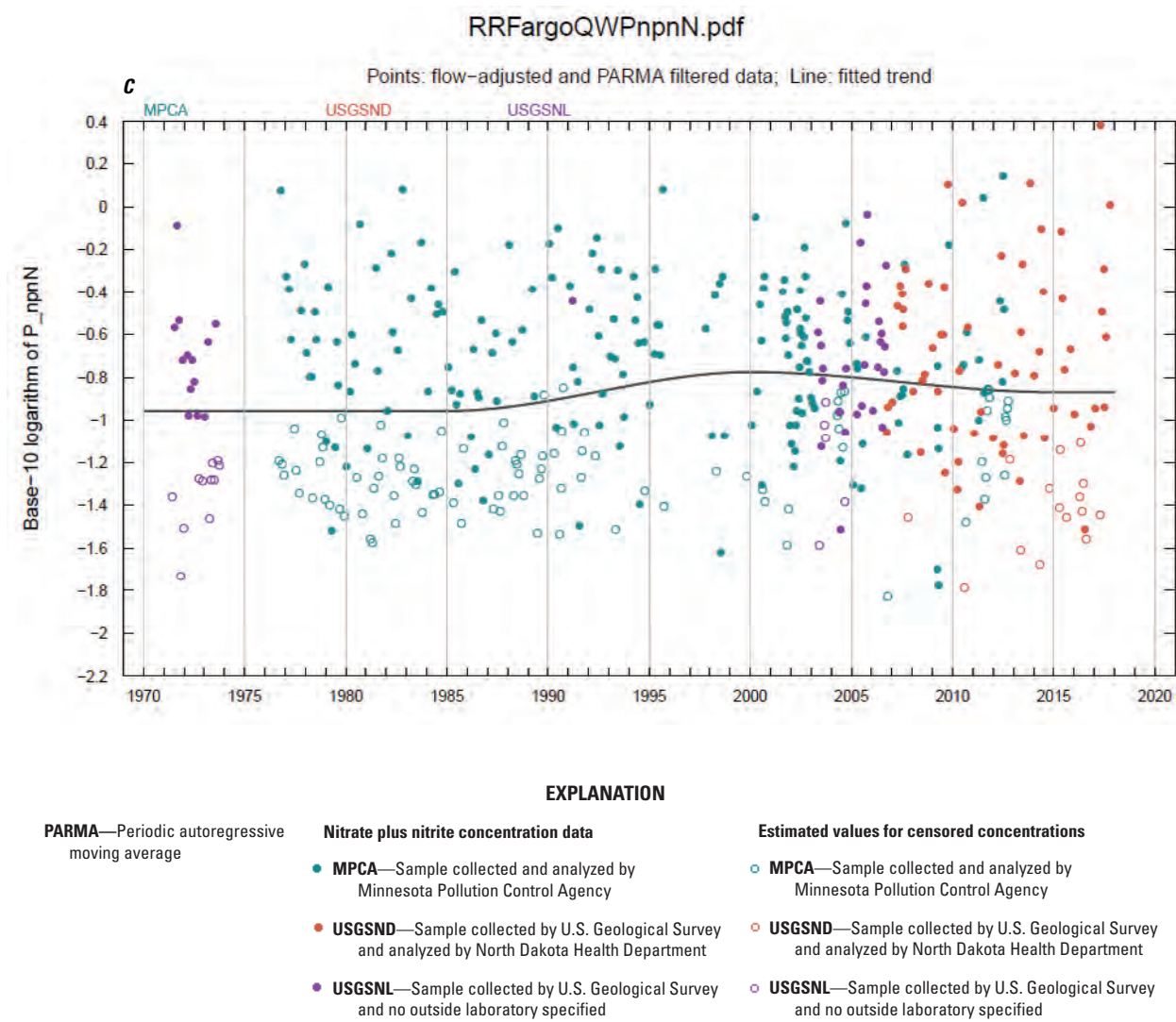
Figure 12. Full graphical output from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the model with two piecewise monotonic trends. *A*, page 1; *B*, page 2; *C*, page 3; *D*, page 4; *E*, page 5; *F*, page 6; *G*, page 7; *H*, page 8; *I*, page 9; *J*, page 10.—Continued
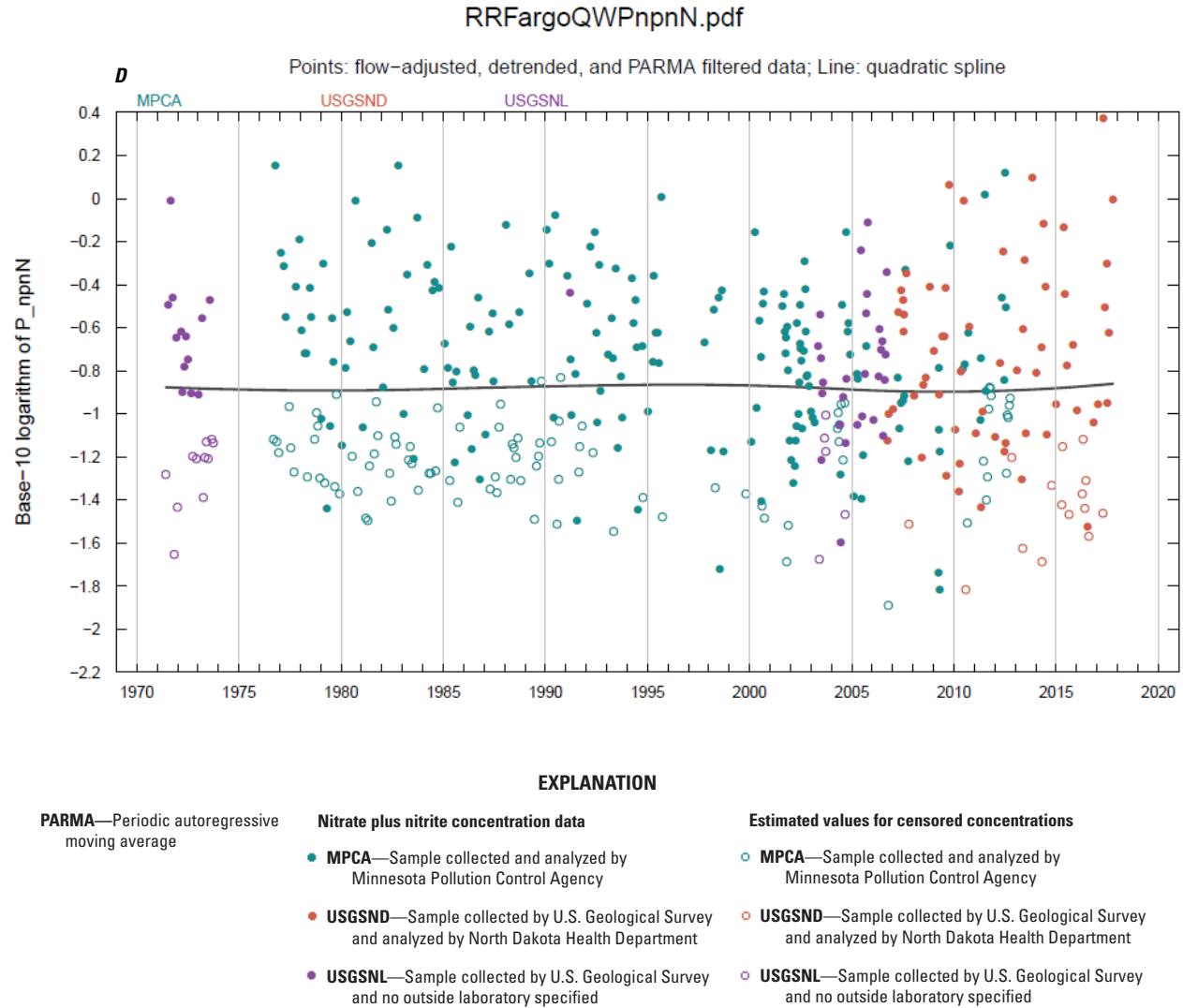
Figure 12. Full graphical output from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the model with two piecewise monotonic trends. *A*, page 1; *B*, page 2; *C*, page 3; *D*, page 4; *E*, page 5; *F*, page 6; *G*, page 7; *H*, page 8; *I*, page 9; *J*, page 10.—Continued

