APPENDIX C - COMPILATION OF THE PRIMARY GROUNDWATER LEVEL DATA SET

APPENDIX C
Compilation of the Primary Groundwater-Level Data Set with Emphasis on Data-Filling Techniques

by Nathan Johnson, P.E.

INTRODUCTION:

The following is a description of methods employed for compiling and then expanding the available set of groundwater-level observations for use in the calibration of the NFSEG groundwater model. First, water level data was gathered from data sources, i.e., various governmental agencies. Second, quality assurance methods were developed to ensure data integrity. Third, since data collection takes a large amount of resources, statistical methods were developed to leverage current knowledge to impute additional statistically derived groundwater level data. With more comprehensive groundwater data, groundwater models may increase in accuracy and robustness to inform decision makers about water resources in the state of Florida.

METHODS:

DATA CLEANING AND AQUIFER CLASSIFICATION:

Groundwater level data were gathered from agency sources including United States Geological Survey (USGS), St Johns River Water Management District (SRJWMD), Southwest Florida Water Management District (SFWMD), Suwannee River Water Management District (SRWMD), South Florida Water Management District (SFWMD), and Northwest Florida Water Management District (NWFWMD). Margit Crowell provided the groundwater data for SWFWMD using the Microsoft Access format. Megan Weatherington from SRWMD provided the data in Microsoft Access format. Nathan Johnson compiled the SJRWMD data from the internal Hydrstra database. USGS data was gathered from the NWIS database internet retrieval system. The vertical datum was standardized to NAVD88 using Corpscon6. Median monthly value were developed using all existing daily values.

A database of monitoring well metadata was developed. Wells were assigned an aquifer classification in the aquiferFinal field in the database based on a hierarchical classification system. The methods used to determine the aquifer classification were reported in the field “aquiferSource”. There were two general methods to describe the source including 1) agency classification and 2) hydrostratigraphic aquifer classification. Agency classification was found in the metadata from the source agency. The hydrostratigraphic aquifer classification method was developed by SJRWMD to determine which aquifer respective wells were open to. Where casing depth and total depth were reported by the agency, the hydrostratigraphic unit was discerned. If greater than 70% of the well open hole was available to a single aquifer, the hydrostratigraphic aquifer classification would identify the respective aquifer otherwise would be classified as “Multiaquifer” or “check”. If only total depth was available, then the aquifer classification at this depth was recorded as “Bottom”.

Several discrete aquifers were identified and combined based on literature, geophysical data, and modeling layer assignments. The discrete aquifer categories were defined in Table 1.

aquiferFinal	Aquifer Full Name
APPZ	Avon Park Permeable Zone (UFA)
AVPK	Avon Park (UFA)
Biscayne Aquifer	Biscayne aquifer
Bottom Aquf	Below the FAS
Brunswick Aquifer	Brunswick aquifer
check	undefined aquifer
FAS	Floridan aquifer system
UFA	Upper Floridan aquifer
FPZ	Fernandina permeable zone
ICU	Intermediate confining unit
LFA	Lower Floridan Aquifer
LSCU	Lower semi-confining unit
MCU	Middle confining unit
MultiAquifer	Multiple aquifers
noClass	no aquifer information
OLPZ	Ocala low permeable zone (UFA)
OPZ	Ocala permeable zone (UFA)
Other	Other aquifer
Sandstone aquifer	Sandstone aquifer
SAS	Surficial aquifer system
SECPA	Southeastern coastal plain aquifer
ULFA	Upper/Lower Floridan aquifer
UZLFA	Upper zone of lower Floridan aquifer
Valley and Ridge Aquifer	Valley and ridge aquifer

Table C-1. Aquifer final and corresponding aquifer full name used in aquifer classification.

The two sources of information underwent a hierarchical classification to determine the most defensible aquifer classification. The first part of the aquiferSource identifies the final aquifer classification source (aquiferFinal) while the posterior part displays more information about the alternative method. If the two sources disagree, this will be stated in the second field as a prefix “dis”. The aquiferSource classification was described in Table 2. Wells that were not classified or contained a non-specific classification such as Floridan Aquifer System (FAS) were assigned aquifer classification based on hydrostratigraphy. This will be refined further in future iterations.

aquiferSource	Derivation
Agency/Strat	aquiferFinal = Agency, Stratigraphy agree
Strat/Agency	aquiferFinal = Stratigraphy, Agency general
Agency/disStrat	aquiferFinal = Agency, Stratigraphy disagree
Strat	aquiferFinal = Stratigraphy, No Agency
Bottom	aquiferFinal = Bottom, No Agency
Agency/disBottom	aquiferFinal = Agency, Bottom disagree
Agency/Bottom	aquiferFinal = Agency, Bottom agree
noClass	aquiferFinal = noClass, No Agency, No Casing Depth, No total Depth
Agency	aquiferFinal = agency, No Casing Depth, No total Dpeth
Bottom/Agency	aquiferFinal = Bottom, Agency general

Table C-2. Description of the well aquifer source in the field aquifer source

Well data was combined if various agencies reported data for the same physical well. Many agencies have assimilated well data from the USGS and have distinct naming conventions. Agencies sometimes reported USGS IDs in addition to the agency unique name. There were many cases where agencies annexed USGS wells and did not incorporate the previously recorded USGS data. Data from the same well was combined and given a common Name based on the following hierarchical order: USGS, SJRWMD, SWFWMD, SFWMD, SRWMD, NWFWMD. Occasionally, reported USGS IDs from agencies did not exist within the NWIS database and the USGS ID was skipped for common Name assignment.

QUALITY ASSURANCE:

The statistical software R was used to screen data using normalized agglomerative cluster analysis to identify wells that exhibited irregular patterns. Since wells have varying periods of record, cluster analysis was performed on five-year periods allowing for 20% missing within the period. When individual wells were identified as a single cluster, they were examined and culled for outliers, shifts, below threshold values, etc. This process is proficient at selecting outliers where relatively continuous data is present over several years however other data may not meet these criteria and were left unaltered.

REGRESSION IMPUTATION/FILL:

Since monitoring wells contain varying periods of record and continuity, data gaps were examined and partially imputed using robust and scientifically defensible methods. Initially, linear regression models were built between selected original wells and wells within +/- 0.5 degrees latitude and longitude. In this case, the original independent well would be the explanatory variable and the adjacent wells were the response variables. Non-linear regression methods using transformations of variables were initially examined, however linear methods were most parsimonious. The best linearly correlated well within the adjacent area was used to create a simple linear regression model and fill gaps where data exists for the original well. Since autocorrelation exists within the well time series, several thresholds were set to reduce spurious correlation. The regression relationship must have ten matching pairs on corresponding dates and extend over three years so that the effects of autocorrelation are reduced when building statistical models. The regression must have a coefficient of determination (R^2) greater than 0.90 to ensure that the independent well explains 90% of the variability of the fill well. The fill well must contain at least three non-corresponding dates and must have at least one level after the year 1999.

Summary of the well selection thresholds is as follows:

Regression metric $R^2 > 0.90$

1. Original and fill well must overlap by three years
2. Original and fill well must have ten matched pairs on corresponding dates
3. Fill well must contain three non-corresponding data points
4. Fill well must contain data post 2000-01-01

When the thresholds were met, the statistical model was used to impute/fill data for the original wells. This methodology vastly expanded the amount of data available for the models by leveraging the relationship between highly correlated wells. This process was repeated twice so that the maximum number of wells could be filled using the simple linear regression method. The first iteration was labeled "first filled" and the second iteration using the results from the first iteration were called "second filled"

PRINCIPLE COMPONENT IMPUTATION/FILL:

When well data was insufficient to meet the thresholds for the linear regression imputation method, another method was developed that leveraged the time series signal of spatial regions to inform and fill well time series. First, agglomerative cluster analysis was selected to group wells into clusters based on their normalized Euclidean distance. The method starts with all wells in their own cluster and merges wells using the Euclidean distances based on the Wards linkage. The number of clusters was optimized by merging clusters until a unique spatial grouping pattern was formed in addition to bootstrapped clustering distance convergence.

Once clusters were identified, principle component analysis (PCA) was performed to calculate the orthogonal eigenvectors that explained the variance within the group. The first principle component was required to describe greater than 85% of the variance of the wells within the cluster. If the first principle component explained less than 85% of the variance, then more clusters were added and the process repeated. Next, linear regression was executed with the first principle component as the explanatory variable and wells with little data as the response variables. The PCA regressions were given thresholds to ensure non-spurious models, however with small degrees of freedom, this imputation method should only be used in areas where other imputation methods do not produce sufficient data, data is very limited, clusters are spatially grouped, and PCA explains > 85% of the variance within a cluster.

RESULTS:

The original dataset for the total domain contained 18,977 well points and 1,061,673 median monthly values and spatially shown in Figures 1 and 2 over the period 1950-2012 and 2000-2012 respectively. The use of the three different methods augmented the total monthly median values to 1,507,917. This increased the amount of data by nearly 50%. The filled data categorized by imputation method produced 357,622 first filled, 115,141 second filled, and 11,810 PCA filled monthly values. The summary of quantity of stations and monthly values by fill type is given in Table C-3 and quantity of stations separated by aquifer in Table C-4.

