USGS
science for a changing world

National Water-Quality Assessment Program

waterData—An R Package for Retrieval, Analysis, and Anomaly Calculation of Daily Hydrologic Time Series Data, Version 1.0

Open-File Report 2012–1168

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

waterData—An R Package for Retrieval, Analysis, and Anomaly Calculation of Daily Hydrologic Time Series Data, Version 1.0

By Karen R. Ryberg and Aldo V. Vecchia

National Water-Quality Assessment Program

Open-File Report 2012–1168

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

U.S. Department of the Interior
KEN SALAZAR, Secretary

U.S. Geological Survey
Marcia K. McNutt, Director

U.S. Geological Survey, Reston, Virginia: 2012

Suggested citation:
Ryberg, K.R, and Vecchia, A.V., 2012, waterData—An R package for retrieval, analysis, and anomaly calculation of daily hydrologic time series data, version 1.0: U.S. Geological Survey Open-File Report 2012–1168, 8 p.

Contents

Tables

Conversion Factors

Inch/Pound to SI

Multiply	By	To obtain
	Area	
square mile (mi^2)	259.0	hectare (ha)
square mile (mi^2)	2.590	square kilometer (km^2)
	Flow rate	
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)

SI to Inch/Pound

Multiply	By	To obtain
	Area	
square kilometer (km^2)	247.1	acre
square kilometer (km^2)	0.3861	square mile (mi^2)
	Flow rate	
cubic meter per second (m^3/s)	35.31	cubic foot per second (ft^3/s)

waterData—An R Package for Retrieval, Analysis, and Anomaly Calculation of Daily Hydrologic Time Series Data, Version 1.0

By Karen R. Ryberg and Aldo V. Vecchia

Abstract

Hydrologic time series data and associated anomalies (multiple components of the original time series representing variability at longer-term and shorter-term time scales) are useful for modeling trends in hydrologic variables, such as streamflow, and for modeling water-quality constituents. An R package, called **waterData,** has been developed for importing daily hydrologic time series data from U.S. Geological Survey streamgages into the R programming environment. In addition to streamflow, data retrieval may include gage height and continuous physical property data, such as specific conductance, pH, water temperature, turbidity, and dissolved oxygen. The package allows for importing daily hydrologic data into R, plotting the data, fixing common data problems, summarizing the data, and the calculation and graphical presentation of anomalies.

Introduction

U.S. Geological Survey (USGS) daily hydrologic data can be used to identify trends in the hydrologic variables themselves, used as exogenous variables in trend models for water-quality data (Helsel and Hirsch, 2002), and divided into multiple components, or anomalies, representing variability over longer-term and shorter-term time scales. Those components can be used as multiple exogenous variables in multiple regression models. The R package **waterData** was developed to provide functions to import daily hydrologic time series data, perform data checks, fix data problems, plot the data, and calculate and plot anomalies. The plot functions are intended for exploratory data analysis and not for final publication purposes.

A complete example of the process of importing daily data, summarizing it, performing data checks, fixing common data problems, plotting the data, and calculating anomalies is provided in the vignette, or tutorial, in Appendix 1. Additional detail for each function, including the arguments and returned values, is provided in Appendix 2.

Description of waterData

This collection of functions is written as a package for R (*http://www.r-project.org/*, R Development Team, 2012c), an open source language and a general environment for statistical computing and graphics that runs on a variety of operating systems including Linux®, Mac OS®, UNIX®, and Windows®. R can be extended for additional functionality using packages. Additional information on the installation and administration of R and packages that extend it is available in the R Installation and Administration manual (*http://streaming.stat.iastate.edu/CRAN/doc/manuals/R-admin.pdf*, R Development Team, 2011).

In many hydrologic trend studies, a large part of the work involves importing, checking, and exploratory data analysis before trend models can be used. The R package, **waterData**, described in this report was created to standardize and streamline this process. Table 1 lists the functions in **waterData** and provides a brief description of each. For more details on these functions and help preparing data sets for analysis, see the Vignette and R Documentation in the Appendixes. Help files, which contain the same information as Appendix 2, also are available in R once the package has been installed.

