Appendix 1. Statistical Reanalysis of Medina Lake Stage Data and Groundwater Outflows from Medina/Diversion Lake System, San Antonio Area, Texas

Contents

Figures

Table

Suggested citation:
Asquith, W.H., and Slattery, R.N., 2017, Appendix 1. Statistical reanalysis of Medina Lake stage data and groundwater outflows from Medina/Diversion Lake system, San Antonio area, Texas, *in* Slattery, R.N., and Miller, L.D., 2017, A water-budget analysis of Medina and Diversion Lakes and the Medina/Diversion Lake system, with estimated recharge to Edwards aquifer, San Antonio area, Texas (ver. 1.1, February 2017): U.S. Geological Survey Scientific Investigations Report 2004–5209, 41 p., https://doi.org/10.3133/sir20045209.

Appendix 1. Statistical Reanalysis of Medina Lake Stage Data and Groundwater Outflows from Medina/Diversion Lake System, San Antonio Area, Texas

By William H. Asquith and Richard N. Slattery

Introduction

This appendix to Slattery and Miller (2017) provides a statistical reanalysis of Medina Lake stage data and groundwater outflows (GW_{out}) from Medina Lake and Diversion Lake (hereinafter referred to as the “Medina/Diversion Lake system”). The GW_{out} term was computed as the residual of all other terms from a detailed water-budget analysis (Slattery and Miller, 2017, eq. 6 and p. 6–14). The data used in the original water-budget analysis of the Medina/Diversion Lake system are depicted in Slattery and Miller (2017, fig. 11 and table 5) and are available in machine-readable form (Asquith and Slattery, 2016).

Different regression methods were tested in an effort to identify a regression equation that more accurately models the relation between GW_{out} from the Medina/Diversion Lake system and Medina Lake stage (water-surface elevation in feet [ft] above the National Geodetic Vertical Datum of 1929) than the original ordinary least-squares regression equation described in Slattery and Miller (2017, eq. 4 and fig. 11).

Statistical Reanalysis

The statistical reanalysis of Medina Lake stage and GW_{out} data (Slattery and Miller, 2017, fig. 11 and table 5) was done by using the R statistical software (version R-3.3.2) (R Core Team, 2016). Three regression equations to model the relation between Medina Lake stage and GW_{out} were evaluated: (1) a linear-linear ordinary least-squares (OLS) regression equation; (2) a linear-linear weighted least-squares (WLS) regression equation; and (3) a log-log weighted least-squares (log-log WLS) regression equation (Helsel and Hirsch, 2002). The three regression equations are depicted graphically (fig. 1–1). The abbreviations applicable to mathematical functions and statistical terms from the R regression output are provided (fig. 1–2), along with the R regression outputs of the three regression equations (figs. 1–3 through 1–5). The adjusted coefficient of determination (adjusted R-squared) is a useful diagnostic statistic for evaluating how well regression equations fit the data they model (fig. 1–1 and figs. 1–3 through 1–5; table 1–1). Unlike the unadjusted coefficient of determination (R-squared) that increases with each additional explanatory variable added to a regression equation, the adjusted R-squared increases when the addition of an explanatory variable improves the model more than would be expected by chance. As a result, the adjusted R-squared tends to be a better indicator of the predictive capability of the regression equation compared to the unadjusted R-squared (Helsel and Hirsch, 2002). The preferred regression equation was identified as the one with the highest adjusted R-squared value and lowest residual standard error. The 75-percent and 90-percent

Table 1–1. Summary of the regression residuals and regression-fit statistics for regression equations used to model the Medina/Diversion Lake system near San Antonio, Texas.