Figure 12. Full graphical output from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the model with two piecewise monotonic trends. *A*, page 1; *B*, page 2; *C*, page 3; *D*, page 4; *E*, page 5; *F*, page 6; *G*, page 7; *H*, page 8; *I*, page 9; *J*, page 10.—Continued
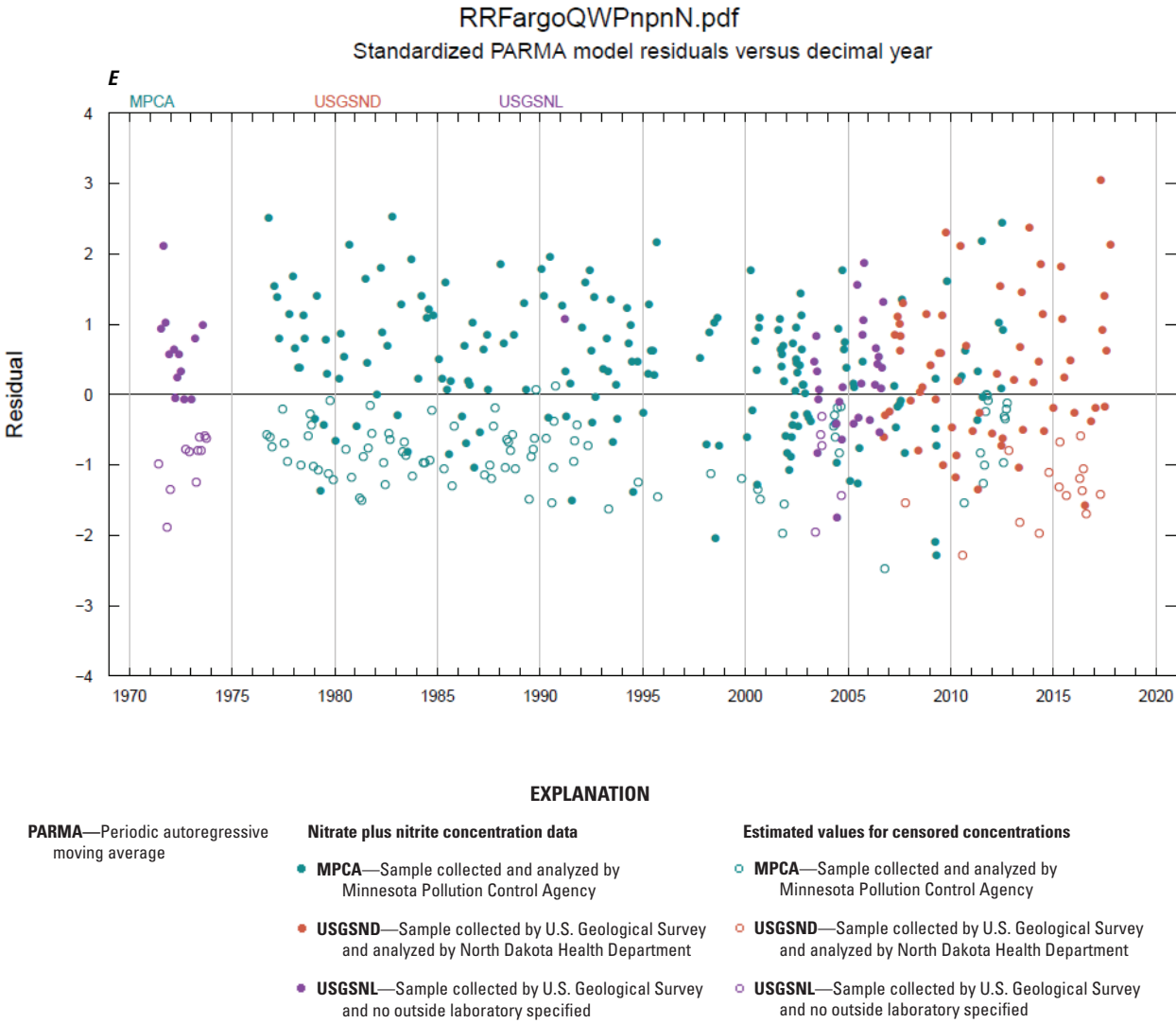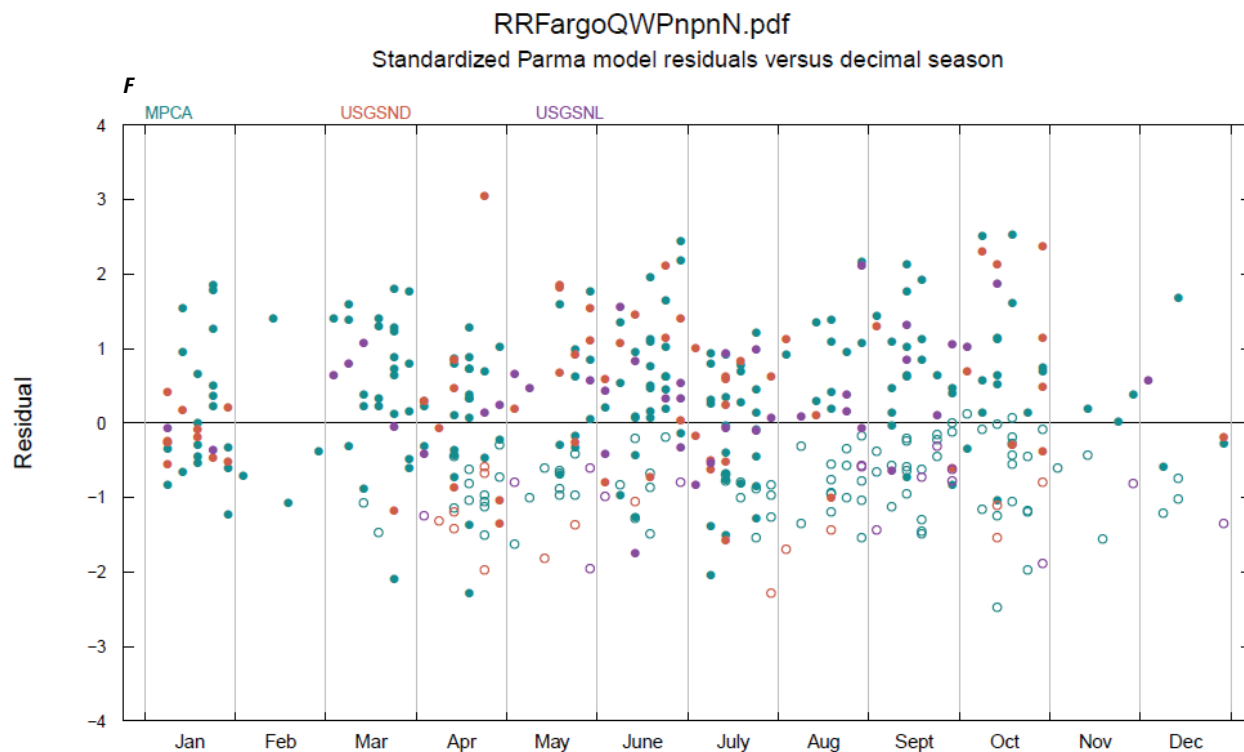
Figure 12. Full graphical output from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the model with two piecewise monotonic trends. *A*, page 1; *B*, page 2; *C*, page 3; *D*, page 4; *E*, page 5; *F*, page 6; *G*, page 7; *H*, page 8; *I*, page 9; *J*, page 10.—Continued

RRFargoQWPnpnN.pdf

Points: flow–adjusted and PARMA filtered data
Line: fitted trend

EXPLANATION

PARMA—Periodic autoregressive moving average

Figure 12. Full graphical output from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the model with two piecewise monotonic trends. *A*, page 1; *B*, page 2; *C*, page 3; *D*, page 4; *E*, page 5; *F*, page 6; *G*, page 7; *H*, page 8; *I*, page 9; *J*, page 10.—Continued

Figure 12. Full graphical output from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the model with two piecewise monotonic trends. *A,* page 1; *B,* page 2; *C,* page 3; *D,* page 4; *E,* page 5; *F,* page 6; *G,* page 7; *H,* page 8; *I,* page 9; *J,* page 10.—Continued

