Hydrologic Record Extension of Water-Level Data in the Everglades Depth Estimation Network (EDEN), 1991–99

Scientific Investigations Report 2014–5226

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

Hydrologic Record Extension of Water-Level Data in the Everglades Depth Estimation Network (EDEN), 1991–99

By Paul A. Conrads, Matthew D. Petkewich, Andrew M. O'Reilly, and Pamela A. Telis

Scientific Investigations Report 2014–5226

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

U.S. Department of the Interior
SALLY JEWELL, Secretary

U.S. Geological Survey
Suzette M. Kimball, Acting Director

U.S. Geological Survey, Reston, Virginia: 2014

Contents

Figures

Tables (Tables 1 and 3 are available at *http://pubs.usgs.gov/sir/2014/5226/*)

Appendixes 1–2 (available at *http://pubs.usgs.gov/sir/2014/5226/*)

Conversion Factors

Inch/Pound to SI

Multiply	By	To obtain
Length		
inch (in.)	2.5400	centimeter (cm)
foot (ft)	0.3048	meter (m).
mile (mi)	1.609	kilometer (km)
Hydraulic gradient		
inch per mile (in/mi)	1.5783	centimeters per kilometer (cm/km)
foot per mile (ft/mi)	0.1894	meter per kilometer (m/km)
Area		
square mile (mi^2)	2.590	square kilometer (km^2)

SI to Inch/Pound

Multiply	By	To obtain
Length		
centimeter (cm)	0.3937	inch (in.)
millimeter (mm)	0.03937	inch (in.)
meter (m)	3.281	foot (ft)
kilometer (km)	0.6214	mile (mi)
Area		
square meter (m^2)	10.76	square foot (ft^2)
square kilometer (km^2)	0.3861	square mile (mi^2)

Vertical coordinate information is referenced to the North American Vertical Datum of 1988 (NAVD 88).

Horizontal coordinate information is referenced to the North American Datum of 1983 (NAD 83).

Elevation, as used in this report, refers to distance above the vertical datum.

Abbreviations

ADAM	Automated Data Assurance and Management
ANN	artificial neural network
BCNP	Big Cypress National Preserve
BEP	back error propagation
CERP	Comprehensive Everglades Restoration Plan
EDEN	Everglades Depth Estimation Network
EDENapps	Everglades Depth Estimation Network applications
ENP	Everglades National Park
GUI	graphical user interface
HLN	hidden layer neuron

ME	mean error
MLP	multilayer perceptron
MSE	mean square error
NAVD 88	North American Vertical Datum of 1988
OLS	ordinary least squares
PME	percent model error
R	Pearson correlation coefficient
R^2	coefficient of determination
RMSE	root mean square error
SFWMD	South Florida Water Management District
SISO	Single Input Single Output
SOFIA	South Florida Information Access
USGS	U.S. Geological Survey
V2	EDEN Surface-Water Model Version 2
WCA	Water Conservation Area
WL	water level

Hydrologic Record Extension of Water-Level Data in the Everglades Depth Estimation Network (EDEN), 1991–99

By Paul A. Conrads, Matthew D. Petkewich, Andrew M. O'Reilly, and Pamela A. Telis

Abstract

The real-time Everglades Depth Estimation Network (EDEN) has been established to support a variety of scientific and water management purposes. The expansiveness of the Everglades, limited number of gaging stations, and extreme sensitivity of the ecosystem to small changes in water depth have created a need for accurate water-level and water-depth maps. The EDEN water-surface elevation model uses data from approximately 240 gages in the Everglades to create daily continuous interpolations of the water-surface elevation and water depth for the freshwater portion of the Everglades from 2000 to the present (2014). These maps provide hydrologic data previously unavailable for assessing biological and ecological studies.

Ecologists working in the Everglades expressed a need to the EDEN project team for daily EDEN water-level surfaces from 1990 to 1999. The additional 10 years of surfaces will provide ecologists and resource managers with two decades (1991–2011) of surfaces to analyze hydrologic dynamics. Before 2000, many of the EDEN gages used to generate water surfaces were not in operation. These datasets were extended to provide estimations of hydrologic time-series histories. The general approach to the record extension (hindcasts) was to (1) create a database of available data from 1990 to the present; (2) use dynamic cluster analysis to group stations with similar hydrologic behaviors for subareas of the Everglades with a large number of stations; (3) use results from the cluster analysis to select candidate explanatory variables; (4) develop linear regression or artificial neural network models to extend water-level records; and (5) evaluate record extensions by using model performance statistics and comparison of water-surface maps for similar hydrologic conditions for the hindcasted period (1991–99) and measured period (2000–11).

To hindcast and fill data records, 214 empirical models were developed—189 are linear regression models and 25 are artificial neural network models. The coefficient of determination (R^2) for 163 of the models is greater than 0.80 and the median percent model error (root mean square error divided by the range of the measured data) is 5 percent. To evaluate the performance of the hindcast models as a group, contour maps of modeled water-level surfaces at 2-centimeter (cm) intervals were generated using the hindcasted data. The 2-cm contour maps were examined for selected days to verify that water surfaces from the EDEN model are consistent with the input data. The biweekly 2-cm contour maps did show a higher number of issues during days in 1990 as compared to days after 1990. May 1990 had the lowest water levels in the Everglades of the 21-year dataset used for the hindcasting study. To hindcast these record low conditions in 1990, many of the hindcast models would require large extrapolations beyond the range of the predictive quality of the models. For these reasons, it was decided to limit the hindcasted data to the period January 1, 1991, to December 31, 1999. Overall, the hindcasted and gap-filled data are assumed to provide reasonable estimates of station-specific water-level data for an extended historical period to inform research and natural resource management in the Everglades.

Introduction

The Everglades Depth Estimation Network (EDEN) project was started to provide scientists working on the restoration of the Everglades with spatially continuous quality-assured and quality-controlled hydrologic data at any point within the freshwater part of the Greater Everglades (fig. 1). The Everglades Depth Estimation Network is a network of real-time water-level gaging stations, an integrated ground-elevation model, and a water-surface elevation model that originally was designed to provide scientists, engineers, and water-resource managers with current water-level and water-depth information (2000–14) for the entire freshwater portion of the Greater Everglades (Telis, 2005, 2006; Telis and others, 2014). The Everglades Depth Estimation Network offers a consistent and documented dataset that can be used by scientists and water-resource managers to (1) guide large-scale field operations, (2) integrate hydrologic and ecological data and analysis, and (3) support biological and ecological restoration assessments that measure ecosystem responses to the Comprehensive Everglades Restoration

Plan (CERP; U.S. Army Corps of Engineers, 1999). In addition, EDEN, with the integration of real-time data and models, provides opportunities for real-time evaluation of water-level conditions and water-resource management practices.

The expansiveness of the Everglades, limited number of gaging stations, and sensitivity of the ecosystem to small changes in water depth have created a need for accurate water-level and water-depth maps (Pearlstine and others, 2007). To create daily water-level surfaces for the Everglades, EDEN integrates data from a network of more than 250 water-level gages maintained by multiple agencies, including Big Cypress National Preserve (BCNP), Everglades National Park (ENP), South Florida Water Management District (SFWMD), and U.S. Geological Survey (USGS) (fig. 1). The EDEN surface-water model (Telis and others, 2014) generates daily water-level surfaces gridded to 400-square-meter cells for January 1, 2000, to the present (2014) by using the EDEN grid developed by Jones and Price (2007a). Examples of the daily water-level elevation are shown in figure 2. When these water-level

surfaces are combined with EDEN's digital elevation model for ground surface (Jones and Price, 2007b; Jones and others, 2012; Xie and others, 2011), daily surfaces of water depth can be generated. To assist users in applying the EDEN datasets to their needs, a series of tools, or applications (EDENapps), were developed to view, extract, plot, and manipulate EDEN data to create other derived hydrologic data (Telis and Henkel, 2009).

Scientists using EDEN data requested that daily EDEN water surfaces be created from 1990 to 2000, a period that did not previously have any surfaces. The extended period of hydrologic record would provide a longer time period for ecologists and scientists to evaluate the relation between hydrologic changes and ecological response, and ultimately, assist local, State, and Federal agencies in their efforts to restore the Everglades ecosystem. Unfortunately, before 2000, fewer EDEN gages used to generate water surfaces were in operation (fig. 3). To be able to generate water-surface maps for 1990 to 1999, water-level data for the EDEN stations that were not in operation needed to be extended back to 1990.

Figure 1. Location of the water-level stations used in the Everglades Depth Estimation Network (EDEN) surface-water model to hindcast water surfaces

A. November 1, 2011

B. April 30, 2011
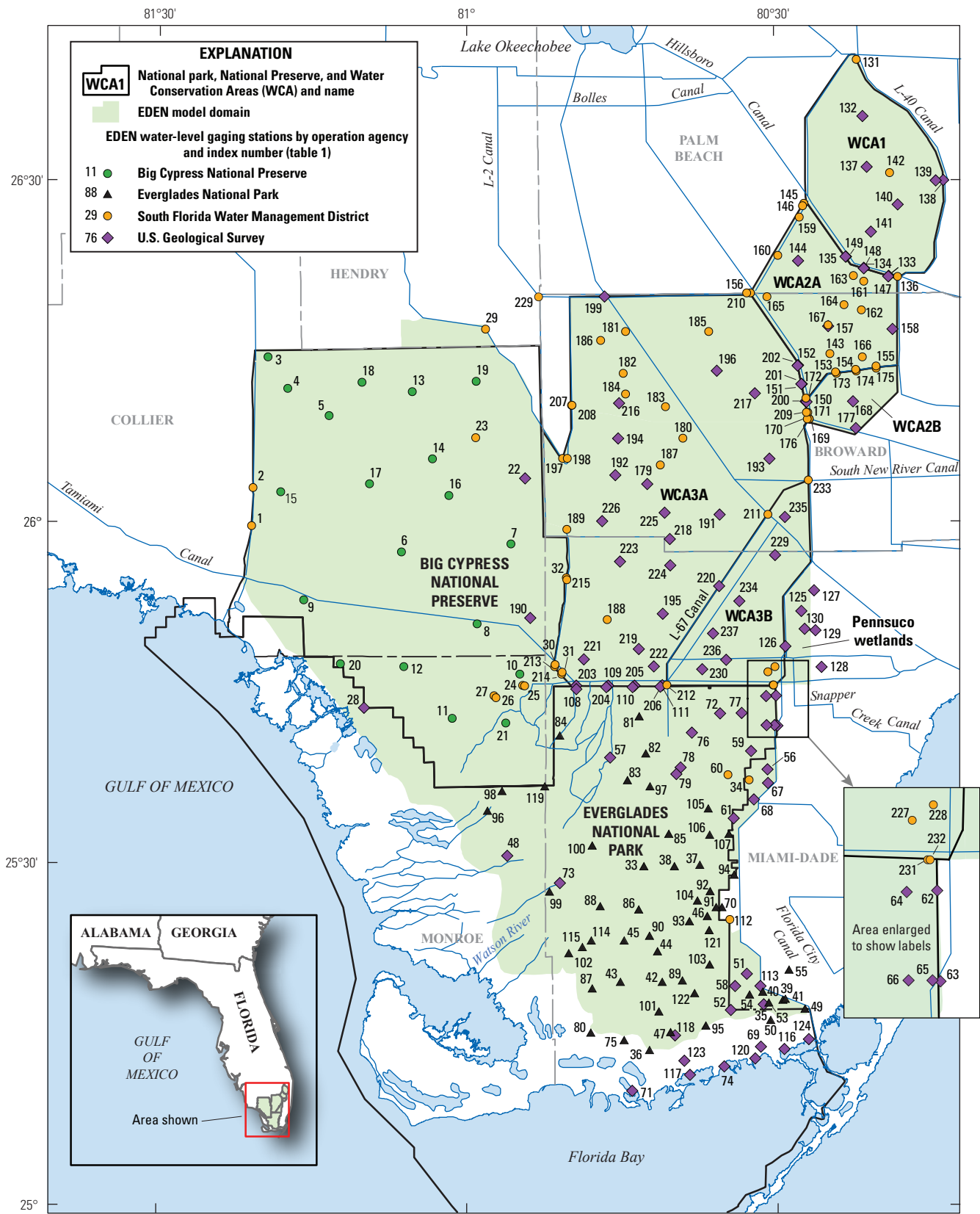
Base from Esri® Data & Maps, 2010
U.S. Geological Survey Everglades Depth Estimation Network, 2011
Universal Transverse Mercator projection, zone 17

