Appendix I
Hydrologic Gap-Filling Technical Memo

1.0 STAGE AND FLOW DATA

The United States Geological Survey (USGS) maintains six gages along the Middle Suwannee River study reach (MSR) (**Table 1**). The gages at Ellaville, Luraville, Branford, Bell, and Wilcox were the focus of the hydrologic data analysis. Data from the Dowling Park gaging station was excluded from use based on its more recent period of record and its placement mainly in response to local flooding issues (John Good, personal communication). **Table 2** shows the percent availability of flow and stage data for the five gages of interest. Based on the availability of USGS approved data presented in **Tables 1** and **2** it was determined that October 1, 1932 thru September 30, 2015 (Water Year, WY 1933 to WY 2015) would be used as the period of record for all analyses. Missing data (primarily for Bell and Luraville) were backfilled using appropriate gap filling methodologies, as described below, and the long-term (WY1933 to WY2015) gap-filled flow time-series were subsequently used to develop reference timeframe (RTF) flow time-series for the USGS gages of interest along MSR (except Dowling Park).

Table 1 - USGS Gage Information along Middle Suwannee River

Site Name	Gage Number	Latitude	Longitude	County	Drainage Area (sq. mi)	Period of Record
SUWANNEE RIVER AT ELLAVILLE	02319500	30.38466	-83.1718	Suwannee	6,970	2/1/1927 - current
SUWANNEE RIVER AT DOWLING PARK	02319800	30.24494	-83.2496	Lafayette	7,190	10/1/1996 - current
SUWANNEE RIVER AT LURAVILLE	02320000	30.09995	-83.1715	Lafayette	7,280	2/1/1927-12/31/1937 9/27/1996 - current
SUWANNEE RIVER AT BRANFORD	02320500	29.95579	-82.9276	Suwannee	7,880	7/1/1931 - current
SUWANNEE RIVER NEAR BELL	02323000	29.79134	-82.9243	Gilchrist	9,390	6/1/1932 - 12/31/1956 8/4/2000 - current
SUWANNEE RIVER NEAR WILCOX	02323500	29.58968	-82.9365	Levy	9,640	10/1/1930 - 9/30/1931 10/1/1941 - current

Note: Period of record is for daily stage and discharge; current is as of report writing time (March 2021)

Table 2 - Summary of Available Data at Gages of Interest along Middle Suwannee River (WY 1933 thru WY 2015)

USGS Gage Station ID	USGS Gage Name	Data Type	Number of Records Available	Number of Records Missing	Percent Data Available
02319500	Ellaville*	Stage	29,878	437	98.6%
		Flow	30,315	0	100.0%
02319800	Dowling Park	Stage	6,707	23,608	22.1%
		Flow	6,937	23,378	22.9%
02320000	Luraville	Stage	8,752	21,563	28.9%
		Flow	8,861	21,454	29.2%
02320500	Branford*	Stage	30,194	121	99.6%
		Flow	30,315	0	100.0%
02323000	Bell	Stage	13,821	16,494	45.6%
		Flow	15,237	15,078	50.3%
02323500	Wilcox**	Stage	24,473	5,842	80.7%
		Flow	27,028	3,287	89.2%

*MFL compliance gage, ** MFL rule previously completed as part of the Lower Suwannee

2.0 DATA REVIEW AND GAP-FILLING

2.1 Gap-Filling Stage Data

The first step in the analysis was to fill data gaps in the stage time-series at the five stations of interest (**Table 2**). The patterns of data gaps were analyzed to determine the best methodology for gap-filling. In general, shorter gaps (~days to weeks) present less uncertainty in the analysis, allowing for comparatively simple methods. Large (~2 month or larger) gaps were, by definition, ignored in the short-gap filling process. **Figure 1** shows the frequency and duration of short gaps in stage data for the five stations. The majority of short gaps in the stage time series were less than 15 days long, and very few gaps existed between 15 and 60 days. Linear interpolation was used to backfill gaps less than 15 days long and a spline interpolation method was applied to gaps between 15 and 60 days. For all larger data gaps, a multiple imputation technique was used (described later).