Figure 12. Full graphical output from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the model with two piecewise monotonic trends. *A,* page 1; *B,* page 2; *C,* page 3; *D,* page 4; *E,* page 5; *F,* page 6; *G,* page 7; *H,* page 8; *I,* page 9; *J,* page 10.—Continued
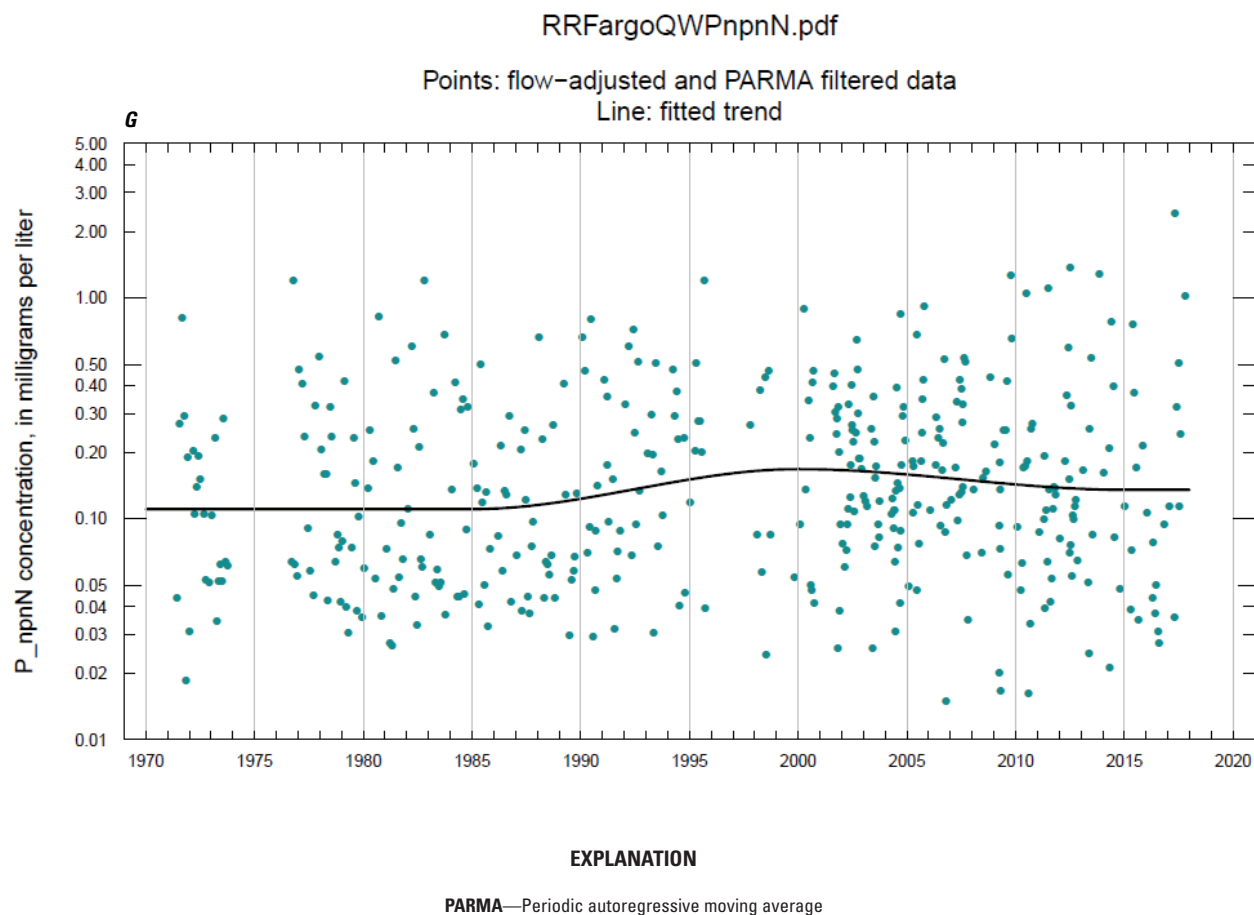
Figure 12.   Full graphical output from runQWmodel for nitrate plus nitrite (npnN) concentration for RRFargoQWP dataset for the model with two piecewise monotonic trends. *A*, page 1; *B*, page 2; *C*, page 3; *D*, page 4; *E*, page 5; *F*, page 6; *G*, page 7; *H*, page 8; *I*, page 9; *J*, page 10.—Continued
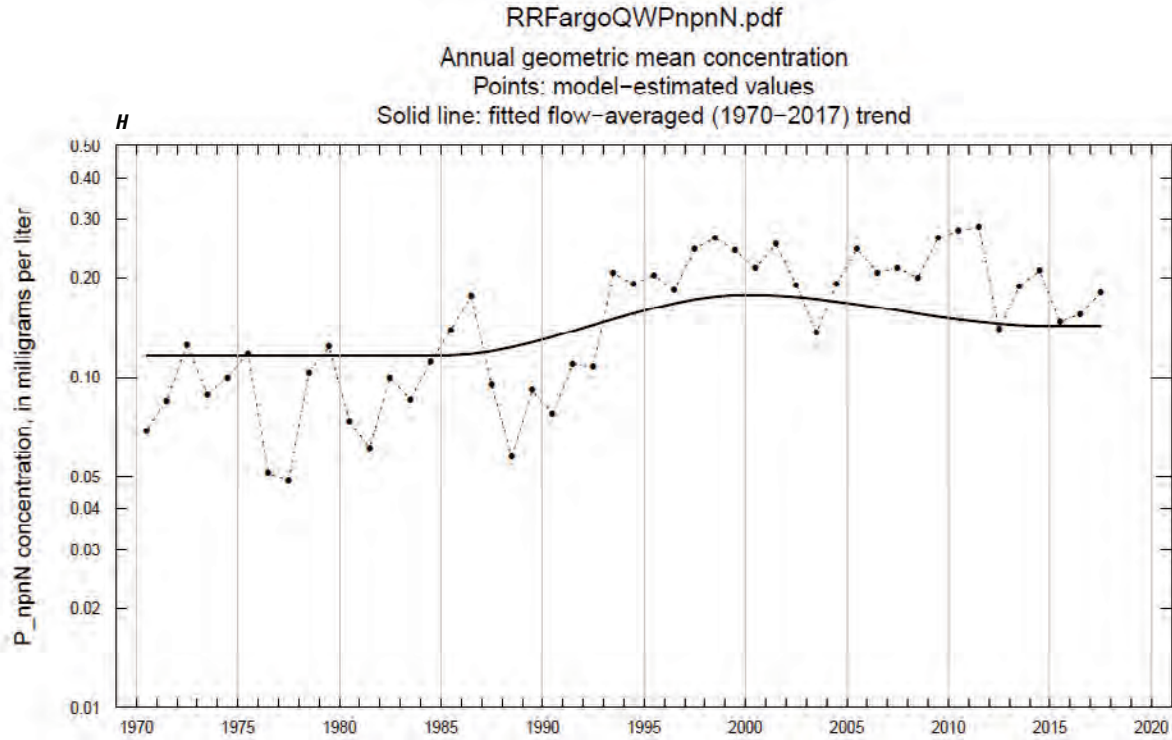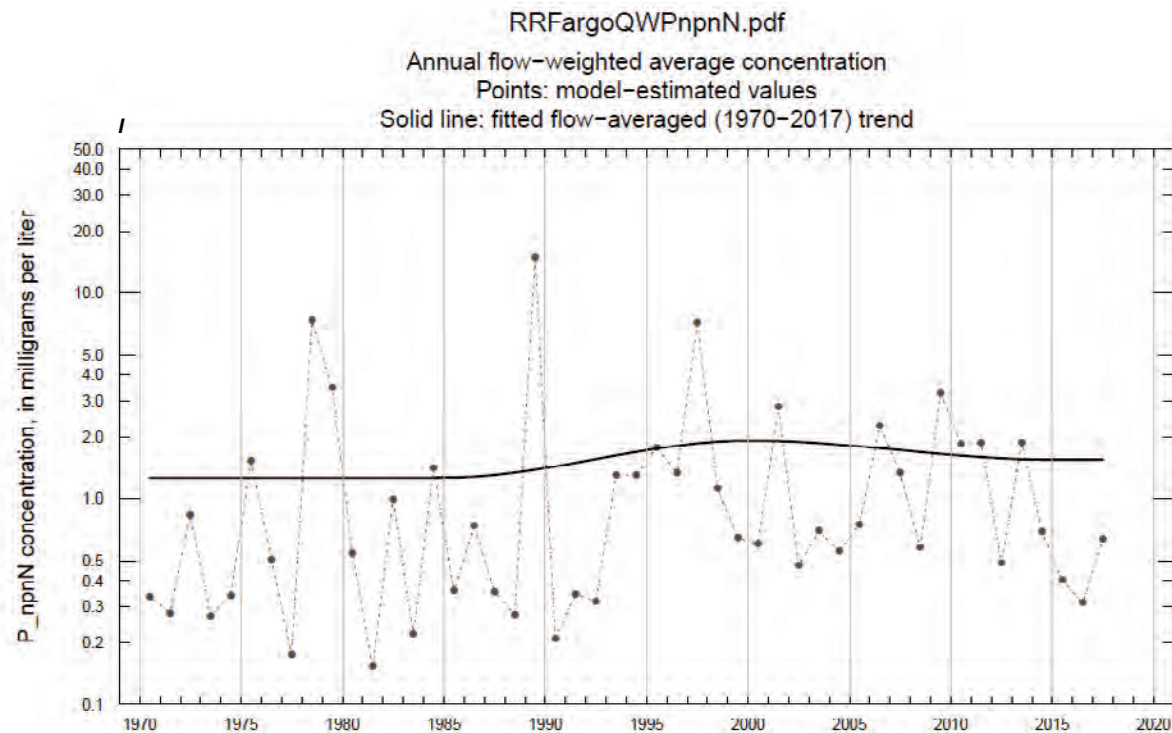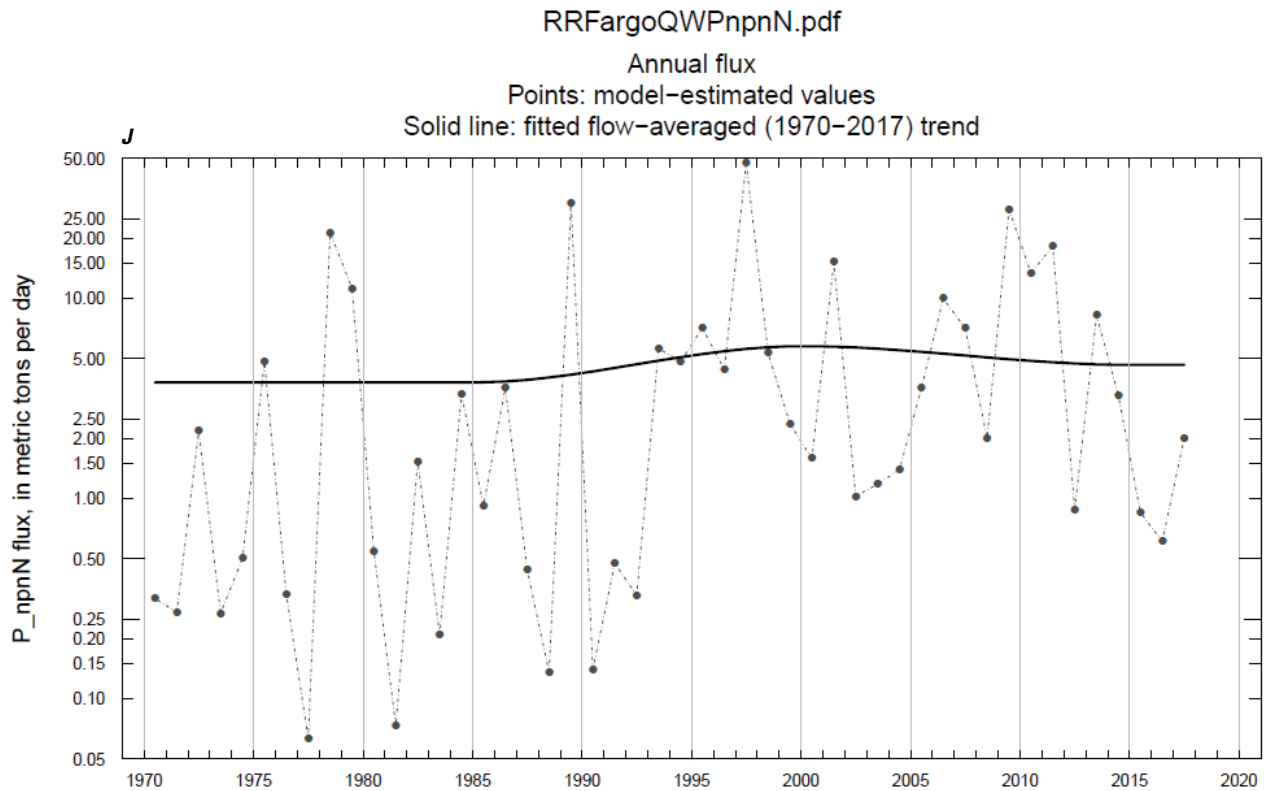
- Page 5: Standardized PARMA model residuals (eq. 14) versus decimal year. The residuals should be approximately independent, standard normal random variables. Obvious nonsymmetrical (skewed) residuals, outliers (values larger than about 3.5 in absolute value), or nonrandom structure (systematic changes in central tendency or variability, excessive clumping, etc.) may indicate that the model results are unreliable.

- Page 6: Standardized PARMA model residuals (eq. 14) versus decimal season. These should be examined in a similar manner to the previous plot (page 5) to ensure that there is no obvious seasonal structure remaining in the residuals.

- Page 7: Same as page 3, but without the error codes and censoring information and with improved y-axis tick marks and labels.

- Page 8: Estimated annual geometric mean concentration along with the fitted trend (eq. 24). The trend is referred to as a "flow-averaged" trend because it is an estimate of the geometric mean concentration under the hypothetical assumption that flow-related variability was the same year after year.

- Page 9: Estimated annual flow-weighted average concentration (eq. 25) along with the fitted flow-averaged trend (eq. 27). Note that, for this example, the annual flow-weighted average concentrations are much higher than the annual geometric mean concentrations (page 8) because npnN concentrations tend to be much higher during high-flow conditions compared to low-flow conditions.

- Page 10: Estimated annual flux (eq. 28) along with the fitted flow-averaged trend (eq. 29). Note the high degree of flow-related variability in the estimated annual flux as indicated by large deviations of the annual values from the fitted trend. Flux tended to be much lower compared to the flow-averaged trend during dry years such as 1977 and 1981 and much higher during wet years such as 1997 and 2009.

Example 2—Parameter PTot for RRFargoQWP

The second example shows how to combine monotonic trends with potential step trends related to differences in sample collection or laboratory analysis methods. Run lines 43–45 of the script file (fig. 2) to obtain the full output (default argument, fullout=T) for the null model and two potential trend models for PTot (total phosphorus) for the RRFargoQWP

```
****************
** -2lnLik = -953.96
**    AIC = -911.96
**   ecode = 1
**  modnum = 3

**  Parameter estimates

        Variable   CV(log)  SE(log)  Pvalue   CV(pct)
             int   -0.8208   0.0198   0.00000      NA
           xcos1   -0.1532   0.0248   0.00011   -29.73
           xsin1   -0.0270   0.0173   0.15064    -6.03
           xcos2   -0.0217   0.0140   0.15358    -4.87
           xsin2    0.0763   0.0156   0.00063    19.21
            falt    0.1314   0.0390   0.00716    11.72
            famt    0.2344   0.0373   0.00009    24.46
          famtsq    0.1492   0.0432   0.00620    27.46
            fast    0.4087   0.0397   0.00000    25.24
         fastxmt    0.2504   0.0882   0.01762    17.78
```

```
****************
** -2lnLik = -984.31
**    AIC = -936.31
**   ecode = 1
**  modnum = 3

**  Parameter estimates

       Variable   CV(log)  SE(log)  Pvalue   CV(pct)
            int   -0.8178   0.0158   0.00000      NA
          xcos1   -0.1589   0.0227   0.00001   -30.64
          xsin1   -0.0300   0.0162   0.08749    -6.67
          xcos2   -0.0193   0.0137   0.18318    -4.35
          xsin2    0.0740   0.0157   0.00040    18.58
           falt    0.2058   0.0391   0.00015    18.96
           famt    0.2433   0.0332   0.00001    25.50
         famtsq    0.1415   0.0411   0.00436    25.87
           fast    0.4293   0.0397   0.00000    26.67
        fastxmt    0.2224   0.0882   0.02549    15.65
       m1975x1985  -0.2023   0.0406   0.00025   -37.24
       m1985x2000   0.0996   0.0425   0.03566    25.78
       m2000x2015  -0.1592   0.0416   0.00210   -30.69
```

```
****************
** -2lnLik = -995.67
**    AIC = -945.67
**   ecode = 1
**  modnum = 3

**  Parameter estimates

       Variable   CV(log)  SE(log)  Pvalue   CV(pct)
            int   -0.8754   0.0234   0.00000      NA
          xcos1   -0.1557   0.0231   0.00001   -30.13
          xsin1   -0.0301   0.0164   0.08867    -6.70
          xcos2   -0.0149   0.0139   0.30025    -3.37
          xsin2    0.0713   0.0156   0.00044    17.84
           falt    0.1967   0.0384   0.00016    18.05
           famt    0.2444   0.0332   0.00000    25.63
         famtsq    0.1403   0.0414   0.00440    25.63
           fast    0.4330   0.0394   0.00000    26.93
        fastxmt    0.2253   0.0871   0.02156    15.87
           rMPCA   0.0892   0.0258   0.00380    22.80
       m1975x1985  -0.2042   0.0400   0.00016   -37.51
       m1985x2000   0.1209   0.0422   0.01251    32.10
       m2000x2015  -0.1034   0.0442   0.03460   -21.19
```

