Can Machine Learning and Geostatistics Overcome Lack of Data in Assessing Recovery of Water Levels and Ecological Conditions at Unmonitored Wetlands and Lakes?

Dan Schmutz, Stephanie K. Garvis, Danny Goodding, and Christopher Shea

Tampa Bay Water is the largest wholesale supplier of water in the state of Florida, serving more than 2.5 million people in the Tampa Bay Area (Figure 1). Some of the wellfields in Tampa Bay Water's system have been pumping groundwater for over 50 years. Groundwater continues to be a vital part of the Tampa Bay region's water supply, with more than 50 percent of the regional supply coming from wellfields (Tampa Bay Water, 2017). Though groundwater pumping has been drastically cut back, previous pumping from these wellfields contributed to lower water levels in some of the region's lakes and wetlands, which led to deleterious ecological changes.

In 1998, Tampa Bay Water gained ownership and control of all of the regional wellfields in the Tampa Bay area. The Southwest Florida Water Management District (SWFWMD) issued a new permit to Tampa Bay Water that consolidated the permits for 11 of these wellfields located in Pasco, northern Hillsborough, and northeast Pinellas counties. This new permit, known as the "consolidated permit," lowered the permitted annual average pumping limit for these 11 wellfields from 192 mil gal per day (mgd) to 90 mgd. Tampa Bay Water currently operates these wellfields as an interconnected system at this lower pumping limit to promote environmental recovery near the wellfields.

Dan Schmutz, M.S., is chief environmental scientist, and Stephanie K. Garvis, M.S., and Danny Goodding, M.S., are environmental scientists with Greenman-Pedersen Inc. in Orlando. Christopher Shea, M.S., P.W.S., is a senior environmental analyst with Tampa Bay Water.

Background

Tampa Bay Water is required by Special Condition 11 of Water Use Permit (WUP) No. 20011771.001 to evaluate "the recovery of water resource and environmental systems attributable

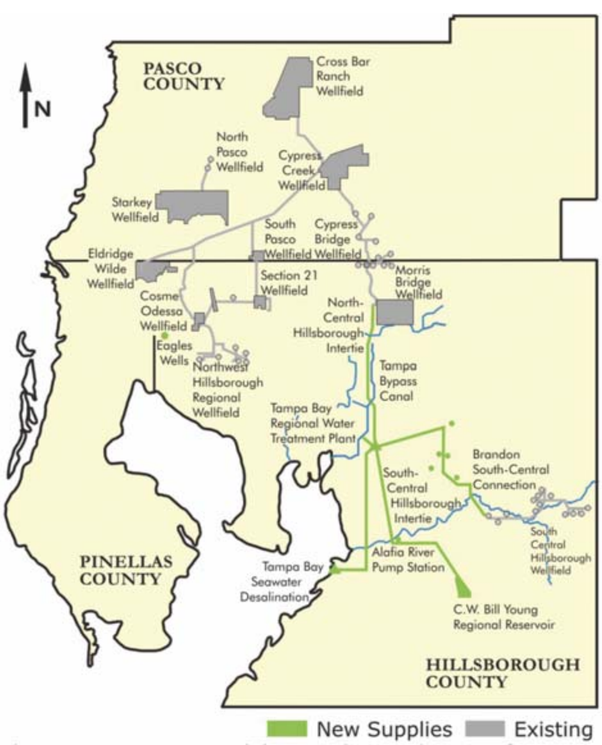

Figure 1. Tampa Bay Water existing groundwater and newer surface water supplies.

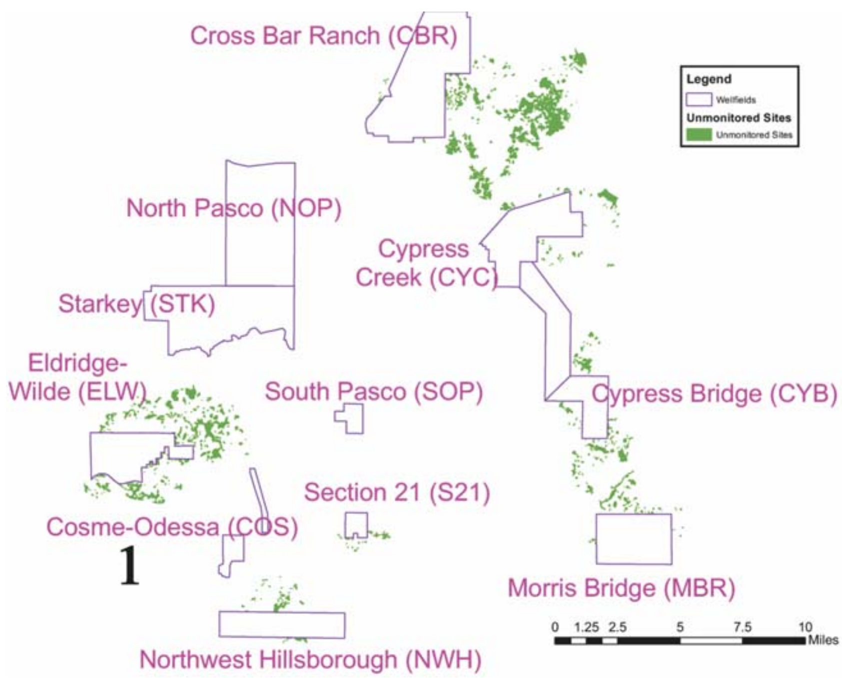

Figure 2. Unmonitored sites of concern and Tampa Bay Water wellfields.

to reduction of…withdrawals…to 90 mgd." As described in the permit recovery assessment work plan and schedule (Tampa Bay Water, 2012), a key issue to be resolved is how "wetland health (or recovery) criteria may be applied to wetlands lacking hydrologic data" (i.e., "unmonitored" wetlands). Tampa Bay Water (2013) previously defined areas of investigation for the recovery analysis using hydrologic modeling output and geographic information systems (GIS) analyses to produce a composite 2-ft surficial aquifer system (SAS) drawdown (DDN), representing the maximum of historical pumpage and several possible future scenarios, both scaled to 90 mgd.

A recovery assessment GIS deliverable previously prepared by Greenman-Pedersen Inc. (GPI), and provided to Tampa Bay Water in January 2016, found 684 unmonitored sites occurring within the 2-ft SAS DDN contour, consisting of 675 wetlands and nine lakes, shown in Figure 2. (Subsequent analyses, based on revised modeling output available after the completion of this study, have increased the number of unmonitored sites to 749, and results for the entire set will be reported in a future publication.) Tampa Bay Water is being assisted by GPI in developing proposed methods for estimating ecological and hydrological conditions at unmonitored sites. This article describes these proposed recovery assessment methods, as well as preliminary results of analyses performed to facilitate method development.

Following approval of the proposed methods, it's anticipated that they will be applied in a future project phase to aid in the assignment of unmonitored sites to appropriate recovery assessment (RA) status bins. These bins document wetland/lake conditions relative to approved recovery metrics, as well as evidence of trends towards recovery, and they consist of the following categories:

- Never impacted
- No cutback, meets metric
- Recovered
- Improved, not fully recovered
- Not fully recovered, continuing wellfield impact
- Impacted due to other causes
- More detailed analysis needed

Conceptually, the problem of assessing unmonitored wetlands is one of statistical interpolation. Specifically, there is a need to develop defensible approaches for transferring information from nearby sites with known recovery assessment statuses to unmonitored sites. The term "nearby" might imply physical proximity, but it could also imply proximity in multivariate space (i.e., statistical "nearest neighbors"). Determination of appropriate spatial support (e.g., "Over what distance is recovery status correlated?") is a subset of the overall problem of determining how recovery varies among sites that are close in a statistical sense. Development of statistical models to allow inference at unmonitored wetlands requires the development of adequate datasets from nearby monitored wetlands collected from an appropriate time period (e.g., the period after wellfield production was reduced to 90 mgd, or postcutback).

In seeking to develop these methods, it's anticipated that the methods might vary by wetland community type (e.g., isolated cypress) or surrounding soil classification (i.e., xeric or mesic) based on previous findings that water levels in wetlands in different soil settings behave differently (GPI, 2016). Methods also were anticipated to be hierarchical in nature, meaning that broad screening tools would be proposed as a first cut to classify, or bin, wetlands using the least amount of new data collection.

The unmonitored wetlands and lakes of concern occur primarily in eight regions within the northern Tampa Bay (NTB) area associated with various wellfields (Figure 2):

- Eldridge-Wilde (ELW)
- Northwest Hillsborough (NWH)
- Section 21 (S21)
- Morris Bridge (MBR)
- Cypress Bridge (CYB)
- Cypress Creek (CYC)
- Cross Bar Ranch (CBR)
- CYC/CBR interwellfield area

More than 400 wetlands, lakes, and connected systems with water-level data (i.e., monitored sites) also occur in and around these areas, providing a potential basis for statistical interpolation to the unmonitored sites. Some ecological data are available for selected unmonitored sites, including wetland health assessment (WHA) data from previous SWFWMD studies. This present study also investigates the utility of several regional datasets believed to have some potential for supporting predictions of wetland water levels at unmonitored sites, including SAS DDN, Upper Floridan aquifer (UFA) DDN, and surrounding soil classification.