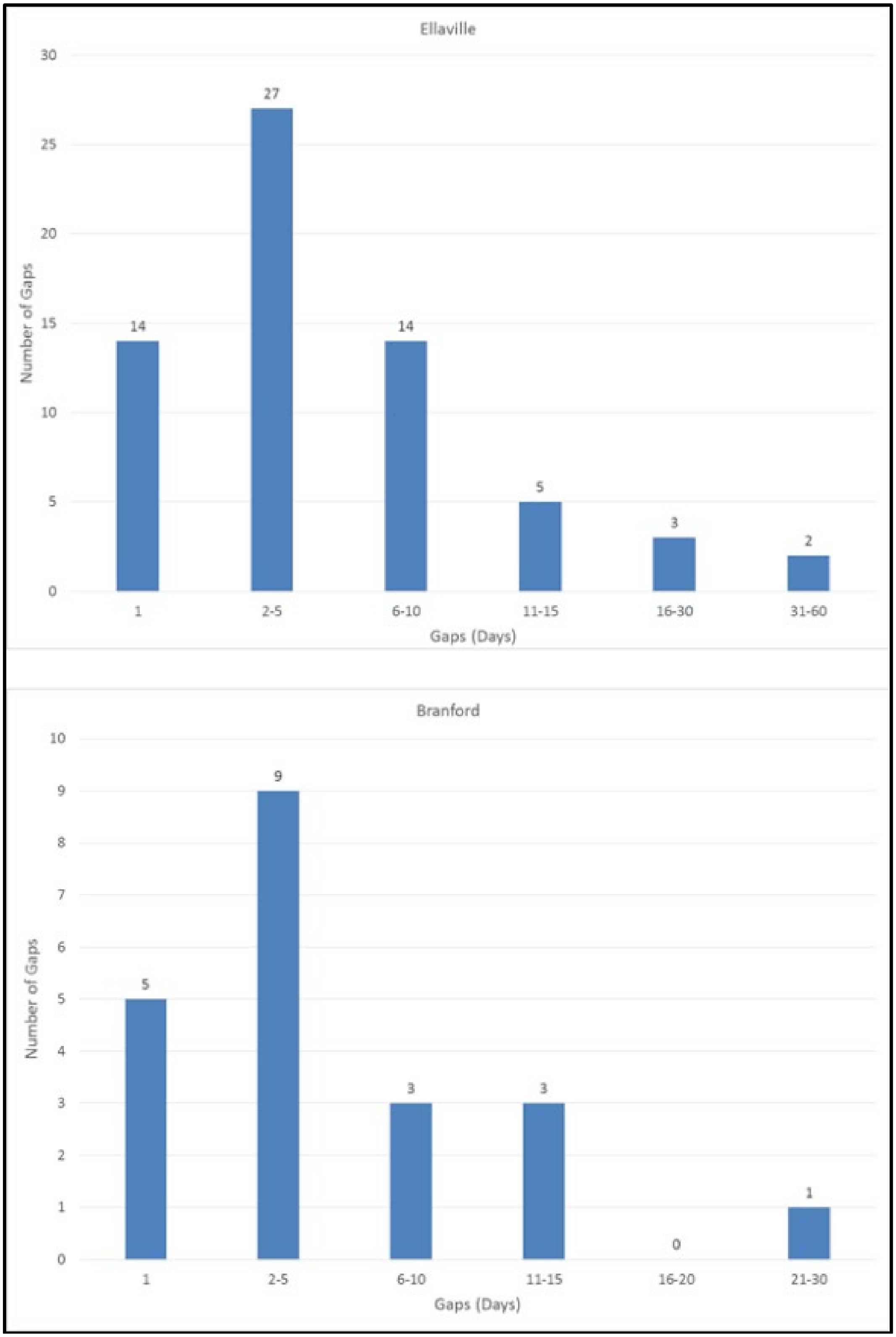

Figure 1(a) - Frequency and Duration of Data Short Gaps (< 60 days) in Stage Data (Ellaville and Luraville)