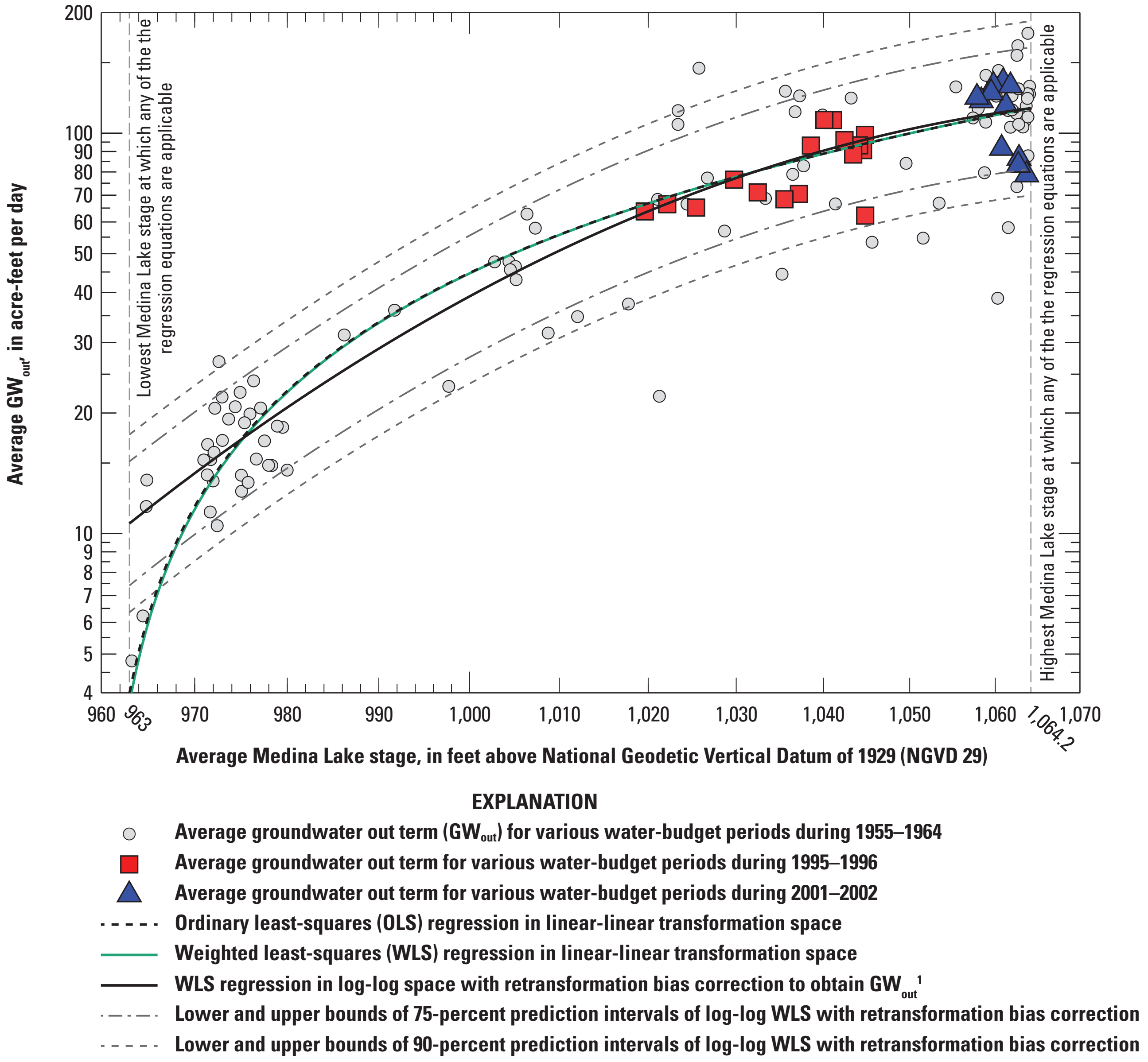
[R^2, coefficient of determination; Adj R^2, adjusted coefficient of determination; RSE, residual standard error; DF, degrees of freedom; p-value, probability value; OLS, ordinary least-squares regression; WLS, weighted least-squares regression; acre-ft/d, acre-feet per day; <, less than]