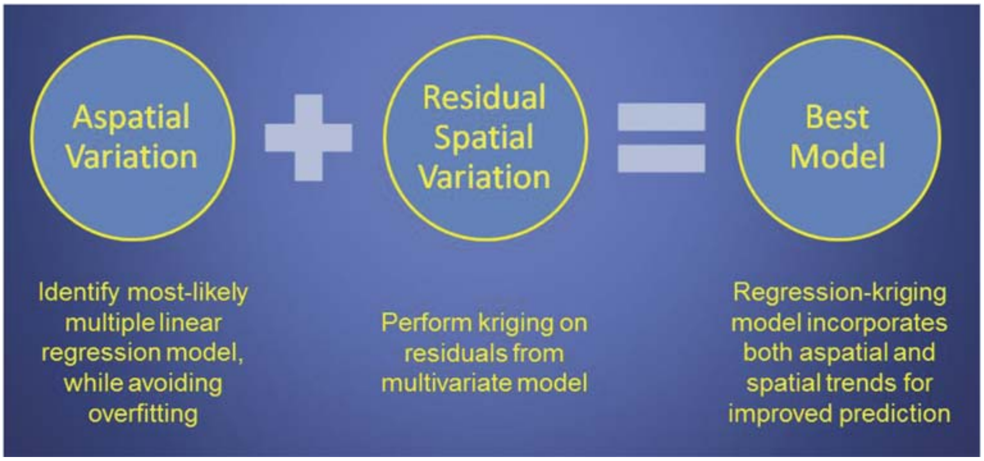

Figure 3. Conceptual model of regression-kriging approach to predicting data at unmonitored sites.

Given the large number of potential variables that might be useful for interpolating water levels at unmonitored sites, the discipline of machine learning, a field of computer science that develops algorithms that can learn from and make predictions on data (Alpaydin, 2009), was used to help develop robust statistical models, which are those that are expected to perform well on new data, meaning they are not "overfit" to unique aspects of a particular dataset. The danger in overfitting is that a model that appears to perform well on past data (or data located in certain spatial areas) will perform poorly in the future (or in other unstudied spatial locations) because the analyst mistakenly fit the model to noise rather than signal in the development dataset (James et al., 2017).

In the past three decades, there has been increasing recognition of the value of hybrid spatial interpolation approaches to predicting values at unmonitored locations. These hybrid techniques combine two conceptually different but complementary techniques: analysis by multiple linear regression (or other machine learning algorithm) to predict based on aspatial auxiliary variables, and spatial interpolations based solely on values of the known points and their spatial autocorrelation characteristics (e.g., kriging).

Typically, these hybrid techniques provide more accurate predictions than either single approach (Hengl et al., 2007). Hybrid spatial interpolation techniques have been applied successfully to problems such as water-table mapping (Desbarats et al., 2002), interpolation of soil properties (Odeh et al., 1994; Hengl et al., 2015), and estimation of rainfall (Pardo-Iguzquiza, 1998), among many others.

Continued on page 38

Continued from page 37

Regression-kriging (RK) is a general term used to describe a hybrid spatial interpolation technique that may involve the separate fitting of an aspatial model (e.g., multiple linear regression, random forest, etc.) and subsequent kriging of residuals from the model (Hengl et al., 2003). Key assumptions of the technique are that residuals are normally distributed with constant variance and the values of the auxiliary variables are known at all locations needed for prediction.

Methods

In this study, RK was used for development of the best model to predict historical median (2008-2014) wetland water levels at unmonitored sites by using a combination of aspatial and spatial information from nearby monitored sites. The technique required development of the best possible multiple linear regression model (i.e., selecting most predictive variables while avoiding overfitting), and then examining the residuals from that model for spatial autocorrelation. If positive spatial autocorrelation was found to be present, then kriging of the residuals would be expected to result in improved predictions for the unmonitored sites (Hengl, 2009). In other words, if there is a tendency for the initial aspatial model to predict median wetland water levels too high or too low in spatial clusters, it's implied that unmeasured but spatially autocorrelated factors—not included in the initial model—are affecting the outcome. Although those factors aren't known, their effect can be modeled at unmonitored sites using the technique of kriging, which is a geostatistical procedure that predicts how similarity in the residuals changes with distance. Kriging estimates deviations at unmonitored sites by a weighted averaging of nearby residuals. With positive spatial autocorrelation, residuals near each other will tend to be more similar than residuals farther apart.

In summary, RK involves spatially interpolating residuals from an aspatial model using kriging and adding the results to the predictions from the aspatial model. Conceptually, the aspatial regression predictions for unmonitored sites will be adjusted up or down based on the residual deviations of nearby sites (Figure 3).

The focus of hydrologic prediction at the unmonitored sites was the median water level for years 2008-2014 relative to a high water mark at each site known as the historical normal pool (HNP). The offset of the median water level relative to the HNP—known as the normal pool offset (NPO)—was selected to represent the most appropriate surface water hydrologic correlate of wetland health, based on the results of scientific investigations performed in the NTB area, including the development of minimum levels for isolated cypress-dominated systems (SWFWMD, 1999). The NPOs represent the "offset" of wetland water levels from a reference elevation of historic inundation (i.e., the normal pool). The HNP offsets can be related to historic conditions and they are not affected by differing wetland depths, unlike hydroperiods. Also, the use of a local wetland reference datum, like HNP, allows for the analysis of a large number of sites occurring across a wide range of absolute elevations on a common scale (i.e., ft below HNP). The seven-year time period of 2008-2014 was considered appropriate because it represented a postcutback and operationally stable pumpage configuration, with a range of rainfall conditions, although this time period may be conservative (i.e., underpredicting long-term median wetland water levels), depending on the time frames involved in the reduction of groundwater pumpage on the local scale, and the subsequent aquifer response.

The interpolation methods involved:

- Identifying all sites (e.g., lakes, isolated wetlands, and connected wetlands) with HNP elevations and water-level data for the period of 2008-2014.
- Excluding or truncating erroneous data.
- Determining appropriate groups for interpolation.
- Determining auxiliary variables potentially useful for aspatial prediction of NPOs.
- Developing the best aspatial multiple linear regression model using an information criterion-based search through all possible models.
- Performing RK to fit a variogram model to explain spatial-autocorrelation in residuals from the aspatial multiple linear regression model.

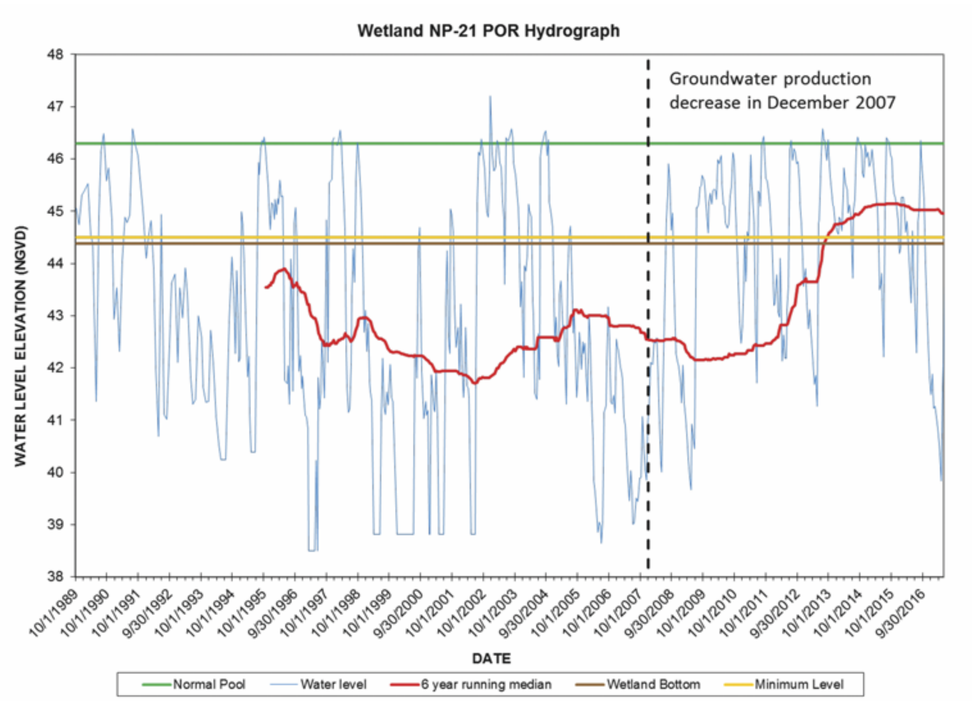

Figure 4. Hydrograph documenting increasing six-year median water levels following production cutbacks.

A small number of sites were excluded due to groundwater augmentation or data problems. There were 309 monitored sites (wetlands and lakes) with adequate data to calculate a median offset for 2008-2014 for use in developing a model to predict water levels at the unmonitored sites. Wetland and lake water levels for monitored sites primarily were obtained from either Tampa Bay Water or SWFWMD using their applications (DataMart and work management information system [WMIS], respectively). Appropriate care was taken to ensure that a common vertical datum was used, as some records were available in both National Geodetic Vertical Datum (NGVD)29 and North American Vertical Datum (NAVD)88 through WMIS. (An Excel spreadsheet of mean monthly lake levels was provided by Brian Ormiston and Claudia Listopad.)

Wetland and lake NPOs were calculated by subtracting HNPs from the median water levels for the period 2008-2014. Medians were calculated using all available data for the period of interest. Visual examination of hydrographs was

Continued on page 40

Table 1. Independent variables evaluated for use in predicting wetland/lake median normal pool offsets.