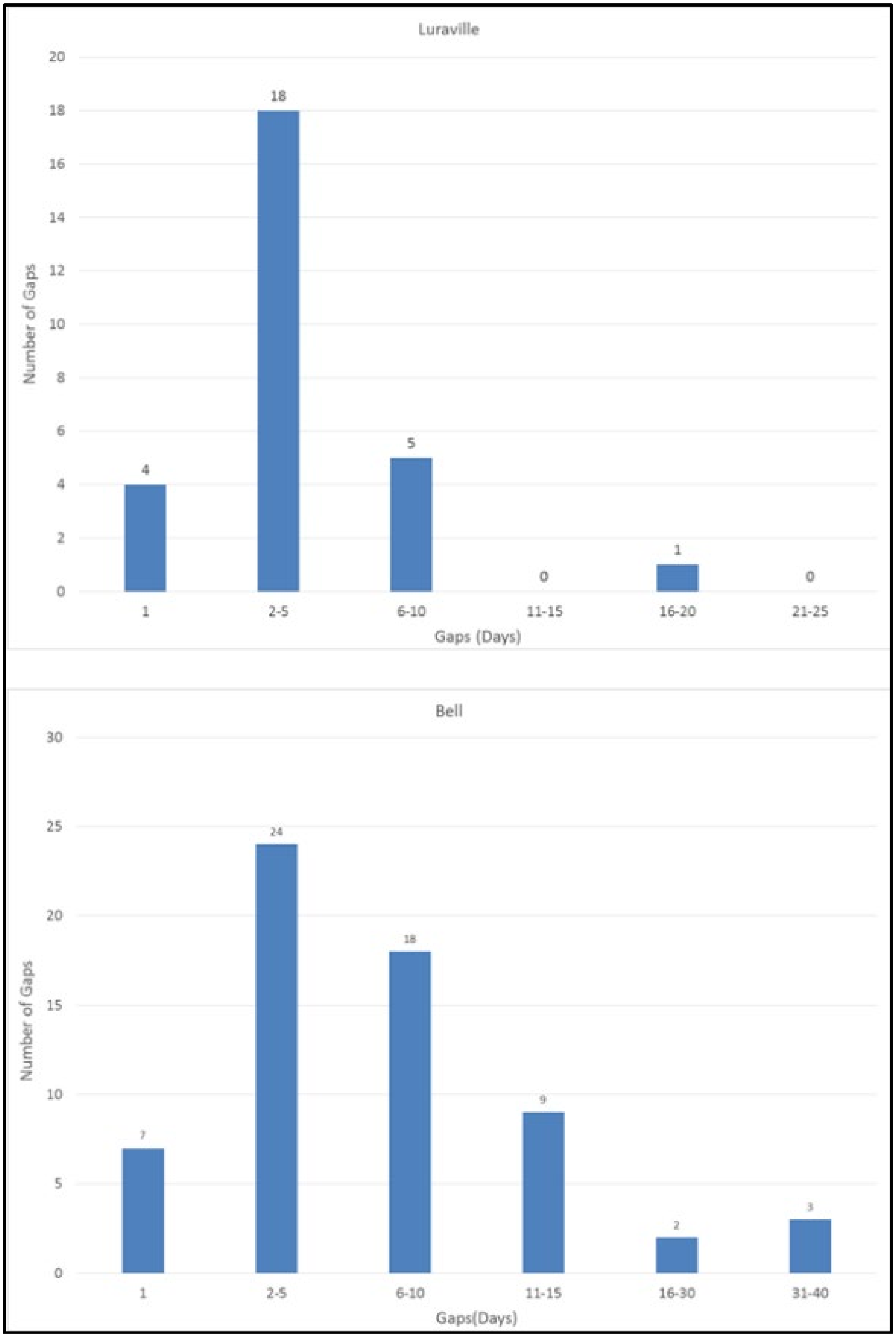

Figure 1(b) - Frequency and Duration of Data Short Gaps (< 60 days) in Stage Data (Branford and Bell)

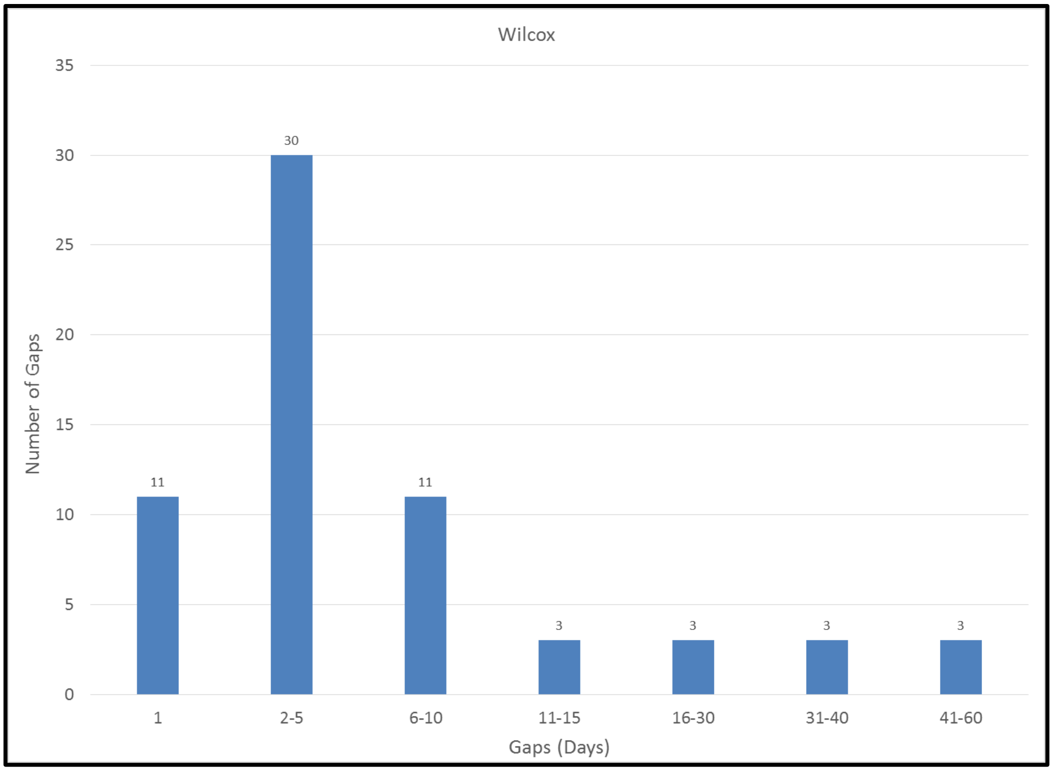

Figure 1(c) – Frequency and Duration of Data Short Gaps (< 60 days) in Stage Data (Wilcox)