EXPLANATION

[Top, null model; bottom left, model M1; bottom right, model M2;
MPCA, sample collected and analyzed by Minnesota Pollution Control Agency]

–2lnLik—Minus two times the natural logarithm of the maximized likelihood function

AIC—Penalized likelihood value (not used)

ecode—Error code from nonlinear optimization program

modnum—Model number (table1)

CV(log)—Estimated coefficient value

SE(log)—Approximate standard error of estimated coefficient

Pvalue—Approximate probability value of coefficient

CV(pct)—Estimated coefficient value expressed as a percentage

int—Intercept

xcos1—Cosine with period 1 year

xsin1—Sine with period 1 year

xcos2—Cosine with period 6 months

xsin2—Sine with period 6 months

falt—Long-term flow anomaly

famt—Midterm flow anomaly

famtsq—Square of midterm flow anomaly

fast—Short-term flow anomaly

fastxmt—Product of short-term and midterm flow anomalies

rMPCA—Step trend for remark code MPCA

m1975x1985—Monotonic trend from 1975 to 1985

m1985x2000—Monotonic trend from 1985 to 2000

m2000x2015—Monotonic trend from 2000 to 2015

Figure 13.   Parameter estimation results from runQWmodel for total phosphorus concentration for RRFargoQWP dataset for the null model and two alternative trend models.
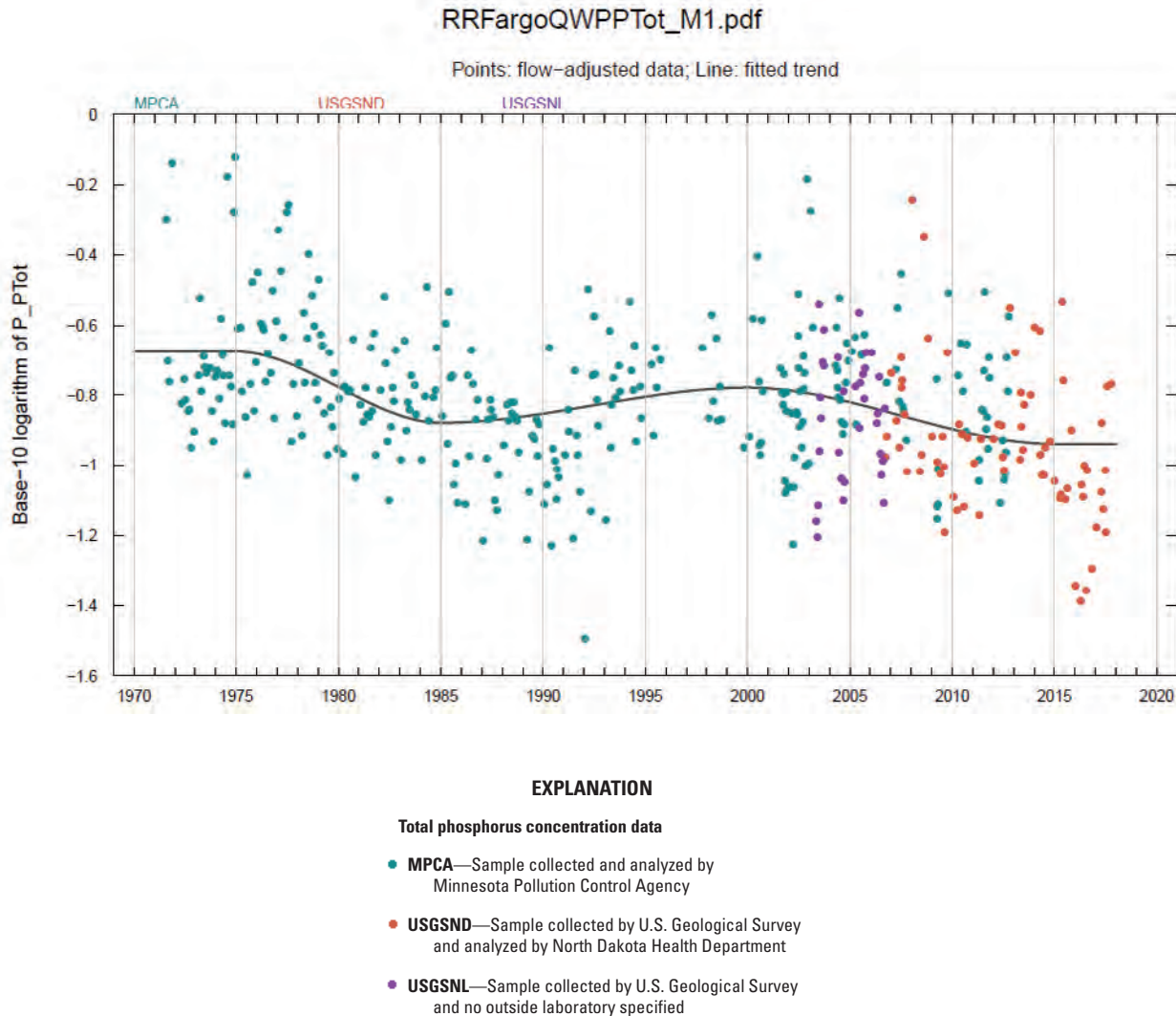
Figure 14.   Flow-adjusted total phosphorus (PTot) concentration for RRFargoQWP dataset for model with piecewise monotonic trend (model M1).

dataset, and examine the parameter estimation results from the files RRFargoQWPPTot_null.txt, RRFargoQWPPTot_M1.txt, and RRFargoQWPPTot_M2.txt (see fig. 13). The optional run-name argument is used in these commands to name the output files. If this command was not used, the graphical output file from each successive model run would have the same name (RRFargoQWPPTot.pdf) and would be overwritten each time the new output is produced. Model M2 ($-2\ln Lik=-995.67$), which includes the piecewise monotonic trends and a step trend for remark code "MPCA," is a much better alternative than model M1 ($-2\ln Lik=-984.31$), which includes just the piecewise monotonic trends. The p-value for the GLR test for comparing model M2 versus M1 ($G=-984.31+995.67=11.36$, $P=1-pchisq(11.36, df=1)=0.00075$) is small ($<0.01$), indicat-ing that the more complex model (M2) is the better alterna-tive. The fitted trend for model M1 (fig. 14) consisted of a highly significant downtrend (-37 percent, p-value=0.00025) from 1975 to 1985, followed by a mildly significant uptrend

($+26$ percent, p-value=0.036) from 1985 to 2000 and another highly significant downtrend (-31 percent, p-value=0.002) from 2000 to 2015. For model M2 (fig. 15), there was a highly significant step trend for remark code MPCA ($+23$ percent, p-value=0.0038), indicating that concentrations with that remark code tend to be about 23 percent higher, on average, compared to the other two remark codes. Whenever one or more step trends are included in the model, there is an extra page of graphical output showing flow-adjusted concentrations with step trends included (top plot in fig. 15) and with step trends removed (bottom plot). Step trends based on remark codes are assumed to apply to the observations, not the true water-quality conditions, and thus need to be removed before analyzing true trends.