Regression equation	Regression residuals[1]					Regression fit statistics[1]					
	Minimum	First quartile	Median	Third quartile	Maximum	R^2	Adj R^2	RSE	F-statistic	DF	p-value
	Units in acre-feet per day										
Linear-Linear OLS	–72.49	–7.37	0.54	7.32	72.15	0.76	0.76	21.8 acre-ft/d	405.1	1 and 125	<0.001
Linear-Linear WLS	–69.39	–7.61	0.71	8.09	77.77	0.77	0.77	21.9 acre-ft/d	415.5	1 and 125	<0.001
	Units in logarithms										
Log-Log WLS[2]	–0.47	–0.07	0.02	0.08	0.35	0.88	0.88	[3]0.13	460.3	2 and 124	<0.001

[1] Summary of rounded values from original output produced by the R statistical software (R Core Team, 2016) (figs. 1–3 through 1–5).

[2] Residuals for log-log WLS cannot be directly retransformed into units of acre-feet per day.

[3] Residual standard error for log-log WLS is unitless.



[1]PREFERRED REGRESSION EQUATION

$${}_{\delta}GW_{out} = \delta \times 10^{\{2{,}413.26\log_{10}(ML) - 397.52[\log_{10}(ML)]^2 - 3{,}660.51\}}$$

$${}_{\delta}GW_{out} = \text{groundwater outflow from Medina/Diversion Lake system, in acre-feet per day}$$

$$ML = \text{Medina Lake stage, in feet above NGVD 29}: \; 963 \leq ML \leq 1{,}064.2 \text{ feet}$$

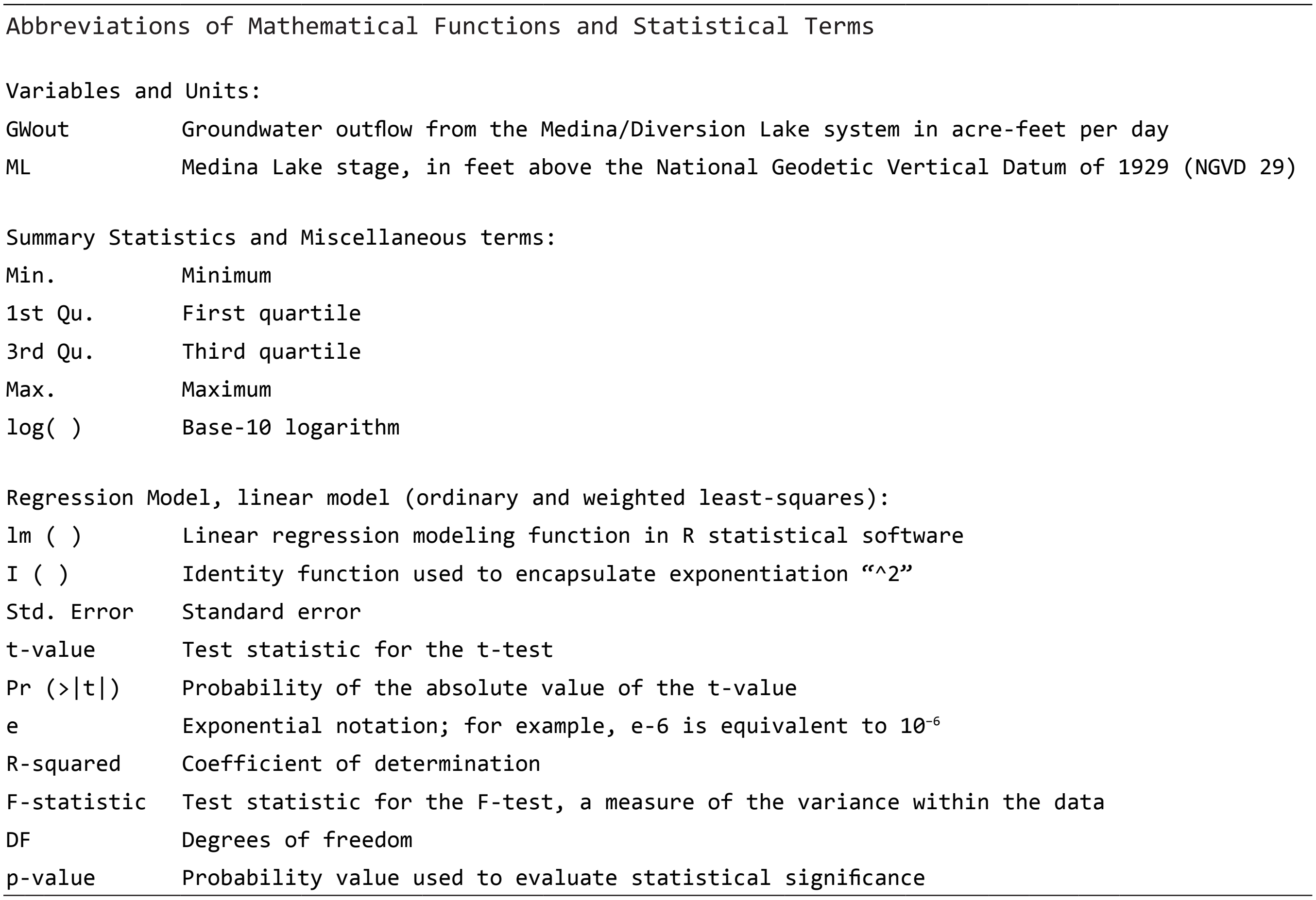
$$\delta = \text{bias correction factor}$$