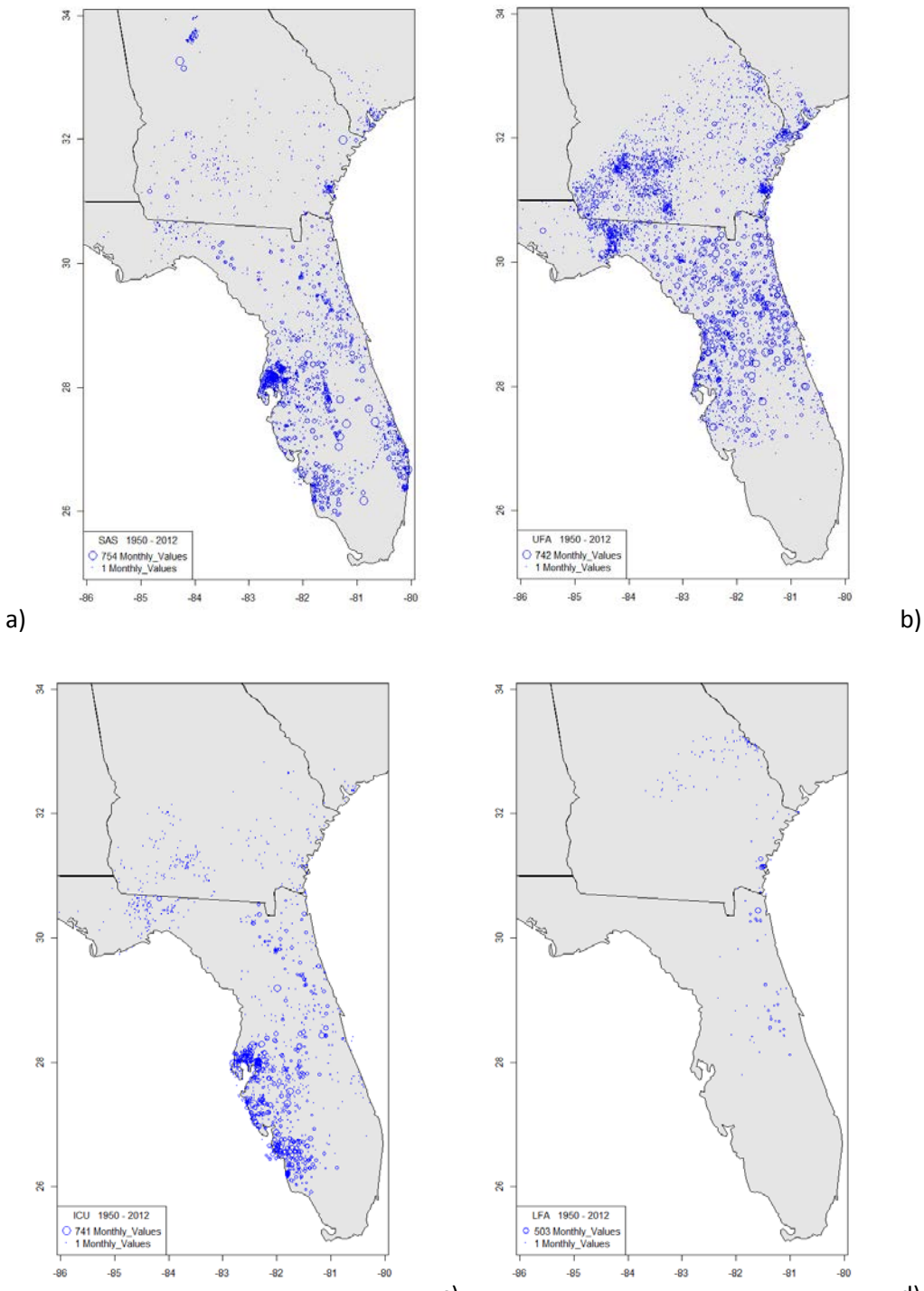


Figure C-1. Monthly groundwater-level data available (1-756) using original data (1950-2012) in a) SAS b) UFA c) ICU d) LFA

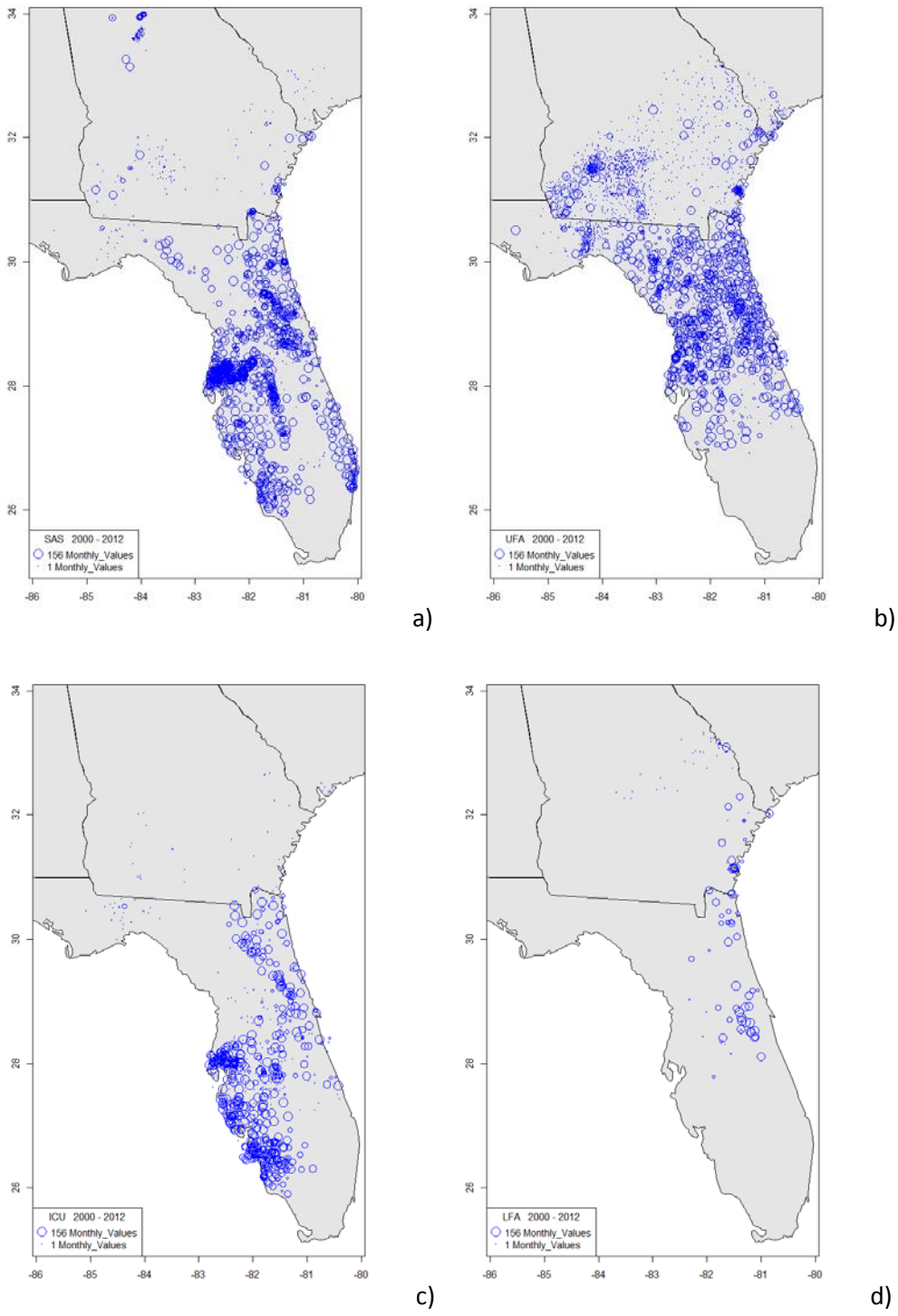


Figure C-2. Monthly groundwater-level data available (1-156) using original data (2000-2012) in a) SAS b) UFA c) ICU d) LFA

Fill Method	Stations	Monthly Values
Original	18977	1061673
First Filled	2891	357622
Second Filled	1725	115141
PCA Filled	67	11810
Total	18977	1546246

Table C-3. Summary of data available separated by data fill type.

AquiferFinal	Data Fill Type			
	Original	First Filled	Second Filled	PCA Filled
Undefined	34	0	0	0
APPZ	113	60	39	0
AVPK	32	21	19	0
Biscayne aquifer	740	132	47	0
Bottom Aquf	299	6	0	0
Brunswick Aquifer	44	10	5	0
check	13	0	0	0
Crystalline Ridge Aquifer	1013	7	0	0
Crystalline Rock Aquifer	1	0	0	0
FAS	1033	199	149	0
FPZ	4	2	1	0
ICU	1845	445	279	0
LFA	199	59	46	0
LSCU	0	0	0	0
MCU	110	20	17	0
MultiAquifer	204	11	6	0
noClass	442	16	12	0
OLPZ	21	15	7	0
OPZ	257	127	71	0
Other	5	1	1	0
Sandstone aquifer	0	0	0	0
SAS	4560	485	191	0
SECPA	988	38	6	0
UFA	6285	1212	816	67
ULFA	8	3	0	0
UZLFA	570	22	13	0
Valley and Ridge Aquifer	157	0	0	0
Total	18977	2891	1725	67

Table C-4. Summary of quantity of stations separated by aquifer and data fill type.

The data was first quality controlled by using cluster analysis of wells over a period of five and ten years. An example of cluster analysis on data that has had no quality analysis is illustrated for the period of 2000-2010 (Figure 3). This cluster identified Clusters 4, 6, 7, 8 and 9 to be examined and data removed if necessary. Wells could exhibit shifts, outlier, below detection limit, and other anomalous behavior (Figure 4). After anomalous data was adjusted, the final cluster analyses contained wells that behaved similarly to one another (Figure 5). This result quality controlled data was used in the remainder of the analysis.

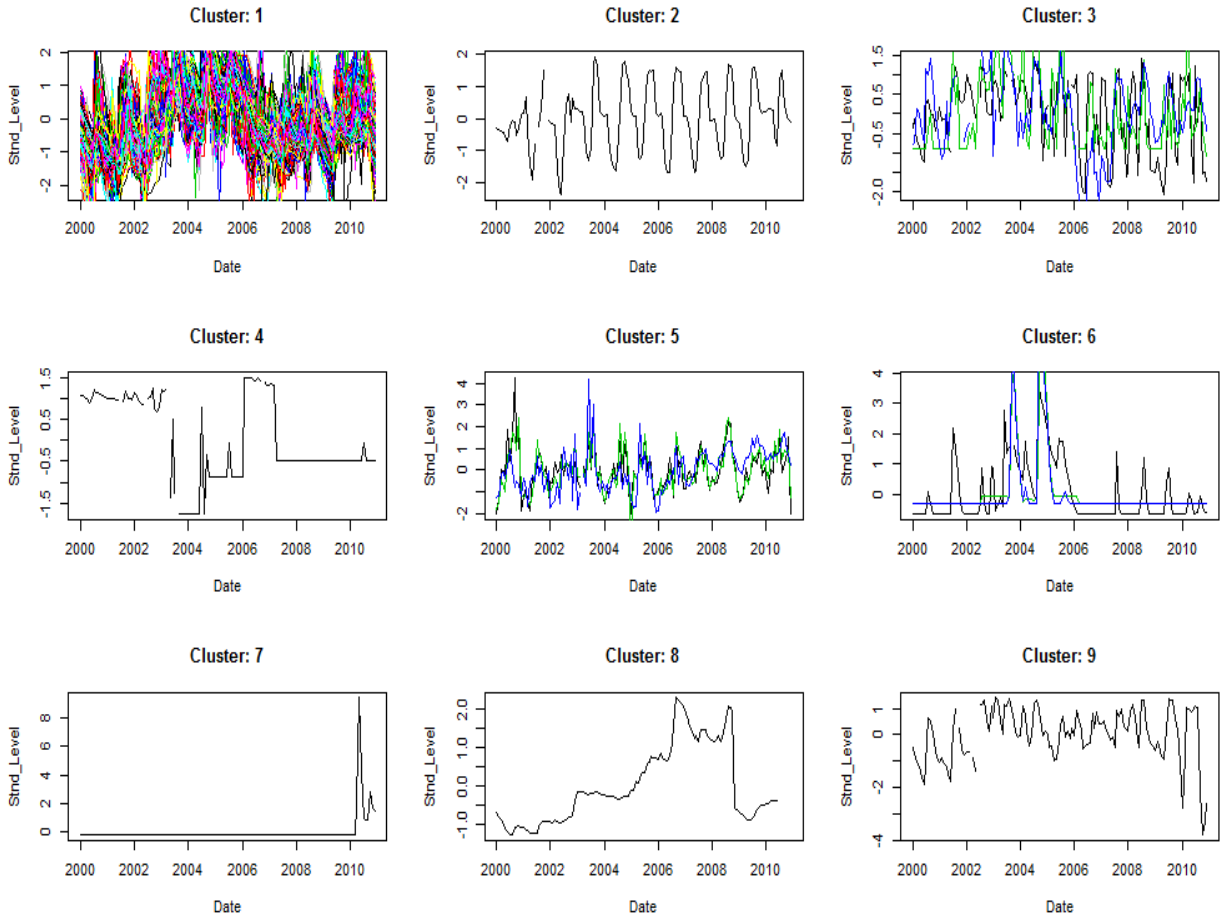
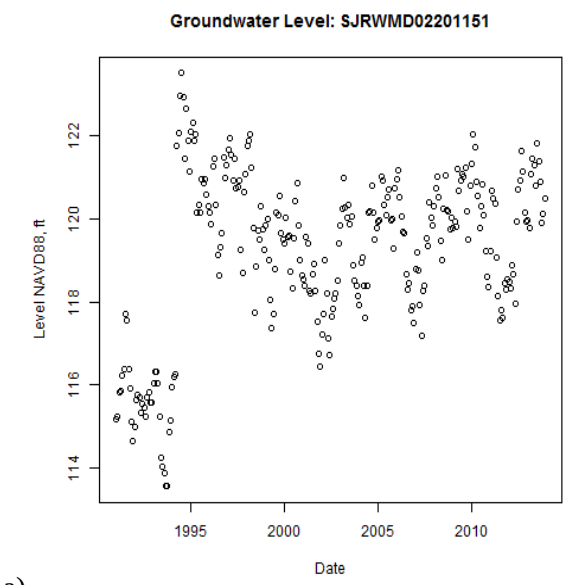
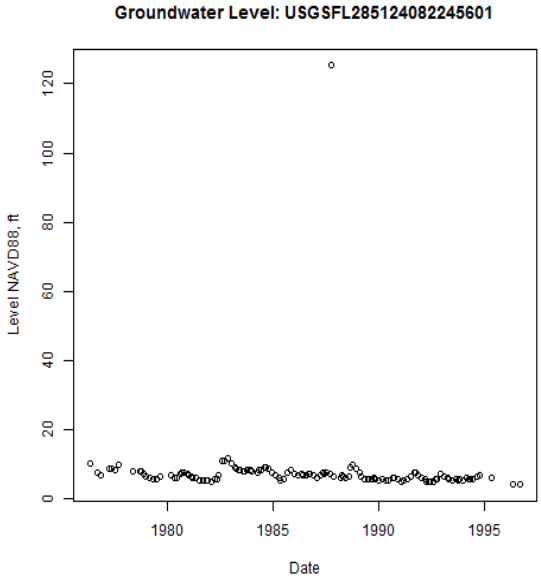