2.2 Gap-Filling Flow Data

As with the stage data, the first step for gap analysis of flow data involved screening and removing periods with at least a year of data gap. Luraville and Bell were the only two gages found to have long data gaps (at least a year in length) in the flow time-series. The Ellaville and Branford flow records had no flow data gaps in the period of record from WY1933 thru WY2015.

Luraville had no short-term flow data gaps, but flow records at Wilcox and Bell exhibited short-term data gaps. The largest short-term data-gap for these two gages occurred at Bell, spanning 6 days. Thus, a linear interpolation method was used to backfill short-term flow data gaps at Bell and Wilcox. A multiple imputation technique was used for longer data gaps at Luraville and Bell.

2.3 Multiple Imputation

Multiple Imputation (MI) methodology involves backfilling of missing data with a set of plausible values and then developing inferential statistics based on the "pooled" data (Rubin 1987). The first step in MI involves developing *a priori* Bayesian (probability based) relationships between available datasets that are parameterized using associations from neighboring stations having data from similar spans when the gap-filled station also had data. Subsequently, the MI method uses Markov chain Monte Carlo techniques to replace the missing values with several data "chains" (or sets of time-series) of plausible values derived from the Bayesian relationships developed earlier. The individual chains of data are subsequently pooled together to form a data cloud (set of several time-series) from which estimates of best-fit backfilled values can be developed (Su et al 2011). Significantly high degree of efficiency in the imputation process can be achieved by relatively smaller number of chains (3-10) (Rubin 1987). For the current study six data chains were developed for backfill analysis.

Values from upstream and downstream gages were used with the available concurrent records in the existing time-series from the station of interest to backfill the missing stage and flow data (data gaps greater than 60 days). Data from Ellaville and Branford were used for the Luraville gage. The Branford and Wilcox gages were used for gap-filling Bell data. Imputations for stage and flow data were conducted independent of each other (i.e. the flow dataset was used to impute flow records and the stage dataset was use to impute stage records) except for Wilcox. For the Wilcox gage, stage imputations were made using flow data at Wilcox and stage data from Bell. The methodology used to implement multiple imputation process follows the step-by-step process described in Goodrich and Kropko (2014). **Figure 2** shows an example of imputed data-chains output from the MI process for a selected date range at the Ellaville gage. Every chain (shown by points) represents a statistically plausible backfilled time-series for the Ellaville gage. Data from all chains are pooled together to create a large dataset and a multiple regression model was developed to estimate the flow values for the corresponding data gaps. The gages used for the multiple regression model were the same gages that were used for conducting MI.

Figures 3 and **4** show the Luraville and Bell gage backfilled datasets. Spot checks were conducted on the backfilled dataset to determine the quality of backfilling process. **Figure 5 (a** and **b)** shows representative examples of backfilled flow at the Bell gage for a missing period of record during WY 2006, and the backfilled stage data for the Wilcox gage. These results illustrate that the backfilling process provided reasonable results and were able to capture anticipated fluctuations and magnitude. To further investigate the method's effectiveness, a verification test on Luraville stage and flow data was conducted. Observed flow and stage values from 1/1/2000 thru 12/31/2004 (5-year period of record) were removed from the flow and stage time-series and MI methodology was applied to simulate the flow and stage data for the removed period of record.

Figure 6 (a and **b)** shows the observed and backfilled stage and flow time-series from the MI methodology. From **Figure 6 (a** and **b)** it is evident that the MI method was able to successfully simulate the observed flow records with a high degree of accuracy, thereby indicating the usefulness of the MI method.