A primary feature of the package is the function, `importDVs`, that imports USGS daily hydrologic time series data from the USGS Daily Values Site Web Service (*http://waterservices.usgs.gov/rest/DV-Service.html*). This function provides a direct link from USGS water services to R, rather than the often-used process with an intermediate step of downloading the data to a file, then importing into R. The result of the function call is a data frame with columns staid (USGS station identification number), val (the value of the time series retrieved), dates (the date of each observation), and qualcode (USGS data qualification codes). The USGS parameter code and statistics code are attached to the data frame as attributes, code and stat, respectively, so later one can verify exactly what data were downloaded (for example, parameter code 00060, streamflow in cubic feet per second, and statistics code 00003, mean, for mean daily streamflow).

Table 1. List of functions in **waterData** package and short description of each.

[USGS, U.S. Geological Survey; URL, Uniform Resource Locator]

Function	Description
cleanUp	Identifies and cleans up, based on user specifications, hydrologic time series data by replacing 0's with very small values, such as in the case of an analysis using logarithms of streamflow, which do not work for zero values. The function also will replace negative values with NA, an indicator of missing values.
compAnom	Calculates short-, medium-, and long-term anomalies from hydrologic time series.
fillMiss	Fills in missing values by estimating them using a structured time series and a time series smoother.
importDVs	Imports daily USGS hydrologic time series data.
plotAnoms	Plots hydrologic anomalies.
plotParam	Plots hydrologic time series data.
siteInfo	Retrieves streamgage site information.
summaryStats	Calculates summary statistics based on a hydrologic time series.
tellMeSiteURL	Provides the USGS Site Information Service URL used to retrieve data in the siteInfo function.
tellMeURL	Provides the USGS Daily Values Site Service URL used to retrieve data in the importDVs function.

The package will retrieve and plot streamflow, gage height, and continuous physical property data (including specific conductance, pH, water temperature, turbidity, and dissolved oxygen) from USGS streamgages. Like streamflow data and anomalies, these continuous physical property data may be used as exogenous variables in water-quality analyses. Examples of the use of specific conductance, pH, water temperature, turbidity, and dissolved oxygen, as well as streamflow, in regression analysis for water-quality monitoring are given by Christensen (2001), Christensen and others (2000 and 2006), and Ryberg (2006 and 2007).

Hydrologic data from other entities, such as the U.S. Army Corps of Engineers, may be imported to R by the user. Once the user structures the time series in the same manner as the data returned by the `importDVs` function, with columns staid (station identification number), val (the value of the time series variable), dates (the date of each observation), and qualcode (data qualification codes), the other functions for plotting, checking, and anomaly calculation in the package may be applied to the hydrologic time series.

Another primary feature of the package is that it contains functions, `cleanUp` and `fillMiss`, that check for negative values, values of 0, and missing data, and in some cases remove or replace those values. Negative values, which some agencies use to represent missing data, are problematic and the `cleanUp` function may be used to change negative values to missing values. Streamflow values of 0 cause problems when calculating anomalies because anomalies are based on logarithms. Zero values can be replaced with 0.1 (or other value supplied by the user) using the function `cleanUp`.

Frequently, in hydrologic time series, there are missing values because of equipment malfunctions or other discontinuities in operation. Missing values reduce the time period over which a complete set of anomalies may be calculated. A function is available to fill in some missing values depending on characteristics of the data set. The function, `fillMiss`, uses the `StructTS` function from the base package of R (R Development Team, 2012a) to fit a structural time series model to the data. Then the data are smoothed, using `tsSmooth` (fixed-interval smoothing on a univariate time series using a state-space model; R Development Team, 2012a), and the filled in values are used to replace the missing values in the original time series.

In addition to the functions, **waterData** provides some sample data sets that are used to illustrate the format of the data sets returned by `importDVs` and to illustrate functionality in the vignette. The sample data sets are listed in table 2.

Examples showing how to use **waterData** and all of the associated functions are provided in the package vignette (appendix 1). Vignettes are documents that contain examples of R code and results of running the code, as well as descriptive text (R Development Team, 2012d). Vignettes can be used as tutorials for the package and the vignette for **waterData** is included in this document to familiarize users with the functions in **waterData**. The help documentation for functions and data sets in the **waterData** package is included in Appendix 2

Table 2. List of sample data sets in the **waterData** package and short description of each.