Variable	Description	Source/Process
SAS DDN	Modeled drawdown in Surficial Aquifer System	12 nearest neighbor inverse distance squared weighted interpolation derived from a point file provided by Tampa Bay Water, representing the maximum of the "Historical Production and Scaled Pumpage" scenarios described in Tampa Bay Water (2013).
UFA DDN	Upper Floridan Aquifer drawdown (change in feet)	Differenced two rasters (recent minus past): 1) pre-development surface from "Digital Surfaces and Hydrogeologic Data for the Floridan Aquifer System in Florida and in Parts of Georgia, Alabama, and South Carolina By Jason C. Bellino" (Bellino 2011), and 2) prepared medians from Jan 2008-Dec 2009 from Lee and Fouad (2014), "Creating a monthly time series of the potentiometric surface in the Upper Floridan aquifer".
IA Thickness	Thickness of Intermediate Aquifer System	File previously prepared for the Florida Aquifer Vulnerability Assessment (Arthur et al. 2005)
Head Difference	Hydraulic head difference between the Surficial Aquifer System (SAS) and the Floridan Aquifer System (FAS)	File previously prepared for the Florida Aquifer Vulnerability Assessment (Arthur et al. 2005)
Xeric Ratio	Percentage of soils supporting xeric vegetation communities in a 500-feet wetland buffer relative to xeric and mesic vegetation-supporting soils combined	A classified soils layer from a xeric wetlands study (GPI 2016) was used to determine Xeric Ratio using the methods described in Berryman & Henigar, Inc. and SDI Environmental Services, Inc. (2000).
Xeric Y/N	Binary designation of xeric-soil association	A binary thresholding of Xeric Ratio in which Xeric Ratios greater than 27% were considered "Y" (i.e., yes they are xeric-associated sites), while those with less than or equal to 27% were considered "N".
Soil Permeability	Permeability values (in/hr)	File previously prepared for the Florida Aquifer Vulnerability Assessment (Arthur et al. 2005)
Distance to Near Well	Distance to nearest production well	Calculated using the Euclidian distance function in the ArcGIS Spatial Analyst extension
Kernel Density of Wells	Density of production wells	ArcGIS was used (with default settings) to calculate a kernel density of production wells.
Area Perimeter Ratio	Ratio of wetland or lake polygon area to its perimeter	Calculated using ArcGIS
Acres	Area of wetland or lake polygon	Calculated using ArcGIS
Rainfall	Mean rainfall for 2008-2014	Calculated based on 11 gap-filled stations using a 10 nearest neighbor inverse distance squared weighted interpolation

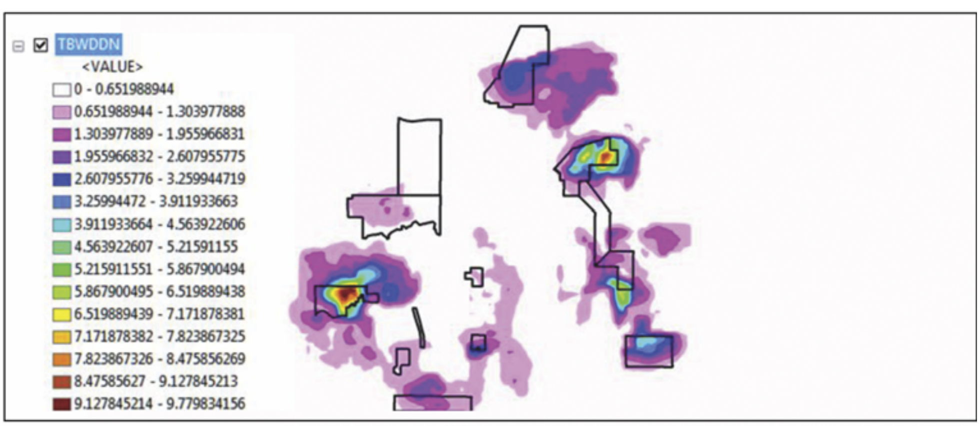

Figure 5. Surficial aquifer system drawdown (ft): 12 nearest-neighbors inverse-distance-squared weighted interpolation.

Continued from page 38

performed to confirm that the calculated medians would be representative (i.e., excessive dry values might prevent calculation of accurate medians). The HNPs were obtained from a variety of sources that had been compiled for the RA GIS (RAGIS) project (GPI and Applied Ecology, 2015). These HNPs were checked against an environmental management plan (EMP) database maintained by Tampa Bay Water.

Specific NPOs are known from past studies (e.g., SWFWMD, 1999; GPI, 2016) to represent recovered conditions for monitored wetlands, depending on the type of wetland (e.g., isolated or connected) and the surrounding soil type (e.g., mesic or xeric). For example, Figure 4 shows that, following the decrease in local groundwater production at the North Pasco Wellfield, the six-year running median water level rose above the surrogate minimum level (calculated as a NPO of -1.8 ft). In other words, when the six-year median is shown to rise above the relevant site-specific threshold NPO, the site is considered "low or no stress" and, therefore, hydrologically recovered.

A decision was made to transform the dependent variable, NPO, in order to improve the normality of residuals from the planned linear model. Specifically, the following transformation was used:

$$LN((NPO * -1) + 1)$$

Prior to applying the transformation, three sites with very small positive HNPs were adjusted to zero. (One interesting consequence of using a nonlinear transformation on the dependent variable is that prediction intervals generated from a fitted statistical model will not be symmetric when the prediction intervals are transformed back into their original scale.)

A variety of independent or auxiliary variables were obtained and prepared to provide the best possible aspatial portion of the RK model. These 12 independent variables are shown in Table 1.

Maps of several of the independent variables prepared and investigated (GPI, 2017) are provided here. The SAS DDN (Figure 5) was based on a 12 nearest-neighbors inverse-distance-squared weighted interpolation derived from a point file provided by Tampa Bay Water, representing the maximum of the historical production and scaled pumpage scenarios described (Tampa Bay Water, 2013). In Figure 5, positive numbers represent ft of predicted surficial aquifer DDN in the NTB area associated with anticipated distributions and rates of groundwater production (scaled to 90 mgd).

A classified soils layer (Figure 6) prepared from U.S. Department of Agriculture Natural

Resources Conservation Service (USDA-NRCS) soil survey geographic (SSURGO) data in a previous xeric wetlands study (GPI, 2016) was used to determine two variables: xeric ratio and xeric yes/no (Y/N). Xeric ratio was calculated using the methods described in Berryman & Henigar Inc. and SDI Environmental Services Inc. (2000). A 500-ft buffer was constructed around each wetland or lake of interest and the ratio of the following areas calculated:

$$\frac{Xeric}{Xeric + Mesic}$$

Xeric Y/Ns representing a binary thresholding of xeric ratios greater than 27 percent were considered "Y" (i.e., yes, they are xeric-associated sites), while those with less than or equal to 27 percent were considered "N" or no. The splitting criterion of 27 percent was based on methods presented in Berryman & Henigar Inc. and SDI Environmental Services Inc. (2000).

Mean rainfall for 2008-2014 was calculated based on 11 gap-filled stations using a 10 nearest-neighbors inverse-distance-squared weighted interpolation (Figure 7). All rainfall data were obtained from Tampa Bay Water, except for 26353 CLERMONT 9 S NWS, which was provided by the National Oceanic and Atmospheric Administration (NOAA). The 11 rainfall station locations used were:

- RN-CBR-CB01
- RN-STK-STK14
- RN-NOP-NOP
- RN-NHW-5
- RN-MBR-3C
- RN-NWH-S21
- RN-ELW-METER_PIT
- RN-CYB-CYB7
- RN-CYC-CC3
- RN-CNR-T3
- 26353 CLERMONT 9 S NWS

A best subsets regression search was undertaken using the Bayesian information criterion (BIC) to avoid overfitting. Minimization of the BIC allowed identification of the model in a set of candidate models that gave the best balance between model fit and complexity, with the intent that whatever variables resulted in the most probable model from a BIC perspective would also yield valid predictions using out-of-sample data in the future. The best subsets regression, performed using the glmulti package in R (Calcagno and de Mazancourt, 2010), consisted of fitting all possible regression models without interactions (more than 4,000 models) using all various combinations of the 12 independent variables. The final selected variables were used in a multiple linear regression, allowing interactions (i.e., the final aspatial model).

Spatial autocorrelation remaining in the residuals from the final aspatial model was modeled using a variogram as part of the RK process. At a conceptual level, kriging represents a family of interpolation techniques that explicitly model the spatial autocorrelation in the data using a technique called the variogram (Figure 8). The variogram is an empirical measure of how differences between pairs of sample points change

Continued on page 42

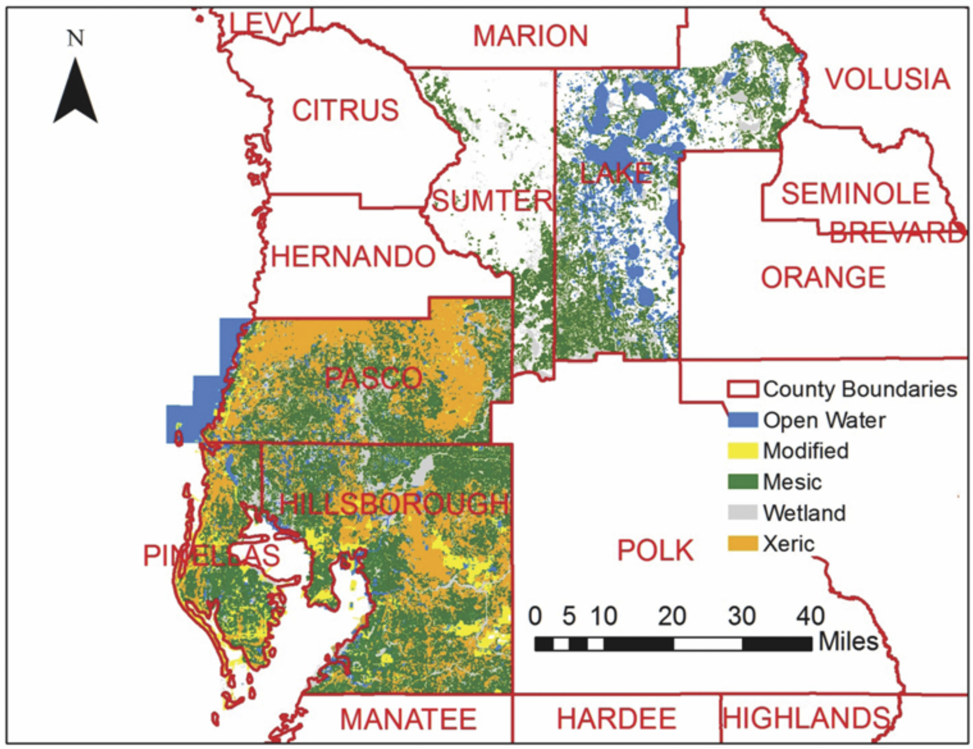

Figure 6. Classified soils layer from previous xeric wetlands research.