0   5   10   15   20 MILES
0   5   10   15   20 KILOMETERS

EXPLANATION

**Water-surface elevation,
in centimeters**

High : 470

Low : −65

Canal/stream
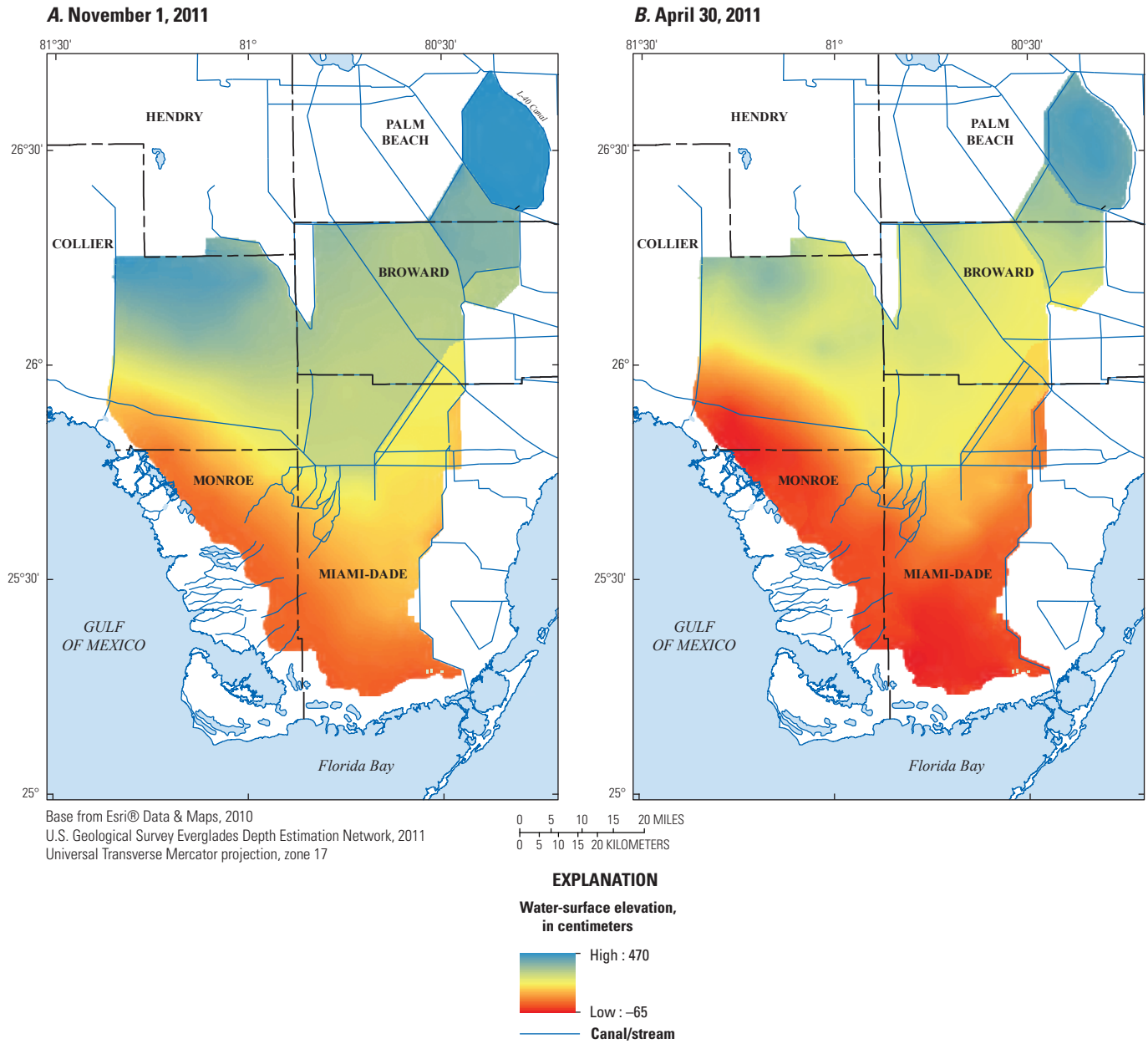
Figure 2.   Example of the Everglades Depth Estimation Network (EDEN) water-surface map for *A*, a wet season day and *B*, a dry season day
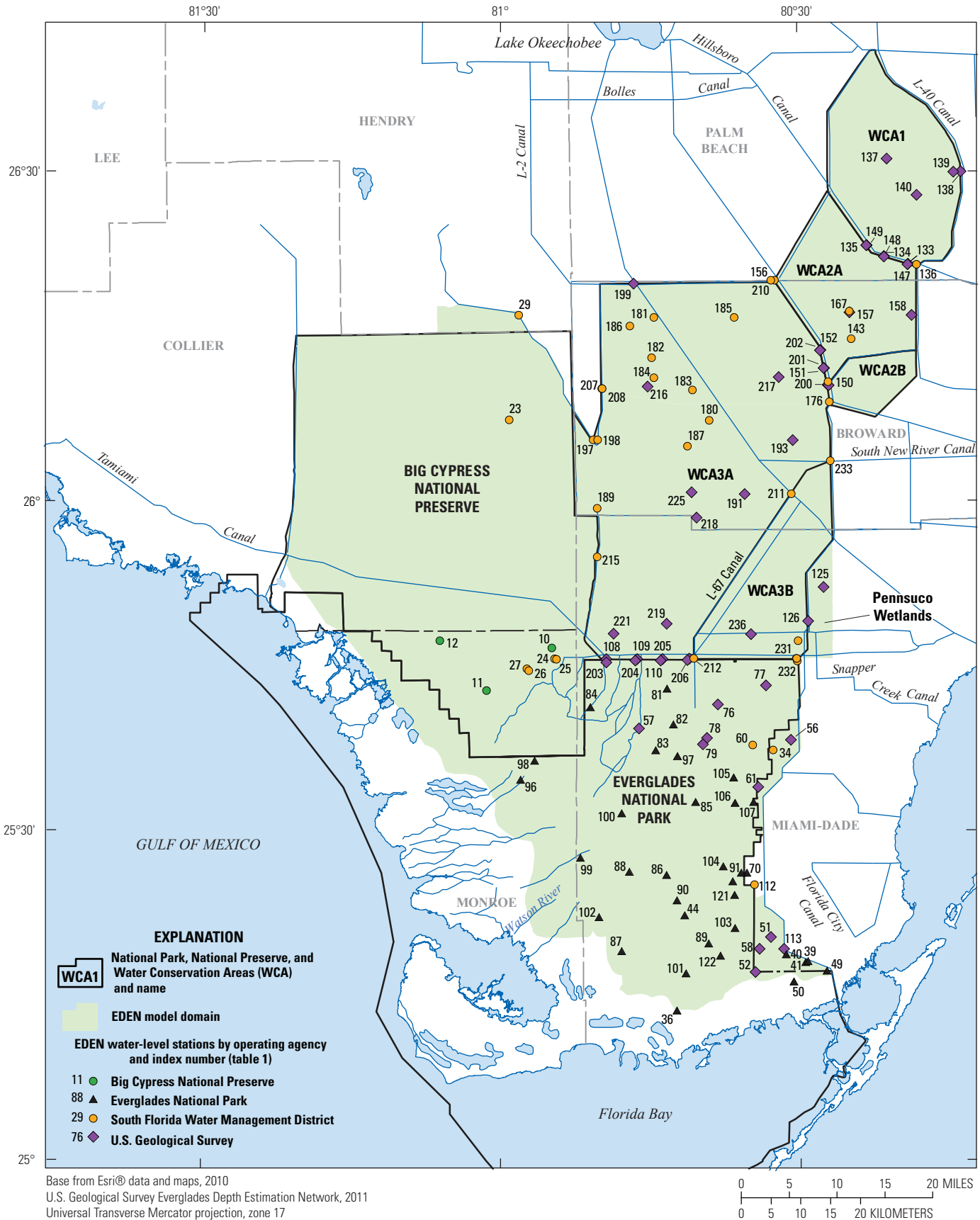
Figure 3. Location of the 106 stations in the Everglades Depth Estimation Network (EDEN) that were operating in 1990.

Purpose and Scope

This report documents the water-level record extensions (hindcasts) of 237 stations in the freshwater portion of the Everglades. The geographical extent of the hindcasts includes gaging stations in Water Conservation Area (WCA) 1, WCA2A, WCA2B, WCA3A, WCA3B, Pennsuco Wetlands, BCNP, and the ENP (fig. 1). An important part of the USGS mission is to provide scientific information for the effective water-resources management of the Nation. To assess the quantity and quality of the Nation's surface water, the USGS collects hydrologic and water-quality data from rivers, lakes, estuaries, and wetlands using standardized methods, and maintains the data from these stations in a national database. One of the critical directions outlined in the USGS science strategy for 2007–17 (U.S. Geological Survey, 2007) is to provide scientific information for understanding ecosystems and predicting ecosystem change. To study the causes and consequences of ecological change, the USGS collects hydrological, biological, and other data using standardized methods, and interprets these data for policymakers to indicate how changes will potentially affect natural resources and the public. The techniques presented in this report demonstrate how valuable information can be extracted from existing USGS databases, integrated with other agencies' data networks, and extended back in time (hindcasted) to evaluate the linkage between hydrology and ecological response.

Description of Study Area

The Everglades is a vast wetland consisting of approximately 2.9 million acres covering much of southern Florida. The Everglades primarily consists of peat soils and tall sawgrass that are interspersed with slightly raised tree islands covered by shrubs and woody vegetation (McPherson and others, 1995). Historically, the Everglades was an uninterrupted wetland that extended from Lake Okeechobee and flowed to the southwestern tip of Florida (fig. 4; Richardson and others, 1990). The annual rainfall in the Everglades generally is between 50 and 60 inches (in.), depending on location, with substantially more rainfall along the eastern edge (Lodge, 1994). The rainfall has a distinct seasonal pattern, with a wet season from May or June through September or October that accounts for about 75 percent of the annual total. Water depths in the freshwater marshes range from 0 to 3 feet (ft) during the wet season. During the annual wet season, water levels rise and inundate most of the land, producing seasonal flows into the Florida Bay and the Gulf of Mexico (fig. 1). Heavy rainfall associated with tropical depressions, storms, and hurricanes have a large effect on water levels. A single such event can increase water levels by 1 ft or more over large parts of the Everglades, and because of the slow runoff rates, this can effect water levels for months (German, 2000). Minimum seasonal water levels generally happen in May before the onset of the wet season. In contrast, during the

dry season, water levels decline to near land surface (McPherson and others, 1995). In particularly dry years, large portions of the Everglades may become dry and are subject to wildfires. This hydrologic pattern of wet and dry seasons helped produce and sustain the unique ecosystem of the Everglades.

The Everglades contains several types of environments, including freshwater marshes, tree islands, pinelands, mangrove swamps, and shallow coastal marine waters. The EDEN project is concerned with freshwater marshes, the predominant Everglades ecosystem. These marshes are characterized by sawgrass stands of varying density and height, ran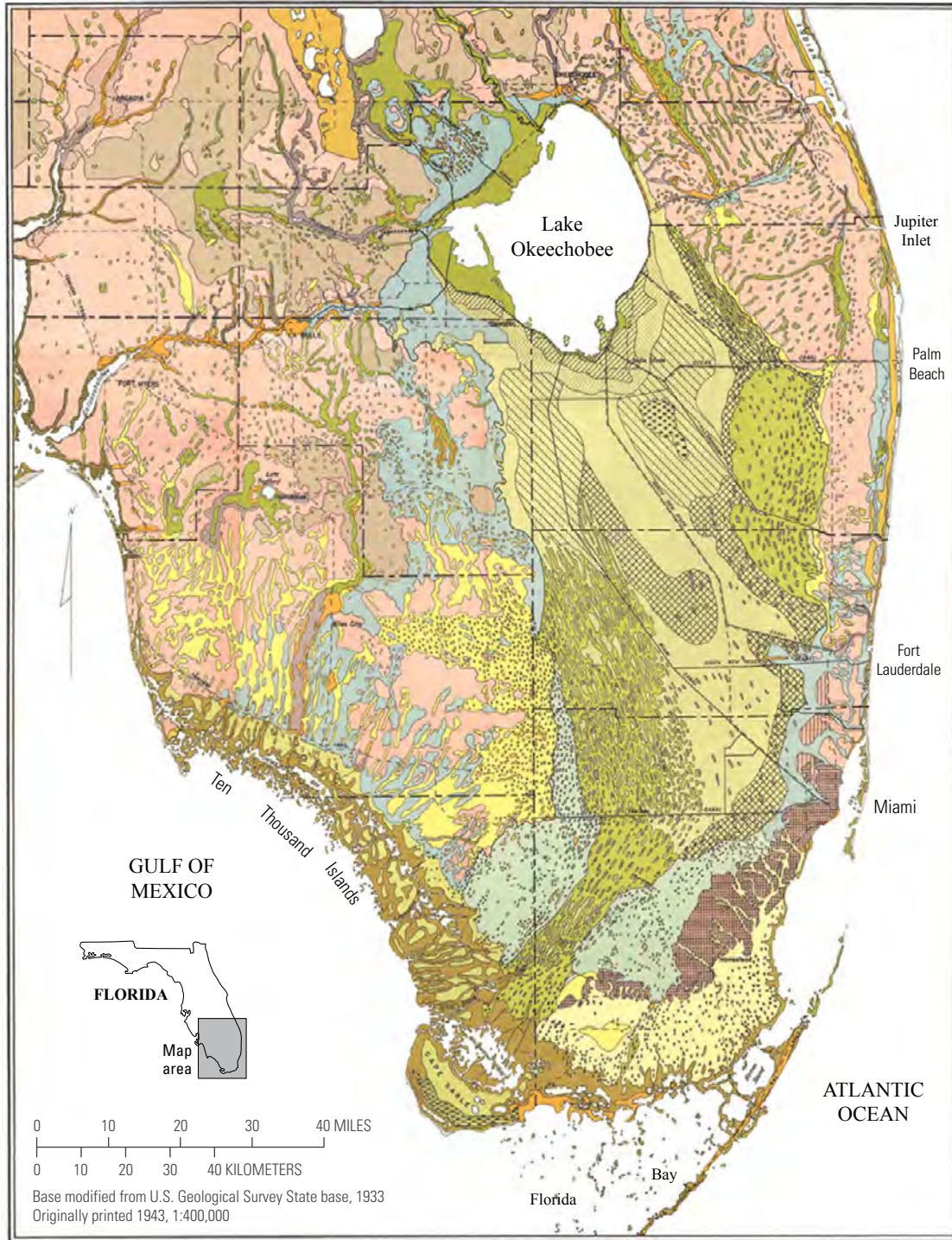ging from 2 to 3 ft above land surface to 9 ft in some northern areas. Other common emergent plants in the freshwater marshes include spike rush, muhly grass, and, in some areas, cattails. Typical topographic and vegetative features include ridge and slough, tree islands, wet prairie, sawgrass, and marl prairie (German, 2000).