Data set	Description
badDataSet	A mean daily streamflow time series with a problematic negative value and some zero values.
misQ05054000	A mean daily streamflow time series with many missing values.
pH05082500	A median daily pH time series with missing values.

of this document. Help features within R are further described in the manual *An Introduction to R* (Venables and others, 2011).

Anomalies

After the data have been prepared, the package can be used to calculate and visualize anomalies. The anomaly concept was first described by Vecchia (2003), and subsequently used and refined in numerous analyses of surface-water quality (Ryberg and others, 2010; Sullivan and others, 2009; Vecchia and others, 2009; Vecchia and others, 2008; Alexander and Smith, 2006; Ryberg and Vecchia, 2006; Vecchia, 2005). The majority of these studies using the streamflow anomaly concept have been related to pesticide concentrations in surface-water samples, but the same concepts may be applied to analyzing nutrients and other chemical constituents.

Anomalies may be calculated over multiple time scales. In an analysis of pesticide concentrations in urban streams (Ryberg and others, 2010), three streamflow (flow) anomalies, daily, 10-day, and 100-day, were included in the time series model (SEAWAVE-Q; Sullivan and others, 2009) to help account for flow-related variability in pesticide concentrations. As an example in this report, daily, 10-day, and 100-day anomalies are computed using the log-transformed daily flow. Other time scales may be used and are available in the package **waterData**.

In the following example, the short-term anomaly represents 1-day to 10-day flow variability, and is defined as

$$STFA(t) = X(t) - X_{10}(t) \quad (1)$$

where

$STFA(t)$ is the short-term anomaly (dimensionless) at time t;

$X(t)$ is the log-transformed daily flow, in cubic meters per second or cubic feet per second; and

$X_{10}(t)$ is the average of log-transformed daily flow for 10 days up to and including time t.

The mid-term anomaly represents 10- to 100-day flow variability and is defined as

$$MTFA(t) = X_{10}(t) - X_{100}(t) \quad (2)$$

where

$MTFA(t)$ is the mid-term anomaly (dimensionless) at time t; and

$X_{100}(t)$ is the average of log-transformed daily flow for 100 days up to and including time t.

The long-term anomaly represents greater than 100-day flow variability and is defined as

$$LTFA(t) = X_{100}(t) - X_{*} \quad (3)$$

where

$LTFA(t)$ is the long-term anomaly (dimensionless) at time t; and

X_{*} is the average of log-transformed daily flow for the specified period.

Unlike *STFA* (equation 1), which tends to affect constituent concentrations in a relatively consistent manner among different sites and pesticides, *MTFA* (equation 2) and *LTFA* (equation 3) can affect the concentrations of chemical constituents in different ways and to different degrees depending on the chemical properties, as well as the climatic and hydrologic properties of the basin. For example, a relatively large basin with substantial nonurban runoff and higher-than-normal seasonal flow conditions (as indicated by a positive value for *LTFA*) can cause decreased constituent concentrations because of dilution from nonurban runoff (Ryberg and others, 2010).

STFA, *MTFA*, and *LTFA* may be used as exogenous variables in multiple regression models to examine trends in water-quality constituents. Numerous published reports serve as examples of the use of streamflow anomalies for trend analysis of water-quality concentration data (Sullivan and others, 2009; Vecchia and others, 2009; Vecchia and others, 2008; Alexander and Smith, 2006; Ryberg and Vecchia, 2006; Vecchia, 2005). The streamflow anomalies also may be useful for understanding the variability of streamflow over varying time scales and for estimating flow at ungaged locations.

Summary

Hydrologic time series data and associated anomalies (multiple components of the original time series representing variability over longer-term and shorter-term time scales) are useful for modeling trends in hydrologic variables, such as streamflow, and for modeling water-quality constituents. An R package, called **waterData**, has been developed for importing daily hydrologic time series data from U.S. Geological Survey (USGS) streamgages into the R programming environment. In addition to streamflow, data retrieval may include gage height and continuous physical property data, such as specific conductance, pH, water temperature, turbidity, and dissolved oxygen. The R package allows for plotting the data, fixing common data problems, summarizing the data, and the calculation and graphical presentation of anomalies. Users may independently import into R hydrologic data from other entities and structure it in the same manner as the USGS data, then use the function to plot, fix, summarize, and calculate anomalies.