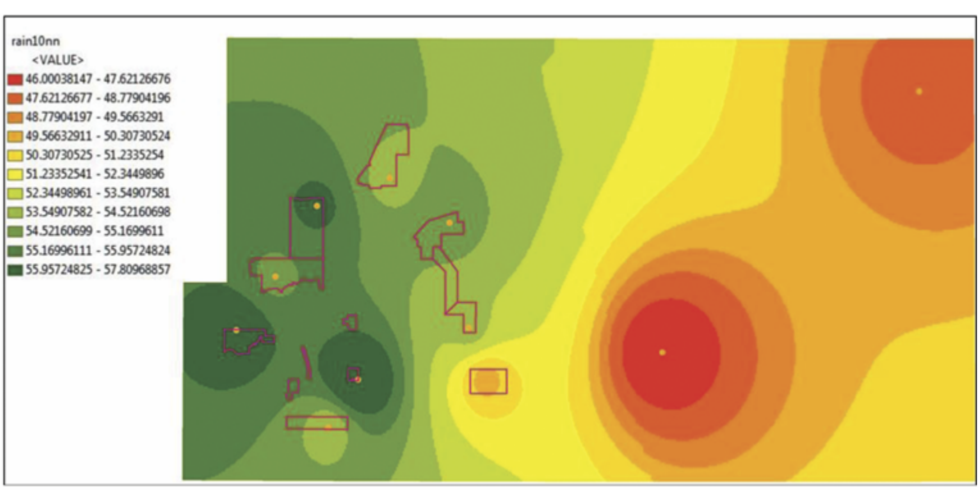

Figure 7. Mean rainfall for 2008-2014 (in.); orange dots represent rain gauges.

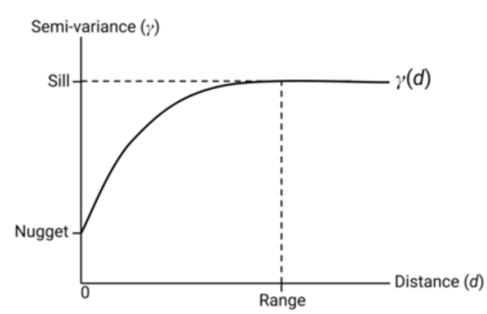

Figure 8. Diagram of an idealized variogram *(Google, 2016)*.

Continued from page 41

with distance. The typical plotted variogram function (the semivariance) rises as distance increases (Figure 8), indicating that the average differences between values at locations in the study area are smaller when the samples are close together and much larger when they are farther apart. The point where the semivariance function flatlines represents the distance (range) at which positive spatial autocorrelation is negligible. Note that a portion of the variance, called the nugget, cannot be explained based on proximity to other samples (Figure 8); it could, however, represent some combination of variability at distances smaller than the typical sample distance, measurement error, or variability due to unmeasured factors. The variogram provides a method for weighting the influence of nearby sites in estimating unknown values (i.e., a data-driven estimate of the spatial covariance between samples).

The focus of ecological condition interpolation was the five-year WHA rating. The WHA program has been performed at more than 400 wetlands in the NTB area in four general time periods: 1997/1998 (Rochow, 1998), 2004/2005 (Bureau Veritas et al., 2006), 2009/2010 (GPI Southeast Inc. et al., 2010), and most recently in 2016. Among other data collected at each site, an overall WHA rating on a 1-3 scale is available. This rating represents a "relative estimation of wetland health" with sites rated 1 being considered "severely stressed," those rated 2 as "moderately stressed," and those rated 3 as "low or no stress."

Although this rating is ultimately based on the professional opinion of the environmental scientist performing the field evaluation, the opinion is supported by detailed information collected, such as which plant species were present at each site and various wetland-quality questions (e.g., "fallen trees": >25, 5-25, or <5 percent). The WHA ratings do not attribute a cause for observed stress, although relevant observations regarding land use, ditching, etc., are included on the field sheets.

Based on availability for this study, a file was prepared representing 2009 conditions to use as the basis for understanding ecological conditions across the study area. The following data were included in the 2009 conditions file: 423 WHA ratings actually collected in 2009/2010 (but 20 of those were outside study area), and 3 ratings only (n=472) from 1997/1998 at those sites where no 2009/2010 data existed. The rationale for including sites rated 3 from 1997/1998 in the 2009 conditions file is that conditions are unlikely to have worsened in most areas due to decreased production. The combined 2009 dataset provided 895 sites with a rating of 1, 2, or 3. In the present study, ecological conditions were interpolated across the NTB area using the inverse-distance-squared weighted interpolation in ArcGIS using the 12 nearest neighbors and then assigned to the unmonitored sites.

Hydrologic Results

Figure 9 represents the information criteria (IC) profile plot, or the ranked BIC values of the very best models evaluated in the all-subsets regression search through all possible combinations of the 12 candidate auxiliary variables. Note that "best" in this context implies that the lowest BIC value represents the most likely model, minimizing overfitting, given this particular dataset. The red line in Figure 9 is placed two BIC units above the very best model, based on the rule of thumb that any models within two BIC units of the best one are worth considering (Calcagno and de Mazancourt, 2010). When faced with multiple "best" models, investigators may wish to explore a model-averaging approach, also known as multimodel inference (Burnham and Anderson, 1998). In this case, only the very best model was chosen to be implemented (lowest BIC).

The best subsets regression search among all possible combinations of the 12 auxiliary variables resulted in just three variables being included in the BIC best model: SAS DDN, xeric ratio, and intermediate aquifer (IA) thickness with a minimum BIC of 392.7 (Figure 9). Although referred to as IA thickness by the creators of the Florida aquifer vulnerability assessment (FAVA) dataset, it's been noted that much of the NTB area lacks an intermediate aquifer, and this variable might be more appropriately reduced to a simple presence/absence of a confining layer.

The final aspatial model selected used the three variables identified by the BIC best subsets search—SAS DDN, xeric ratio, and IA thickness—and allowed for interactions among them. This final model had an adjusted R-squared of 0.33 (Figure 10). Visualizations of the partial residuals were performed using the R package visualization of regression models (visreg) from Breheny and Burchett, 2016. The visualizations documented the modeled effects of each independent variable, while statistically holding other variables constant (either at their median values or two selected values to allow visualization of interactions).

The SAS DDN showed the expected relationship, with larger negative NPOs associated with larger DDNs in Figure 11 (larger negative NPOs were represented as larger positive values on the transformed scale). Higher xeric ratios were also associated with larger negative NPOs

Continued on page 45

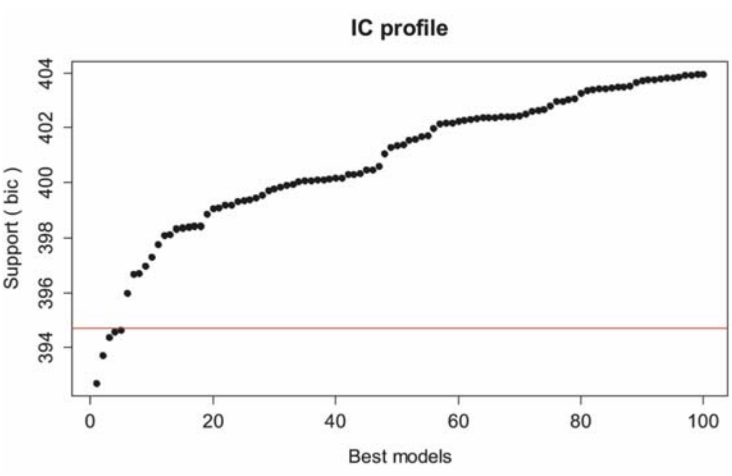

Figure 9. Bayesian information criteria profile plot from R package glmulti output.

```
Call:
lm(formula = npo2008trans ~ 1 + TBWDDN * RAXericRat * IAthicknes,
    data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.12722 -0.30295 -0.04353  0.31175  1.08242

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.818640   0.040503  20.212  < 2e-16 ***
TBWDDN                       0.099453   0.019041   5.223 3.29e-07 ***
RAXericRat                   0.364086   0.130451   2.791 0.005591 **
IAthicknes                   0.001234   0.002601   0.474 0.635608
TBWDDN:RAXericRat            0.370914   0.095028   3.903 0.000117 ***
TBWDDN:IAthicknes            0.003990   0.001465   2.724 0.006822 **
RAXericRat:IAthicknes        0.020237   0.012141   1.667 0.096588 .
TBWDDN:RAXericRat:IAthicknes -0.010934  0.006936  -1.576 0.115963
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4251 on 301 degrees of freedom
Multiple R-squared:  0.3479,   Adjusted R-squared:  0.3327
F-statistic: 22.94 on 7 and 301 DF,  p-value: < 2.2e-16
```

Figure 10. Final aspatial regression model R output.

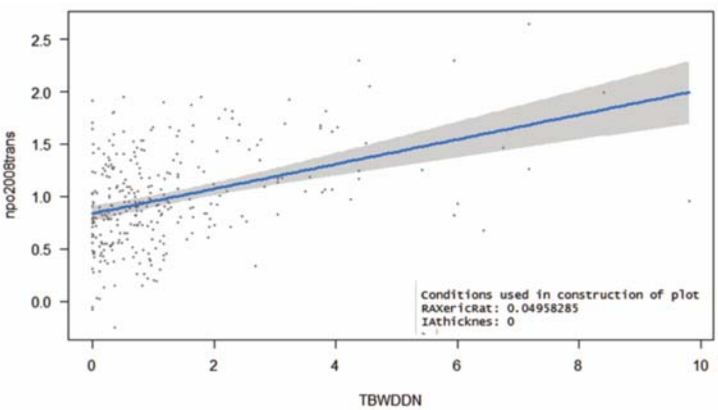

Figure 11. Partial residuals visualization: surficial aquifer system drawdown.