From the mid-1800s to the late 1900s, the flow patterns, and thus the ecosystem, of the Everglades have been substantially altered. Beginning in the mid-1800s, wetlands began to be drained and used for agricultural purposes and urban development such that by the early 1990s, about 50 percent of the historic Everglades had been drained (McPherson and others, 1995). With population growth and increased agricultural production, flood mitigation and water use in South Florida became prominent concerns. To address those concerns, WCA1, WCA2, and WCA3 (fig. 1) were constructed by the U.S. Army Corps of Engineers in the 1940s with the goal of regulating water through an extensive series of levees and canals. In general, the WCAs store water during the wet season and supply water during the dry season. The combined effect of drained wetlands and water regulation introduced during the 1800s and 1900s diverts an estimated 40 percent of the water originally flowing through the Everglades (McPherson and others, 1995). The substantial changes in land use and flow patterns within the Everglades have had adverse environmental effects on the hydrology, water quality, and native plant and animal communities (McPherson and others, 1995). Since the late 1990s, through the collaboration of Federal, State, and private agencies, substantial focus has been given to mitigate these adverse effects and, where possible, to begin restoration of the ecosystem and ecological communities of the Everglades.

The study area is the freshwater portion of the greater Everglades area, which extends from south of Lake Okeechobee to the southern part of the ENP (fig. 1). This area is a wetlands system that is about 50 miles (mi) wide and about 100 mi long. The Everglades is regarded as unique in the world because it is not primarily associated with a natural river system but is itself a wide and shallow "river" that transports water by sheet flow from Lake Okeechobee to the Gulf of Mexico. The land slopes within this shallow "river" generally are less than about 0.2 foot per mile (ft/mi; German, 2000).

EXPLANATION

General vegetation

- Scrub forest
- Hammock forest
- Bay tree forest
- Inland swamp
- Cypress forest
- Cypress heads or domes
- Pine flatwood forest
- High pine forest
- Miami Rockland pine forest
- Miami open pine forest
- Mangrove swamp
- Mangrove swamp and saltwater marsh
- Saltwater marsh
- Salt Prairie (Cape Sable area)
- Freshwater marsh (outside of E.)
- Wet prairies
- Saw palmetto or Dry Prairies
- Southern coast marsh prairie
- Coastal beach and dune vegetation

Main Everglades (E.) vegetation

- Saw-grass marsh (medium dense to sparse)
- Saw-grass marsh (dense)
- Saw-grass marsh (with wax-myrtle thickets)
- Willow and elderberry zone (mostly cultivated)
- Custard-apple zone (mostly cultivated)
- Saw-grass marsh (with abundant ferns and cattails)
- Slough, pond, and lake (with aquatic plants)
- Tree island, bay tree forest
- Tree island, hammock forest
- Marsh prairie, southern E.

Figure 4. Vegetation map of southern Florida circa 1943 (modified from Davis, 1943).

Data-Collection Network

The EDEN database (2014) is composed of hourly water-level data from 312 gaging stations and includes marsh, river, canal, and control structure stations operated by the BCNP, ENP, SFWMD, and the USGS in the freshwater and coastal waters. Water monitoring networks typically are established to meet the mission of the agency, be it regulatory compliance requirements, scientific investigation, or other reasons. The configuration of networks often changes to meet changing agency objectives and funding levels. As with many monitoring networks, the total number of EDEN gages may change from year to year because of funding levels, agency mandates, and gaging logistics. For any given year, the number of input stations used in the EDEN surface-water model may change the network configuration. Of the 312 EDEN stations, 237 are used in the EDEN Surface-Water Version 2 Model (referred to in this report as the "V2 model"; Telis and others, 2014) to generate daily water-surface maps for the year 2000 (table 1). These 237 stations are the subset of EDEN gages that are used to create the hindcasted water-level surfaces for this study. The other stations in the EDEN database are used to evaluate coastal conditions of the Everglades, estimate missing data for the stations in the water-surface model, or are discontinued EDEN stations. In this report, the names of the stations follow the naming convention used by EDEN and are similar to the names used by the agency that maintains the gage.

Previous Studies

The EDEN project has used various techniques for hindcasting data for stations added to the gaging network to provide a concurrent period of water-level data records for the water-surface model and for estimating missing data. Soon after the initiation of the EDEN project and the development of the water-surface model in 2006, 25 real-time water-level gaging stations were added to the original EDEN network of 253 established water-level gaging stations. These stations did not have data back to 2000, the original beginning of the EDEN database. To incorporate the data from the added stations to the 7-year EDEN database in the greater Everglades, the short-term water-level records (generally less than 1 year) needed to be hindcasted to be concurrent with data from the established gaging stations in the database. Conrads and Roehl (2007) used a three-step modeling approach using artificial neural network (ANN) models to estimate the water levels at the new stations. The ANN models used static variables that represent the gaging station location and percent vegetation in addition to dynamic variables that represent water-level data from the established EDEN gaging stations. The final step of the modeling approach was to simulate the computed error of the initial estimate to increase the accuracy of the final water-level estimate by incorporating the computed error term.

The three-step modeling approach for estimating water levels at the new EDEN gaging stations produced satisfactory

results. The coefficients of determination (R^2) for 21 of the 25 estimates were greater than 0.95, and all of the estimates were greater than 0.82. The model estimates indicated good agreement with the measured data. For some new EDEN stations with limited measured data, the hindcasts included periods beyond the range of the data used to train the ANN models. The comparison of the hindcasts with long-term water-level data proximal to the new EDEN gaging stations indicated that the water-level estimates were reasonable. The percent model error (root mean square error divided by the range of the measured data) was less than 6 percent, and for most of the stations (20 of 25), the percent model error was less than 1 percent.

To increase the accuracy of the EDEN daily water-level surfaces, water-level estimation equations were developed to fill missing data for every EDEN station (Conrads and Petkewich, 2009; Petkewich and Conrads; 2013). To minimize the inability of estimating data because of missing data for an input station, a minimum of three linear regression equations were developed for each station using different input stations. Of the 667 water-level estimation equations developed to fill missing data at 223 stations, more than 72 percent of the equations had an R^2 greater than 0.90, and 97 percent have an R^2 greater than 0.70. The estimation equations recently were recomputed to accommodate change in the gaging network and further understanding of the hydrologic response of the system (Petkewich and Conrads, 2013).

Daamen and others (2010) developed the Automated Data Assurance and Management (ADAM) software to allow quick and accurate quality-assurance review of the EDEN data. The ADAM software utilizes user-defined filter settings, linear regression equations (documented in Petkewich and Conrads, 2013), and multivariate regression equations to estimate missing data and quality assure real-time data. The automation process is used to efficiently identify erroneous data and provide accurate replacement estimates. In addition, the tool can be used to manually review previously processed results and modify the estimates, if necessary.

Approach

The general approach for hindcasting water-level data back to 1990 is to identify long-term datasets (measured data back to 1990) that are correlated to the short-term datasets (measured data beginning after 1990). An empirical model is developed that uses the long-term dataset for the explanatory (input) variable to estimate (hindcast) the short-term dataset back to 1990. An example of a hindcasted dataset is shown in figure 5. The measured data for station W14 in WCA3A (map label 224, fig. 1) began operation on January 25, 2006 (table 1). To extend the record to 1990, the long-term station SITE_64 located to the north (map label 218, fig. 1) was used as input to an empirical model (in this case a linear regression model) to estimate the water level for W14 for the period January 1, 1990, to January 24, 2006. It is important to note
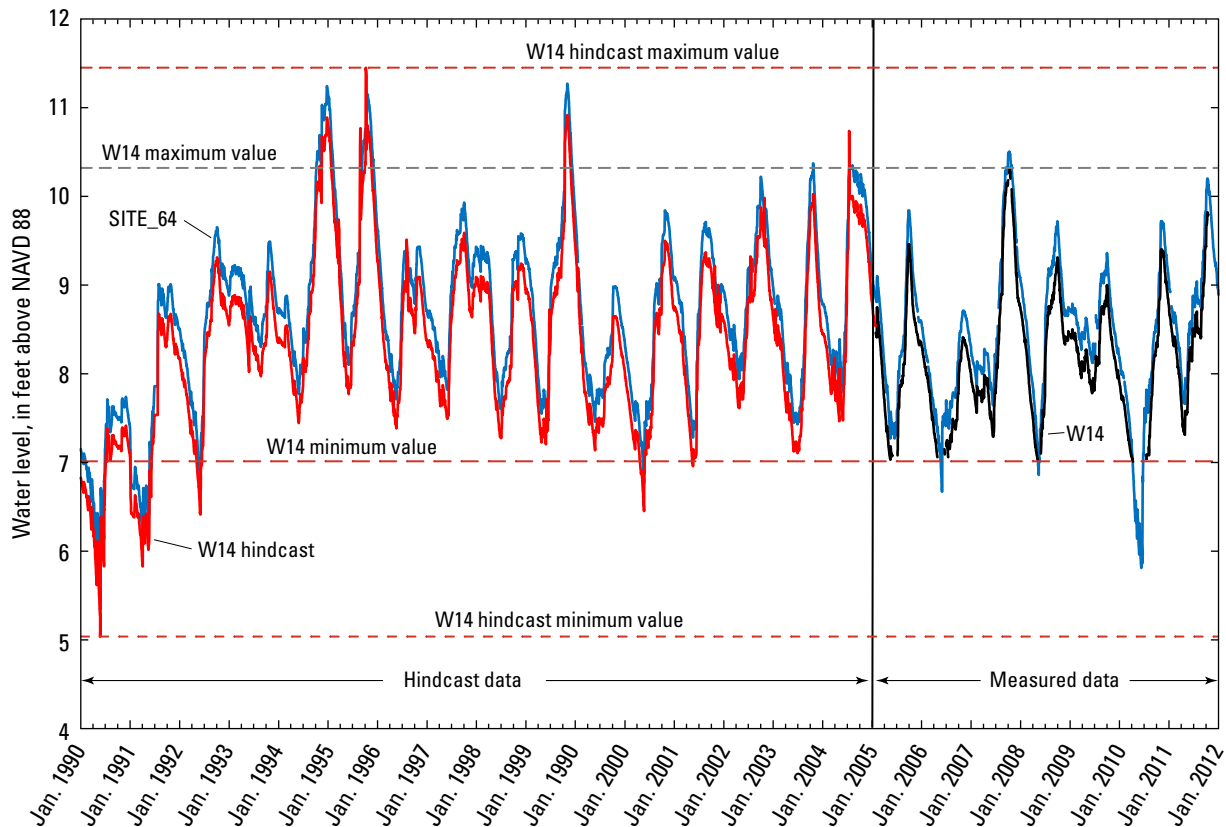
Figure 5.   Plot showing an example of hindcasting a short-term dataset (W14) to a longer term dataset (W14 Hindcast).

the range of the measured and estimated data. Models, both empirical and mechanistic, are more accurate when interpolating within the historical range of the data used to develop the model (for W14, 7.01 to 10.31 ft, fig. 5) than extrapolating beyond the range of the data used to develop the model (for W14 on fig. 5., extrapolation below 7.01 ft to 5.03 and above 10.31 ft to 11.45 ft). Within the range of measured data, the model is interpolating from conditions exhibited in the historical data. Models are less reliable when estimating outside of the range of the data used to develop the model. Under these conditions, the models have to extrapolate to conditions that are not exhibited in the data. To make these extrapolations, the model assumes that the relation between input and output variables remain constant between the measured conditions and the extrapolated conditions. The models do not take into account changes in flow-regulation schedules or meteorological conditions. Depending on the dynamics of the system being modeled, this assumption may or may not be valid. For the example of hindcasting site W14, the model had to extrapolate a little more than 1 ft for high-water conditions and almost 2 ft for low-water conditions. Limited extrapolation often is appropriate. In a shallow wetland system like the Everglades, high-water extrapolations are not as problematic as the low-water extrapolation. During high-water conditions, there is extensive ponding in the WCAs and

high-water extrapolation probably increases the extent of the ponding. During low-water conditions, areas of the Everglades can become dry and surface water becomes shallow groundwater or discontinuous shallow pools or ponds. Because of micro-topographic features, small decreases in surface-water elevations may translate to extensive dry areas. As discussed in the "Evalu-ation of Hindcasts using Modeled Water-Level Surfaces" section later in the report, extrapolating to the extreme low-water level in May 1990 was not appropriate.

The estimation of water levels for EDEN in previous studies relied on the configuration of the water-level station network (fig. 1), which has fairly even spatial distributions of gaging stations. For most of the stations in the network, there are proximal stations that are well correlated to generate accurate water-level data estimates. The spatial distribution of the EDEN stations in 1990 (fig. 3) is uneven and there are large areas without a gage. In 1990, 106 of the EDEN stations used in the V2 model were in operation (table 2 and figs. 3 and 6). From 1990 to 1996, that number grew to 177 stations. Between 1997 and 2001, 27 stations were added to the network, between 2005 and 2006, 31 stations were added to the network, and two stations in 2009 (fig. 6). The most significant increases in stations occurred in WCA2B and BCNP where there was a 1,100- and 460-percent increase in gages, respectively between 1990 and 2000 (table 2).

Table 2.   The number of stations by conservation area in the Everglade Depth Estimation Network (EDEN) Surface Water Model Version 2 in 1990 and 2000.