$$\delta = \begin{cases} 1 & \text{if median } GW_{out} \text{ is desired} \\ 1.042 & \text{if mean } GW_{out} \text{ is desired} \end{cases}$$

$$\text{adjRsq} = 0.88 \quad \text{adjusted R-squared}$$

$$\text{RSE} = 0.13 \quad \text{residual standard error (logarithmic, base-10 [}\log_{10}\text{])}$$

Figure 1–1. Relation between average Medina Lake stage and average groundwater outflow (GW$_{out}$) for various water-budget periods during 1955–1964, 1995–1996, and 2001–2002 (table 5 in Slattery and Miller, 2017) when Medina Lake water-budget data were collected from the Medina/Diversion Lake system modeled with *A,* a linear-linear ordinary least-squares (OLS) regression equation; *B,* a linear-linear weighted least-squares (WLS) regression equation; and *C,* a log-log weighted least-squares (log-log WLS) regression equation, and prediction intervals for the preferred regression equation.



Abbreviations of Mathematical Functions and Statistical Terms

Variables and Units:

GWout	Groundwater outflow from the Medina/Diversion Lake system in acre-feet per day
ML	Medina Lake stage, in feet above the National Geodetic Vertical Datum of 1929 (NGVD 29)

Summary Statistics and Miscellaneous terms:

Min.	Minimum
1st Qu.	First quartile
3rd Qu.	Third quartile
Max.	Maximum
log( )	Base-10 logarithm

Regression Model, linear model (ordinary and weighted least-squares):

lm ( )	Linear regression modeling function in R statistical software
I ( )	Identity function used to encapsulate exponentiation “^2”
Std. Error	Standard error
t-value	Test statistic for the t-test
Pr (>|t|)	Probability of the absolute value of the t-value
e	Exponential notation; for example, e-6 is equivalent to 10^{-6}
R-squared	Coefficient of determination
F-statistic	Test statistic for the F-test, a measure of the variance within the data
DF	Degrees of freedom
p-value	Probability value used to evaluate statistical significance

Figure 1–2. Abbreviations of mathematical functions and statistical terms related to output from the R statistical software (R Core Team, 2016) for the regression reanalysis shown in figures 1–3 through 1–5; detailed discussion of these technical terms is available in Helsel and Hirsch (2002).

prediction intervals associated with the preferred regression equation also were determined. Faraway (2005, 2006) and Helsel and Hirsch (2002) provide detailed descriptions of the methods used to derive the regression equations and prediction intervals contained herein.