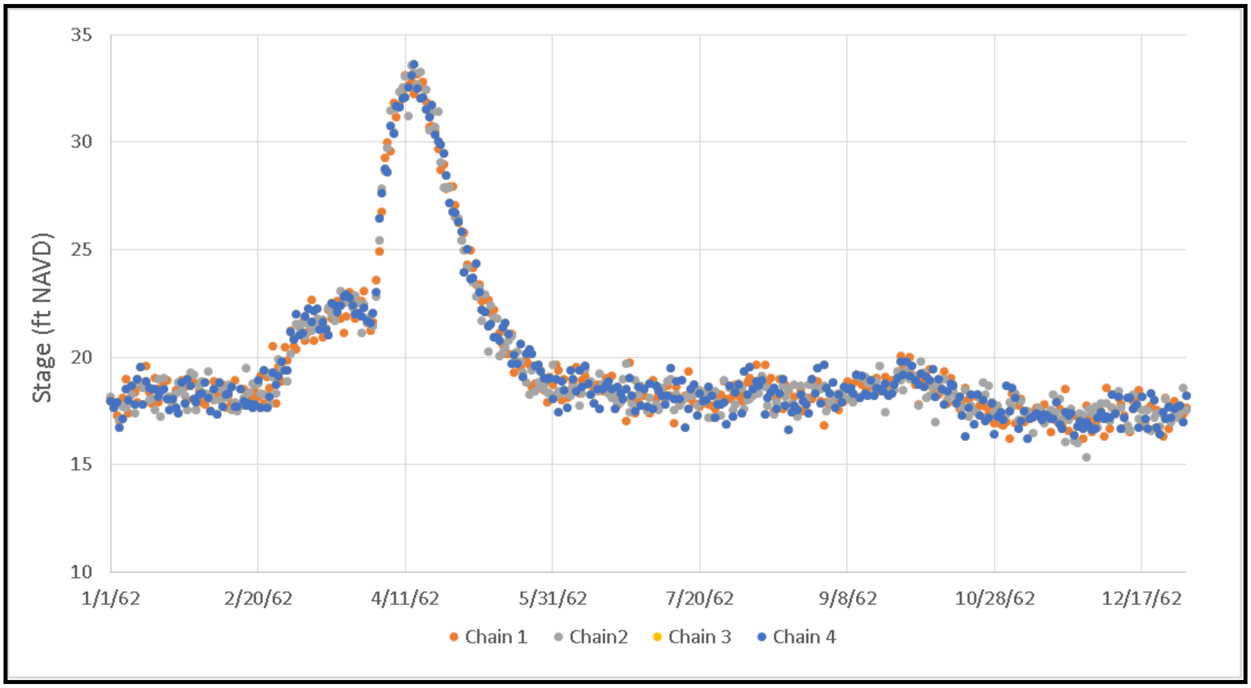

Figure 2 - Sample Imputed Dataset Chains for a Selected Period of Record at Ellaville Gage

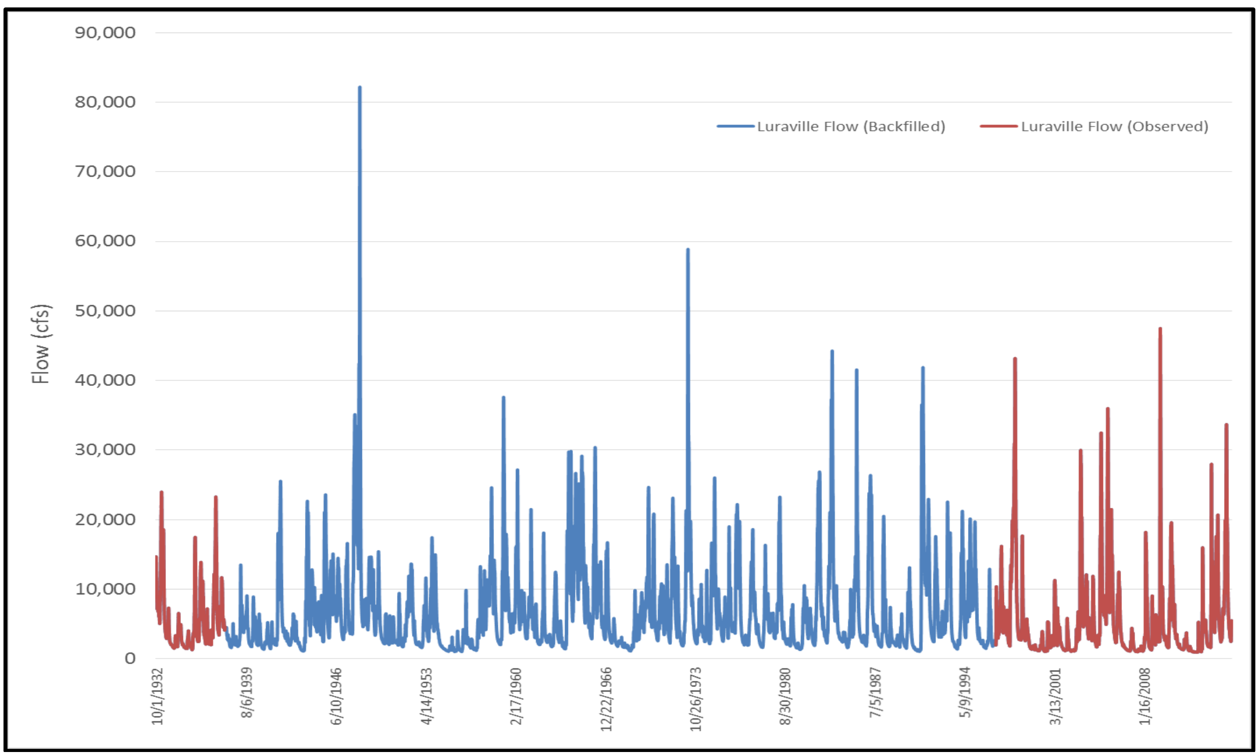

Figure 3 - Backfilled Dataset for Luraville Gage Plotted along with Observed Data