Area	1990	2000
Big Cypress National Preserve	7	32
Everglades National Park	47	92
Pennsuco Wetlands	2	6
Water Conservation Area 1	8	12
Water Conservation Area 2A	9	25
Water Conservation Area 2B	1	11
Water Conservation Area 3A	28	48
Water Conservation Area 3B	4	11
Total	106	237
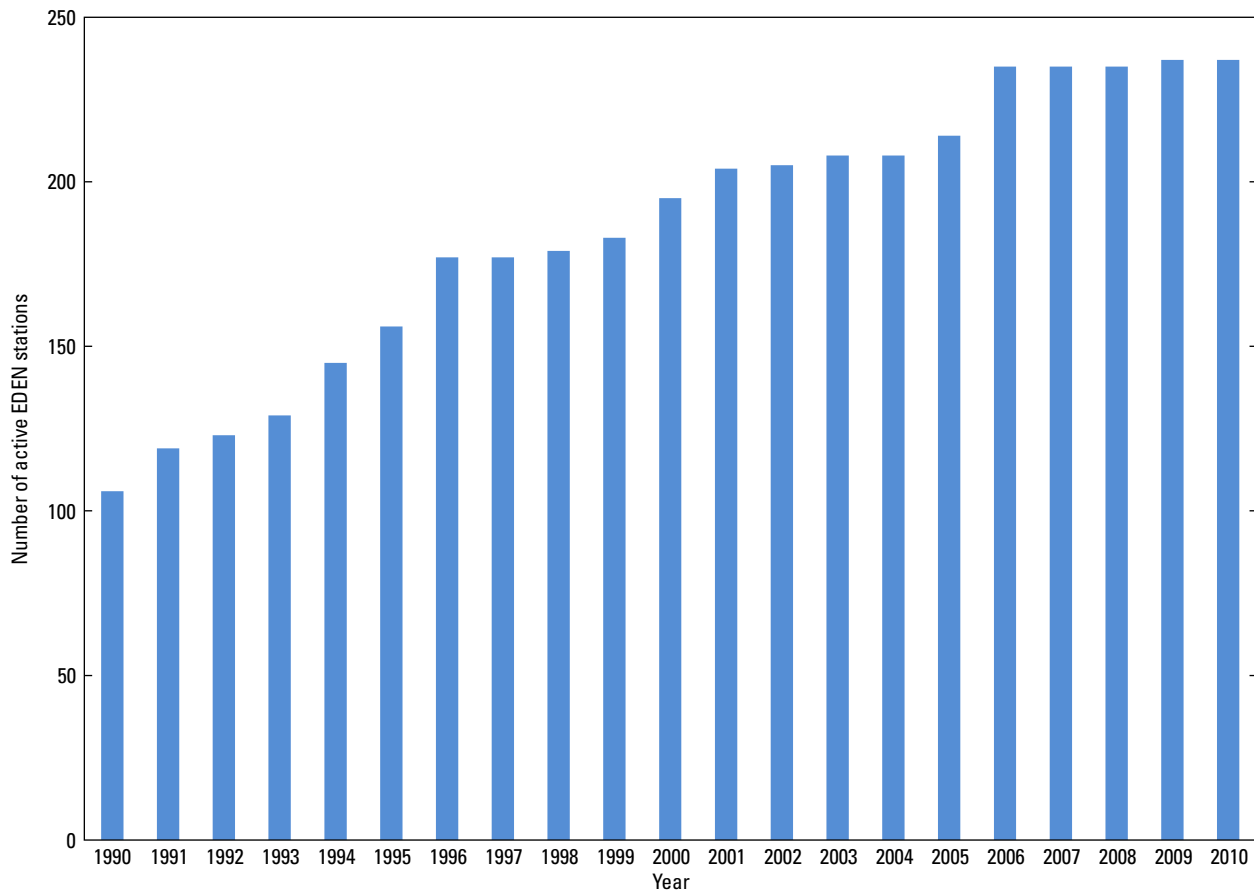
Figure 6.   Bar chart showing the number of Everglades Depth Estimation Network (EDEN) stations in operation for the years 1990 to 2010.

For this study, hindcasts are considered estimated record prior to a gage being in operation and estimates are values to fill data gaps in the operational data record. Of the 237 stations needed for the EDEN V2 model, only 23 of the stations had complete records without any data gaps, 46 stations had data gaps to be filled, and 168 stations started after January 1, 1990, and were hindcasted (table 1). The approach for hindcasting and estimating record is the same for this study. A timeline of the 237 EDEN stations for this study that shows each station and a color-coded timeline of the data type—measured, hindcasted, or estimated—is shown in figure 7. One can quickly see the predominance of measured data after 2000 and the large number of hindcasted sites in the early 1990s.

The challenge for hindcasting stations back to 1990 is that EDEN stations that are well correlated with another EDEN station may not have been operating in 1990. For example, station EDEN_11 in WCA2A (map label 144, fig. 1) is estimated in Petkewich and Conrads (2013) by using station WC2AS1 (map station label 160), WCA2F1 (map station label 163), or WCA2F4 (map station label 164). These three stations are the highest correlated stations in the current (2014) EDEN network. To extend the water-level record for EDEN_11 to 1990, the available measured data are limited to two marsh sites located to the southeast of EDEN_11, stations 2A300 and SITE_17 (map station labels 143 and 157, respectively, on figs. 1 and 3).

Because of the large number of stations to hindcast for this study, most of the hindcasts were estimated using linear regression models. For sites that had unsatisfactory linear regression estimates, ANN models were used. The general process to compute hindcasts to 1990 was to complete the following steps:

- create a database of available data for each Water Conservation Area (WCA), BCNP, ENP, and Pennsuco Wetlands from 1990 to the 2012;

- convert all data to the North American Vertical Datum of 1988 (NAVD 88);

- fill data gaps of 3 days or less with linear interpolation;

- for Water Conservation Areas (WCA) with a large number of stations, perform time-series cluster analysis (Risley and others, 2003; Roehl and others, 2006) to group gages with similar hydrologic behaviors;

- develop linear regression or ANN models to extend water-level records;

- hindcast water-level records and (or) fill periods of missing record;

- apply shifts to the hindcast records, if necessary; and

- evaluate record extensions with model performance statistics and comparison of water-surface maps for similar hydrologic conditions from the hindcasted period (1990–99) and measured period (2000–12).

A more detailed description of the dynamic cluster analysis, linear regression models, ANN models, and the application of shifts is presented in the "Estimating Water Levels" section of the report.

The preference for input data for a model was to use measured data rather than hindcasted or estimated data. The models used to hindcast and fill missing data have inherent errors and using the results of these models as input for another model propagates the error into another dataset. There were stations where, because of their location and (or) period of missing record, the only reasonable input dataset to develop a satisfactory model was a hindcasted or estimated dataset.

Limitation of the Datasets

As with any modeling effort, the reliability of the model is dependent on the completeness of the datasets and on the quality of the data and range of measured conditions used for developing models. The quality of the hindcast models rely on the available period of record and historical range of water level in addition to the spatial distribution of long-term input stations. There are the usual data limitations of measurement accuracy, the physical location of the monitoring station, instrumentation error, and human error. The agencies that collected the EDEN data all have quality-assurance and quality-control plans to minimize the error in their measured water-level data. For this study, there are two limitations of datasets used for hindcasting the EDEN. The first limitation is the poor spatial distribution of the monitoring sites and the availability of measured data in particular areas of the Everglades for estimating accurate water levels back to 1990. For example, the spatial limitation of available data makes it more difficult to estimate water levels in BCNP as compared to ENP (fig. 3). The second limitation is the range of the water levels measured at a particular site that is used for estimating water levels at another site. If the range of data inputs to a model is small, the estimation model may have to extrapolate more to estimate the full range of water level at the output station and there will be less certainty in the extrapolated values.

Estimating Water Levels

The dynamic variability of the water level at a particular location in the Everglades is caused by a combination of meteorological factors; for example, rainfall, evapotranspiration, and releases through water-control structures to move water through the various conservation areas to ENP. The water-level inputs at one location that are used for inputs for another location incorporate the localized meteorological effects and control structure releases. The water releases through control structures are regulated through water delivery schedules that are designed to minimize the detrimental ecological effects of water releases. There have been three Water Delivery Schedule Plans in operation during the study
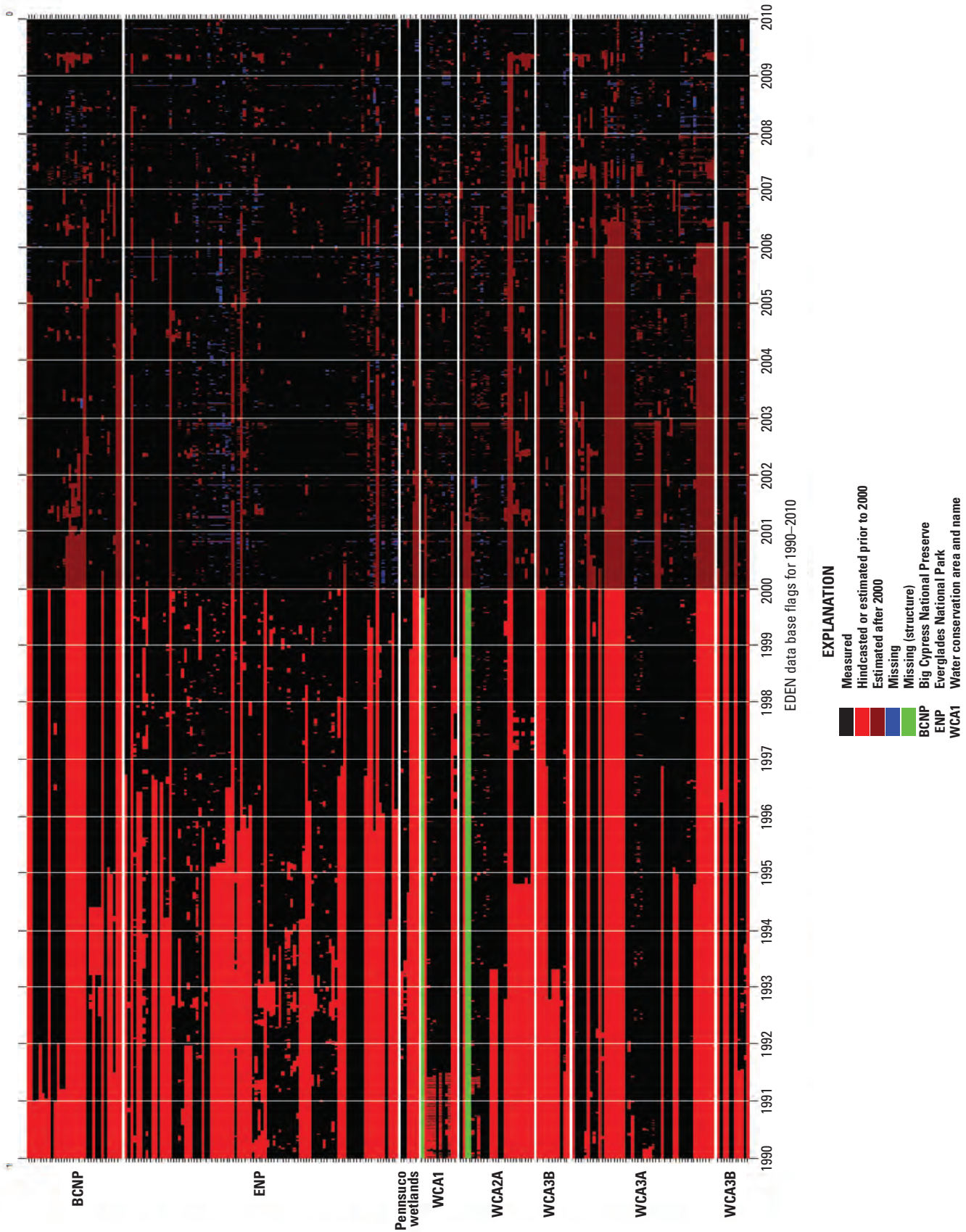
Figure 7.   Timeline from 1990 to 2010 showing data types for the 237 stations used in the study. (Due to the small size of the rows, station labels are not shown. Stations are in the same order as they are listed in table 1).

period: the Experimental Program of Water Deliveries to ENP (1984–99), The Interim Structural and Operational Plan (1999–2002), and the Interim Operational Plan (2002–12) (South Florida Natural Resource Center, 2005; U.S. Army Corps of Engineers, 2011). Although these Water Delivery Schedules have the overall effect of altering the volume and timing of flows to ENP, for the purposes of this study it is assumed that the changes in the operation schedules do not have a substantial effect on the localized water-level relations (correlation) between the nearby input stations used to estimate water levels.

Estimation models generally are in one of two categories: deterministic (or mechanistic) or empirical. Deterministic models are created from first-principles equations, whereas empirical models are created directly from data and adapt generalized mathematical functions to fit a line or surface through data from one or more variables. Because of the large number of sites to hindcast or fill, sites were first estimated with ordinary least squares (OLS) regression models. If the estimates were not satisfactory, ANN models were developed to improve the accuracy of the models. For areas with a large number of stations to fill or hindcast, time-series cluster analysis was used to group stations by similar hydrologic behavior. These groupings facilitated the selection of input (predictor) sites. Below are descriptions of the time-series clustering technique, the empirical models used in the study, statistical measures of model accuracy, and the shifting of estimated water-level data.

Time-Series Cluster Analysis

For areas with a large number of sites, such as the ENP and BCNP south of the Tamiami Canal and WCA3A (fig. 1), a time-series clustering algorithm was applied to divide sites into classes (groups) having similar behaviors (Roehl and others, 2006; Stewart and others, 2006; Risley and others, 2003). Next, the central tendency (low-frequency variability) was removed individually for each site by computing the difference from a "standard" time series (average of all the time-series data for sites in that class), leaving the residuals that accentuate differences between signals. Finally, a cross-correlation matrix (Pearson R) of the residuals was constructed as a measure of the dynamic similarity among sites. In the matrix, each cell represented the behavioral similarity between the site listed for that row header and that column header. The k-means clustering algorithm was applied to the cells in the matrix (Weiss and Indurkhya, 1998). For k-number of groups, the k-means algorithm optimizes which members of the overall group should be in groups 1 through k. The optimal number of groups was determined by using the root mean square error (RMSE) as a measure of the difference in distance between each member and the mean of the group. In some cases, the optimal number of groups can be selected at the inflection point between a sharp vertical decline in RMSE and a horizontal plateau. In other cases, a more gradual reduction in RMSE with an increasing number of groups happens.

An example of the time-series clustering is shown for 27 sites in WCA3A (fig. 8). The sites were clustered into three classes of behaviors (Groups 1–3). The 7-day average water level for each group is shown on figure 9. The breakout of the three groups generally follows the ground elevation gradient with higher water levels to the north and lower water levels to the south. Group 1 sites are located in the upper region of WCA3A (fig. 8) and have the highest average water levels as compared to the other two groups (fig. 9). The Group 2 sites are located in the lower region near the Tamiami Canal. The Group 3 sites generally are located between the Group 1 and 2 sites. The water levels for the Group 2 and 3 sites indicate there are often ponding conditions at the Group 2 sites that cause the water levels to be higher than the Group 3 sites during high-water conditions, for example January to March of 2007 and 2008 (fig. 9). During the dry season, Group 2 and 3 water levels are similar, whereas during the height of some wet seasons (for example, in October 2006, November 2008, and September 2009), the water level for the Group 3 sites is higher than the Group 2 sites (fig. 9).