Figure C-3. Cluster analysis with non-quality assured data (2000-2010)

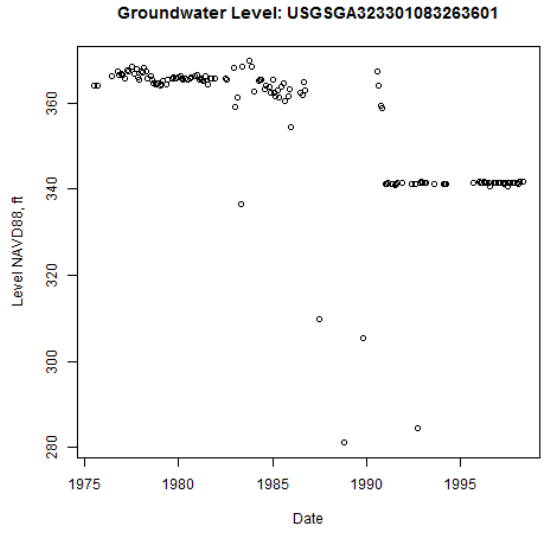


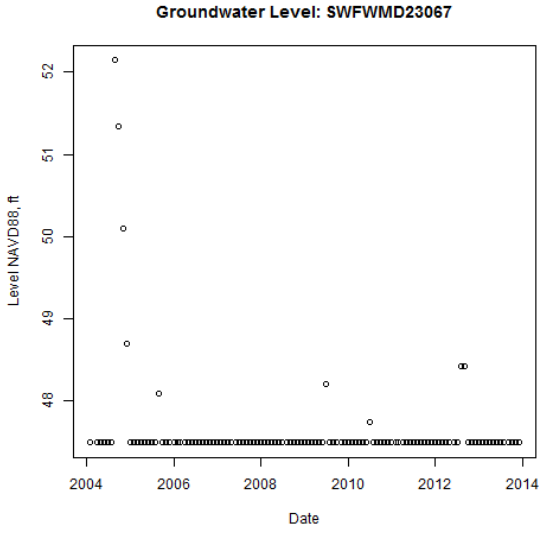
a)



b)

c)





d)

Figure C-4. Well demonstrating a) shift b) outlier c) undetermined error d) below detection limit

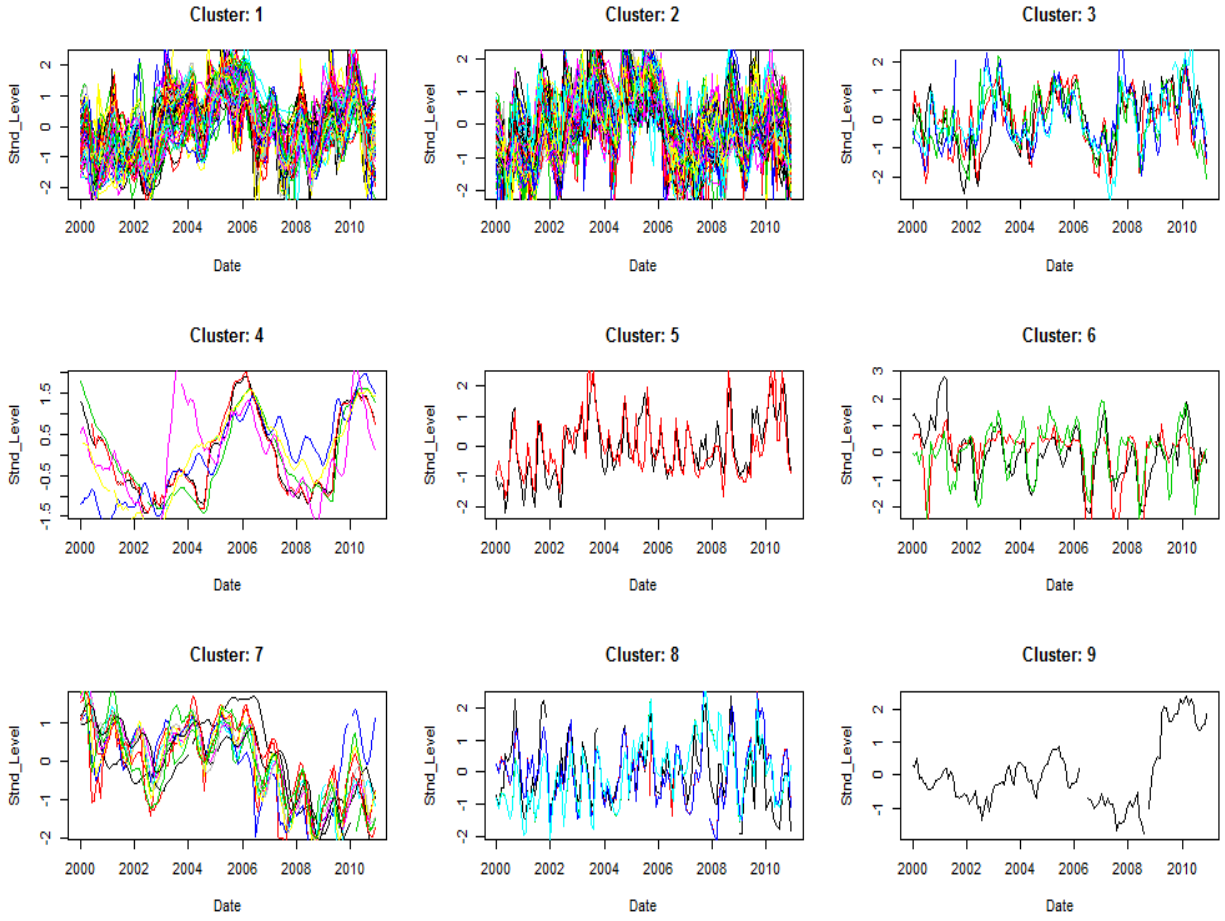


Figure C-5. Cluster analysis with quality assured data (2000-2010).

Once the quality controlled dataset was developed, the data was filled using linear regression according to the thresholds set for the data gap imputation. For example, the explanatory well SWFWMD25162 (UFA) was filled for data prior to 2005 using the adjacent response well SWFWMD24802 (ICU). The linear regression summary statistics included R^2 = 0.989, degrees of freedom (DF) of 61, and root mean square error (RMSE) of 0.558 ft (Figure 6). The same well was second filled using response well SWFWMD17974 (OLPZ) to add an addition four months of data. The linear regression summary statistics were R2= 0.988, DF = 219, and RMSE = 0.756 (Figure C-7). The locations of both the independent wells and dependent wells are shown in the figures as well to illustrate a spatial context for the filling wells and used for visual examination (Figures C-6 and C-7).

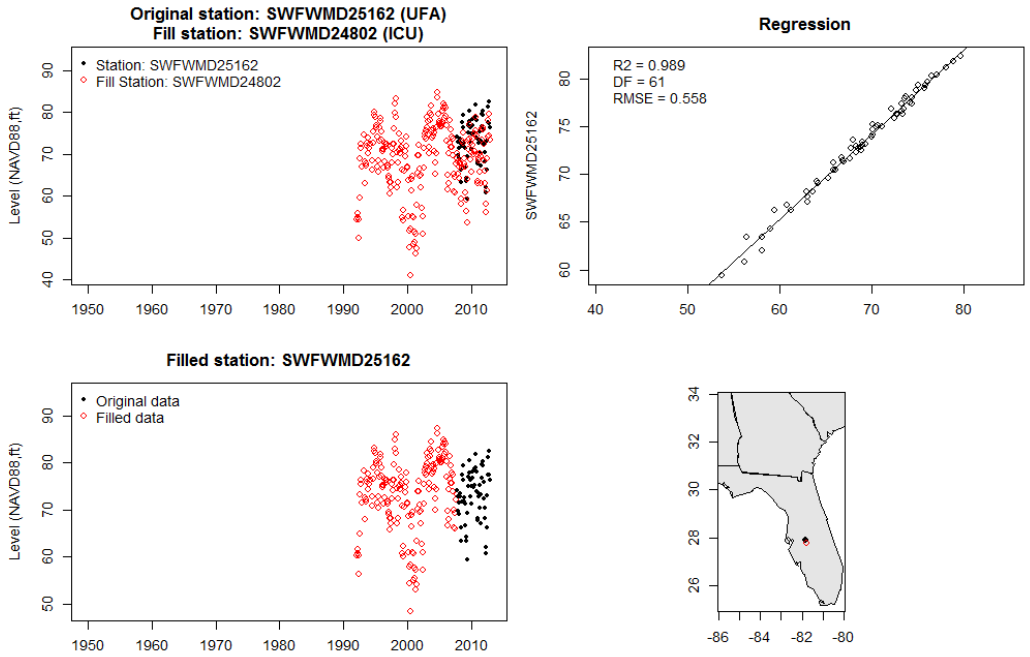


Figure C-6. Linear regression analysis showing the original and fill wells (top left), linear regression (top right), resulting dataset (bottom left), and locations of both wells (bottom right).