With multiple remark codes, there may be several alterna-tive representations for the step trends that cannot readily be distinguished from one another based on the observations alone and, thus, may require expert judgement or further
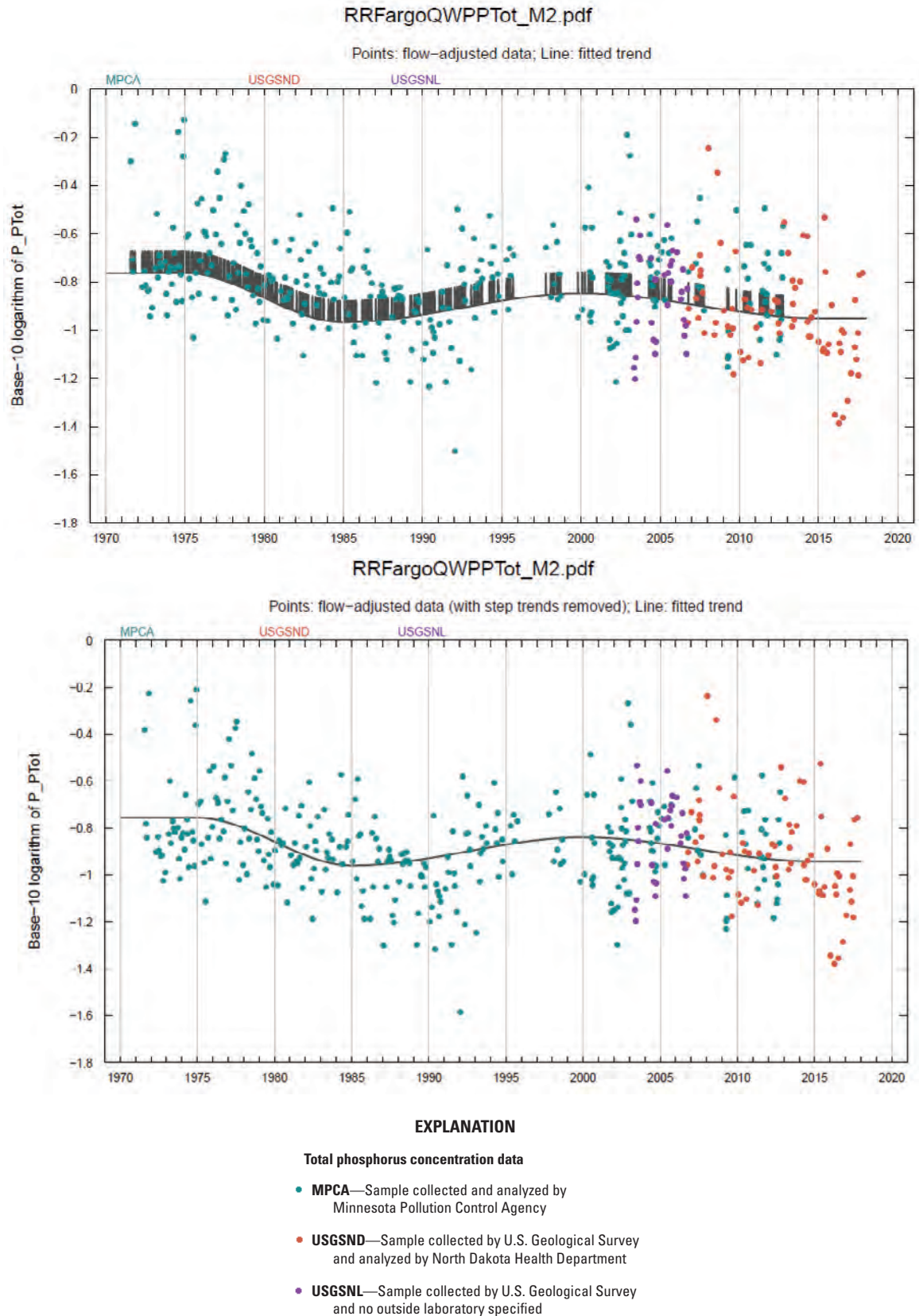
Figure 15. Flow-adjusted total phosphorus (PTot) concentration for RRFargoQWP dataset for model with piecewise monotonic trend and step trend for remark code MPCA (model M2).
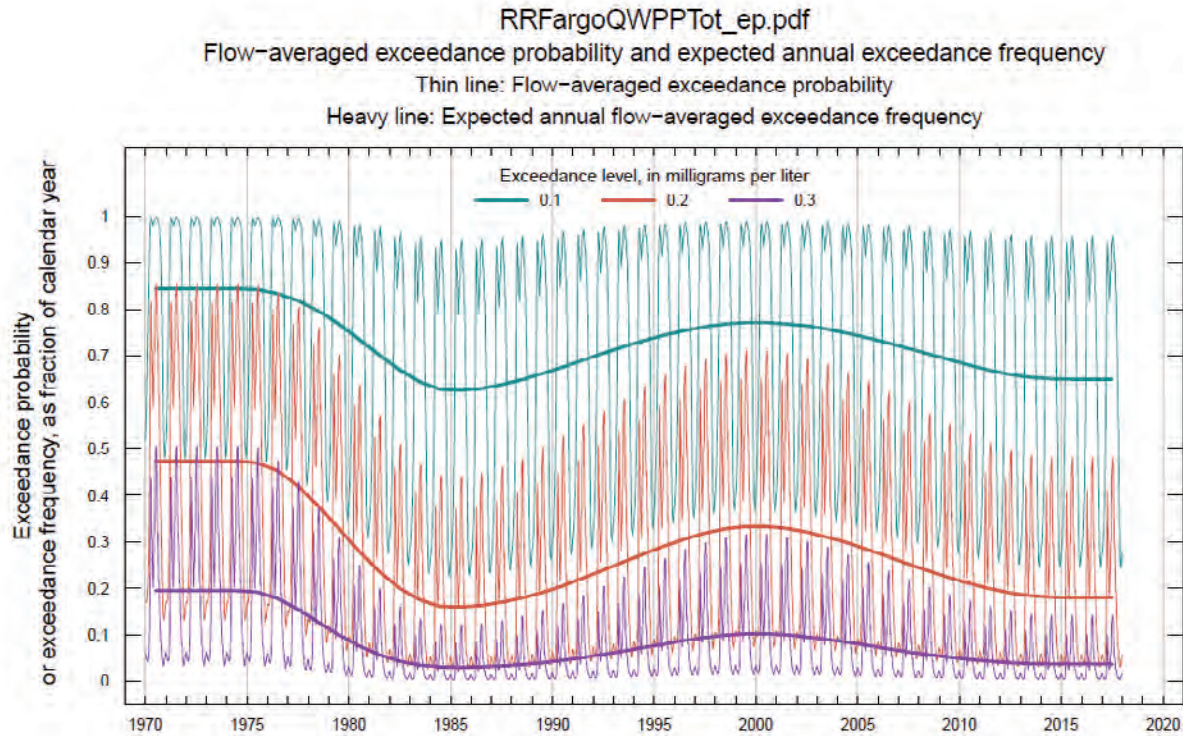
Figure 16.   Flow-averaged exceedance probabilities and expected annual exceedance frequencies for total phosphorus (PTot) concentration for RRFargoQWP dataset for model M2.

investigation before accepting the model. In this example, the "true" concentrations were assumed to be represented by the USGSNL and USGSND remark codes and concentrations with remark code "MPCA" were adjusted down by 23 percent to be consistent with the other observations. Another essentially equivalent alternative would be to assume observations with remark code MPCA represent "true" concentrations and adjust concentrations with the other remark codes up to be consistent with the MPCA data. This alternative would be completed by specifying remxx=c("USGSND," "USGSNL"). Deciding which alternative is best may require more investigation into sample collection and preservation methods and laboratory analysis methods used by the various agencies.

Exceedance probabilities for this example can be examined by running line 48 of the script file (fig. 2), which fits the same model used previously (M2) but with three specified exceedance levels of 0.1, 0.2, and 0.3 milligram per liter (mg/L). Exceedance levels are specified using a numeric vector for the exlev argument, in this case, exlev=c(0.1,0.2,0.3). After running line 48, examine the graphical output file (RRFargoQWPPTot_ep.pdf) and note the extra page of output (fig. 16) showing the flow-averaged exceedance probabilities (eq. 32) and expected annual flow-averaged exceedance frequencies (eq. 33).

Example 3—Parameter PTot for RRHalstadQWP

Example 3 is used to reinforce some of the concepts illustrated in the previous examples. In addition, generic dataframes with detailed numerical output are described and the methodology for including user-defined ancillary trend variables is illustrated.

Run lines 51 and 52 of the script (fig. 2) to fit the same trend models used in the previous example for this dataset and examine the parameter estimation results from the text files RRHalstadQWPPTot_M1.txt and RRHalstadQWPPTot_M2.txt (fig. 17). Similar to the previous example, there was a highly significant step trend (p-value<0.001) for remark code MPCA, with those observations about 45-percent higher, on average, compared to the other remark codes. A comparison of the piecewise monotonic trends with and without the step trend (fig. 18) indicates substantial differences in the two fitted trends. In particular, with the step trend included, the fitted trend line was shifted down overall and the monotonic trends were smaller and less significant compared to the model without the step trend. Compared with example 2 (see fig. 15), for this site there was an earlier time interval from 1978 through 1995 during which Minnesota Pollution Control Agency and USGS data were collected, which resulted in an even larger and more significant step trend for this site. With no step trend, the standardized PARMA model residuals for this site (fig. 19, top plot) clearly indicated two distinct populations of residuals, especially during the earlier period, with a much higher

```
****************
** -2lnLik = -408.54
**     AIC = -360.54
**   ecode = 1
**  modnum = 3

**  Parameter estimates

        Variable  CV(log)  SE(log)  Pvalue   CV(pct)
             int  -0.4904   0.0135  0.00000       NA
           xcos1  -0.2126   0.0188  0.00000   -38.71
           xsin1  -0.0777   0.0170  0.00053   -16.38
           xcos2   0.0096   0.0131  0.47388     2.24
           xsin2  -0.0246   0.0145  0.11416    -5.51
            falt  -0.1291   0.0405  0.00714    -9.08
            famt  -0.1584   0.0339  0.00044   -13.48
          famtsq   0.1818   0.0434  0.00106    29.17
            fast   0.3152   0.0382  0.00000    21.03
         fastxmt  -0.1826   0.0822  0.04468   -10.14
       m1975x1985 -0.1783   0.0321  0.00009   -33.67
       m1985x2000  0.1776   0.0371  0.00036    50.52
       m2000x2015 -0.1363   0.0406  0.00519   -26.94
```

```
****************
** -2lnLik = -495.94
**     AIC = -445.94
**   ecode = 1
**  modnum = 3

**  Parameter estimates

        Variable  CV(log)  SE(log)  Pvalue   CV(pct)
             int  -0.5821   0.0158  0.00000       NA
           xcos1  -0.1893   0.0178  0.00000   -35.33
           xsin1  -0.0775   0.0162  0.00028   -16.34
           xcos2   0.0079   0.0120  0.52172     1.84
           xsin2  -0.0241   0.0134  0.09327    -5.40
            falt  -0.1336   0.0372  0.00297    -9.39
            famt  -0.1459   0.0320  0.00044   -12.49
          famtsq   0.1836   0.0401  0.00043    29.50
            fast   0.2952   0.0345  0.00000    19.57
         fastxmt  -0.1373   0.0734  0.08255    -7.72
            rMPCA   0.1606   0.0160  0.00000    44.74
       m1975x1985 -0.1062   0.0304  0.00354   -21.69
       m1985x2000  0.1119   0.0344  0.00580    29.39
       m2000x2015 -0.0690   0.0378  0.08953   -14.69
```

EXPLANATION

[MPCA, sample collected and analyzed by Minnesota Pollution Control Agency]

–2lnLik—Minus two times the natural logarithm of the maximized likelihood function

AIC—Penalized likelihood value (not used)

ecode—Error code from nonlinear optimization program

modnum—Model number (table 1)

CV(log)—Estimated coefficient value

SE(log)—Approximate standard error of estimated coefficient

Pvalue—Approximate probability value of coefficient

CV(pct)—Estimated coefficient value expressed as a percentage

int—Intercept

xcos1—Cosine with period 1 year

xsin1—Sine with period 1 year

xcos2—Cosine with period 6 months

xsin2—Sine with period 6 months

falt—Long-term flow anomaly

famt—Midterm flow anomaly

famtsq—Square of midterm flow anomaly

fast—Short-term flow anomaly

fastxmt—Product of short-term and midterm flow anomalies

rMPCA—Step trend for remark code MPCA

m1975x1985—Monotonic trend from 1975 to 1985

m1985x2000—Monotonic trend from 1985 to 2000

m2000x2015—Monotonic trend from 2000 to 2015

Figure 17. Parameter estimation results for total phosphorus concentration for RRHalstadQWP dataset for piecewise monotonic trend model with no step trends (left) and with step trend for remark code MPCA (right).

proportion of negative residuals for remark code USGSNL and positive residuals for remark code MPCA. With the step trend included (fig. 19, bottom plot), the residual are much more homogeneous. In this case, further examination of the data for this site indicated that the step trend was likely related to a difference between the water-quality sampling locations for the Minnesota Pollution Control Agency data versus the USGS data.