The time-series clustering analysis is valuable for selecting candidate input sites for the prediction models. The area between two groups indicates that there is a change in water-level dynamics and which stations belong to each group. For example, 3AS3W1 and EDEN_8 (fig. 8, map label 188 and 195, respectively) would make better inputs for predicting SITE_65 (fig. 8, map label 219) than W5 (fig. 8, map label 222) because of their dynamic similarity of the sites in Group 3 even though W5, in Group 2, is closer to SITE_65.
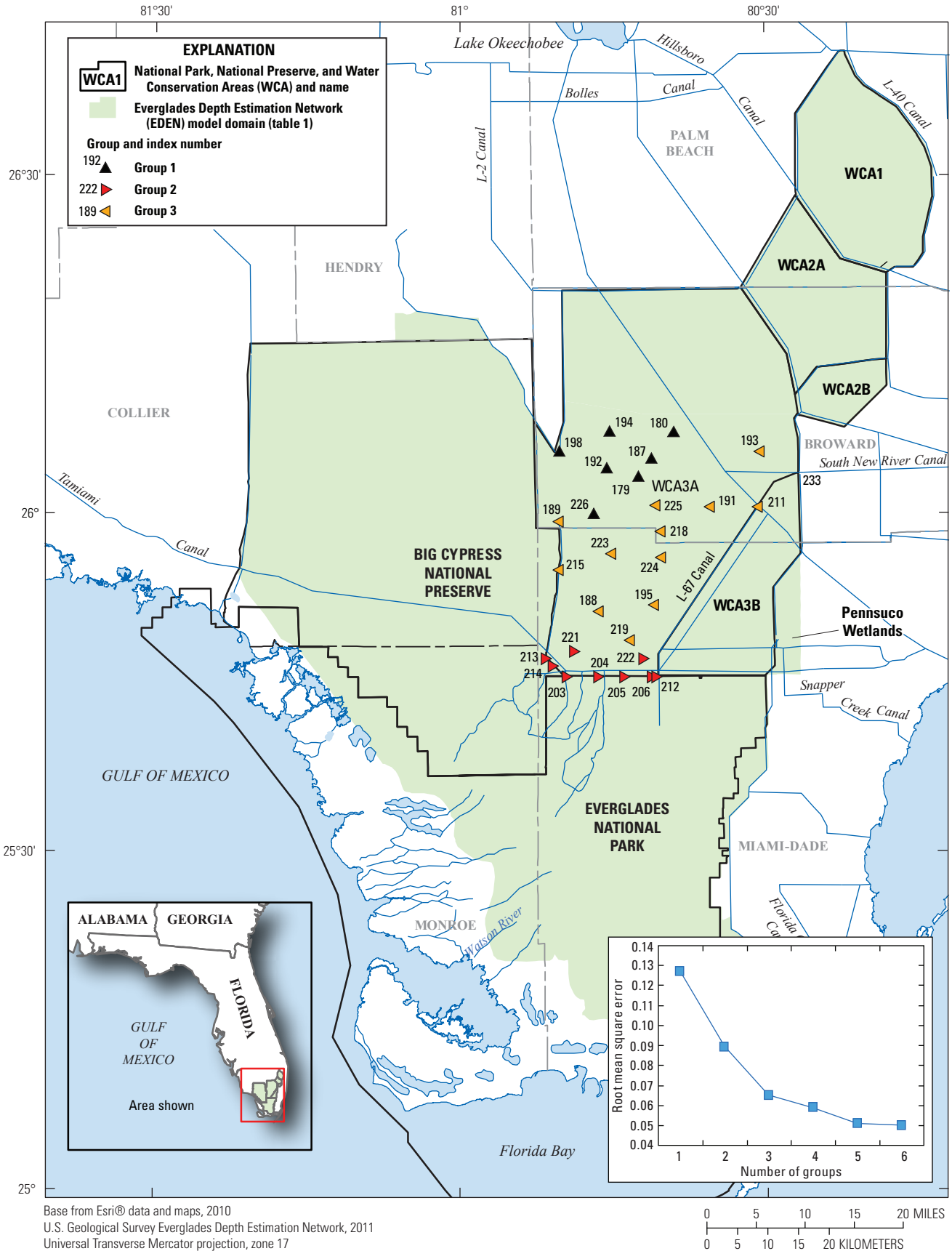
Figure 8. Map showing the time-series cluster analysis results for Water Conservation Area 3A.
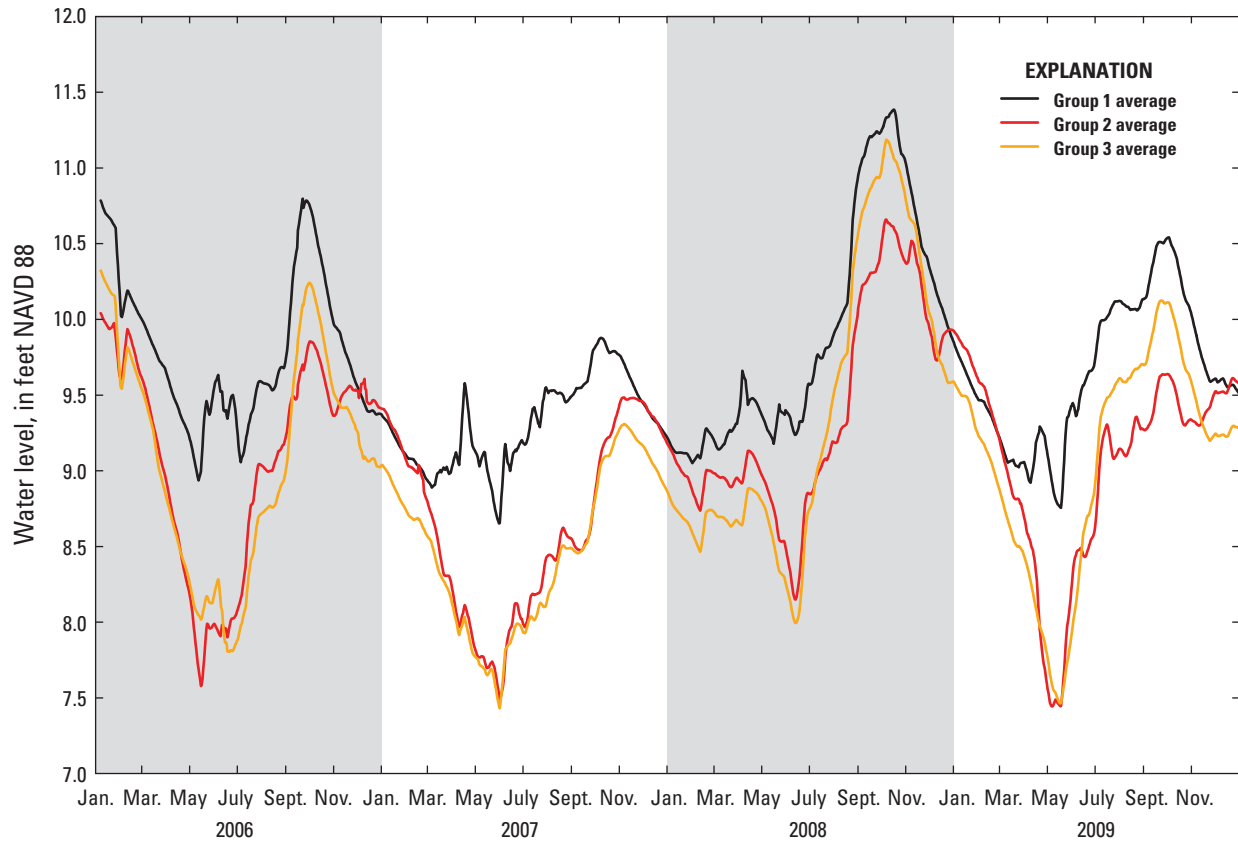
Figure 9.   Plot showing the 7-day average water levels for the Group 1, Group 2, and Group 3 gages in Water Conservation Area 3A South for the period January 1, 2006 to December 31, 2009.

Empirical Estimation Models

The most common empirical modeling approach is OLS, which relates variables using straight lines, planes, or hyper-planes, whether the actual relations are linear or not. Artificial neural network models also are empirical models, are developed directly from data, and are a flexible mathematical structure capable of describing complex nonlinear relations between input and output datasets. Calibration of either type of model attempts to optimally synthesize a line or surface through the measured data. The principal advantages of empirical models, such as ANN models, over mechanistic models are that they can be developed faster and are more accurate, provided the modeled systems are well characterized by data. Empirical models, however, are prone to problems when poorly applied. Overfitting and multicollinearity caused by correlated input variables can lead to invalid mappings, or relations, between input and output variables (Roehl and others, 2003). Of the 214 water-level estimation models, 189 are linear regression models and 25 are ANN models (table 3).

Linear Regression Models

Linear regression models often are used to estimate values of one variable based on another variable that has more data (Helsel and Hirsch, 1995). Linear regression models for water-level estimation equations are in the form of $y=mx+b$, where y is the estimated value, m is the slope, x is the value from the input ("predictor") station, and b is the y-intercept. All the linear regression models in the study used only one input station. For some sites, there are periods when the input station is missing data and no estimates are made using that model. For these cases, a second model was developed to estimate the data when the first input is missing. There is one site, CV5NR (fig. 1, map label 41), where a third model needed to be developed because of periods of missing data from the first two input sites. The input variables for the models are listed in table 3. A description of the variables is provided in table 1-1 of the appendix.

Artificial Neural Network Models

In most cases, water-level data at the input and output stations are well correlated and linear regression models produced satisfactory results. For stations where the linear regression estimates were not adequate, usually evaluated by visual inspection of measured and estimated data, an ANN model was developed. ANN models are a flexible mathematical structure capable of describing complex nonlinear relations between input and output datasets. The structure of ANN models is loosely based on the biological nervous system with interconnections of neurons and synapses (Hinton, 1992). Although numerous types of ANN models exist, the most commonly used type of ANN model is the multilayer perceptron (Rosenblatt, 1958), which is used in this study and described in detail by Jensen (1994). Multilayer perceptron ANNs can synthesize functions to fit multidimensional, nonlinear data. Multilayer perceptron ANNs are constructed from layers of interconnected processing elements called neurons that execute a simple "transfer function" (fig. 10). All input layer neurons are connected to all hidden layer neurons, and all hidden layer neurons are connected to all output layer neurons. Multiple hidden layers are possible, but a single layer is sufficient for most problems.

Typically, as well as for this study, linear transfer functions are used to scale input values from the input layer to the hidden layer, yielding values that generally are within the range that corresponds to the most linear part of the s-shaped sigmoid nonlinear transfer functions used from the hidden layer to the output layer (fig. 10). Each connection has a weight (w_i) associated with it, which scales the output received by a neuron from a neuron in an antecedent layer. The output of a neuron is a simple combination of the values it receives through its input connections and the associated weights, as well as the neuron's transfer function.

An ANN is "trained" by iteratively adjusting its weights to minimize the error by which it maps inputs to outputs for a dataset composed of input/output vector pairs. Prediction accuracy during and after training can be measured by a number of metrics, including R^2 and RMSE. An algorithm that commonly is used to train multilayer perceptron ANNs is the back error propagation (BEP) training algorithm (Rumelhart and others, 1986). The algorithm optimally minimizes the error in the objective function by adjusting the weights into and out of the hidden layer of the model (fig. 10).

Experimentation with various ANN model architectural and training parameters is a typical part of the modeling process. For correlation analysis or predictive modeling applications, a number of potential ANN models are trained and evaluated for their statistical accuracy and their representation of dynamics of the system. Interactions between combinations of variables are considered, in addition to the selection of the training dataset from the overall dataset. In developing ANN models, it is customary to set aside a subset of the data to provide an independent evaluation of model performance. Typically, datasets are divided into "training" and "testing" datasets. There are many strategies for partitioning data into training and testing datasets, but the most common is random selection of a specified percentage of the total population of measurements.

The models were calibrated by using the training dataset and evaluated with the testing dataset. For models with a large dataset with good representation over the range of historical behaviors, a small percentage of the dataset (10–25 percent) may be selected for the training dataset. For models with limited data, a larger percentage (75–100 percent) may be used in the training dataset. To mitigate the extrapolation and sparseness issues, the ANN models were conservatively trained using a method referred to as "stop training" to both fit the data and extrapolate in a minimally nonlinear and, therefore, predictable fashion. Stop training simply means stopping the training process before the ANN has fit the data to the maximum extent possible. Architectural and training parameters allow the modeler to control the geometric complexity of the surface that the ANN fits to the data. The ANN models in this report were developed using the iQuest data-mining software1[1] (Version 2.03C DM Rev31). The software writes R^2 and RMSE to the graphical user interface (GUI) during training, and an inflection in the rate of change in these parameters indicates a transition from a generally linear, multivariate surface fit to a progressively nonlinear fit. This inflection point was used to trigger stop training.