The average Medina Lake stage and average GW_{out} data (referred to as mean GW_{out} in Slattery and Miller [2017, table 5]) are depicted for various water-budget periods (table 5 in Slattery and Miller, 2017) when Medina Lake water-budget data were collected during 1955–1964, 1995–1996, and 2001–2002 (fig. 1–1). The range of average Medina Lake stage and GW_{out} data from 1955–1964 is much larger compared to the range of 1995–1996 and 2001–2002 data. For the statistical reanalysis, data from the three time periods are grouped into a single dataset, and results of this reanalysis are based on the entire range of data.

Linear-Linear Ordinary Least-Squares

The regression equation shown in Slattery and Miller (2017, eq. 4, fig. 11) was an OLS regression in linear-linear transformation space. For comparison purposes with alternative regression equations, the OLS regression described in Slattery and Miller (2017) was reproduced by using the R statistical software (figs. 1–1 and 1–3). The reproduced OLS regression differs slightly from the original OLS regression because of subtle computational differences in the software used to make the original and reproduced versions. The OLS regression equation as reproduced in figures 1–1 and 1–3 has an intercept of about –1,057 ft, a slope of about 1.102, an adjusted R-squared value of about 0.76, and a residual standard error of about 21.8 acre-feet per day (table 1–1).

Linear-Linear Weighted Least-Squares

Because the average GW_{out} data values are based on water-budget periods of unequal sizes with unequal variability (Slattery and Miller, 2017, table 5), the use of a WLS method of regression analysis could potentially explain more of the variability in the data compared to the OLS regression (Helsel and Hirsch, 2002). To compute a weighted least-squares regression, each of the GW_{out} data points was multiplied by a “weight factor” for use in the WLS regression. The weight factor was determined as the number of days in the water-budget period (sample size) (Slattery and Miller, 2017, table 5; Asquith and Slattery, 2016) divided by the coefficient of variation. The coefficient of variation is the standard deviation of the GW_{out} for each water-budget period divided by the average GW_{out} for each water-budget period (Slattery and Miller, 2017, table 5). As a result, GW_{out} values with greater variability and smaller sample sizes were assigned smaller weight factors, whereas GW_{out} values with lesser variability and larger sample size were assigned larger weight factors (Helsel and Hirsch, 2002).

Once the weight factors were assigned, a WLS regression equation representing the relation between Medina Lake stage and average GW_{out} was developed. The results of the WLS regression (green line, fig. 1–1; fig. 1–4) are similar to the results of the OLS regression (black dashed line, fig. 1–1; fig. 1–3). The intercept determined from WLS regression is about –1,060 ft, the slope is about 1.105, the adjusted R-squared value is about 0.77, and the residual standard error is about 21.9 acre-feet per day (table 1–1; fig. 1–4). Although the WLS regression equation is similar to the OLS regression equation, the incorporation of weight factors represents a more robust method of statistical analysis (Helsel and Hirsch, 2002, p. 248).