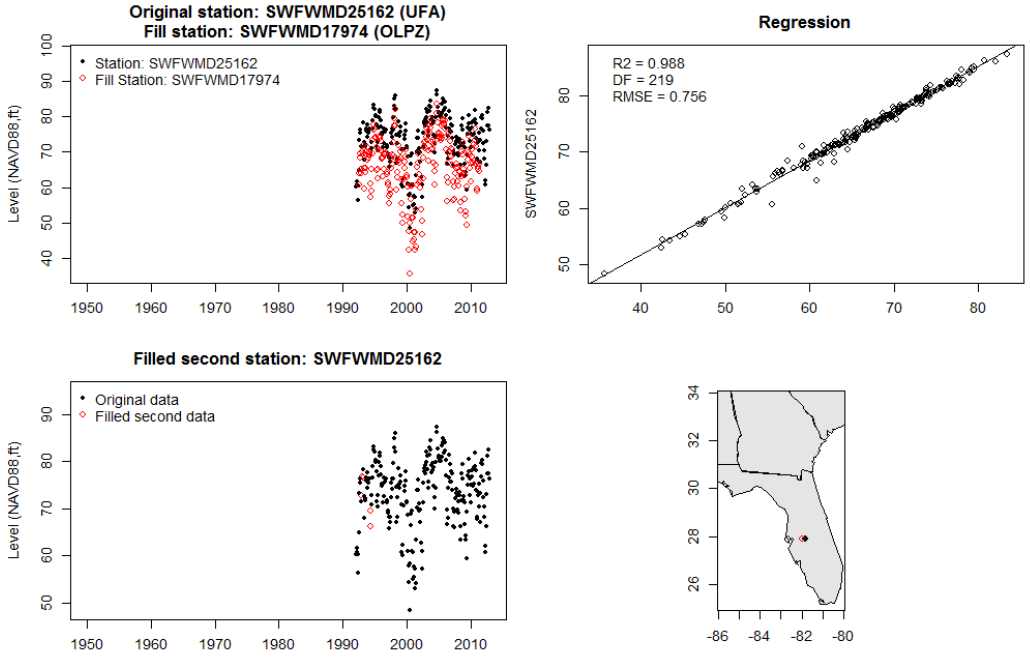
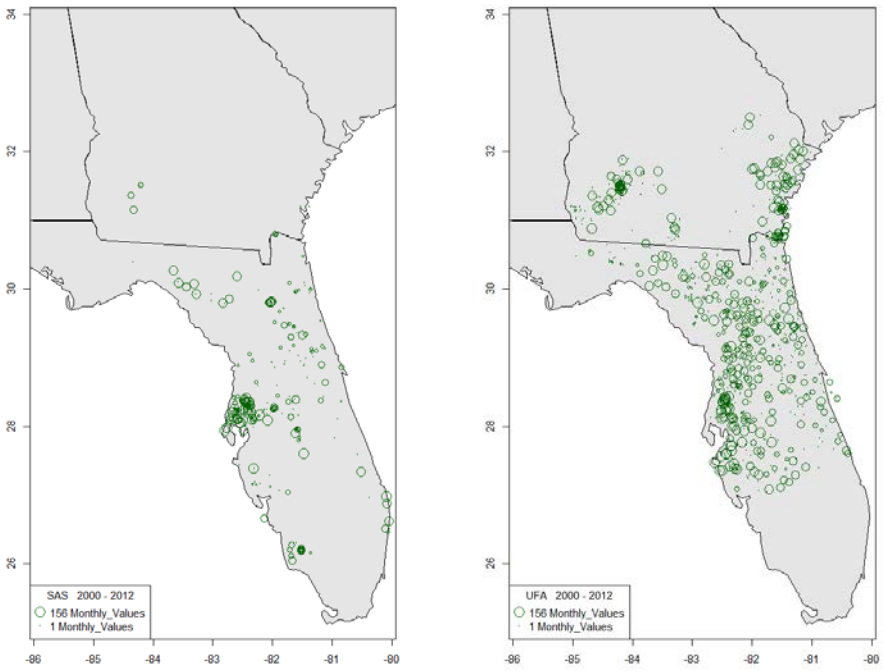


Figure C-7. Second linear regression analysis showing the original and fill wells (top left), linear regression (top right), resulting dataset (bottom left), and locations of both wells (bottom right).

After filling in data using the R script, the final data and linear regression models were presented in spreadsheet files. The original data was designated “original” and the filled data is designated “first filled” and “second filled”. Overall, 2,892 wells and 357,622 monthly groundwater levels were filled using the first filling method and 1,725 wells and 115,141 monthly groundwater levels were filled using the second filled method. Figure 8 spatially illustrates the quantity of first filled data that is available in the SAS, UFA, ICU and LFA over the period 2000-2012. Figure 9 spatially illustrates the quantity of second filled data that is available in the SAS, UFA, ICU and LFA over the period 2000-2012. A majority of the stations that were filled were UFA stations. Nearly 33% of first filled stations were UFA and nearly 50% of second filled stations were UFA (Table 4). Additionally, a summary of model metrics (RMSE, R2, degrees of freedom) was provided in Figure 10 for each filling method. All models provide a summary statistic R2 of greater than 0.90 since it is a threshold with the model. Most models have an RMSE of less than 2 feet however there are several linear models in both the first and second fill that have a greater than 2 feet RMSE indicating a poorer model fit. In additional iterations, this may be included as a model threshold to remove some of the uncertainty. The degrees of freedom in the models were generally skewed left as was expected since many wells have not been monitored over extensive periods.



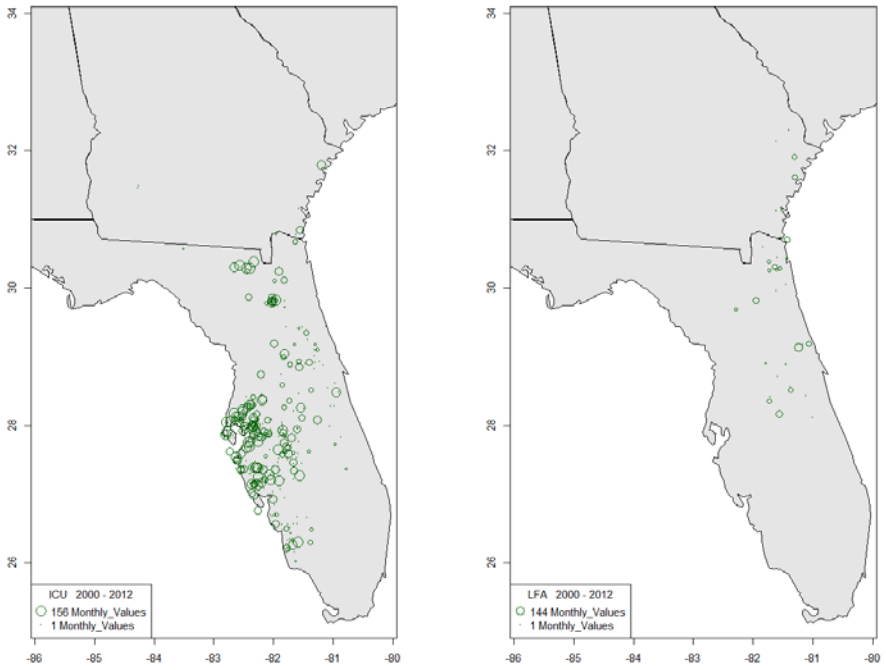
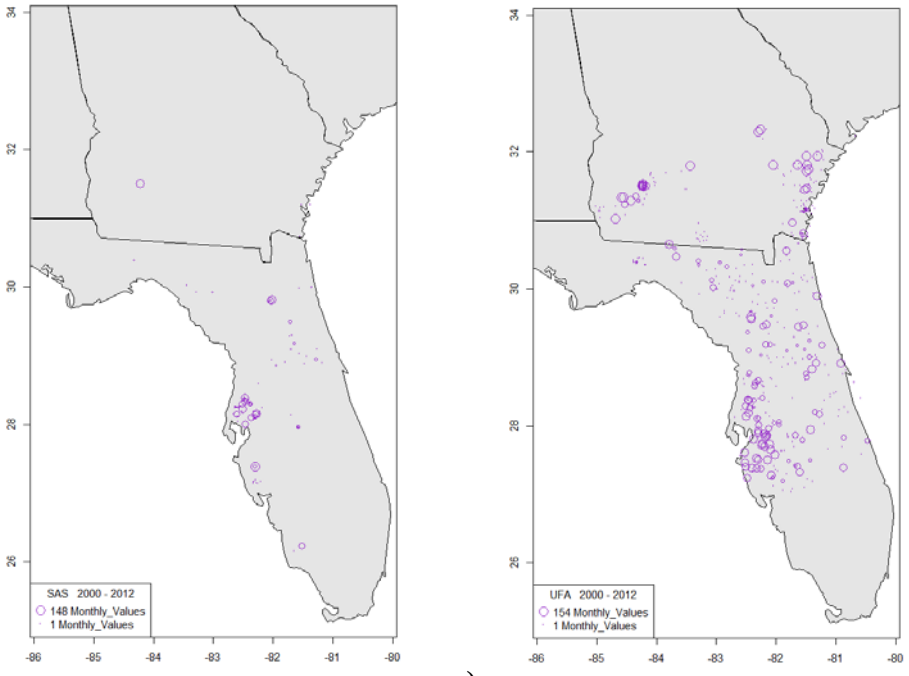
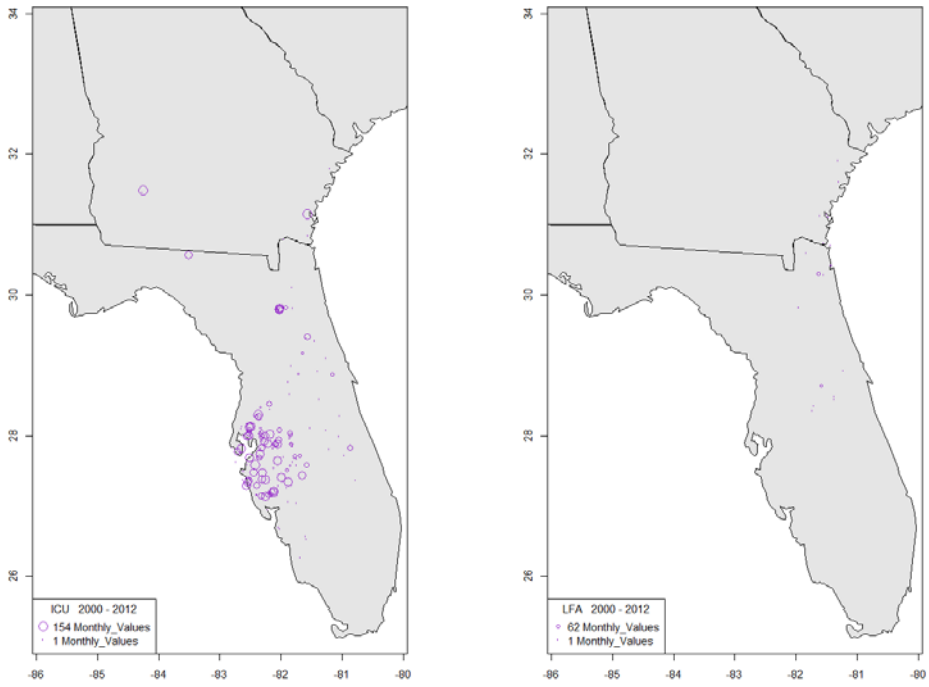