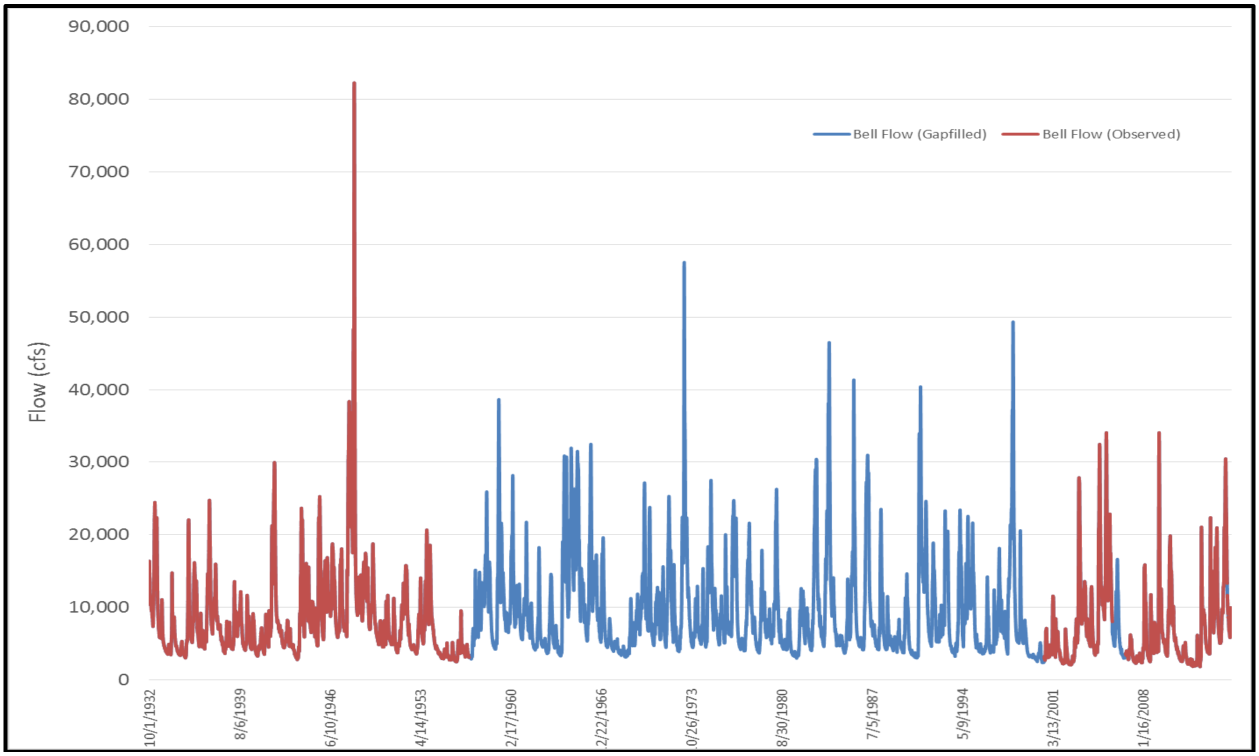

Figure 4 - Backfilled Dataset for Bell Gage Plotted along with Observed Data

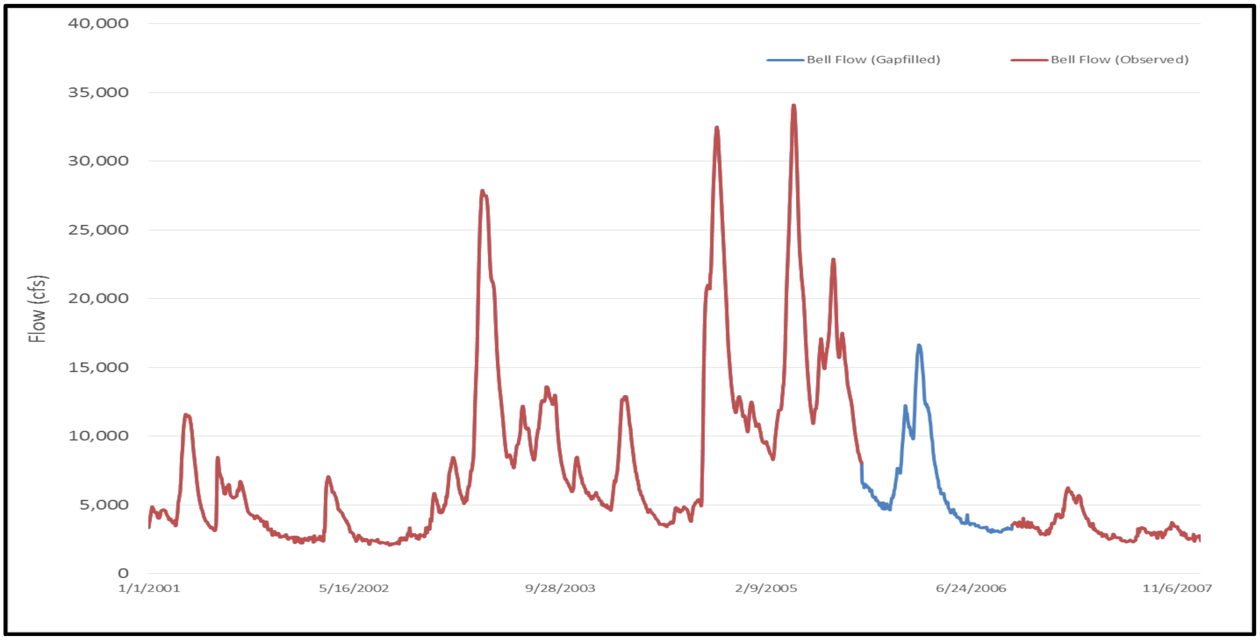

Figure 5(a) - Selected Backfilled