Log-Log Weighted Least-Squares

Next, different log-log WLS regression analyses were developed. Base-10 logarithmic transformations of Medina Lake stage and GW_{out} were calculated, and different log-log WLS regression equations were developed by using the weight factors described in the “Linear-Linear Weighted Least-Squares” section of this appendix. Preliminary computations in R with the Medina Lake stage and GW_{out} data were done by using generalized additive modeling (GAM) published by Wood (2016a, b). The depicted log-log WLS regression equation (black line, fig. 1–1) includes a quadratic term based on Medina Lake stage (a squared value of a predictor variable of lake stage is part of the equation). The addition of this squared predictor variable was done following interpretation of the preliminary GAM regression methods (Wood, 2006). The resulting log-log WLS regression equation represented the visual curvature of the data better than either the linear-linear WLS regression equation or the OLS regression equation (fig. 1–1). The adjusted R-squared value of about 0.88 for the log-log WLS regression equation (fig. 1–5) is notably larger than the adjusted R-squared value of about 0.77 for the linear-linear

```
Call:
lm(formula = GWout ~ ML)
Residuals:
   Min.  1st Qu.  Median  3rd Qu.     Max.
-72.487   -7.371   0.538    7.315   72.147
Coefficients:
               Estimate  Std.Error  t-value  Pr(>|t|)
(Intercept) -1.057e+03  5.614e+01   -18.83    <2e-16
ML           1.102e+00  5.475e-02    20.13    <2e-16
---
Residual standard error: 21.79 on 125 degrees of freedom
Multiple R-squared:  0.7642,   Adjusted R-squared:  0.7623
F-statistic: 405.1 on 1 and 125 DF,  p-value: < 2.2e-16
```

Figure 1–3. Ordinary least-squares regression dependent on untransformed data as produced in output by the R statistical software (Linear-Linear OLS) (R Core Team, 2016) using data from Slattery and Miller (2017, fig. 11 and table 5). Residual standard error is measured in acre-feet per day.

```
Call:
lm(formula = GWout ~ ML, weights = W)
Weighted Residuals:
   Min.  1st Qu.  Median  3rd Qu.    Max.
-69.389   -7.609   0.710    8.093  77.773
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.060e+03  5.563e+01  -19.06   <2e-16
ML           1.105e+00  5.419e-02   20.38   <2e-16
---
Residual standard error: 21.9 on 125 degrees of freedom
Multiple R-squared:  0.7687,   Adjusted R-squared:  0.7669
F-statistic: 415.5 on 1 and 125 DF,  p-value: < 2.2e-16
```

Figure 1–4. Weighted least-squares regression dependent on untransformed data as produced by the R statistical software (Linear-Linear WLS) (R Core Team, 2016) using data from Slattery and Miller (2017, fig. 11 and table 5). Residual standard error is measured in acre-feet per day.

WLS, indicating that the log-log WLS explains more variability than the linear-linear WLS, and thus is likely a preferable expression of the relation between Medina Lake stage and GW_{out}.

The retransformation of GW_{out}, from the log-log WLS regression equation back into a linear-linear space, represents the median estimate of GW_{out} (fig. 1–1). To obtain the average GW_{out}, a retransformation bias-correction factor is needed. The Duan smearing estimator (Helsel and Hirsch, 2002, p. 256–257) was used to estimate a retransformation bias-correction factor. The bias-correction factor (δ) was computed as the mean of the retransformed residuals of the log-log WLS regression and is applied as a coefficient to the log-log WLS regression equation (fig. 1–1). The bias-correction factor, estimated as 1.042, indicates that an uncorrected retransformed prediction from the log-log WLS regression equation (figs. 1–1 and 1–5) will underestimate the GW_{out} by about 4 percent. The prediction intervals for the log-log WLS regression equation (fig. 1–1) were also corrected by using the Duan smearing factor.

Comparison of Statistical Reanalysis Methods

A visual comparison of the linear-linear WLS and log-log WLS regression equations (fig. 1–1) indicates that during Medina Lake stages greater than about 1,030 ft, both regression equations predict similar GW_{out} values. The linear-linear WLS and log-log WLS regression equations diverge between about 976 and 1,030 ft of Medina Lake stage; within this range, the log-log WLS regression estimates less GW_{out} compared to the linear-linear regression. The greatest relative differences between the linear-linear WLS and log-log WLS regression equations are evident for lake stages ranging from about 963 to 970 ft; within this range, the log-log WLS regression estimates more GW_{out} than the linear-linear WLS. The distribution of the substantial number of data points in the range of 972–980 ft tends to support the use of the log-log WLS regression equation. The log-log WLS regression equation is applicable for Medina Lake stages between 963 and 1,064.2 ft—about the same as the minimum (963.27 ft) and maximum (1,064.04 ft) stages represented in the data. The stage value of 1,064.2 ft also represents the approximate altitude of the Medina Lake spillway.