Figure C-8. First-filled quantity of median monthly groundwater level data available (1-156) using only first-filled data (2000-2012) in the a) SAS b) UFA c) ICU d) LFA

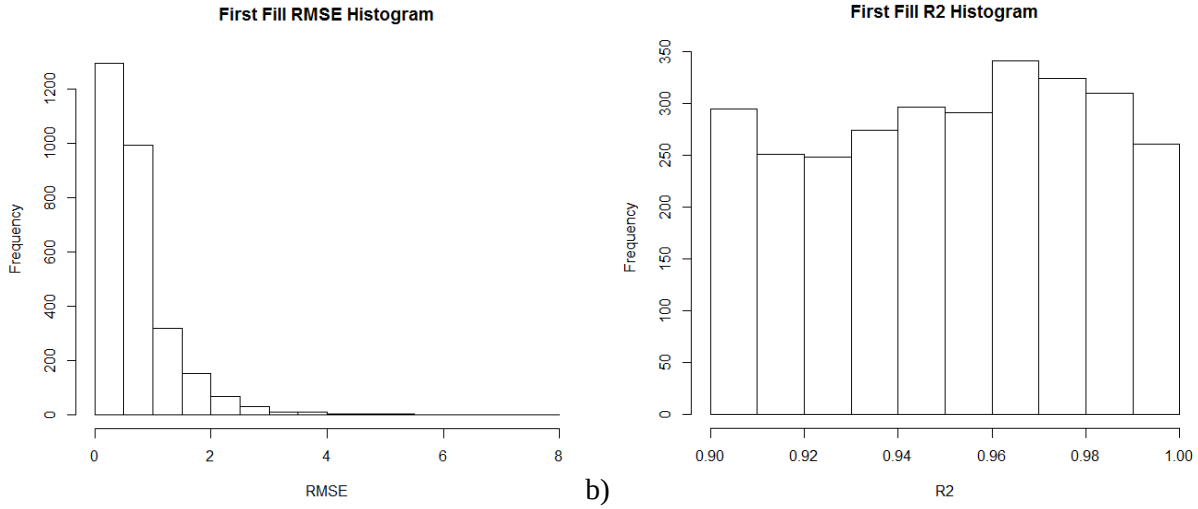


a) b)



c) d)

Figure C-9. Second-filled quantity of median monthly groundwater level data available (1-156) using only second filled data (2000-2012) in the a) SAS b) UFA c) ICU d) LFA



b)

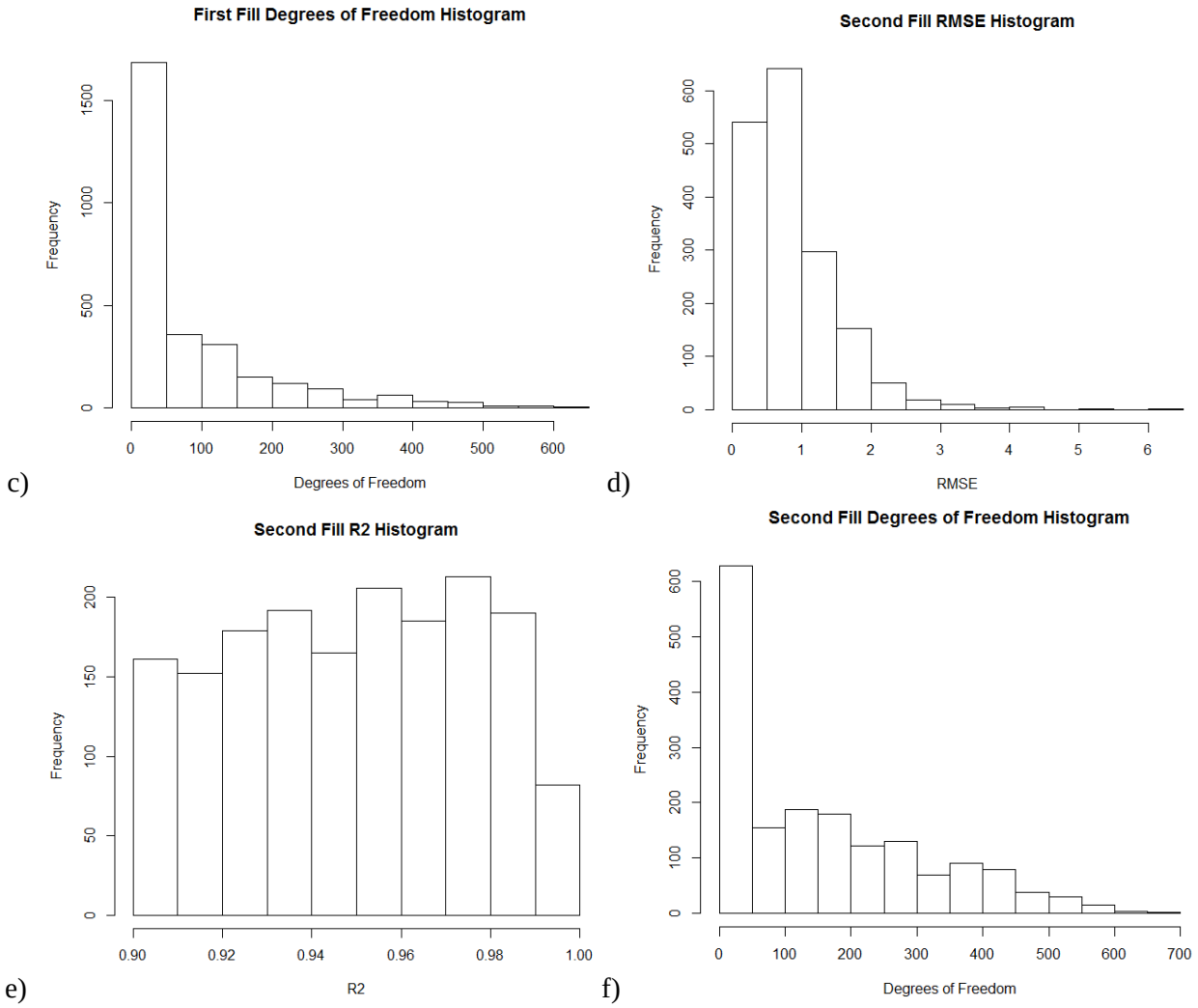


Figure C-10. Summary statistics for first fill a) RMSE b) R2 c) degrees of freedom and second fill d) RMSE e) R2 f) degrees of freedom linear regression models.

Once data was filled using both linear regression filling methods, several large spatial gaps existed within Georgia and the northern part of Florida in the UFA. The UFA for 2001, 2009 and 2010 all illustrate a large spatial gaps in Georgia (Figure C-11). This area was filled using cluster analysis combined with principle component analysis. Cluster analysis over the period 1982-2010 binned the UFA wells in the region into twenty-four groups to optimize the spatial grouping (Figure C-12). The period 1982-2010 was selected since many wells have level data in the UFA in 1982. Each well was normalized and plotted in its respective cluster (Figure C-13) to illustrate the respective cluster signal.

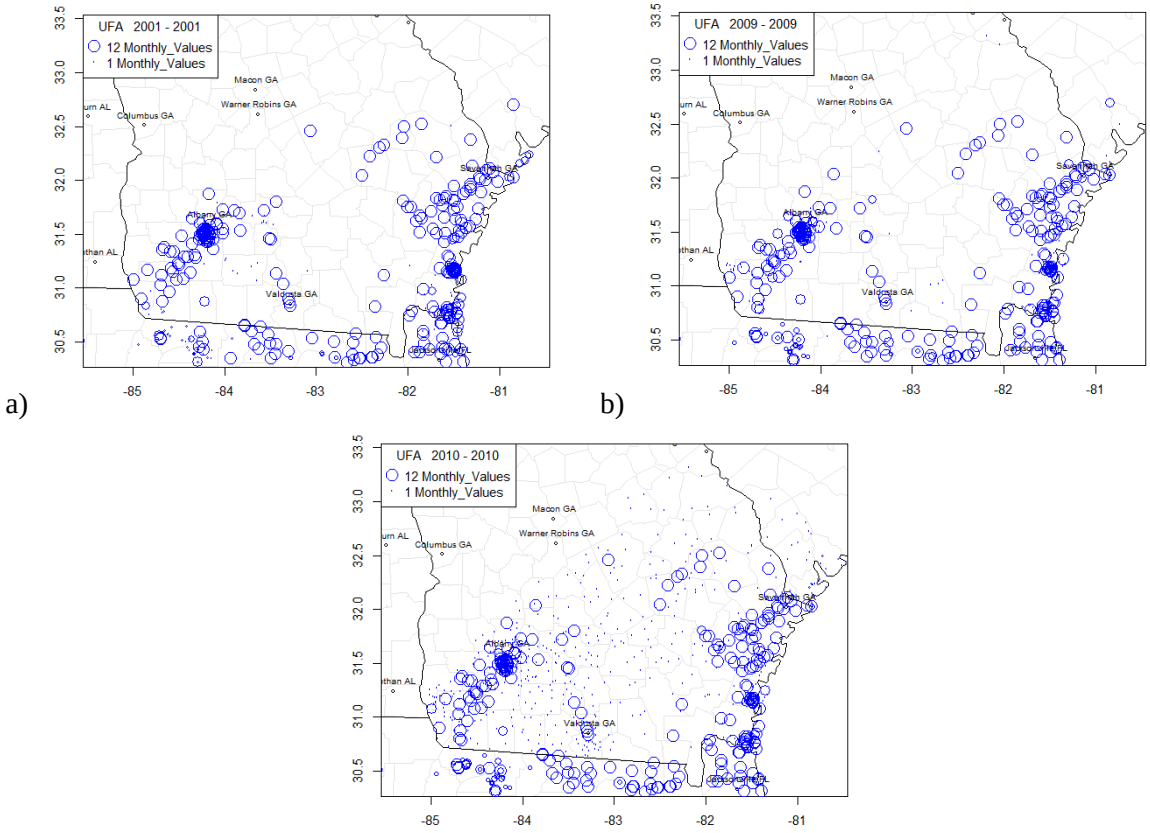


Figure C-11. Quantity of median monthly groundwater levels after first and second filling for years a) 2001 b) 2009 c) 2010.

Several clusters contained only one well including: Cluster 3 (SJRWMD27234872), Cluster 5 (USGS301852081234201), Cluster 6 (USGS302416081522601), Cluster 10 (USGS305235084125101), Cluster 13 (USGS311009084495502), Cluster 14 (USGS31633081324101), Cluster 18 (USGS312853084275101), Cluster 20 (USGS313808084093601), Cluster 21 (USGS314330084005402), Cluster 22 (USGS315228084100601), Cluster 24 (USGS322652083033001). These clusters identified wells that represented outliers for general signals of a region. Most likely these outlier clusters are due to pumping centers, aquifer misclassification, representation of a unique region, etc. Other clusters including Cluster 1, Cluster 12 and Cluster 15 contain many wells that span over larger regions. Cluster 1 contains a region that surrounds Jacksonville and extends north. Cluster 12 contains the northern part of the UFA below the fall line and the middle of the part of the state north of Valdosta GA. Cluster 15 contained areas south of Savannah GA and extends westward.