---

[1] The iQuest software is exclusively distributed by Advanced Data Mining, LLC, http://advdmi.com/Services.aspx
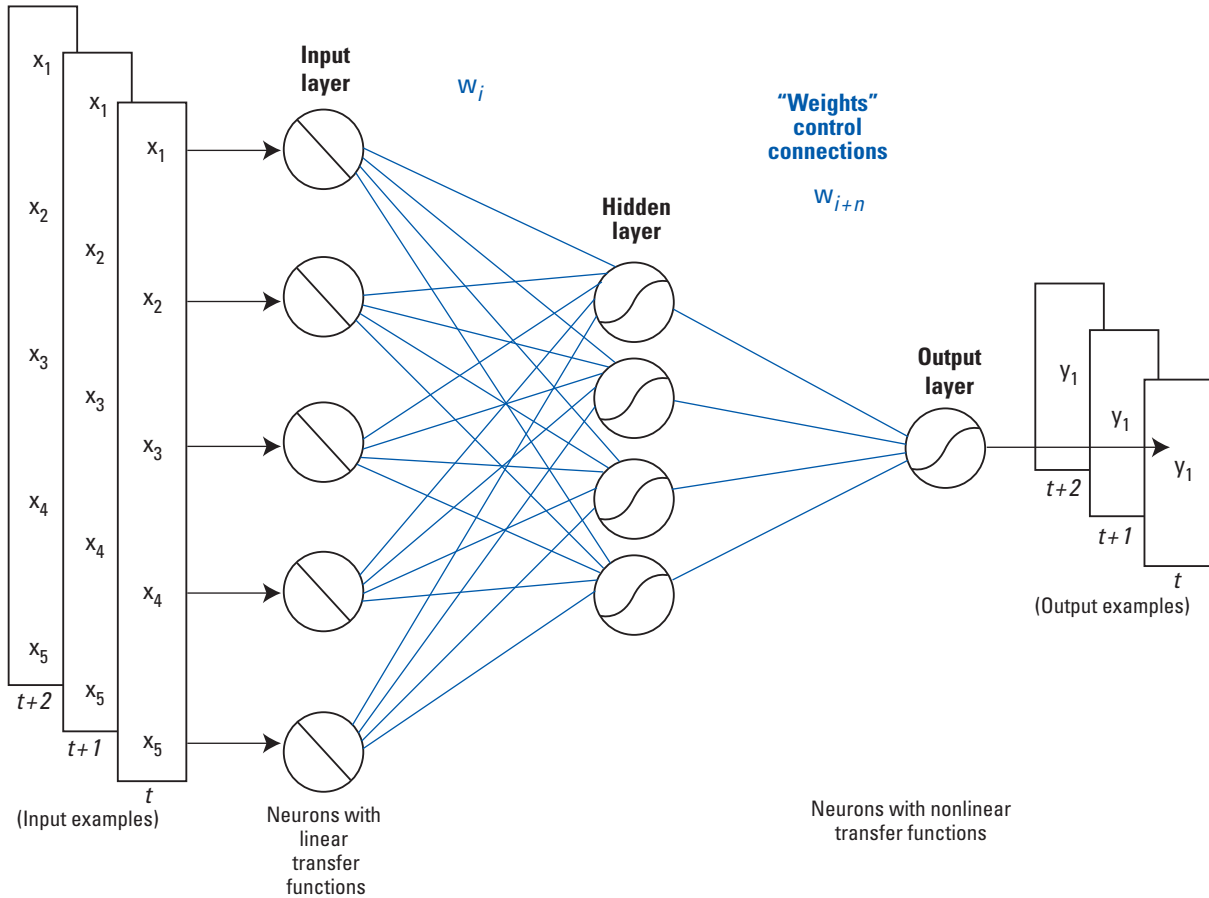
Figure 10.   Multilayer perceptron artificial neural network architecture.

Decorrelation of Input Variables

Many of the ANN models are multivariate models using more than one input variable. Usually, input variables share information about the behavior of an output variable. It is difficult, if not impossible, to understand the individual effects of these input variables (sometimes known as confounded or correlated variables), on an output variable. Empirical models have no notion of process physics or the nature of interrelations between input variables. To clearly analyze the effects of confounded variables, the unique informational content of each variable must be determined by "de-correlating" the confounded variables. For example, the three long-term stations, SITE_64, SITE_65, and SITE_69W (fig. 1, map labels 218, 219, and 220, respectively), are

highly correlated. If the time-series data from these stations are used as input to a model they would need to be decorrelated from each other. Decorrelation can be done by using a number of cascading models where the output from one model is used as input to a subsequent model. For example, SITE_65 and SITE_69W can be decorrelated from SITE_64 in two steps (fig. 11). The first step is to generate a Single Input Single Output (SISO) ANN model using SITE_69W as input and SITE_64 as output. The residual error (the difference between predicted and measured values) is the "unshared" information between the two signals and the decorrelated signal for SITE_69W. The second step is to build an ANN model using SITE_65 and the decorrelated signal for SITE_69W to predict SITE_64. The residual of this model is the decorrelated signal for SITE_65.

WL SITE_64$_{measured}$

First decorrelation model

| WL SITE_64$_{pred1}$ = F$_1$[WL SITE_69W] |

$\longrightarrow \ominus \longrightarrow$

| WL SITE_69W$_{decorrelated}$ |

WL SITE_64$_{measured}$

Second decorrelation model

| WL SITE_64$_{pred}$ = F2[WL SITE_65 and WL SITE_69W$_{decorrelated}$] |

$\longrightarrow \ominus \longrightarrow$

| WL SITE_65$_{decorrelated}$ |

Figure 11.   Schematic showing the decorrelation of water levels (WL) for SITE_64, SITE_65, and SITE_69W.

Statistical Measures of Water-level Estimation Model Prediction Accuracy

Statistical measures of prediction accuracy were computed for the water-level estimation models. Each statistic measures a different aspect of the accuracy of the prediction equations. Estimation accuracy commonly is reported in terms of R^2 and is interpreted as the goodness-of-fit of an equation or model. An interpretation of the statistic may determine how much information one variable, or a group of variab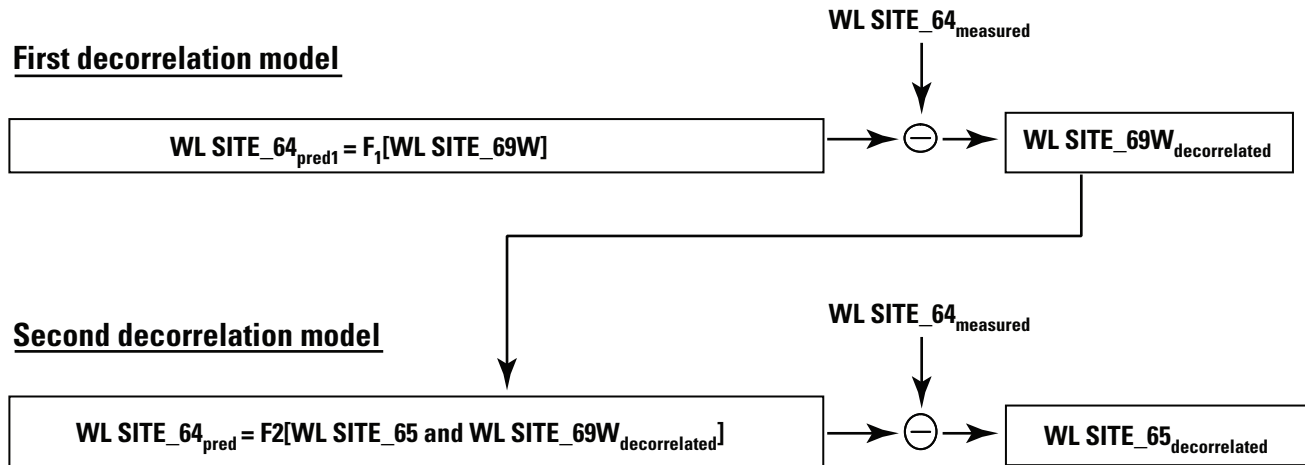les, provides about the behavior of another variable. In this context, an R^2=0.6 might be unsatisfactory. Another interpretation is that R^2 is merely an accounting of how much information is shared by the variables being used. The standard error is the measure of the scatter of the actual observations about the regression line and is the standard deviation of the error of the predicted values in the regression. The standard error can be used to compute confidence intervals for the predictions. The mean error (ME) and RMSE statistics provide a measure of the prediction accuracy of the estimation equations. The ME is a measure of the bias of model predictions—whether the model over- or under-predicts the measured data. The ME is the overall adjustment of the estimated values required to equal the measured values; therefore, positive and negative MEs indicate an over- or under-prediction bias by the model, respectively. MEs near zero may be misleading because negative and positive discrepancies in the simulations can cancel each other. RMSEs address the limitations of ME by computing the magnitude, rather than the direction (sign) of the discrepancies. The units of the ME and RMSE statistics are the same as the variable simulated by the model. ME and percent model bias (ME divided by the mean observed value) were mostly zero for all equations and, therefore, are not included in the goodness-of-fit statistics. ME and percent model bias equal zero because

simple linear regression equations, such as those described in this report, inherently minimize the sum of the residuals (error) between the estimates and the measured data. The sum of the residuals, and therefore, ME, are zero for all stations. The accuracy of the models, as given by RMSE, should be evaluated with respect to the range of the output variable. A model may have a low RMSE, but if the range of the output variable is small, the model may only be accurate for a limited range of conditions and the model error may be a relatively large percentage of the model response. Likewise, a model may have a large RMSE, but if the range of the output variable is large, the model error may be a relatively small percentage of the total model response. The percent model error (PME) is computed by dividing the RMSE by the range of the measured data.

The performance statistics for the water-level estimation models are listed in table 1-2 of the appendix. The period used to compute the model statistics is January 1, 1991, to December 31, 2010. The statistics for the individual ANN model development (table 4) are not necessarily indicative of the quality of the final water-level estimates, which are based on multiple models. The number of data points used to train and test the ANN models, including the decorrelation models, are listed in table 4 along with the number of hidden layer neurons (HLNs). Many of the ANN models are multivariate models that use more than one explanatory variable. The statistics for the ANN models in table 1-2 in the appendix are based on all the water-level estimates, and not computed separately by the training and (or) testing datasets. Overall, the water-level estimation model performance statistics are satisfactory. Ninety percent of the water-level estimation models have PMEs of 10 percent or less (fig. 12) with the median value as 5 percent (table 5). The median R^2 for the models is 0.91 and the median standard error is 0.19 ft.

Table 4. Summary of the performance statistics for the training and testing ANN models.

[EDEN, Everglades Depth Estimation Network;HLN, hidden layer neruons; Min, minimum; ft, feet; Max, maximum; N, number of data points; SSE, sum of squared error; RMSE, root mean square error; MSE, mean square error; R^2, coefficient of determination; PME, percent model error;_f, filled; _dec, decorrelated]

Model name	EDEN station name	Map label number	HLN	Min, ft	Max, ft	N_{train}	SSE_{train}, ft	R^2_{train}	$RMSE_{train}$, ft	MSE_{train}, ft	PME_{train}, percent	N_{test}	SSE_{test}, ft	R^2_{test}	$RMSE_{test}$, ft	MSE_{test}, ft	PME_{test}, percent
								Big Cypress National Preserve									
bca2.n32	BCA2	4	2	8.66	14.63	357	197.28	0.77	0.74	0.55	12.45	6,913	3,290.5	0.69	0.69	0.48	5.80
bca3.n32	BCA3	5	3	8.98	13.52	56	4.29	0.91	0.28	0.08	6.10	6,977	758.2	0.83	0.33	0.11	5.47
bca4.n32	BCA4	6	2	5.62	10.61	101	11.17	0.93	0.33	0.11	6.66	7,269	1,665.6	0.79	0.48	0.23	7.31
bca5_1.n32	BCA5	7	3	6.29	10.36	362	30.33	0.92	0.29	0.08	7.12	6,952	620.0	0.92	0.30	0.09	4.25
bca12_1	BCA12	13	3	10.30	14.63	371	48.52	0.89	0.36	0.13	8.35	6,610	799.7	0.84	0.35	0.12	4.23
bca13.n32	BCA13	14	4	7.31	12.39	101	5.75	0.96	0.24	0.06	4.70	7,243	905.6	0.84	0.35	0.13	7.62
bca14_1.n32	BCA14	15	2	4.14	9.97	369	152.68	0.83	0.64	0.41	11.02	6,856	1,988.6	0.78	0.54	0.29	5.08
bca15_1.n32	BCA15	16	1	0.63	7.89	296	15.03	0.87	0.23	0.05	3.10	3,327	147.0	0.84	0.21	0.04	6.89
bca16_1.n32	BCA16	17	2	8.11	12.15	286	46.53	0.85	0.40	0.16	9.97	3,156	347.3	0.85	0.33	0.11	3.38
bca8_1.n32	BCA17	18	3	12.05	14.41	300	8.31	0.94	0.17	0.03	7.04	3,311	104.2	0.91	0.18	0.03	2.53
bca17_1.n32	BCA18	19	3	8.25	13.63	296	44.41	0.90	0.39	0.15	7.20	3,333	446.3	0.88	0.37	0.13	5.19
eden6_1.n32	EDEN_6	22	3	8.55	10.87	53	0.75	0.97	0.12	0.01	5.14	1,169	26.9	0.90	0.15	0.02	2.96
s343a_t_3.n32	S343A_T	30	4	6.70	10.26	292	8.34	0.94	0.17	0.03	4.75	5,374	133.9	0.92	0.16	0.02	3.34
s343B_T_r1.n32	S343B_T	31	3	6.49	10.11	4,080	195.48	0.82	0.22	0.05	6.05	1,402	67.9	0.81	0.22	0.05	3.67
								Water Conservation Area 2A									
eden_11_m1.n32	EDEN_11	144	4	10.59	13.57	1081	56.42	0.85	0.23	0.05	7.67	377	22.3	0.79	0.24	0.06	3.20
wca2an1.n32	WC2AN1	159	2	8.77	14.13	273	61.11	0.67	0.47	0.22	8.83	99	24.3	0.72	0.50	0.25	5.76
wca2as1.n32	WC2AS1	160	2	11.03	12.91	262	7.32	0.88	0.17	0.03	8.89	95	2.8	0.90	0.17	0.03	1.95
wca2art_m1.n32	WCA2RT	165	2	9.32	14.23	3951	130.01	0.95	0.18	0.03	3.69	1,313	44.0	0.95	0.18	0.03	5.00
								Water Conservation Area 2B									
s144t_m1.n32	S144_T	173	2	8.32	10.37	1252	36.30	0.88	0.17	0.03	8.31	1,188	34.2	0.89	0.17	0.03	2.05
s145t_m1.n32	S145_T	174	3	8.20	10.36	1252	40.35	0.84	0.18	0.03	8.28	1,188	37.7	0.86	0.18	0.03	2.16
s146t_m1.n32	S146_T	175	2	8.37	10.59	1231	54.25	0.82	0.21	0.04	9.46	1,161	56.2	0.80	0.22	0.05	2.34
								Water Conservation Area 3A									
eden1_2.n32	EDEN_1	190	2	7.14	8.60	44	0.28	0.97	0.08	0.01	5.44	1,090	5.5	0.95	0.07	0.01	1.31
s343b_h-1.n32	S343B_H	214	2	7.58	11.66	11	0.05	1.00	0.06	0.00	1.57	5,482	28.1	0.99	0.07	0.01	4.58
W2_r2.n32	W2	221	2	6.68	9.58	698	4.49	0.98	0.08	0.01	2.77	724	5.0	0.98	0.08	0.01	3.00
W18_r2.n32	W18	226	2	7.92	10.41	559	9.08	0.96	0.13	0.02	5.12	600	8.5	0.96	0.12	0.01	2.33