The package is now available in the free, public Comprehensive R Archive Network, *http://cran.r-project.org/* (R Core Team, 2012b). The appendixes of this document provide an example of how to use the R package and document the functions.

Disclaimer

This package was written by U.S. Federal government employees in the course of their employment and is therefore in the public domain, which means it is not copyrighted and use is unlimited; however, some of the functions depend on other R-packages, which, although free and open source, have more restrictive licensing. Those packages are **lattice [**GNU (Gnu's Not Unix) GPL (General Public License) ≥ (greater than or equal to version) 2], **latticeExtra (**GPL ≥ 2), **XML (**Berkeley Software Distribution, BSD). R itself is released under the free software license GNU GPL, either Version 2, June 1991, or Version 3, June 2007. Additional information on licensing is available at *http://www.r-project.org/Licenses/* and *http://www.gnu.org/licenses/license-list.html#SoftwareLicenses*.

Although this software package has been used by the U.S. Geological Survey (USGS), no warranty, expressed or implied, is made by the USGS or the U.S. Government as to the accuracy and functioning of the program and related program material nor shall the fact of distribution constitute any such warranty, and no responsibility is assumed by the USGS in connection therewith. This software and related material (data and documentation) are made available by the USGS to be used in the public interest and the advancement of science. Users may, without any fee or cost, use, copy, modify, or distribute this software, and any derivative works thereof, and its supporting documentation, subject to the USGS Software User's Rights Notice, *http://water.usgs.gov/software/help/notice/*.

Acknowledgments

The authors thank the time and effort of the U.S. Geological Survey personnel that reviewed this report and tested the R-package and vignette, and contributed suggestions to improve the package. The testers were Eric Winiger, North Dakota Water Science Center; Max Post van der Burg, Northern Prairie Wildlife Research Center; Jason Fisher, Idaho National Laboratory; and Dave Lorenz, Minnesota Water Science Center.

References Cited

Alexander, R.B., and Smith, R.A., 2006, Trends in the nutrient enrichment of U.S. rivers during the late 20th century and their relation to changes in probable stream trophic conditions: Limnology and Oceanography, v. 51, no. 1, part 2, p. 639–654. (Also available at *http://www.jstor.org/stable/4499617*.)

Christensen, V.G., 2001, Characterization of surface-water quality based on real-time monitoring and regression analysis, Quivira National Wildlife Refuge, south-central Kansas, December 1998 through June 2001: U.S. Geological Survey Water-Resources Investigations Report 01–4248, 28 p. (Also available at *http://pubs.water.usgs.gov/wri014248*.)

Christensen, V.G., Graham, J.L., Milligan, C.R., Pope, L.M., and Ziegler, A.C., 2006, Water quality and relation to taste-and-odor compounds in the North Fork Ninnescah River and Cheney Reservoir, south-central Kansas, 1997–2003: U.S. Geological Survey Scientific Investigations Report 2006–5095, 43 p. (Also available at *http://pubs.usgs.gov/sir/2006/5095/*.)

Christensen, V.G., Jian, Xiaodong, and Ziegler, A.C., 2000, Regression analysis and real-time water-quality monitoring to estimate constituent concentrations, loads, and yields in the Little Arkansas River, south-central Kansas, 1995–99: U.S. Geological Survey Water-Resources Investigations Report 00–4126, 36 p. (Also available at *http://pubs.water.usgs.gov/wri004126*.)

Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 522 p. (Also available at *http://pubs.usgs.gov/twri/twri4a3/*).

R Development Team, 2011, R Installation and administration, Version 2.14.1, 2011-12-22, accessed January 13, 2012, at *http://streaming.stat.iastate.edu/CRAN/doc/manuals/R-admin.pdf*.

R Development Team, 2012a, R – A language and environment for statistical computing: Vienna, Austria, R Foundation for Statistical Computing, [ISBN 3-900051-07-0]. (Also available at *http://www.R-project.org/*.)