In addition to the graphical output and parameter estimation files produced by runQWmodel, several generic dataframes are produced with detailed numerical information that can be used for custom applications. These dataframes are named QWMODOUT5D (5-day model output), QWMODOUTANN (annual model output), QWFLUXOUT5D (5-day flux output), and QWEXPROBOUT5D (5-day exceedance

probabilities, produced only if argument exlev is specified). These are overwritten each time a new model run is completed, so the user needs to rename or save the dataframes if they are needed for custom applications. For the most recent run completed (line 52, fig. 2), the first few lines of these dataframes are shown in figures 20A–D.

The generic dataframe QWMODOUT5D (fig. 20A) contains the input data used to fit the model (in this case, these data are from RRHalstadQWP) along with model output for each 5-day time step. Columns include the flow-related variability (frvar), flow-adjusted concentration data (fadat), concentration trend (tnd), flow-adjusted concentration data with step trends removed (fadat2, same as fadat if there are no step trends), concentration trend with step trends removed (tnd2, same as tnd if there are no step trends), and the standardized
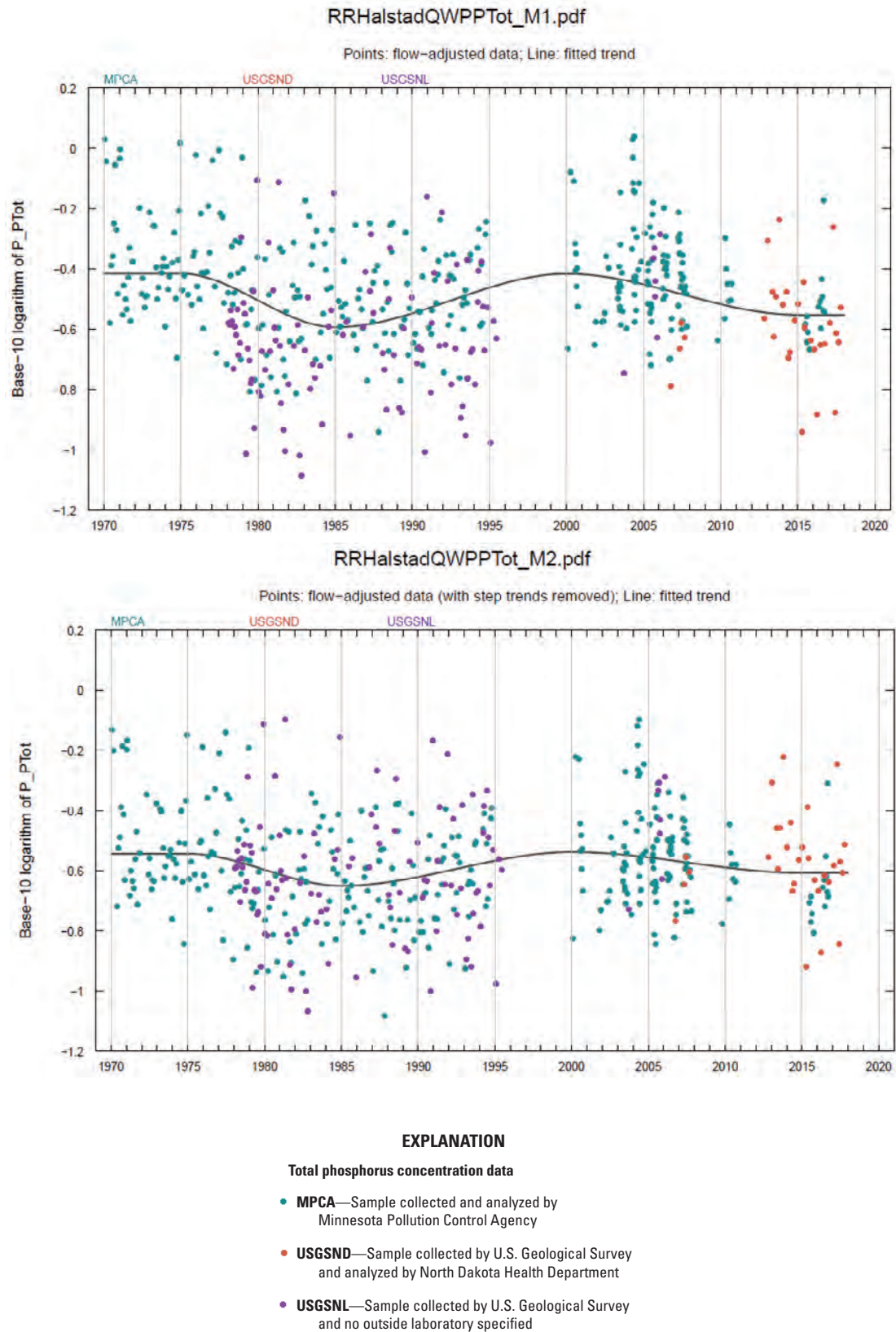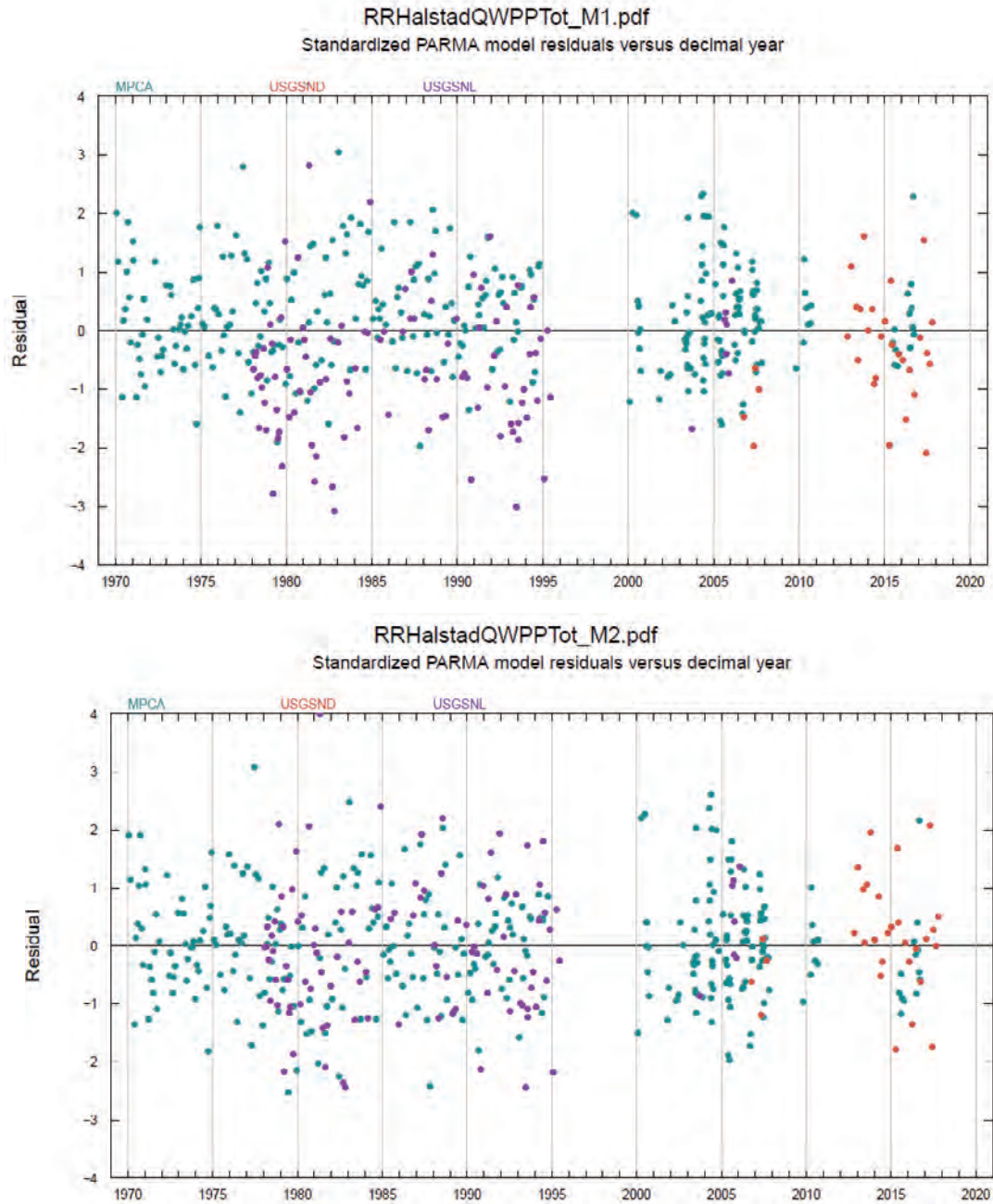
Figure 18. Flow-adjusted total phosphorus (PTot) concentration for RRHalstadQWP dataset for piecewise monotonic trend model with no step trends (model M1, top) and with step trend for remark code MPCA (model M2, bottom).

Figure 19. Standardized periodic autoregressive moving average (PARMA) model residuals for total phosphorus (PTot) concentration for RRHalstadQWP dataset for piecewise monotonic trend model with no step trends (model M1, top) and with step trend for remark code MPCA (model M2, bottom).

A

	yr	mo	da	flow	P_PTot	R_PTot	dyr	logc	frvar	fadat	tnd	fadat2	tnd2	fapfdat	spmres
3457	1970	1	3	295.0	NA	NA	1970.007	NA	-0.686	NA	-0.558	NA	-0.558	NA	NA
3458	1970	1	8	310.0	NA	NA	1970.021	NA	-0.690	NA	-0.558	NA	-0.558	NA	NA
3459	1970	1	13	343.0	NA	NA	1970.035	NA	-0.684	NA	-0.558	NA	-0.558	NA	NA
3460	1970	1	18	364.0	0.82	MPCA=	1970.049	-0.086	-0.682	0.014	-0.397	-0.147	-0.558	-0.248	1.898
3461	1970	1	23	357.0	NA	NA	1970.062	NA	-0.691	NA	-0.558	NA	-0.558	NA	NA
3462	1970	1	28	348.3	NA	NA	1970.076	NA	-0.699	NA	-0.558	NA	-0.558	NA	NA
3463	1970	2	3	336.6	NA	NA	1970.090	NA	-0.706	NA	-0.558	NA	-0.558	NA	NA

EXPLANATION

[PARMA, periodic autoregressive moving average]

yr—Calendar year

mo—Month

da—Day

flow—Streamflow, in cubic feet per second

P_PTot—Concentration data for parameter PTot

R_PTot—Remark code for parameter PTot

dyr—Decimal year for midpoint of 5-day time interval

logc—Logarithmically transformed concentration (base-10 logarithm)

fadat—Flow-adjusted concentration data

tnd—Concentration trend

fadat2—Flow-adjusted concentration data with step trends removed

tnd2—Trend with step trends removed

fapfdat—Flow-adjusted and PARMA filtered concentration data

spmres—Standardized PARMA model residual

MPCA—Sample collected by Minnesota Pollution Control Agency

Figure 20. Generic numerical model output for total phosphorus (PTot) concentration for RRHalstadQWP dataset for model M2. *A*, model output for 5-day time step, QWMODOUT5D; *B*, model output for annual time step, QWMODOUTANN; *C*, flux estimates for 5-day time step, QWFLUXOUT5D; *D*, exceedance probabilities for 5-day time step, QWEXPROBOUT5D.