| Decorrelation model name	Decorrelated variable	Correlated variable	HLN	Min, ft	Max, ft	N_{train}	SSE_{train}, ft	R^2_{train}	$RMSE_{train}$, ft	MSE_{train}, ft	PME_{train}, percent	N_{test}	SSE_{test}, ft	R^2_{test}	$RMSE_{test}$, ft	MSE_{test}, ft	PME_{test}, percent
S7_tdecor.n32	S7_T_dec	EDEN11_f	3	7.83	14.47	5,736	3,102	0.69	0.74	0.54	11.08	1,928	938	0.71	0.70	0.49	6.62
S7_t_dec-site17.n32	Site17_dec-S7T	S7_T	1	7.73	13.65	5,732	1,170	0.75	0.45	0.20	7.63	1,925	372	0.76	0.44	0.19	5.92
S190-t_dec.n32	S190-T_dec	L28GAP	2	7.84	12.07	3,785	454	0.82	0.35	0.12	8.19	3,679	445	0.82	0.35	0.12	4.31
bca14_dec.n32	BCA14_dec	BCA2_f	2	9.30	14.20	11	0	0.98	0.21	0.04	4.28	7,853	3,948	0.68	0.71	0.50	16.74
loop1_h_dec.n32	LOOP1_H_dec	S343A_T	3	6.49	10.45	2,737	67	0.93	0.16	0.02	3.97	2,637	67	0.92	0.16	0.03	4.06
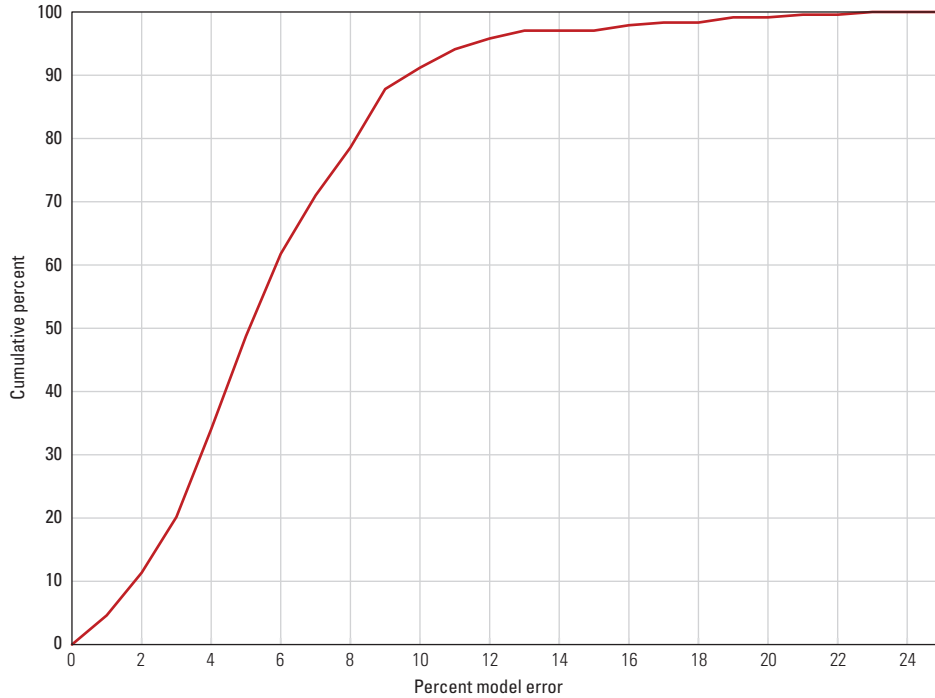
Figure 12.   Frequency distribution of percent model error of the water-level estimation models.

Table 5.   Minimum, median, and maximum values for the summary statistics for 237 water-level estimation models.

[R², coefficient of determination; RMSE, root mean square error]

Statistic	Standard error, in feet	R²	Mean error, in feet	RMSE, in feet	Percent model error
Minimum	0.00	0.26	−0.92	0.03	0.50
Maximum	0.85	1.00	0.88	1.11	22.32
Median	0.19	0.91	0.00	0.20	5.04

Shifting Water-Level Estimates

Once water levels are estimated, hydrographs of each station are evaluated to identify potential errant data and to evaluate the fit of the water-level estimates. For some stations, estimated data are a good representation of the missing data and can be used without any adjustment. Some of the water-level estimates follow the trend of the measured data but are offset from the measured data. At these sites, the relative water-level estimates are good but the absolute estimates are in error. In the Everglades, differences in water level of only 2 or 3 in. can be important to the ecology. To improve the quality of the water-level estimates, shifting techniques are applied to the estimates similar to gage-height correction techniques used for computing continuous water-level data records (Rantz and others, 1982). The applications of shifts are applied for periods of missing water-level data and therefore do not affect the model performance statistics.

The most straight-forward type of data correction is one where a uniform (constant) correction value is applied to the estimates. Prorated corrections are applied over time to improve the quality of the water-level estimates. An example of applying constant and prorated shifts to water-level estimates for NP46 (map label 87, fig. 1) are shown in figure 13. The water-level estimation model of the site captures the overall trend of the data but the estimates are over-predicted by 0.2 – 0.3 ft. Shifts were applied to improve the absolute prediction of the estimates while maintaining the dynamic variability of the model. The shifts were determined by subtracting the modeled water level from the measured water level for the day before and after the data gap. For the short data gaps from September 25 to October 15, 1993, a shift of –0.31 ft was started on September 25 and prorated to –0.24 ft on October 15. For the large data gap from November 11, 1993, to January 20, 1994, a shift of –0.26 ft was applied on November 11 and prorated to –0.22 ft on January 20.
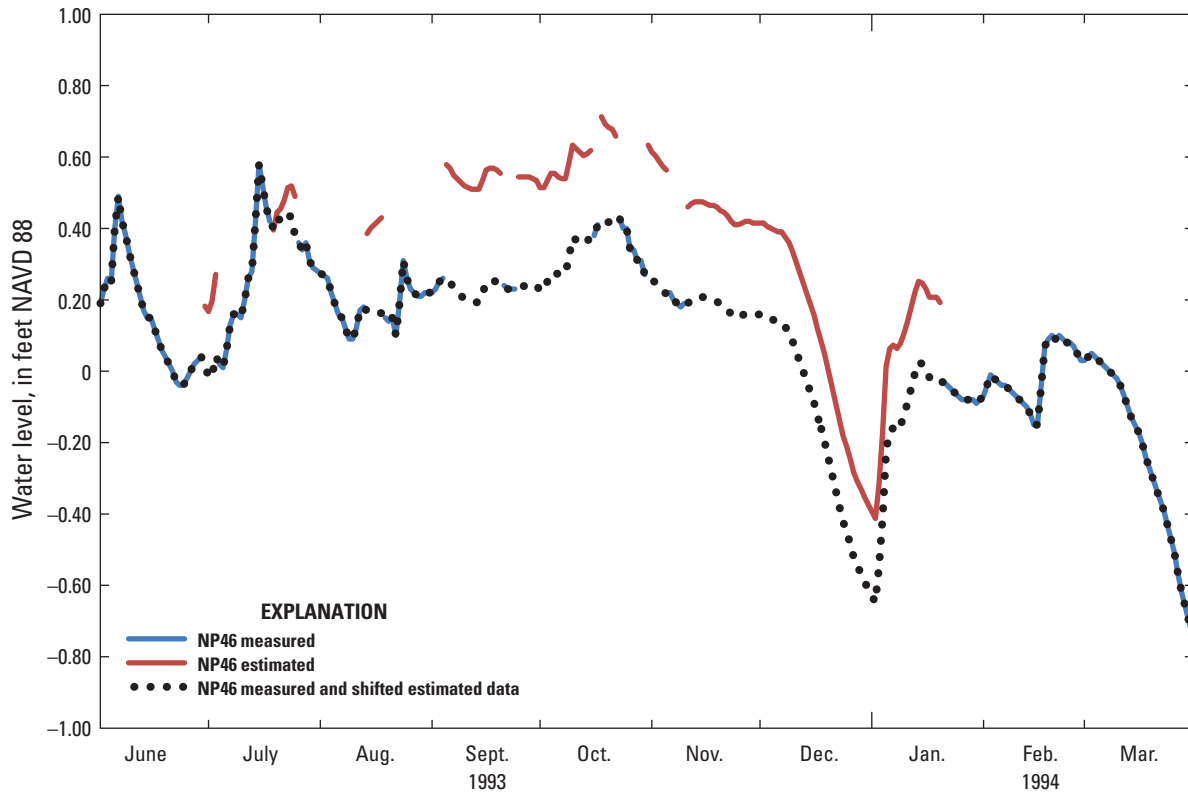
Figure 13. Example of measured, estimated, and shifted estimated data at NP46 over the period June 1993 to March 1994.

Evaluation of Hindcasts using Modeled Water-Level Surfaces

The statistical evaluation of the water-level estimation models quantifies the predictive ability of the model to capture the water-level dynamics for the individual station on a temporal basis. For generating water-level surfaces for the Everglades using the V2 model, the hindcasted water-level estimation models must perform spatially as a group to create water-surface maps that are credible. Individual models may be accurate but model error between sites may cause erroneous water-surface gradients between stations. To evaluate the performance of the hindcast models as a group, contour maps of modeled water-level surfaces at 2-centimeter (cm) intervals were generated using the hindcasted data. The 2-cm contour maps were examined for selected days to verify that modeled water surfaces from the V2 model were consistent with the input data. The high resolution provided by the 2-cm contour intervals highlights areas of the model domain where input data may be incorrect and where water-level gradients may not be consistent with current knowledge of sheet flow in the marshes.

Rather than evaluate water-surface maps for each of the 3,650 days of the hindcasted period, which was time prohibitive, the water-level surfaces generated by the hindcasted dataset (1990–99) were compared with water-level surfaces generated with the measured dataset (2000–10) for a few high- and low-water days. Daily hydrographs were generated using measured long-term data from BCNP, ENP, and the WCAs for 1990 to 2010. The 21-year hydrographs were used to select dates for representative water conditions in the measured period (2000–10) and hindcasted period (1990–99). Because of the natural interannual climate variability in addition to operational differences between the WCAs and BCNP and ENP, it is difficult to identify periods of similar water-level conditions (within +/– 0.3 ft) for the domain of the V2 model. Priority was set on finding similar conditions between ENP and WCA3A. For high-water conditions, the hindcasted surfaces for September 26 and November 19, 1995, were compared with the measured water-level surfaces for October 2, 2008. For low-water conditions, the hindcasted surfaces of March 3, 1994; April 24, 1997; and May 21, 1999 were compared with the measured water-level surface for March 2, 2008.

Overall, the hindcasted 2-cm contours for the high-water condition were similar to the contours generated with the measured data. There were areas on the hindcasted surfaces where potential problems with the hindcasted water-level data were evident along the Tamiami Canal, along the L-67 Canal near SITE_69W, and in central and northern areas of WCA2A. The issues with the contours are not only issues with the water-level estimations of the hindcasts but also are caused by issues with the data. For example, the irregularities along the Tamiami Trail were caused by the datum correction from the National Geodetic Vertical Datum of 1929 to NAVD 88 not being applied to S12B_T (map label 109, fig. 1). S12B_T was not a hindcasted site but data gaps in the record were filled. For the irregularities near SITE_69W, the water-level data were compared with SITE_64, SITE_71, and SITE_76 and no apparent data problems were observed. There are two stations at the locations along the L67 canal: SITE_69W in the canal and SITE_69E (not shown in fig. 1) in the marsh of WCA3B to the east of the canal levee. The water-level data for SITE_69E was mistakenly used in the model rather than SITE_69W and the lower water levels caused the irregularities in the contours. Correcting the datum conversion at S12B_T and substituting data for SITE_69W for modeling WCA3B produced water-level surfaces consistent with adjacent station data.

The issues in the northern area of WCA2A (fig. 14) were a result of a limited number of stations in the areas to use for input stations for the water-level estimation models (fig. 3). The initial linear regression models for WC2AN1, WC2AS1, and EDEN_11 used measured data from the S7_T (for the WC2AN1 and EDEN_11 models) and SC11C_T (for the WC2AS1 model). Individually, the initial linear regression water-level estimates for WC2AN1, WC2AS1, and EDEN_11 seemed satisfactory, but plotted together (fig. 15), the water levels did not follow the ground elevation slope. The topographic slope in WCA2A generally is to the south and southwest (Jones and others, 2007b, 2012). If water levels generally were the same depth, the water levels at the three hindcasted sites would be similar and water levels at a station to the south, such as SITE_17 would be lower. The four sites would have a general downward gradient from WC2AN1 to WC2AS1 to EDEN_11 to SITE_17. The initial water-level estimation models did not follow the general gradient with water levels because WC2AN1 was similar to SITE_17 farther to the south (fig. 15A). The three sites were re-modeled using ANN models and different input stations. Rather than only using measured input datasets, hindcasted and filled datasets were used for input stations. EDEN_11 was modeled with the hindcasted and filled dataset for WCA2F1 (map label 163, fig. 1, table 3). WC2AS1 was then modeled with a multivariate model using the EDEN_11 predictions and S7_T that was decorrelated from the EDEN_11 predictions (table 3). WC2AN1 was then modeled using the WC2AS1 predictions (table 3). The final water-level estimations (fig. 15B) follow the general ground elevation gradient.