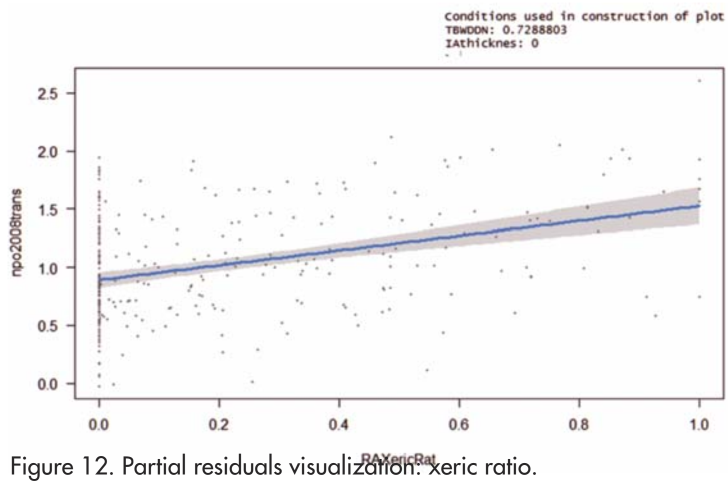

Figure 12. Partial residuals visualization: xeric ratio.

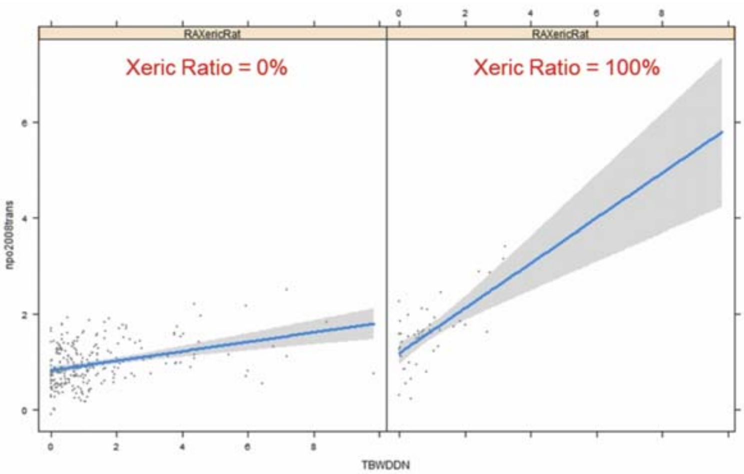

Figure 13. Partial residuals visualization: surficial aquifer system drawdown by xeric ratio.

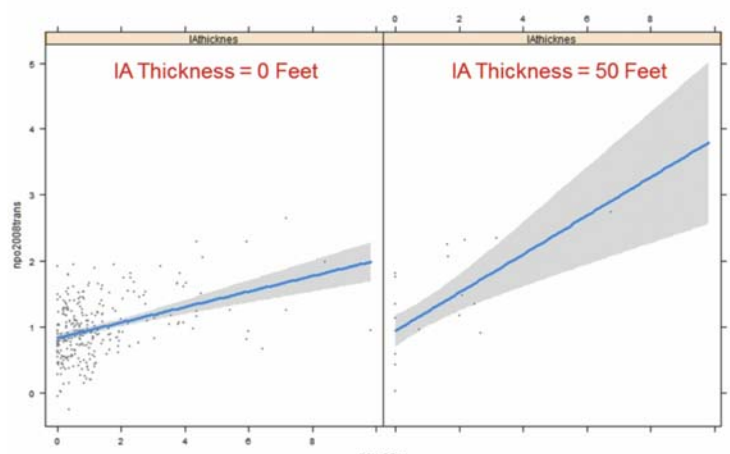

Figure 14. Partial residuals visualization: surficial aquifer system drawdown by intermediate aquifer thickness.

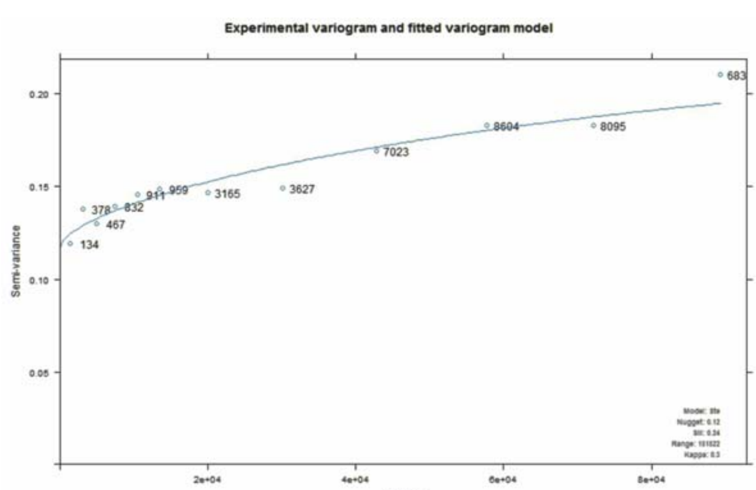

Figure 15. Experimental residual variogram and variogram model.

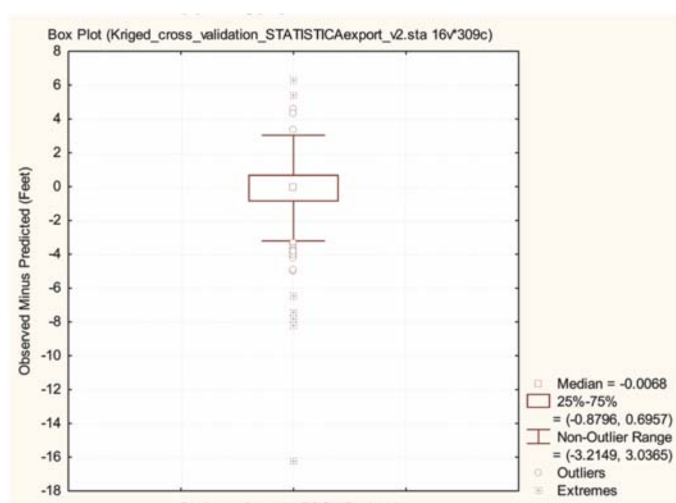

Figure 16. Cross-validated back-transformed regression-kriging residuals.

Continued from page 42
(Figure 12). The relationship between SAS DDN and NPOs showed an interaction with xeric ratio, so that sites with 100 percent xeric ratios tended to have greater negative NPOs at the same magnitude of SAS DDN (Figure 13). The relationship between SAS DDN and NPOs also showed an interaction with IA thickness: sites with IA thickness greater than zero showed greater negative NPOs at the same magnitude of SAS DDN (Figure 14).

The RK was used to model spatial autocorrelation in the residuals remaining after the final selected aspatial model. More specifically, an experimental variogram and variogram model were prepared from the residuals from the final aspatial model (Figure 15) using an autofitting procedure available in the R package automap (Hiemstra et al., 2009). The variogram shows a sill with a semivariance of 0.24 and a nugget of 0.12. The implication of this is that about half of the residual semivariance after the regression model may be explained by spatial autocorrelation. If this residual variogram had been just a flat line, the residuals would not have been kriged, as there would be no improvement in prediction over the final aspatial model.

As expected, the RK predictions showed a higher correlation with the actual NPOs than the simple aspatial model, as the adjusted R^2 for the selected aspatial mode (Figure 10) was about 33 percent. The squared correlation coefficient for the RK approach (i.e., the coefficient of determination) was found to be 52 percent, representing an improvement from the aspatial model; however, the performance on data not used in the development dataset could be worse. To understand future expected performance on unsampled areas, a more realistic estimate of future performance was prepared for the RK model by performing a leave-one-out cross-validation (LOOCV). The LOOCV technique involves refitting the model while leaving out one observation, then deriving an estimate of that one observation as an out-of-sample case.

The procedure is performed for all observations in turn, resulting in an estimate of error closer to that expected for a completely novel future dataset (or in this case, hypothetical performance at unmonitored sites located between monitored sites). Although the LOOCV R^2 was lower (36 percent) than the overly optimistic model (52 percent), and some predictions were far from the observed data (Figure 16), the majority of LOOCV residuals were within 1 ft. Specifically, 50 percent of the time the RK model is expected to predict median water levels at unmonitored wetlands within a range of 0.70 ft lower than the actual value to 0.88 ft higher than

Continued on page 46

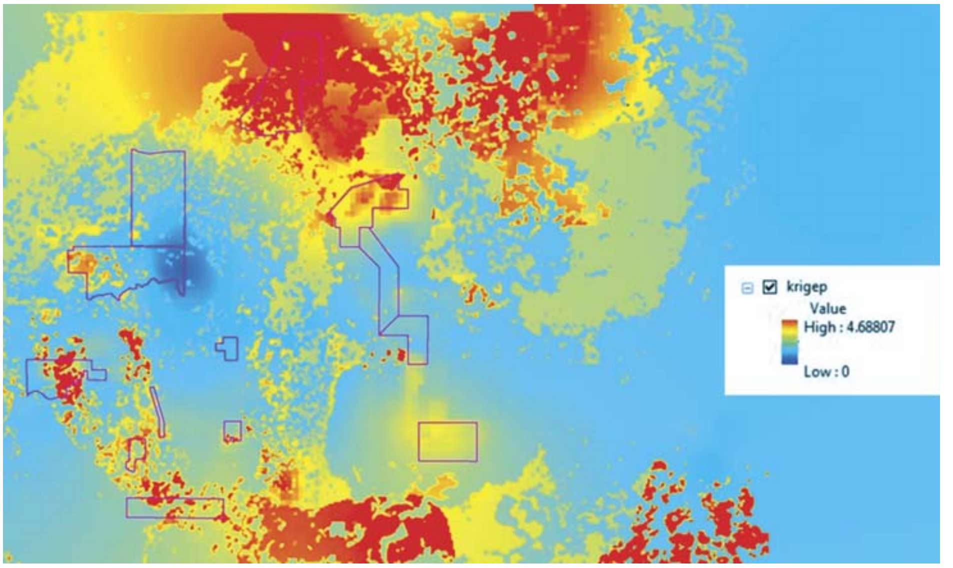

Figure 17. Regression-kriging predictions: transformed units.