Preferred Regression Equation

The preferred regression equation for defining the relation between Medina Lake stage and GW_{out} is the depicted log-log WLS (figs. 1–1 and 1–5); of the regression equations that were evaluated, it has the highest adjusted R-squared value (about 0.88) and the lowest residual standard error (about 0.13). The preferred regression equation is depicted graphically and in algebraic form (fig. 1–1). Also shown in figure 1–1 are the 75-percent and 90-percent prediction intervals for the preferred regression equation. For a given new stage value, the 75-percent and 90-percent prediction intervals represent the upper and lower range of probable GW_{out} values where the actual GW_{out} is predicted to be within the specified prediction interval either 75 or 90 percent of the time (Helsel and Hirsch, 2002).

Uncertainties not addressed by the statistical reanalysis may be associated with the data themselves. For example, few data are available for certain ranges of Medina Lake stage, particularly for stages less than 971 ft, and for stages between 982 and 1,016 ft and between 1,048 and 1,057 ft; the lack of data for certain ranges of stage is more pronounced for the more recent data collected during 1995–1996 and 2001–2002 when relatively little variation in lake stage was recorded compared to the range in stage recorded during 1955–1964 (fig. 1–1). Differences in data collection techniques are an additional source of data uncertainty. The water-budget data from earlier times (1955–1964) were collected using different techniques compared to the more recent data, so the accuracy of the GW_{out} data may have varied over time. For example, to compute evaporation, the earlier period (1955–1964) relied on Texas Water Development Board (2016) evaporation tables, whereas the more recent data from 1995–1996 and 2001–2002 relied on USGS-operated meteorological stations temporarily installed on the lake.

Data from 1995–1996 and 2001–2002 show a similar distribution compared to data from 1955–1964 for the same

```
Call:
lm(formula = log10(GWout) ~ log10(ML) + I(log10(ML)^2), weights = W)
Weighted Residuals:
    Min.   1st Qu.   Median  3rd Qu.     Max.
-0.47010  -0.06519  0.02120  0.07690  0.35067
Coefficients:
               Estimate  Std.Error  t-value  Pr(>|t|)
(Intercept     -3660.51     703.84   -5.201  7.95e-07
log10(ML)       2413.26     468.13    5.155  9.72e-07
I(log10(ML)^2)  -397.52      77.84   -5.107  1.20e-06
---
Residual standard error: 0.1298 on 124 degrees of freedom
Multiple R-squared:  0.8813,   Adjusted R-squared:  0.8794
F-statistic: 460.3 on 2 and 124 DF,  p-value: < 2.2e-16
```

Figure 1–5. Weighted least-squares regression dependent on logarithmically transformed data as produced in output by the R statistical software (log-log WLS) (R Core Team, 2016) using data from Slattery and Miller (2017, fig. 11 and table 5).

lake stages. It is reasonable, therefore, to assume that the more recent water-budget analyses based on the more recent data would produce results for lower stages (less than about 1,030 ft), consistent with the results produced using the data from 1955–1964.