R Development Team, 2012b, The comprehensive R archive network, accessed September 14, 2012, at *http://cran.r-project.org/*.

R Development Team, 2012c, The R project for statistical computing, accessed September 14, 2012, at *http://www.r-project.org/*.

R Development Team, 2012d, Writing R extensions, Version 2.14.1, 2012-06-22, accessed September 14, 2012, at *http://cran.r-project.org/doc/manuals/R-exts.pdf*.

Ryberg, K.R., 2006, Continuous water-quality monitoring and regression analysis to estimate constituent concentrations and loads in the Red River of the North, Fargo, North Dakota, 2003-05: U.S. Geological Survey Scientific Investigations Report 2006–5241, 35 p. (Also available at *http://pubs.water.usgs.gov/SIR20065241*.)

Ryberg, K.R., 2007, Continuous water-quality monitoring and regression analysis to estimate constituent concentrations and loads in the Sheyenne River, North Dakota, 1980–2006: U.S. Geological Survey Scientific Investigations Report 2007–5153, 22 p. (Also available at *http://pubs.usgs.gov/sir/2007/5153/*.)

Ryberg, K.R., and Vecchia, A.V., 2006, Water-quality trend analysis and sampling design for the Devils Lake Basin, North Dakota, January 1965 through September 2003: U.S. Geological Survey Scientific Investigations Report 2006–5238, 64 p. (Also available at *http://pubs.usgs.gov/sir/2006/5238/*.)

Ryberg, K.R., Vecchia, A.V., Martin, J.D., and Gilliom, R.J., 2010, Trends in pesticide concentrations in urban streams in the United States, 1992–2008: U.S. Geological Survey Scientific Investigations Report 2010–5139, 101 p. (Also available at *http://pubs.usgs.gov/sir/2010/5139/*.)

Sullivan, D.J., Vecchia, A.V., Lorenz, D.L., Gilliom, R.J., and Martin, J.D., 2009, Trends in pesticide concentrations in corn-belt streams, 1996–2006: U.S. Geological Survey Scientific Investigations Report 2009–5132, 75 p. (Also available at *http://pubs.usgs.gov/sir/2009/5132/*.)

Vecchia, A.V., 2003, Relation between climate variability and stream water quality in the continental United States: Hydrological Science and Technology, v. 19, no. 1, p. 77–98.

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. Survey Scientific Investigations Report 2005–5224, 54 p. (Also available at *http://pubs.usgs.gov/sir/2005/5224/*.)

Vecchia, A.V., Gilliom, R.J., Sullivan, D.J., Lorenz, D.L., and Martin, J.D., 2009, Trends in concentrations and use of agricultural herbicides for Corn Belt Rivers, 1996–2006: Environmental Science and Technology, v. 43, p. 9,096–9,102. (Also available at *http://water.usgs.gov/nawqa/pubs/es902122j.pdf*.)

Vecchia, A.V., Martin, J.D., and Gilliom, R.J., 2008, Modeling variability and trends in pesticide concentrations in streams: Journal of the American Water Resources Association, v. 44, no. 5, p. 1,308–1,324. (Also available at *http://dx.doi.org/10.1111/j.1752-1688.2008.00225.x*.)

Venables, W.N., Smith, D.M., and the R Development Team, 2011, An Introduction to R, Version 2.13.1, 2011-07-08, accessed August 11, 2011, at *http://cran.r-project.org/doc/manuals/R-intro.pdf*.

Appendixes 1–2

Vignettes are the established R community method for providing examples of how to use the package.

Appendix 1. Vignette

The pdf file can be accessed at *http://pubs.usgs.gov/ofr/2012/1168/downloads/appendix1.pdf*.

Appendix 2. R Documentation

The pdf file can be accessed at *http://pubs.usgs.gov/ofr/2012/1168/downloads/appendix2.pdf*.

Publishing support provided by:
Rolla Publishing Service Center

For more information concerning this publication, contact:
Director, USGS North Dakota Water Science Center
821 East Interstate Avenue
Bismarck, North Dakota 58503
(701) 250–7400

Or visit the North Dakota Water Science Center Web site at:
http://nd.water.usgs.gov/