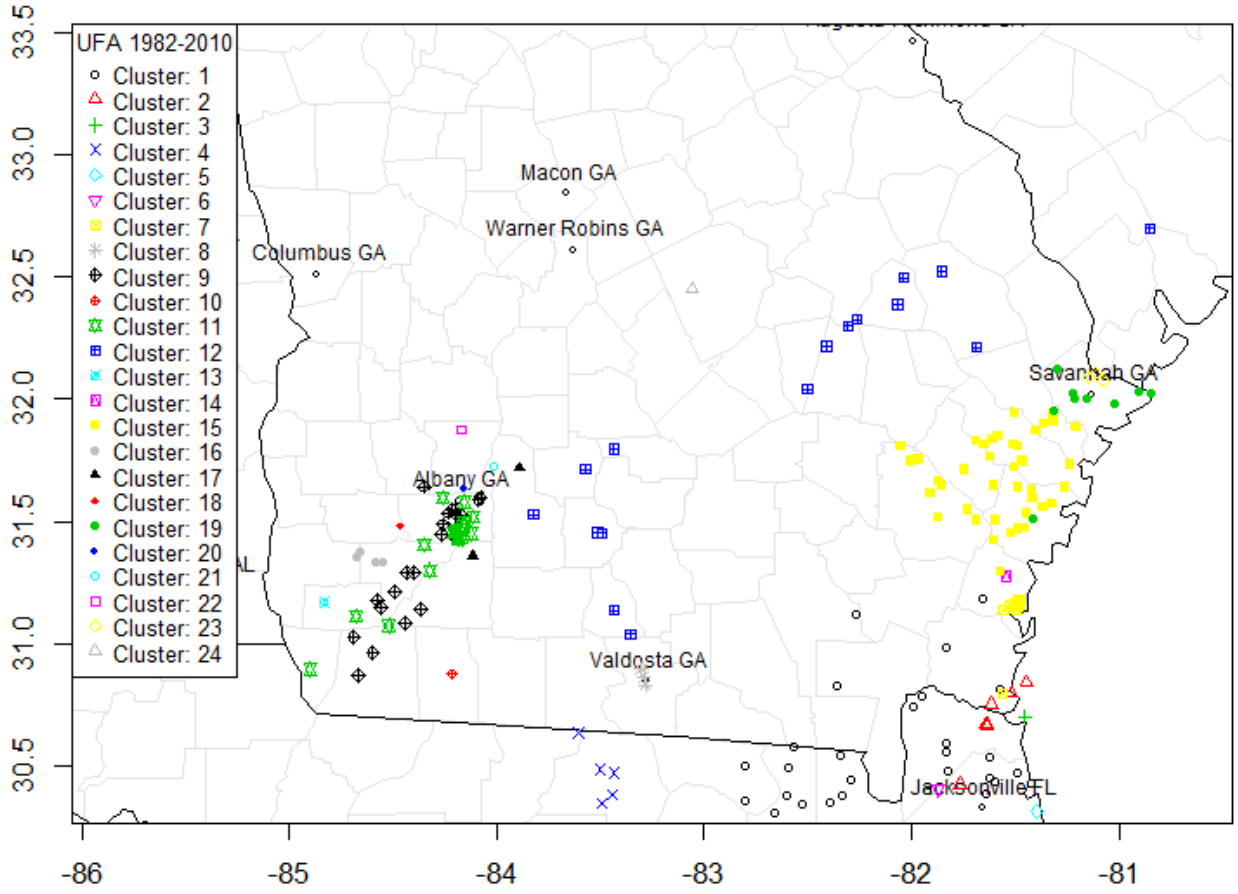


Figure C-12. Map of UFA clusters in Georgia, South Carolina and North Florida (1982-2010)

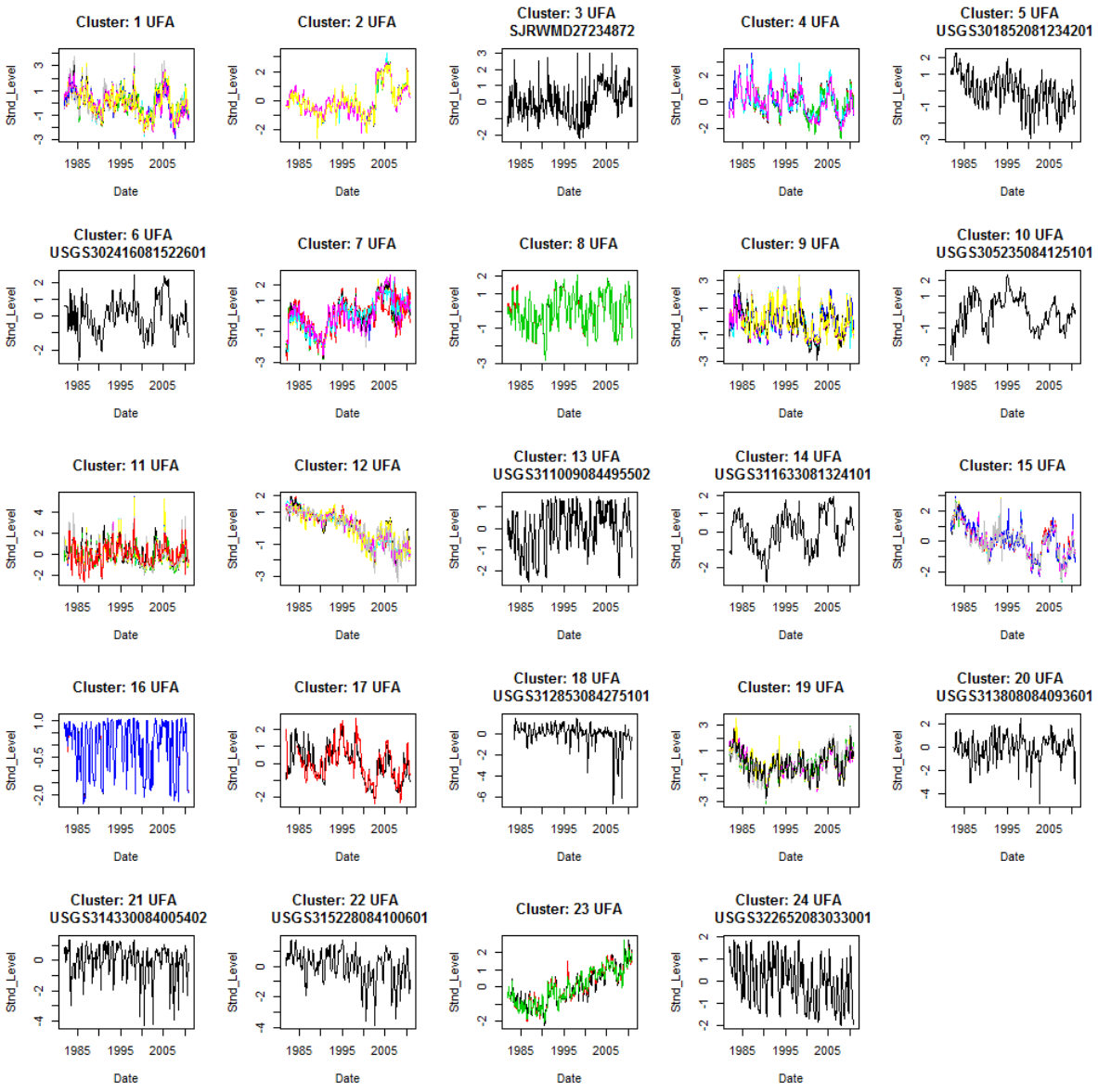


Figure C-13. Normalized UFA well time series grouped by cluster.

Once areas of distinct temporal patterns formed spatial clusters, principle component analysis was applied to each cluster. The first principle component was generated for each cluster and used to fill wells with limited data. PCA analysis can only be performed if there were no gaps in the data. In order to accommodate for this, wells with missing data were removed. PCA was not performed on clusters that had less than two wells. The first principle component for each cluster was illustrated in Figure 14. The proportion of variance explained by the first principle component had to exceed 0.85 as illustrated below each figure in Figure 14. Wells with greater than two data points were filled using linear regression against the first principle component. This process was illustrated in Figures 15 and 16. The location of the of the well and the various adjacent principle components clusters are shown in the top left. A spatial summary of the total wells filled using PCA is illustrated in Figure 17 over the period 2000-2010.

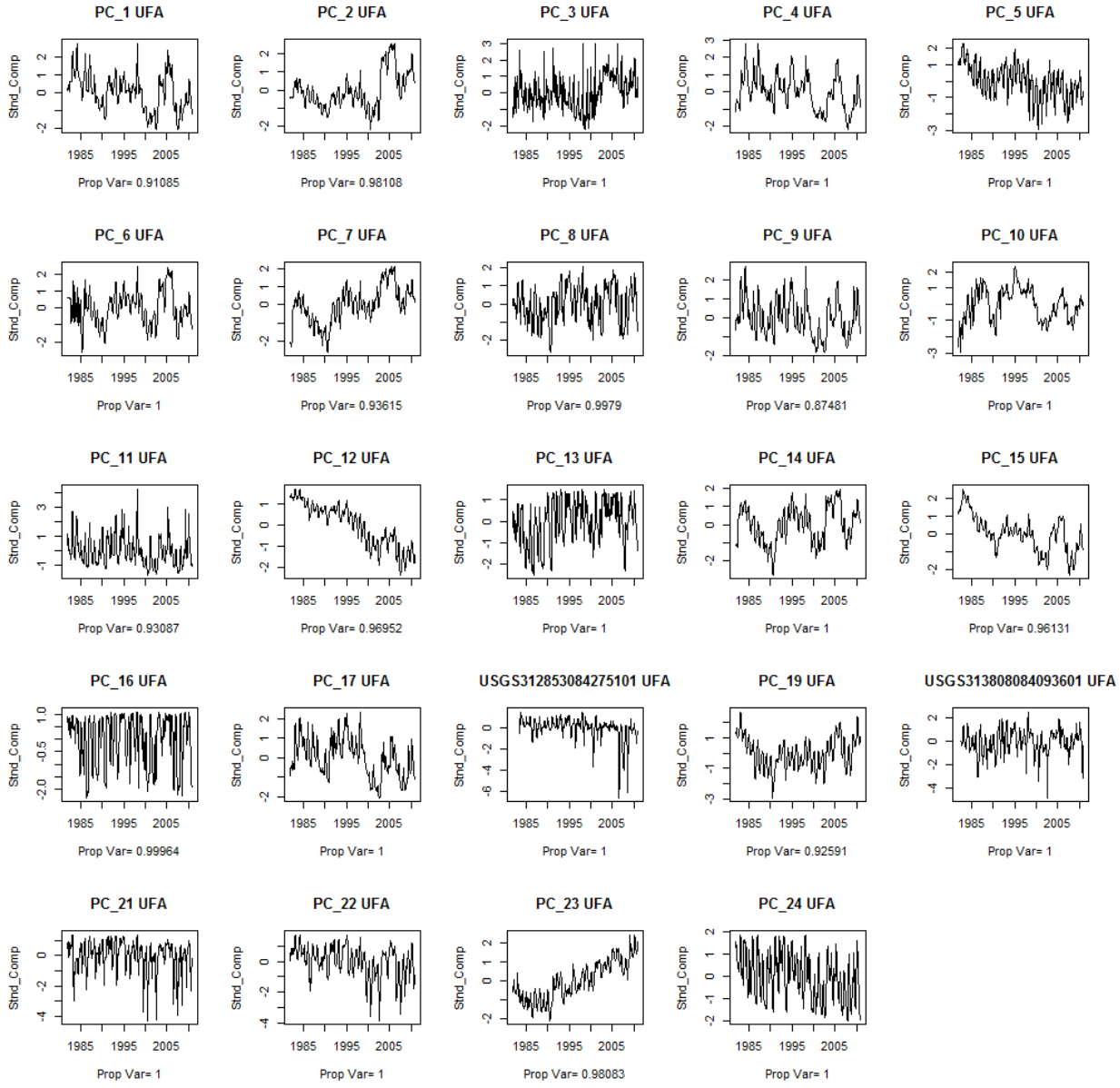


Figure C-14. First principle component for respective UFA cluster wells. Below each graphic reports the proportion of variance described by the first principle component.