B

	dyr	agmc	tagmc	frvP10	frvP90	amflow	afwac	tafwac	aflux	taflux
1	1970.49	0.335	0.297	0.218	0.477	1499.487	0.436	0.381	1.600	2.591
2	1971.49	0.321	0.297	0.222	0.466	1197.911	0.384	0.381	1.125	2.591
3	1972.49	0.282	0.297	0.168	0.374	2282.244	0.338	0.381	1.889	2.591
4	1973.49	0.318	0.297	0.220	0.515	995.699	0.359	0.381	0.876	2.591
5	1974.49	0.295	0.297	0.177	0.421	1856.283	0.366	0.381	1.661	2.591

EXPLANATION

dyr—Decimal year for midpoint of calendar year

agmc—Estimated annual geometric mean concentration, in milligrams per liter

tagmc—Trend in agmc

frvP10—10th percentile of flow-related variability

amflow—Annual mean streamflow, in cubic feet per second

afwac—Annual flow-weighted average concentration, in milligrams per liter

tafwac—Trend in afwac

aflux—Annual flux, in metric tons (1,000 kilograms) per day

Figure 20. Generic numerical model output for total phosphorus (PTot) concentration for RRHalstadQWP dataset for model M2. *A*, model output for 5-day time step, QWMODOUT5D; *B*, model output for annual time step, QWMODOUTANN; *C*, flux estimates for 5-day time step, QWFLUXOUT5D; *D*, exceedance probabilities for 5-day time step, QWEXPROBOUT5D.—Continued

C

	dyr5d	flow5d	flux5d
1	1970.007	295.0	0.18621
2	1970.021	310.0	0.19130
3	1970.035	343.0	0.21130
4	1970.049	364.0	0.22192
5	1970.062	357.0	0.20994

EXPLANATION

dyr5d—Decimal year for midpoint of 5-day time interval

flow5d—Streamflow, in cubic feet per second

flux5d—Flux, in metric tons (1,000 kilograms)

Figure 20.   Generic numerical model output for total phosphorus (PTot) concentration for RRHalstadQWP dataset for model M2. *A*, model output for 5-day time step, QWMODOUT5D; *B*, model output for annual time step, QWMODOUTANN; *C*, flux estimates for 5-day time step, QWFLUXOUT5D; *D*, exceedance probabilities for 5-day time step, QWEXPROBOUT5D.—Continued

D

	dyrout	exlev1	exprfav1	exprann1	exlev2	exprfav2	exprann2	exlev3	exprfav3	exprann3
1	1970.007	0.1	0.878	0.975	0.2	0.489	0.799	0.3	0.235	0.527
2	1970.021	0.1	0.885	0.975	0.2	0.483	0.799	0.3	0.221	0.527
3	1970.035	0.1	0.896	0.975	0.2	0.481	0.799	0.3	0.208	0.527
4	1970.049	0.1	0.910	0.975	0.2	0.482	0.799	0.3	0.195	0.527
5	1970.062	0.1	0.925	0.975	0.2	0.484	0.799	0.3	0.183	0.527

EXPLANATION

[Trailing integers (1, 2, 3) for column names designate three different specified exceedance levels]

dyrout—Decimal year

exlev—Exceedance level, in milligrams per liter

exprfav—Flow-averaged exceedance probability

exprann—Flow-averaged annual exceedance frequency, as expected fraction of time during dyrout–0.5 to dyrout+0.5 with concentration above the exceedance level

Figure 20.   Generic numerical model output for total phosphorus (PTot) concentration for RRHalstadQWP dataset for model M2. *A*, model output for 5-day time step, QWMODOUT5D; *B*, model output for annual time step, QWMODOUTANN; *C*, flux estimates for 5-day time step, QWFLUXOUT5D; *D*, exceedance probabilities for 5-day time step, QWEXPROBOUT5D.—Continued

PARMA model residuals (spmres). For example, the bottom plot in figure 18 can be reproduced by plotting the decimal year versus fadat2 (points) and the decimal year versus tnd2 (line) and the bottom plot in figure 19 can be reproduced by plotting the decimal year versus spmres.

The dataframe QWMODOUTANN (fig. 20B) contains annual model output used for producing plots such as those shown in figure 12H–J. Columns include the estimated annual geometric mean concentration (agmc, the points in fig. 12H) and associated trend (tagmc, the line in fig. 12H), the estimated annual flow-weighted average concentration (afwac, the points in fig. 12I) and associated trend (tafwac, the line in fig. 12I), and the estimated annual flux (aflux, the points in fig. 12J) and associated trend (taflux, the line in fig. 12J).

The dataframe QWFLUXOUT5D (fig. 20C) contains flux estimates for each 5-day time step and can be used for custom applications that require more detained flux estimates than the annual values provided in QWMODOUTANN. Variables include streamflow (flow5d, in cubic feet per second) and estimated flux (flux5d, in metric tons).

The dataframe QWEXPROBOUT5D (fig. 20D) contains information on exceedance probabilities used to produce plots such as figure 16. Variables include, for each specified exceedance level, the flow-averaged exceedance probability (exprfav, used to produce the thin lines in fig. 16) and the flow-averaged annual exceedance frequency (exprann, used to produce the heavy lines in fig. 16, where the annual values are plotted for the midpoint of each decimal year).

The next part of this example shows how to include user-specified trend variables using the optional argument userxx in the runQWmodel() function:

userxx=list(nu, ancxx, vnms)

where

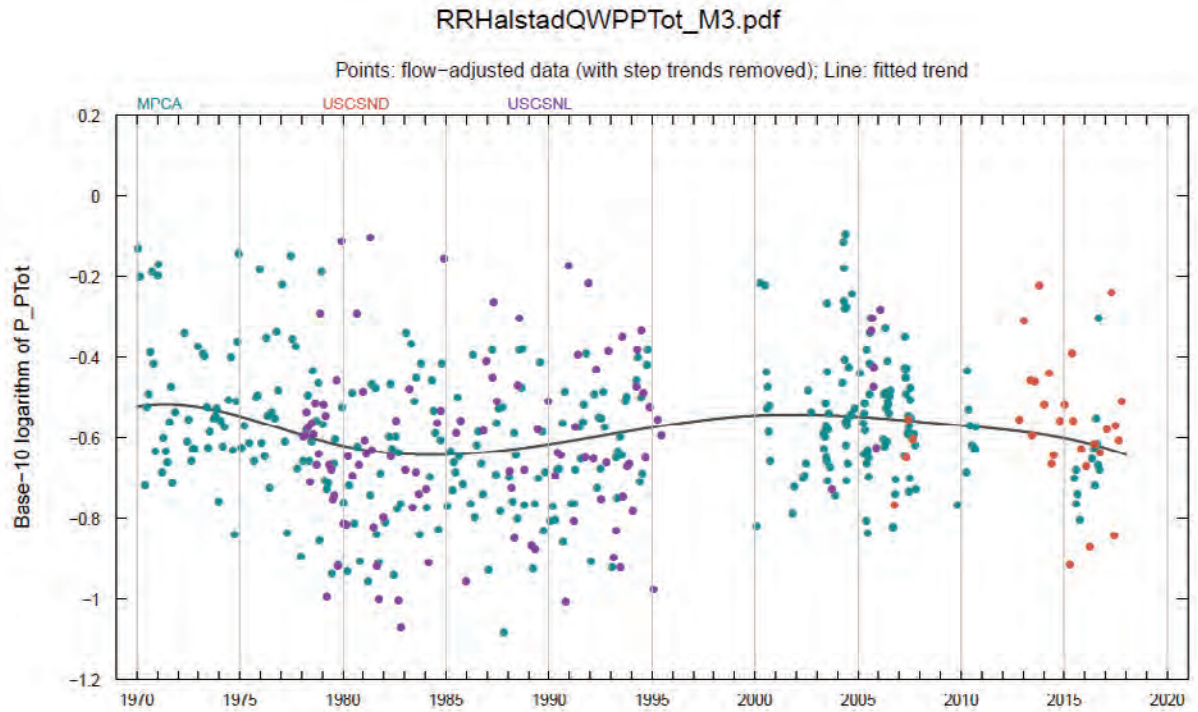|	
nu	is the number of user-specified trends,
ancxx	is a dataframe with a row dimension identical to the input dataframe (XXXQWP) and one or more columns with user-specified ancillary trend variables, and
vnms	is a character vector of length nu specifying the column names of the trend variables to select from ancxx.

The columns of ancxx corresponding to ancillary trend variables must be numeric and have no missing values, and the dates of the rows must correspond to the "date" column of XXXQWP.

Run lines 55–56 of the script file (fig. 2) to create a dataframe called ancxx with 8 columns, where the first column is the decimal year and the remaining 7 columns are basis functions for a cubic spline with the specified knots. Then, run line 57 to fit a trend model with user-specified trend variables consisting of the seven basis functions and step trend for remark code MPCA (model M3). The graphical output for model M3 (see RRHalstadQWPPTot_M3.pdf, fig. 21) was similar to the output for model M2 (fig. 18, bottom plot) and the value of $-2\ln Lik$ for model M3 (-497.56, see RRHalstadWQWPPTot_M3.txt) was less than the value for M2 (-495.94). However, using the GLR principal, the p-value for judging if the more complicated model (M3) is better than the less complicated model (M2) is $P=1-pchisq(1.62, df=4)=0.805$. The large p-value indicates that the less complicated model (M2) is preferred over M3.

Occasionally, when runQWmodel() is computing the maximum likelihood parameter estimates, a runtime error can occur. The most likely cause for the error is a misspecified trend model; for example, in line 59 of the script file (fig. 2), there are 3 monotonic trends specified, but the first two are identical. When running this command, an error window will appear (fig. 22). When this happens, simply close the error window, correct the trend model, and continue. On rare occasions, a similar error window may appear even if the trend model is not misspecified. This usually means that either the trend model, PARMA model, or both are too complicated to allow the maximum likelihood estimates to be computed. If the error cannot be avoided by either simplifying the trend model or lowering the PARMA model number (for example, trying modnum=2 or modnum=1 rather than the default, modnum=3), the data cannot be analyzed using R–QWTREND.