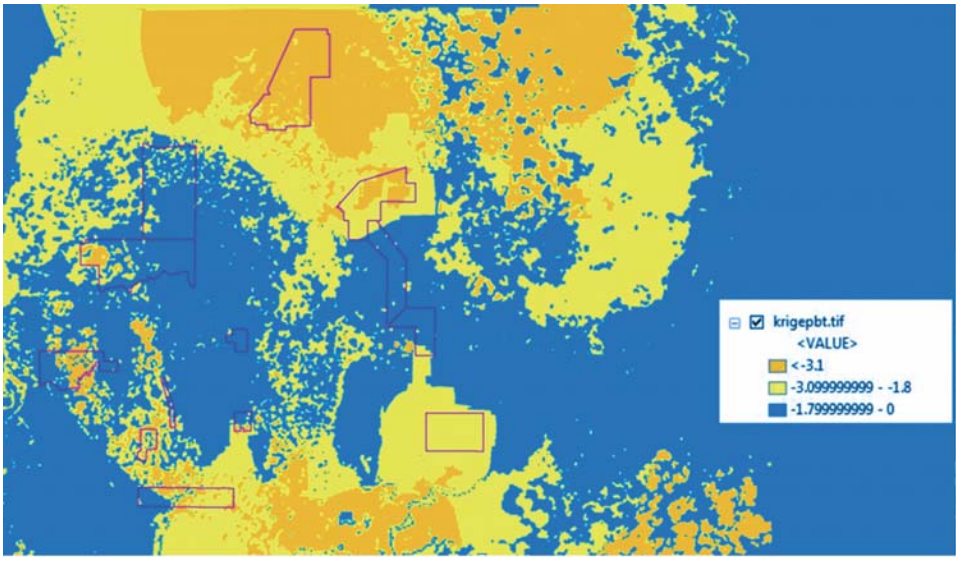

Figure 18. Regression-kriging predictions: back-transformed to feet.

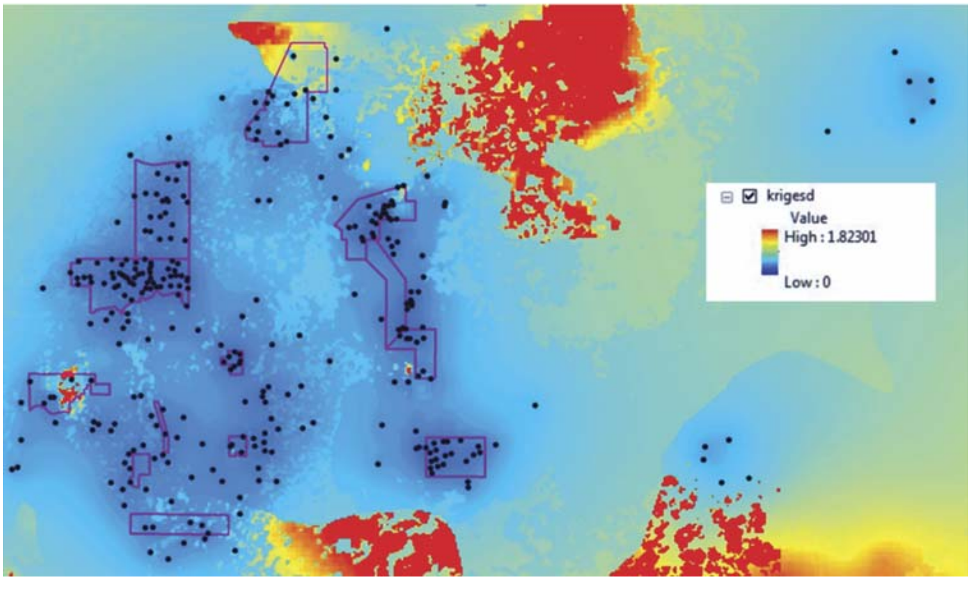

Figure 19. Regression-kriging standard deviations: transformed units (with overlay of monitored site locations).

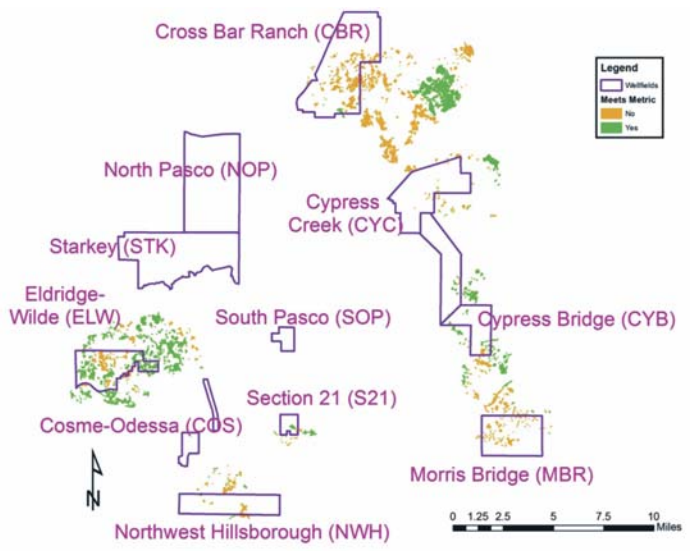

Figure 20. Regression-kriging predictions for 2008-2014 median wetland water levels transferred to unmonitored sites.

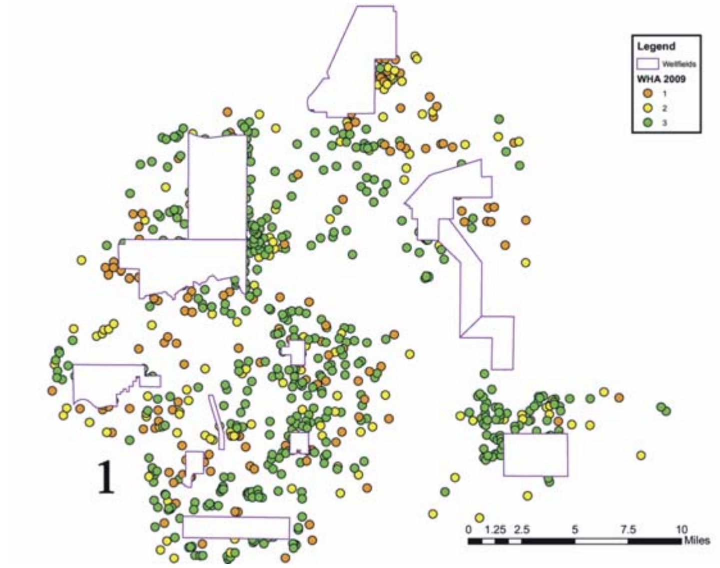

Figure 21. Wetland health assessment data: 2009 conditions as points.

Continued on page 45
the actual value. This level of accuracy is considered useful for wetland recovery planning purposes.

The RK predictions for the 2008-2014 median NPOs are presented in map form in transformed (Figure 17) and back-transformed (ft) units (Figure 18). The back-transformed prediction map shows areas in blue predicted to meet both the isolated cypress standard threshold of 1.8 ft below HNP, as well as the xeric wetlands screening criterion of 3.1 ft below HNP (e.g., much of CYB and South Pasco, and portions of Eldridge-Wilde). Other areas shown in yellow represent those areas predicted to meet the xeric wetlands criteria, but not the isolated cypress standard (e.g., Morris Bridge and portions of CYC). Areas shown in orange were predicted to have unmonitored site median water levels lower than 3.1 ft below HNP, including most of Cross Bar Ranch and parts of CYC and Eldridge-Wilde (i.e., areas with water levels not predicted to meet either threshold).

A map of RK prediction standard deviations is also provided from the RK analysis (Figure 19). Certain areas had very high uncertainty due to few sample sites nearby (shown in red), including some areas on the Eldridge-Wilde and Cross Bar Ranch wellfields. One of the advantages of the RK approach is that a spatially referenced measure of uncertainty is provided. In future implementations, it may be useful to screen out certain areas of high uncertainty from presentations of results, particularly those areas distant from the wellfields that have very few nearby development dataset observations.

When the appropriate thresholds by soil type are considered, overall 253 out of the 684 unmonitored sites (37 percent) are predicted to have met their site-specific thresholds based on RK predictions for the 2008-2014 period (Figure 20). Certain areas showed a high degree of heterogeneity in recovery predictions, such as the areas around the Northwest Hillsborough Wellfield and the southern part of CYB, while other areas tended to be more uniform in showing nearly all unmonitored sites as not recovered

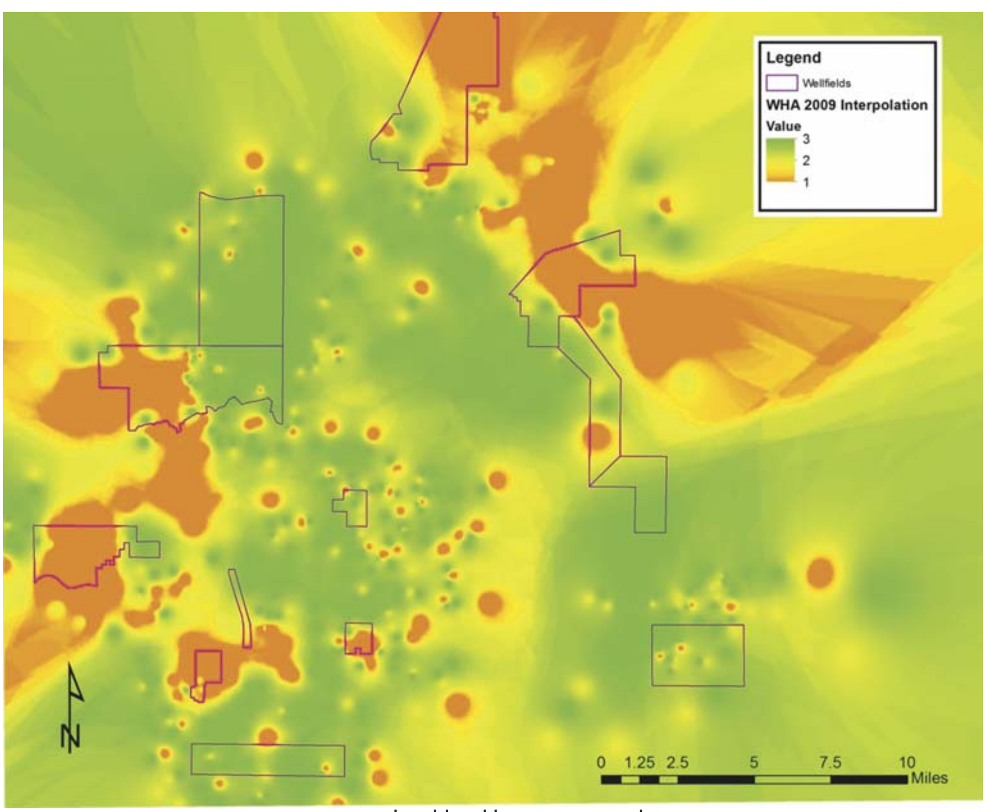

Figure 22. Wetland health assessment data: 2009 12 nearest-neighbors inverse-distance-squared interpolation.