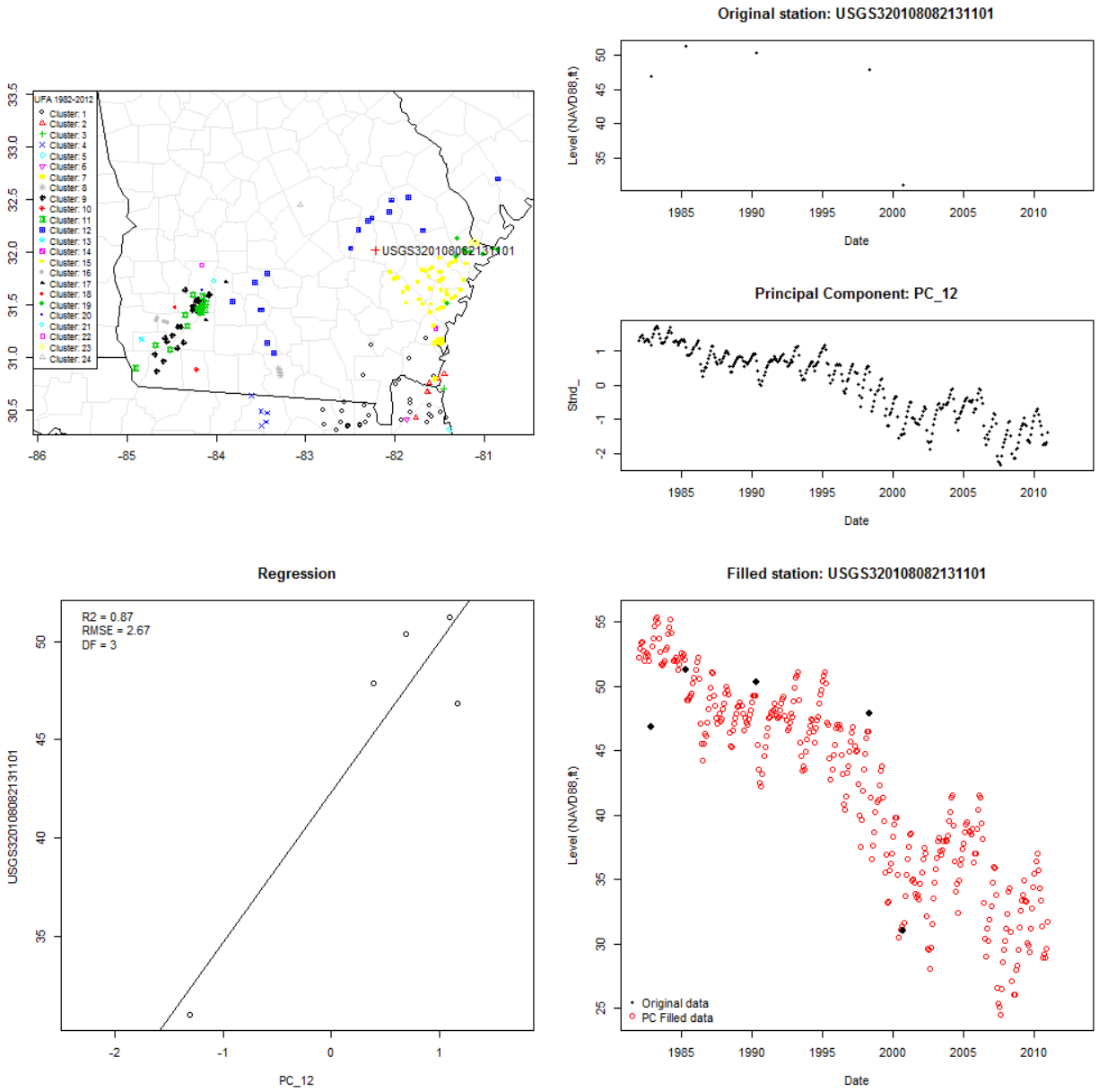


Figure C-15. Locations of dependent well and clusters (top left) dependent well and the selected cluster first principal component (top right), linear regression (bottom left), resulting dataset (bottom right).

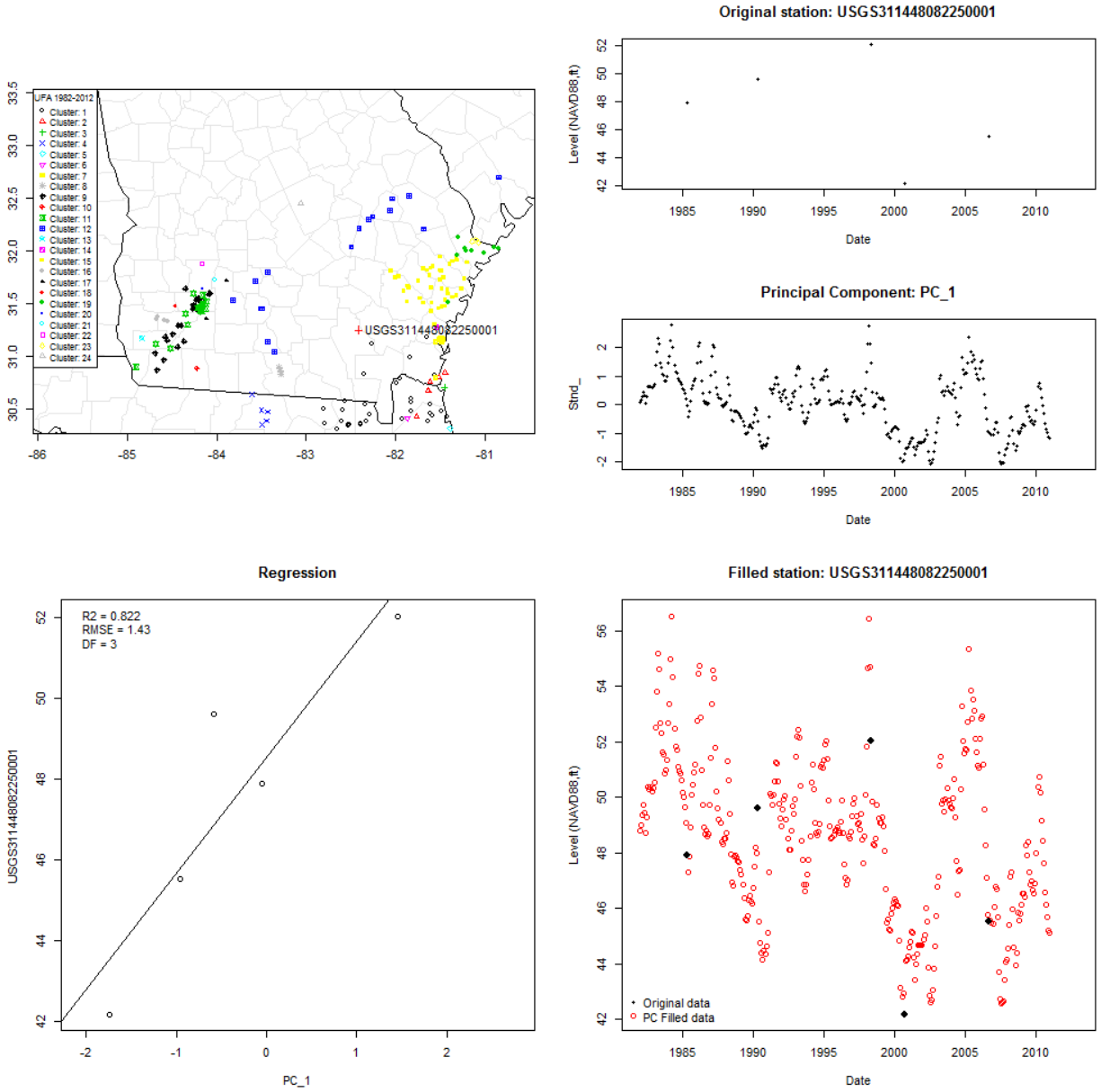


Figure C-16. Locations of dependent well and clusters (top left) dependent well and the selected cluster first principal component (top right), linear regression (bottom left), resulting dataset (bottom right).