Figure 21. Graphical output for total phosphorus (PTot) concentration for RRHalstadQWP dataset for model with user-specified trend (model M3).
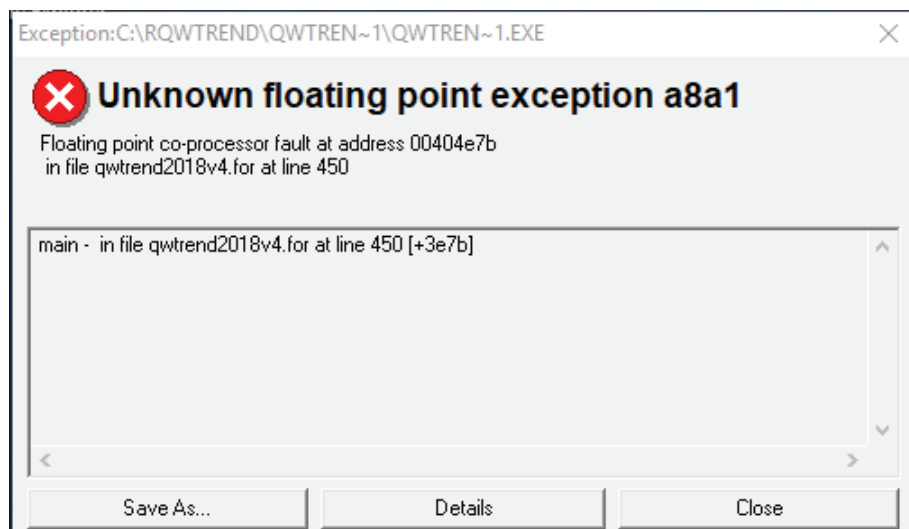

Figure 22. Runtime error window produced by runQWmodel when an error is detected during the maximum likelihood estimation.

Summary

As part of a U.S. Geological Survey water-quality study started in 2018, in cooperation with the International Joint Commission, North Dakota Department of Environmental Quality, and Minnesota Pollution Control Agency, a publicly available software package called R–QWTREND was developed for analyzing trends in stream-water quality. The R–QWTREND package is a collection of functions written in R, an open source language and a general environment for statistical computing and graphics. The package uses a parametric time-series model to express logarithmically transformed concentration in terms of flow-related variability, trend, and serially correlated model errors. Flow-related variability captures natural variability in concentration on the basis of concurrent and antecedent streamflow. The trend identifies systematic changes in concentration in terms of potential step trends, piecewise monotonic trends, or user-specified trends. Maximum likelihood estimation is used to estimate model parameters and determine the best-fit trend model.

This report describes the time-series model and statistical methodology behind R–QWTREND and provides formal documentation for installing and using the package. This report, along with the accompanying software package, practice datasets, and examples, provides all of the necessary materials and documentation for using R–QWTREND to analyze and interpret trends in stream-water quality based on long-term (10 or more years) datasets on constituent concentration from discrete stream-water samples collected multiple times per year (quarterly or more frequent sampling) and for which the stream-water sampling location is colocated with a streamflow-gaging station from which a complete record of daily mean streamflow is available.

References Cited

Cohn, T.A., DeLong, L.L., Gilroy, E.J., Hirsch, R.M., and Wells, D.K., 1989, Estimating constituent loads: Water Resources Research, v. 25, no. 5, p. 937–942. [Also available at https://doi.org/10.1029/WR025i005p00937.]

Dennis, J.E., Jr., and Schnabel, R.B., 1996, Numerical methods for unconstrained optimization of nonlinear equations: Englewood Cliffs, N.J., Society for Industrial and Applied Mathematics, 378 p. [Also available at https://doi.org/10.1137/1.9781611971200.]

Galloway, J.M., Vecchia, A.V., Vining, K.C., Densmore, B.K., and Lundgren, R.F., 2012, Evaluation of water-quality characteristics and sampling design for streams in North Dakota, 1970–2008: U.S. Geological Survey Scientific Investigations Report 2012–5216, 304 p. [Also available at https://doi.org/10.3133/sir20125216.]

Giorgino, M.J., Cuffney, T.F., Harden, S.L., and Feaster, T.D., 2018, Trends in water quality of selected streams and reservoirs used for water supply in the Triangle area of North Carolina, 1989–2013: U.S. Geological Survey Scientific Investigations Report 2018–5077, 67 p., accessed June 2019 at https://doi.org/10.3133/sir20185077.

Graybill, F.A., 1976, Theory and application of the linear model: Duxbury Press, 704 p.

Helsel, D.R., and Hirsch, R.M., 1992, Statistical methods in water resources: Elsevier, 529 p.

Jimenez, C., McLeod, A.I., and Hipel, K.W., 1989, Kalman filter estimation for periodic autoregressive moving average models: Stochastic Hydrology and Hydraulics, v. 3, no. 3, p. 227–240. [Also available at https://doi.org/10.1007/BF01543862.]

Johnson, H.O., Gupta, S.C., Vecchia, A.V., and Zvomuya, F., 2009, Assessment of water quality trends in the Minnesota River using non-parametric and parametric methods: Journal of Environmental Quality, v. 38, no. 3, p. 1018–1030. [Also available at https://doi.org/10.2134/jeq2008.0250.]

Jones, G., and Armstrong, N., 2001, Long-term trends in total nitrogen and total phosphorus concentrations in Manitoba streams: Environment Manitoba, Manitoba Conservation Report no. 2001–07, 175 p.

Metropolitan Council, 2014, Comprehensive water quality assessment of selected metropolitan area streams—Introduction and methodologies: St. Paul, Minn., Metropolitan Council, 46 p., accessed November 7, 2019, at https://metrocouncil.org/Wastewater-Water/Services/Water-Quality-Management/Stream-Monitoring-Analysis/Introduction-and-Methodologies.aspx.

National Water Quality Monitoring Council, 2019, Water quality portal: National Water Quality Monitoring Council web page, accessed October 31, 2019, at https://www.waterqualitydata.us.

Paquette, C., 2011, Statistical analysis of trends in the Red River over a 45-year period: University of Manitoba, UM Project 33898, 126 p.

R Development Core Team, 2019, R—A language and environment for statistical computing: Vienna, Austria, R Foundation for Statistical Computing, accessed June 12, 2019, at https://www.r-project.org.

Risch, M.R., Bunch, A.R., Vecchia, A.V., Martin, J.D., and Baker, N.T., 2014, Water quality in Indiana—Trends in concentrations of selected nutrients, metals, and ions in streams, 2000–10: U.S. Geological Survey Scientific Investigations Report 2014–5205, 47 p. [Also available at https://doi.org/10.3133/sir20145205.]

Salas, J.D., Tabios, G.Q., III, and Bartolini, P., 1985, Approaches to multivariate modeling of water resources time series: Water Resources Bulletin, v. 21, no. 4, p. 683–708. [Also available at https://doi.org/10.1111/j.1752-1688.1985.tb05383.x.]

Sando, S.K., Clark, M.L., Cleasby, T.E., and Barnhart, E.P., 2015, Water-quality trends for selected sites in the Boulder River and Tenmile Creek watersheds, Montana, based on data collected during water years 1997–2013: U.S. Geological Survey Scientific Investigations Report 2015–5008, 46 p., accessed June 2019 at https://doi.org/10.3133/sir20155008.

Sando, S.K., Vecchia, A.V., Lorenz, D.L., and Barnhart, E.P., 2014a, Water-quality trends for selected sampling sites in the upper Clark Fork Basin, Montana, water years 1996–2010: U.S. Geological Survey Scientific Investigations Report 2013–5217, 162 p., with appendixes, accessed June 2019 at http://dx.doi.org/10.3133/sir20135217.

Sando, S.K., Vecchia, A.V., Barnhart, E.P., Sando, T.R., Clark, M.L., and Lorenz, D.L., 2014b, Trends in major-ion constituents and properties for selected sampling sites in the Tongue and Powder River watersheds, Montana and Wyoming, based on data collected during water years 1980–2010: U.S. Geological Survey Scientific Investigations Report 2013–5179, 123 p., accessed June 2019 at https://doi.org/10.3133/sir20135179.

Therneau, T.M., 2019, survival—A package for survival analysis in S: R package version 3.5.1, accessed June 12, 2019, at https://cran.r-project.org/web/packages/survival/index.html.

U.S. Geological Survey, 2019, USGS water data for the Nation: U.S. Geological Survey National Water Information System database, accessed October 31, 2019, at https://nwis.waterdata.usgs.gov/nwis. [Also available at https://doi.org/10.5066/F7P55KJN.]

Vecchia, A.V., 1985, Periodic autoregressive moving average (PARMA) modeling with applications to water resources: Water Resources Bulletin, v. 21, no. 5, p. 721–730. [Also available at https://doi.org/10.1111/j.1752-1688.1985.tb00167.x.]

Vecchia, A.V., 2000, Water-quality trend analysis and sampling design for the Souris River, Saskatchewan, North Dakota and Manitoba: U.S. Geological Survey Water-Resources Investigation Report 2000–4019, 77 p. [Also available at https://doi.org/10.3133/wri004019.]

Vecchia, A.V., 2003, Water-quality trend analysis and sampling design for streams in North Dakota: U.S. Geological Survey Water-Resources Investigation Report 2003–4094, 73 p. [Also available at https://doi.org/10.3133/wri034094.]

Vecchia, A.V., 2005, Water-quality trend analysis and sampling design for streams in the Red River of the North Basin, Minnesota, North Dakota, and South Dakota, 1970–2001: U.S. Geological Survey Scientific Investigations Report 2005–5224, 54 p. [Also available at https://doi.org/10.3133/sir20055224.]

Appendix 1   R–QWTREND Software Package

A folder, QWTREND2018V4, can be downloaded at https://doi.org/10.3133/ofr20201014. The folder contains the files described in table 1.1.

Although the files and data have been processed successfully on a computer system at the U.S. Geological Survey (USGS), no warranty expressed or implied is made regarding the display or utility of the data for other purposes, nor on all computer systems, nor shall the act of distribution constitute any such warranty. The USGS or the U.S. Government shall not be held liable for improper or incorrect use of the data and files described and (or) contained herein.

This software has been approved for release by the USGS. Although the software has been subjected to rigorous review, the USGS reserves the right to update the software as needed pursuant to further analysis and review. No warranty, expressed or implied, is made by the USGS or the U.S. Government as to the functionality of the software and related material nor shall the fact of release constitute any such warranty. Furthermore, the software is released on condition that neither the USGS nor the U.S. Government shall be held liable for any damages resulting from its authorized or unauthorized use.

Table 1.1.   Name and description of files in QWTREND2018V4 folder.

File name	Description
Readme.doc	Word document with instructions for using R–QWTREND.
prepQWdataV4.txt	Text file containing R code for creating function prepQWdata().
runQWmodelV4.txt	Text file containing R code for creating function runQWmodel().
plotQWtrendV4.txt	Text file containing R code for creating function plotQWtrend().
qwtrend2018v4.exe	Windows executable file used for computing maximum likelihood parameter estimates.
salflibc.dll	Dynamic link library required by qwtrend2018v4.exe.
QWTrendV4_practice.RData	R workspace with example datasets.
StartQWTrendV4.R	R script for installing and running R–QWTREND.