Table 2. Three-group classification matrix for 5 percent held-out sample for 12 nearest-neighbors inverse-distance-weighting interpolation.

		Predicted		
		1	2	3
Actual	1	5	3	3
	2	1	9	4
	3	0	8	18

Table 3. Two-group classification matrix for 5 percent held-out sample for 12 nearest-neighbors inverse-distance-weighting interpolation.

		Predicted	
		S	NS
Actual	S	18	7
	NS	8	18

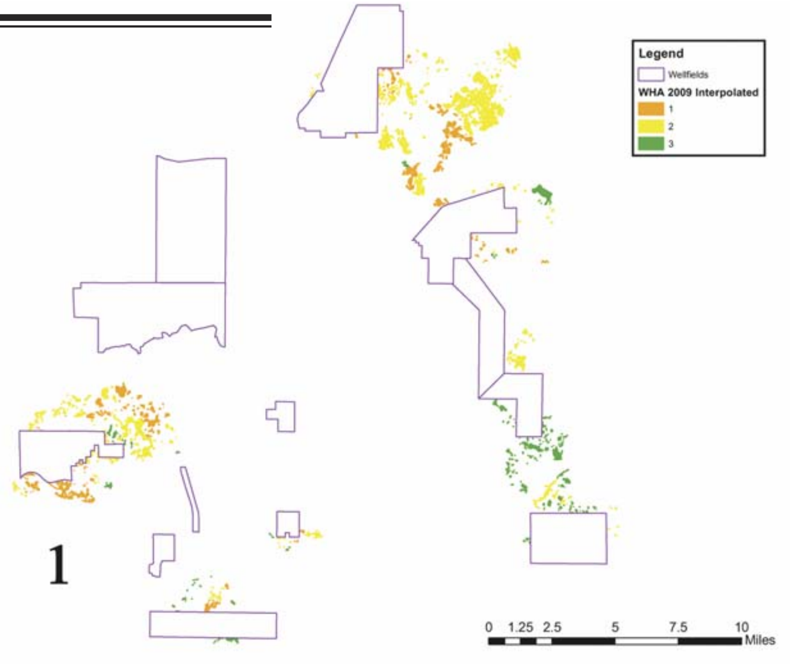

Figure 23. Interpolated wetland health assessment data from 2009 conditions assigned to unmonitored sites.

(Morris Bridge and much of Cross Bar Ranch) or nearly all as recovered (central CYB). The Eldridge-Wilde Wellfield showed a radial pattern of predicted recovery around the periphery of the wellfield.

Ecological Results

Ecological conditions as expressed by the WHA three-point rating for 2009 are presented for 868 wetland points located in the NTB area in the vicinity of the 11 central system Tampa Bay Water wellfields in Figure 21 (with the points representing geographic centroids, constrained to be included within the shape of the original polygon).

The use of a relatively simple spatial interpolation algorithm was investigated to reproduce WHA 2009 values for unmonitored wetlands—inverse distance weighting (IDW). The IDW algorithm calculates unknown locations as a weighted average of nearby points, with the weighting function being a function of the distance. Application of a 12 nearest-neighbors-squared IDW algorithm to the points in Figure 21 (WHA 2009 conditions) yielded Figure 22, where certain areas appear to be dominated by severely stressed conditions (orange color), such as western and central ELW, while other areas appear dominated by low- or no-stress conditions (green color), such as eastern Starkey (STK) and southern CYB. Some areas in between the wellfields show a speckled appearance, reflecting abrupt transitions in site WHA ratings for sites relatively near each other.

In order to evaluate how useful the IDW approach might be for assessing WHA values at completely unmonitored sites, an out-of-sample evaluation was done. The evaluation was accomplished by randomly selecting 51 of the known data points (approximately 5 percent) to be excluded from an interpolation, and then evaluating how well those missing points could be interpolated from their 12 nearest neighbors. Table 2 is a classification matrix, also referred to as a confusion matrix, representing a comparison of the actual and predicted classifications of the 51 out-of-sample points.

The overall accuracy of the classification is based on summing the diagonal elements of the table (i.e., 5, 9, and 18) and dividing by the total count of 51. Therefore, the 12 nearest-neighbors IDW achieves an overall accuracy of 63 percent (i.e., 32 divided by 51). A more useful classification measure, however, would be the ability of the approach to identify sites as either stressed (WHA of 1 or 2) or nonstressed (WHA of 3). The classification matrix for this simpler two-group case is presented in Table 3. The 12 nearest-neighbors IDW accurately classified the out-of-sample points 71 percent of the time (36 divided by 51).

This expected out-of-sample performance of the stressed/nonstressed WHA interpolation of 71 percent represents a useful level of accuracy. As an example, based on the 868 WHA 2009 points occurring in the NTB Area, 327 were considered stressed, for a background rate of stress of 38 percent (327 divided by 868). If there were no other information about an unmonitored site, it could be concluded that there was a 38 percent chance it would be rated stressed in 2009; however, if the simple 12 nearest-neighbors IDW spatial interpolation model indicates that the site is stressed, the positive predicted value can be calculated as the ratio of true positives (18) divided by the sum of the true positives and the false positives (i.e., the number of predicted positives or 26), or 69 percent. This distance-weighted model has substantially increased the chance of correctly classifying a site as stressed (from a "chance alone" rate of 38 percent to an "interpolated predicted" rate of 69 percent).

It's possible that optimization could be used to further improve the accuracy of the IDW approach. Specifically, there are two primary parameters required by the IDW algorithm (the number of neighbors and the distance-weighting). This example shows the result of using 12 nearest neighbors with a weighting factor of inverse distance squared. Cross-validation could be used to select optimal IDW parameters by iteratively holding out individual known observations and estimating them as "test" observations. Evaluation of the cross-validated errors for these temporary test observations could be used to select minimal test errors out of a matrix of parameters investigated.

It's uncertain the extent to which the IDW model accuracy could be improved through optimization, but it's unlikely that the values selected by professional opinion are the most optimal, so 71 percent accuracy may be considered the minimum binary classification accuracy achievable through IDW. An alternative to optimizing the IDW parameters would be using an RK approach similar to that used for the hydrologic analysis

Continued on page 48

Continued from page 47
presented in the previous section, and this is what is intended for future, similar investigations.

Interpolated values from the 12 nearest-neighbors IDW models were spatially assigned to the unmonitored sites, and the results are presented in Figure 23. Actually, 81 of the 684 unmonitored sites (12 percent) had an existing WHA for 2009, so these ratings superseded the interpolated value. A review of the resulting map suggests that nonstressed unmonitored wetlands were concentrated in south CYB, much of MB and NWH, and portions of eastern ELW and western CBR. Other areas in 2009 showed moderate to severe stress conditions.

Summary and Conclusions

A machine-learning approach was effective in screening a large number of variables and developing a spatially explicit estimate of median wetland water levels at unmonitored wetlands and lakes in the NTB area. A hybrid spatial interpolation technique, RK was investigated for interpolating wetland water levels at unmonitored sites. Hybrid spatial interpolation techniques tend to provide more accurate predictions than either individual approach by itself (i.e., regression or kriging).

To demonstrate the RK approach, an information theoretic-driven best-subsets regression was first used to develop the best possible aspatial regression model to predict historical median (2008-2014) NPOs at 309 monitored sites with appropriate data. Three variables useful for predicting water levels were selected from the 12 evaluated: SAS DDN, xeric ratio, and IA thickness (i.e., presence/absence of confining unit). Once the best aspatial regression model was developed, residuals from the model were kriged to statistically characterize their spatial autocorrelation.

A cross-validation approach was used to assess the likely performance of the RK model on future datasets. Performance of the final model was considered useful, with cross-validated residuals indicating that at least half the time the combined RK model is expected to predict water levels at unmonitored wetlands within a range of 0.70 ft lower than the actual value to 0.88 ft higher than the actual value.

When the appropriate thresholds by soil type are considered, overall 253 out of the 684 unmonitored sites (37 percent) met the appropriate hydrologic metric, based on RK predictions for the 2008-2014 period. Certain areas showed a high degree of heterogeneity in median wetland water-level predictions, such as the areas around the Northwest Hillsborough Wellfield and the southern part of CYB, while other areas tended to be more uniform in showing nearly all unmonitored sites as not meeting the metric (Morris Bridge and much of Cross Bar Ranch), or nearly all as meeting the established metric (central CYB). The Eldridge-Wilde Wellfield showed a radial pattern of predicted recovery around the periphery of the wellfield.

The RK framework provides the flexibility to include other data-mining algorithms in the "regression" portion. For example, Appelhans et al. (2015) studied 14 machine-learning algorithms and kriging for interpolating air temperature on Mt. Kilimanjaro and found that numerous tree-based modern data-mining algorithms outperformed both linear regression models and universal kriging, including stochastic gradient boosting, cubist, and random forest. Ultimately, the authors elected to use an RK framework with the cubist model in place of regression (Appelhans et al., 2015). It was recommended that one or more modern data-mining algorithms be tested in a future study for comparison to the multiple linear regression approach used in this study. Although prediction using a different algorithm may be improved, it's likely that interpretability of the effects of the various auxiliary variables will suffer, but prediction is more important than interpretability for the application of estimating NPOs at the unmonitored sites. If multiple predictions are available from independent techniques, each with their own measure of error, they may be combined by weighting their predictions for each cell by the inverse of their errors (Hengl, 2009). Intuitively, diverse algorithms that perform relatively poorly in different geographic areas could be combined to yield overall improved predictions throughout the region of interest.