Flow Dataset for Bell Gage Plotted along with Observed Data

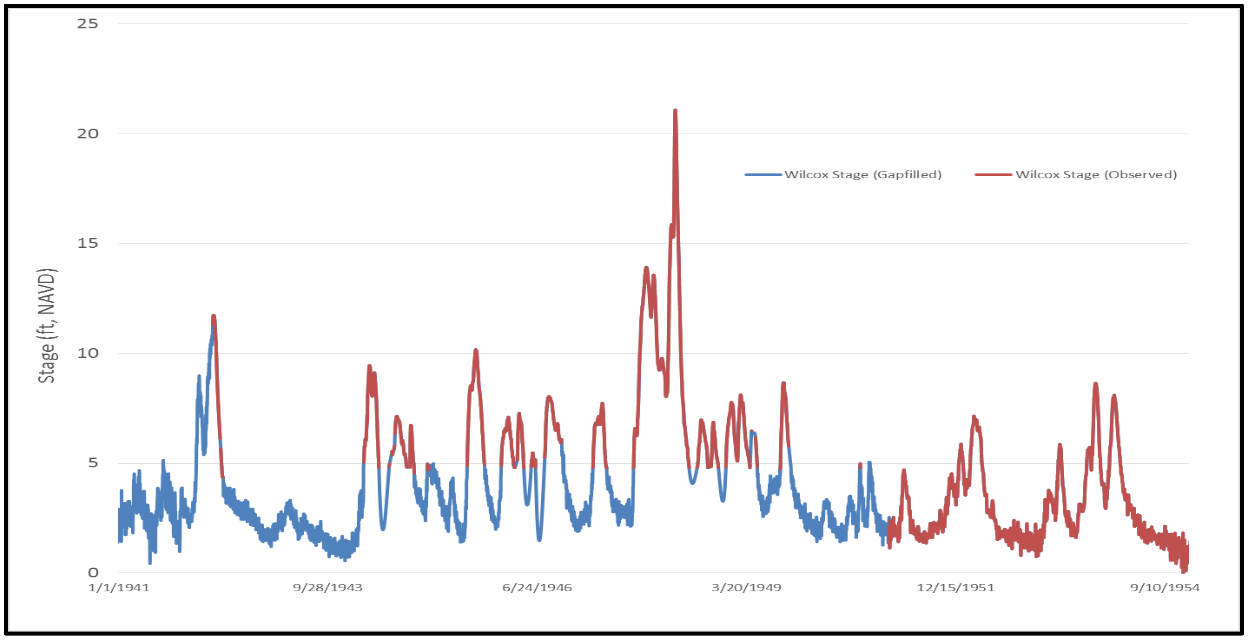

Figure 5(b) - Selected Backfilled Stage Dataset for Wilcox Gage Plotted along with Observed Data