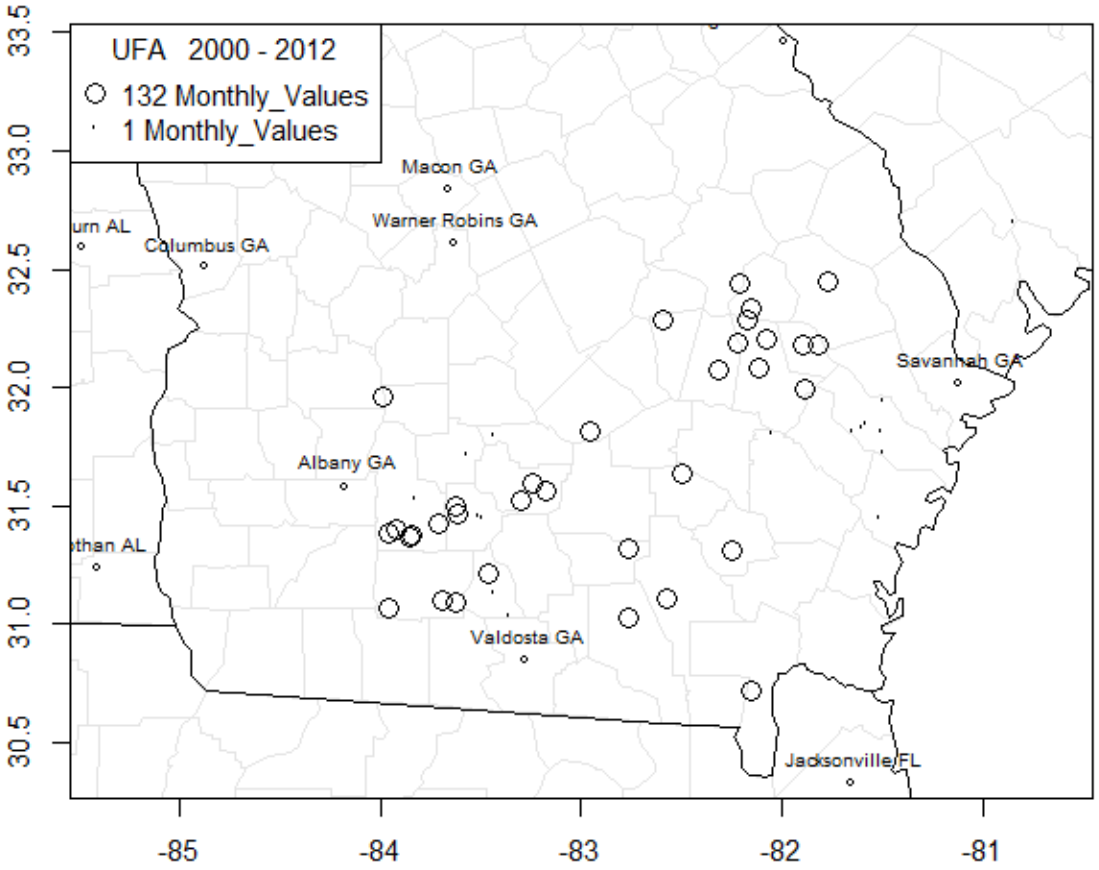


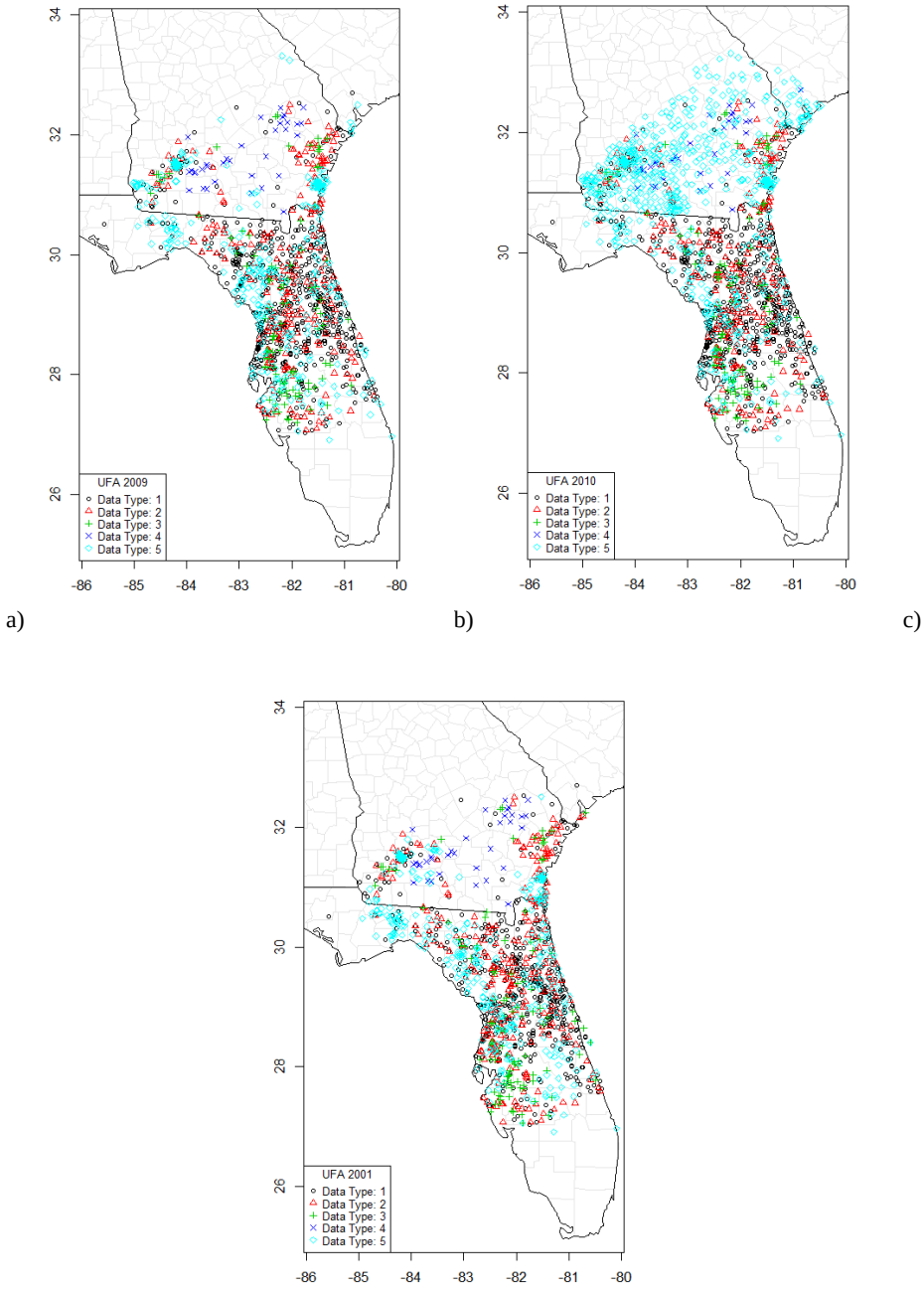
Figure C-17. PCA filled median monthly groundwater level data available (1-132) using only PCA method (2000-2010)

The resulting product included a database of well information with aquifer classification and other metadata. The resulting time series well data was given in five data fill types including 1) original 2) first filled 3) second filled and 4) PCA filled. The data was aggregated into annual median values and assigned a data filling type for steady state groundwater models calibration. Individual wells were given a data fill type for each year based on data available from various filling methods according to the hierarchical list:

Data Fill Type	Description
1	Original data > 6 months
2	Filled data > 6 months
3	Filled second data > 6 months
4	PCA data > 6 months
5	Any data < 6 months

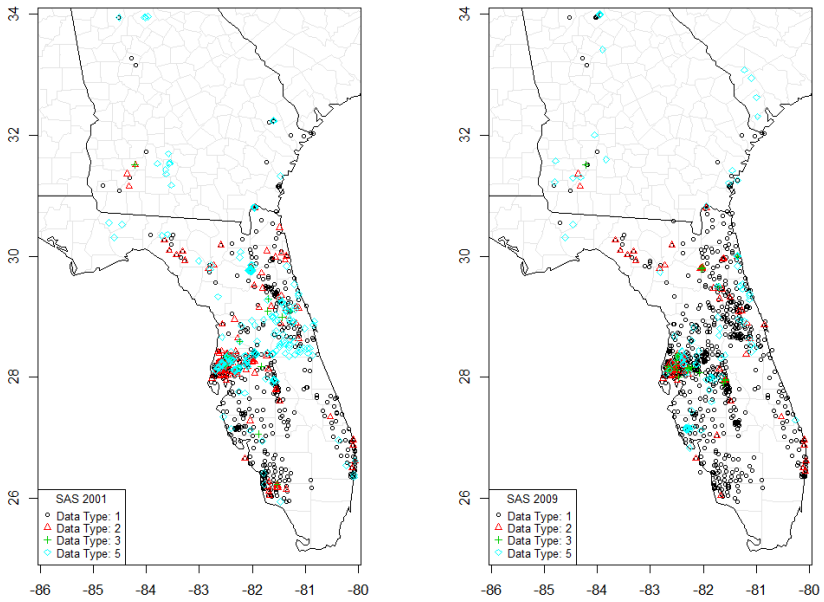
Table C-5. Data fill type and description

Figures 18-20 illustrated the spatial distribution of the different data types in the UFA, SAS, and LFA for the years 2001, 2009, 2010. The median annual value will be used for model calibration targets and weighted during calibration based on the data fill type.

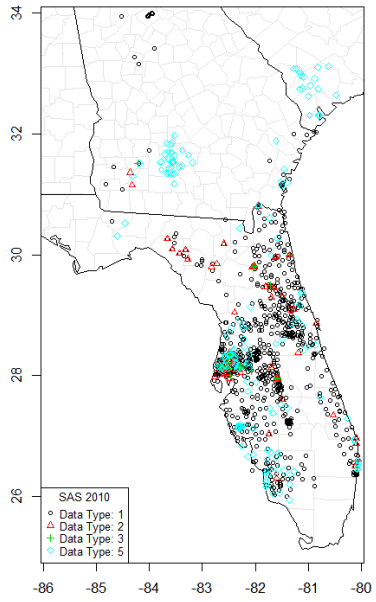


a) b) c)

Figure C-18. Data fill type based on filling method from Table 5 for aquifer UFA in years a) 2001 b) 2009 c) 2010

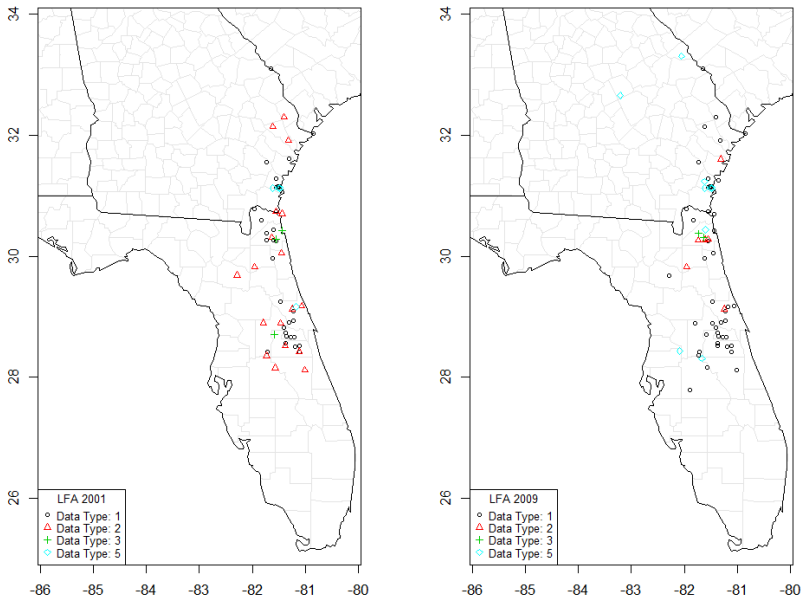


a) b)

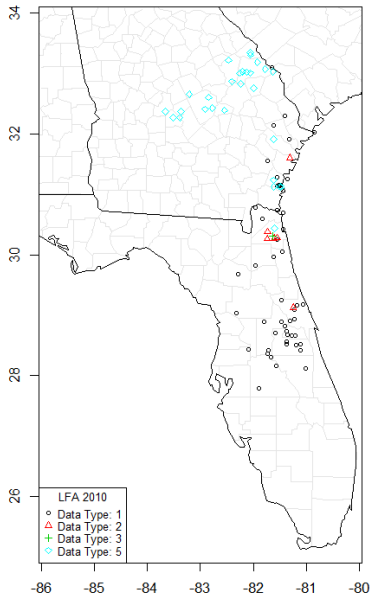


c)

Figure C-19. Data fill type based on filling method from Table 5 for SAS in years a) 2001 b) 2009 c) 2010.



a) b)



c)

Figure C-20. Data fill type based on filling method from Table 5 for LFA in years a) 2001 b) 2009 c) 2010