The five-year WHA program provides a spatially and temporally rich dataset able to support interpolation of wetland conditions, and potentially, changes in conditions over time. The WHA scores representing 2009 conditions documented at 868 sites (and including nonstressed sites from 1997/1998) were used with an IDW algorithm to successfully interpolate ecological conditions at unmonitored sites. Using an out-of-sample analysis, the distance-weighted model substantially increased the chance of correctly classifying a site as stressed (from a "chance alone" rate of 38 percent to an "interpolated predicted" rate of 69 percent).

The availability of multiple sampling events separated by approximate five-year periods also would allow a metric calculation of ecological change in a future study. An update of the analysis presented here will be done to include data collected in 2016, which are expected to be more representative of postcutback ecological conditions because certain wellfields (e.g., Starkey) were not able to reduce groundwater production until the end of 2007. In addition, a time lag is anticipated between the time of production cutbacks and ecological recovery, so the additional years between the 2009 conditions and 2016 conditions are expected to reveal a more nearest-neighbor accurate picture of the extent of recovery.

It's anticipated that both hydrologic and ecological condition interpolations will provide valuable evidence for assigning unmonitored sites to appropriate RA status bins. It's recommended that geostatistical predictions of hydrologic parameters be given precedence over ecologic predictions because hydrologic changes may precede ecological changes in wetlands, and hydrologic data measurements are inherently more precise than ecological WHA ratings. Geostatistical predictions of NPO may be used to quickly identify those sites believed to be "recovered," or meeting their site-specific median water level threshold. Sites predicted to not meet their site-specific thresholds will be categorized as either "improved, not fully recovered" (INFR), "not fully recovered, continuing wellfield impact" (NFRC), or impacted due to other causes (e.g., surface drainage alterations). Distinguishing between the first two cases may be primarily based on how far below predicted water levels are relative to the HNP, or possibly, the wetland bottom. (INFR sites might be near their site-specific thresholds, while NFRC sites are expected to be still relatively far below their thresholds.) The WHA predictions (and predicted changes in WHA conditions) are expected to provide additional evidence to guide categorization.

The results of this study represent a "proof of concept" only. The results suggest RK provides a useful method for recovery analyses of unmonitored wetlands in the area of investigation. The study will be repeated with updated datasets and using other machine-learning approaches (such as random forest). In addition, Tampa Bay Water is using a "weight of evidence" approach to recovery analysis. The RK results will be one parameter considered in the final assessment of recovery status; other hydrological and ecological analyses and professional opinions will factor in the final decisions regarding recovery status. An attempt will be made to assess the relative accuracy of the RK predictions for wetlands in developed landscapes (e.g., suburban areas), as the water budgets of these wetlands may be affected by surface drainage or land-use influences not reflected in the RK model.

It's recommend that the more time-intensive tools available to assess unmonitored sites—aerial imagery analysis and field visits—

be reserved for validating assignments based on the statistical binning process; that is, representative sites may be visited to confirm, for example, that recent hydrological and ecological field indicators point to a particular group of sites being below their threshold, but making sufficient improvement to be considered INFR. In addition to visiting representative sites of well-defined groups, it's likely there will be some unusual or uncertain sites that require field visits to confirm unusual circumstances, such as suspected drainage alterations, excessive soil subsidence, or other site-specific factors.

In conclusion, to answer the question posed by this article's title, it's believed that machine-learning techniques, combined with geostatistical methods, show great promise for assessing recovery of water levels and ecological conditions at unmonitored wetlands and lakes. More broadly, the application of a hybrid statistical method can allow water managers to make better predictions by accounting for known variables believed to influence water levels, as well as unknown, spatially autocorrelated factors.

References

- Alpaydin, E. 2009. Introduction to Machine Learning, Second Edition. MIT Press.
- Applehans, T., E. Mwangomo, D.R. Hardy, A. Hempd, T. Naussa. 2015. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spatial Statistics 14:91-113.
- Arthur, J.D., A.E. Baker, J.R. Cichon, A.R. Wood, and A. Rudin. 2005. Florida Aquifer Vulnerability Assessment (FAVA) Contamination Potential of Florida's Principal Aquifer Systems. A report submitted to the Division of Water Resource Management Florida Department of Environmental Protection. <http://publicfiles.dep.state.fl.us/FGS/WEB/fava/fava_final_dep_report/FAVA_REPORT_MASTER_DOC_3-21-05.pdf> Accessed 12/7/2016.
- Bellino, Jason C. 2011. Digital surfaces and hydrogeologic data for the Floridan aquifer system in Florida and in parts of Georgia, Alabama, and South Carolina. U.S. Geological Survey Data Series 584. <http://pubs.usgs.gov/ds/584/> Accessed 12/7/2016.
- Berryman & Henigar and SDI Environmental Services. 2000. Candidate Sites Evaluation Study. Final report to Tampa Bay Water. Clearwater, Fla.
- Breheny, P. and W. Burchett. 2016. Visualization of Regression Models using visreg. <http://myweb.uiowa.edu/pbreheny/publications/visreg.pdf> Accessed 12/12/2016.
- Bureau Veritas/Berryman & Henigar, PBS&J, and EHI. 2006. Five Year Wetland Assessment. Final report to the Southwest Florida Water Management District.
- Burnham, K.P. and D.R. Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer-Verlag. New York.
- Calcagno, V. and C. de Mazancourt. 2010. glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models. Journal of Statistical Software 34(12):1-29.
- Desbarats, A.J., C.E. Logan, M.J. Hinton, and D.R. Sharpe. 2002. On the kriging of water table elevations using collateral information from a digital elevation model. Journal of Hydrology 255:25-38.
- GPI (Greenman-Pedersen, Inc.). 2017. Development of Proposed Methods for Recovery Assessment Bin Assignment of Unmonitored Sites. Final Report to Tampa Bay Water.
- GPI (Greenman-Pedersen, Inc.). 2016. Development of a Water Level Recovery Metric for Xeric-associated Wetlands in the Northern Tampa Bay Area. Final Report for Tampa Bay Water.
- GPI (Greenman-Pedersen, Inc.). 2015. Recovery Assessment Geodatabase: Needs Assessment, Draft Schema, and Recommendations. Final Report to Tampa Bay Water.
- GPI Southeast, Inc., University of South Florida, and Vanasse, Hangen, Brustlin, Inc. 2010. Five Year Wetland Assessment. Draft report to the Southwest Florida Water Management District.
- Google. 2016. (Diagram of idealized variogram.) Copied without changes from https://developers.google.com/earth-engine/interpolation (accessed 1/15/2017) under Creative commons 3.0 Unported license (https://creativecommons.org/licenses/by/3.0/).
- James, G., D. Witten, T. Hastie, and R. Tibshirani. 2017. An Introduction to Statistical Learning with Applications in R. Springer Science+Business Media. On-line resource, retrieved September 27, 2017, from http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf
- Hengl, T. G.B.M. Heuvelink, B. Kempen, J.G.B. Leenaars, M.G. Walsh, K.D. shepherd, A. Sila, R.A. MacMillan, J. Mendes de Jesus, L. Tamene, J. E. Tondoh. 2015. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. PLoS One 10(6): e0125814. Published online 2015 Jun 25. doi: https://dx.doi.org/10.1371%2Fjournal.pone.0125814.
- Hengl, T. 2009. A Practical Guide to Geostatistical Mapping. Published electronically and available under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 license: http://spatial-analyst.net/book (Accessed 1/11/2017).
- Hengl, T. G.B.M. Heuvelink, D.G. Rossiter. 2007. About regression-kriging: From equations to case studies. Computers & Geosciences 33:1301-1315.
- Hengl, T., G.B.M. Heuvelink, A. Stein. 2003. Comparison of kriging with external drift and regression-kriging. Technical note, ITC, Available on-line at http://www.itc.nl/library/Academic_output
- Hiemstra , P.H., E.J. Pebesma, C.J.W. Twenhofel, G.B.M. Heuvelink. 2009. Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network. Computers & Geosciences 35(8):1711-1721.
- Lee, T.M., and Fouad, G.G. 2014. Creating a monthly time series of the potentiometric surface in the Upper Floridan aquifer, Northern Tampa Bay area, Florida, January 2000–December 2009: U.S. Geological Survey Scientific Investigations Report 2014–5038, 26 p., <http://dx.doi.org/10.3133/sir20145038> Accessed 12/7/2016. ISSN 2328-0328.
- Odeh, I.O.A., A.B. McBratney, D.J. Chittleborough. 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63:197-214.
- Pardo-Iguzquiza, E. 1998. Comparison of geostatistical methods for estimating the areal average climatological rainfall mean using data on precipitation and topography. International Journal of Climatology 18:1031-1047.
- Rochow, T.F. 1998. The Effects of Water Table Level Changes on Fresh-water Marsh and Cypress Wetlands in the Northern Tampa Bay Region: A Review. Environmental Section Technical Report 1998-1. Prepared by the Southwest Florida Water Management District. Brooksville, Fla.
- SWFWMD (Southwest Florida Water Management District). 1999. Northern Tampa Bay Minimum Flows & Levels White Papers. Southwest Florida Water Management District. Brooksville, Fla.
- Tampa Bay Water. 2017. Recovery Assessment. On-line resource, retrieved September 27, 2017, from https://www.tampabaywater.org/documents/Recovery-Assessment.pdf.
- Tampa Bay Water. 2013. Defining Areas of Investigation for Recovery Analysis. Memorandum to the Southwest Florida Water Management District. Clearwater, Fla.
- Tampa Bay Water. 2012. Permit Recovery Assessment Work Plan and Schedule. Memorandum submitted to the Southwest Florida Water Management District. Clearwater, Fla.