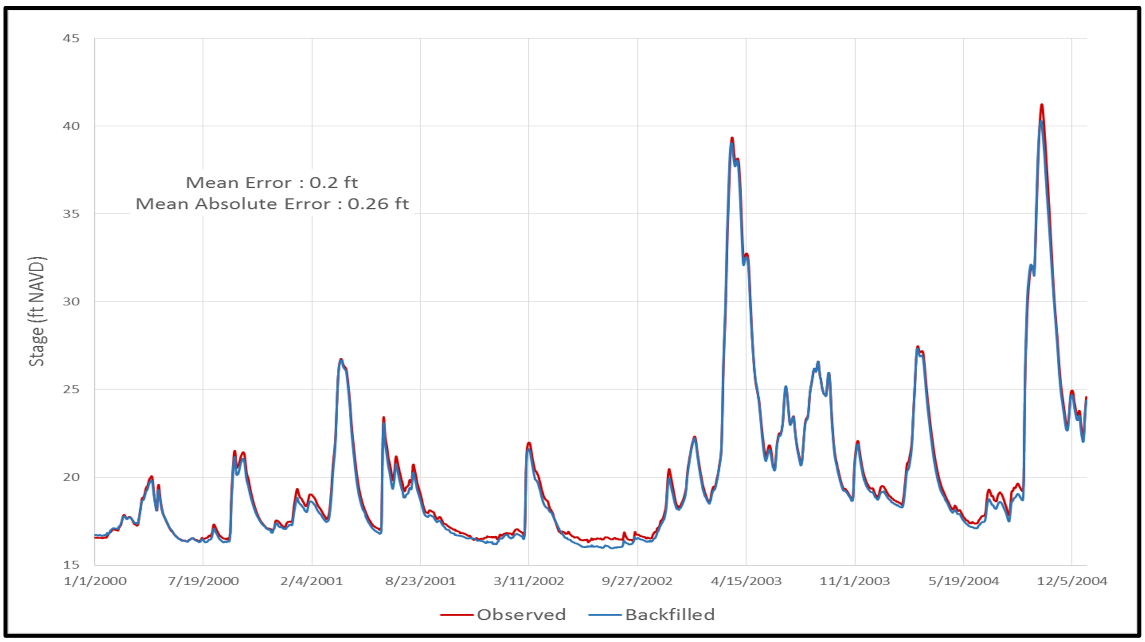

Figure 6(a) – Verification of Backfilled versus Observed Stage Data at Luraville Gage

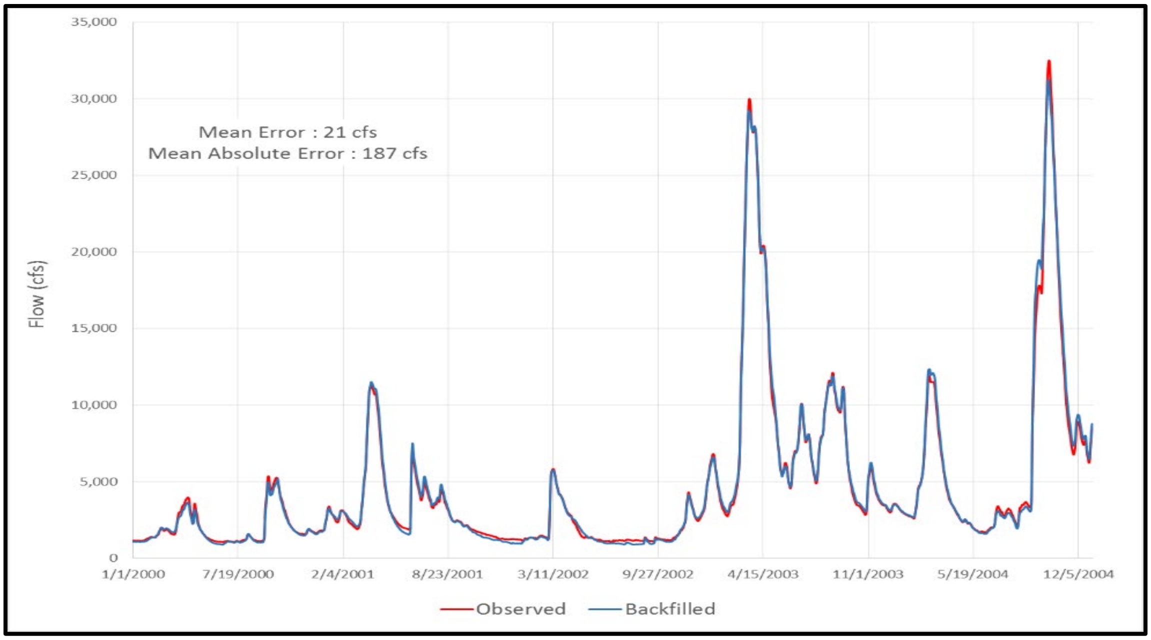

Figure 6(b) – Verification of Backfilled versus Observed Flow Data at Luraville Gage

3.0 REFERENCES CITED

Goodrich, B. and J. Kropko. 2014. An Example of mi Usage. The Comprehensive R Archive Network (CRAN). [*https://cran.r-project.org/web/packages/mi/vignettes/mi_vignette.pdf*]

Rubin, D.B. 1987. Multiple Imputation for Nonresponse in Surveys, New York: John Wiley and Sons Inc.

Su.Y, A.Gelman,J. Hill, and M.Yajima. 2011. Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. Journal of Statistical Software 45(2)1-31.