Additional sources of uncertainty that were not addressed in this regression reanalysis include the possibilities of serially correlated data, unaccounted-for antecedent hydrological conditions, and differences in hydrogeology at different lake stages that might affect GW_{out}. Serial correlation is the dependence or correlation between residuals for values collected in a time series (Helsel and Hirsch, 2002). For Medina Lake stage data, the possibility of serial correlation means that stage values collected consecutively over time might not be independent values. Instead of varying independently, the data might change similarly in response to monthly and annual scale changes in hydrometeorological processes, and are therefore not randomly distributed in time (Helsel and Hirsch, 2002). Accounting for the effects of serial correlation would be difficult and would require more data than available in this study. The antecedent hydrological conditions in the area surrounding the Medina/Diversion Lake system are also not represented in the regression equation. Also not specifically accounted for in the regression equation are differences in hydrogeology that might affect GW_{out} at different stages in either Medina or Diversion Lake (Lambert and others, 2000).

Digital Files

Digital files available in Asquith and Slattery (2016) include:

- `MedinaLakeGWoutReanalysis.R` (the R script file);
- `MedinaLakeGWoutReanalysis.csv` (the R input data file); and
- `MedinaLakeGWoutReanalysis.xml` (the metadata).

The file `MedinaLakeGWoutReanalysis.R` is a ".R" script file compatible with at least version R-3.3.2 of the R programming language (R Core Team, 2016). The file `MedinaLakeGWoutReanalysis.csv` is a plain-text, comma-delimited file that contains all the data used in this analysis. The file `MedinaLakeGWoutReanalysis.xml` contains information designed to assist users in the understanding and reuse of the `MedinaLakeGWoutReanalysis.csv` file by providing details about where the data were collected and how they were used. The .R script file and the .csv data file are in the Unicode (UTF-8) encoding with a carriage-return, new-line character line-ending convention to facilitate use across computer operating systems. Executing the .R script file in the R programming language processes the data from the .csv file and reproduces the regression equations depicted in figure 1–1 and the statistical outputs depicted in figures 1–3 through 1–5. A general operational knowledge of the R statistical software language is required for additional data manipulation and analysis.

References Cited

Asquith, W.A., and Slattery, R.N., 2016, Data and R script pertaining to Medina Lake stage and groundwater outflow, Medina and Diversion Lake system near San Antonio, Texas, 2016: U.S. Geological Survey data release, https://doi.org/10.5066/F7ZS2TNF.

Faraway, J.J., 2005, Linear models with R: Boca Raton, Fla., Chapman and Hall, CRC press, 240 p.

Faraway, J.J., 2006, Extending the linear model with R—Generalized linear, mixed effects and nonparametric regression models: Boca Raton, Fla., Chapman and Hall, CRC press, 331 p.

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, accessed November 4, 2016, at http://pubs.usgs.gov/twri/twri4a3.

Lambert, R.B., Grim, K.C., and Lee, R.W., 2000, Hydrology, hydrologic budget, and water chemistry of the Medina Lake area, Texas: U.S. Geological Survey Water-Resources Investigations Report 00–4148, 53 p.

R Core Team, 2016, R—A language and environment for statistical computing: Vienna, Austria, R Foundation for Statistical Computing, accessed November 4, 2016, at http://www.r-project.org/.

Slattery, R.N., and Miller, L.D., 2017, A water-budget analysis of Medina and Diversion Lakes and the Medina/Diversion Lake system, with estimated recharge to Edwards aquifer, San Antonio area, Texas (ver. 1.1, February 2017): U.S. Geological Survey Scientific Investigations Report 2004–5209, 41 p., https://doi.org/10.3133/sir20045209.

Texas Water Development Board, 2016, Precipitation and lake evaporation—Quadrangle 809, accessed November 14, 2016, at http://www.twdb.texas.gov/surfacewater/conditions/evaporation/.

Wood, S.N., 2006, Generalized additive models—An introduction with R: Boca Raton, Fla., Chapman and Hall, CRC Press, 392 p.

Wood, S.N., 2016a, Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation: R package version 1.8-12, accessed November 4, 2016, at http://cran.r-project.org/package=mgcv.

Wood, S.N., 2016b, Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation, accessed November 4, 2016, at https://cran.r-project.org/web/packages/mgcv/mgcv.pdf.