In addition to comparing the 2-cm contoured water-level maps for the selected high-water and low-water days, 240 2-cm contour water-level maps were generated for the 1st and 15th of each month for January 1, 1990, to December 31, 1999. Irregularities in contours were evaluated and issues involving data inputs and (or) water-level estimation models were resolved. The biweekly 2-cm contour maps did indicate a higher number of issues in the 1990s. In May 1990, the lowest water levels of the 21-year dataset used for the hindcasting study happened in the Everglades (for example, see long-term measured water levels for SITE_64 shown in fig. 5). To estimate these low-water levels, the water-level estimation models had to extrapolate beyond the range of measured data. As discussed in the "Approach" section of this report, the models assume that the relation between input and output variable remains constant between the measured conditions and the extrapolated conditions. For the extreme low-water level during 1990, this assumption is not valid. For the low water in May 1990, many of the models would have to be extrapolated about 2 ft lower than the measured data used to develop the models. In addition, models that use a hindcasted dataset for input would have the errors for two models compounded to extrapolate to the extreme low-water levels. Additionally, the effort to resolve the water-level estimation model extrapolation is challenging because of the lack of measured data in 1990 for many areas of the EDEN domain (fig. 3). For these reasons, it was decided to limit the hindcasted data to the period January 1, 1991, to December 31, 1999.

An example of the hindcasted and measured water-surface maps is shown in figure 16. The two maps were generated for days with similar high-water conditions. The map on the left was generated using the hindcasted data for September 26, 1995, and the map on the right was generated using the measured data for October 2, 2008. The high-water conditions for ENP and WCA3A, as represented by NP62 and SITE_64, respectively (fig. 1, map labels 88 and 218) were within 0.13 ft at both gages for the two dates. Overall, the hindcasted water-elevation surface for the high-water condition is similar to the surface generated with the measured data. Overall, the hindcasted and gap-filled data are believed to provide reasonable estimates of station-specific water-level data for an extended historical period to inform research and natural resource management in the Everglades. The datasets with the record extensions are available on the EDEN Web site using the Explore and View EDEN (EVE) graphical interface (*http://sofia.usgs.gov/eden/eve/index.php*).
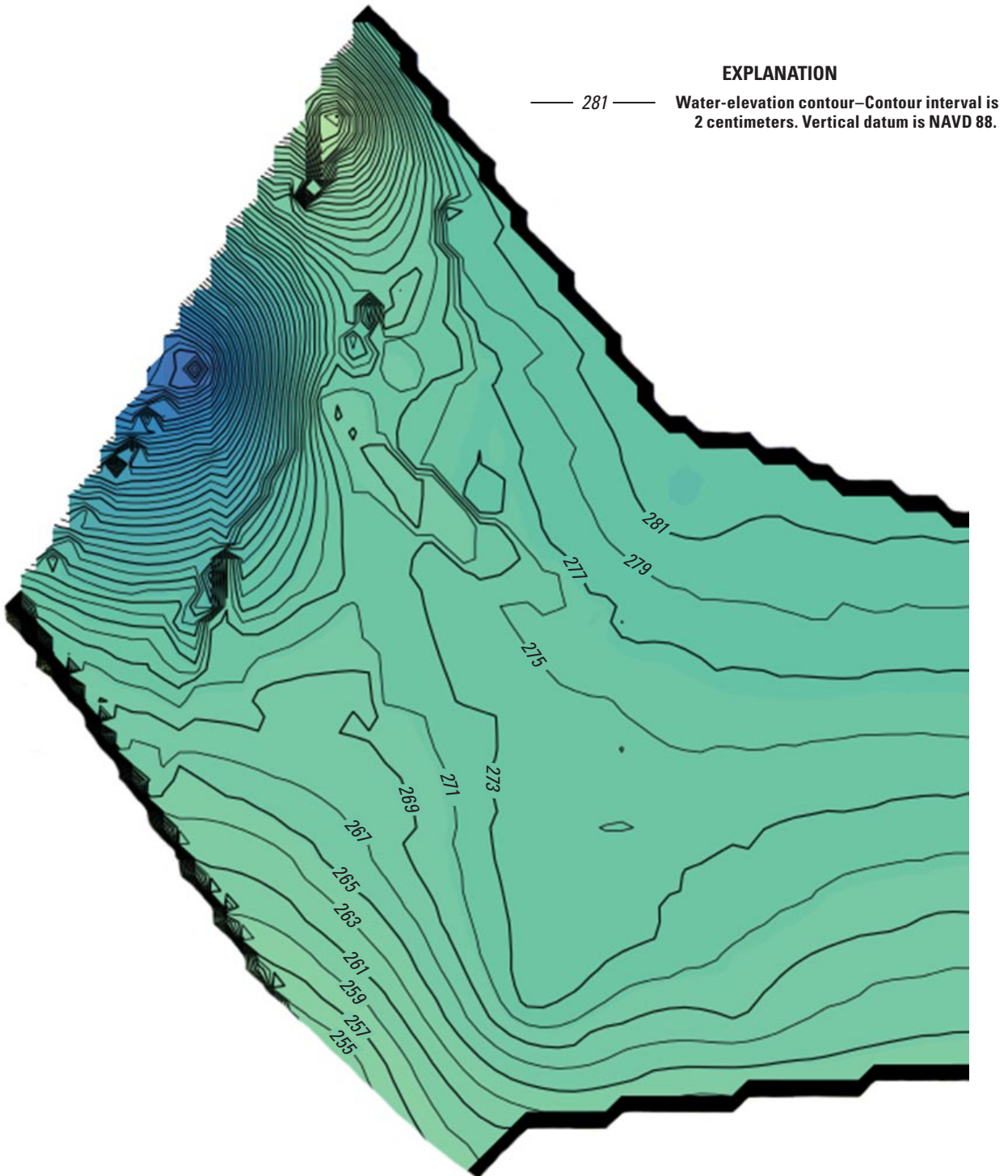
Figure 14.   Example of 2-centimeter contour water-surface elevation maps for high-water conditions on September 26, 1995 used to evaluate the preliminary hindcasted data.
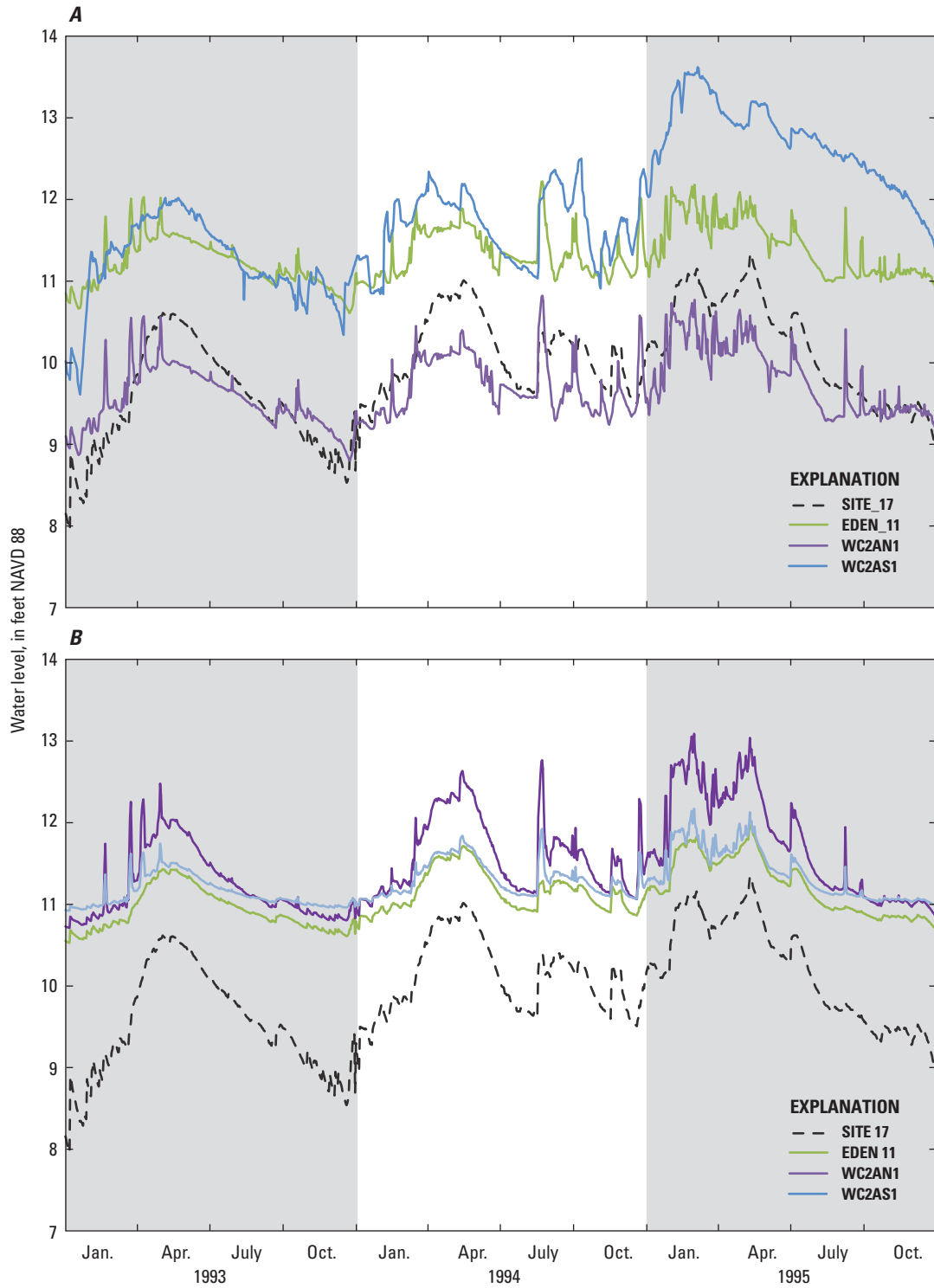
Figure 15. Measured data for SITE_17 and hindcasted water-level data for EDEN_11, WC2AN1, and WC2AS1 for *A*, preliminary water-level estimation models and *B*, final water-level estimation models.

Water elevation,
in centimeters (NAVD 88)

A

B

405

305

205

105

5
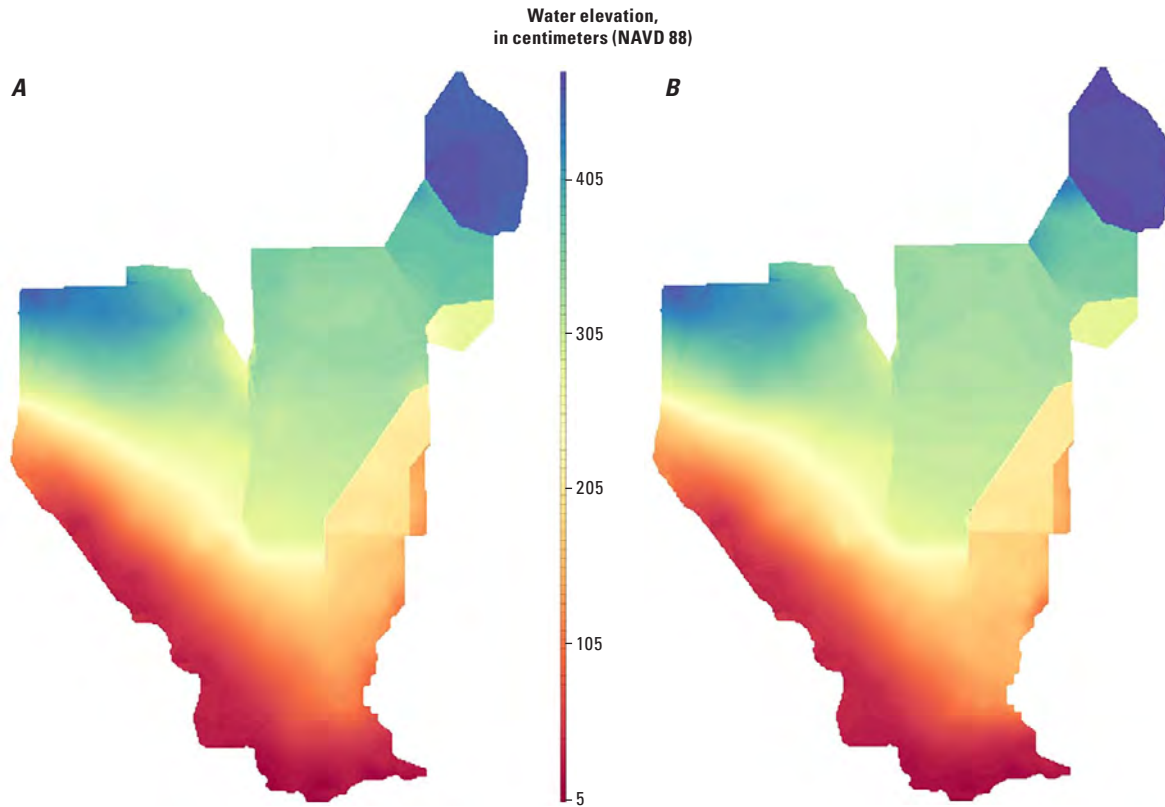
Figure 16.   Example of water-surface elevation maps for high water conditions using *A*, hindcasted data for September 26, 1995 and *B*, measured data for October 2, 2008.

Summary

The Everglades Depth Estimation Network (EDEN) water-surface elevation model uses data from approximately 240 gages in the Everglades to create daily continuous interpolations of the water-surface elevation for the freshwater portion of the Everglades from 2000 to the present (2014). To provide a longer period for assessing biological and ecological impacts, the EDEN team decided to hindcast (extend water-level data records back in time) to 1990. Before 2000, many of the EDEN gages used to generate water surfaces were not in operation. Of the 237 stations used in the EDEN water-surface model in 2000, only 106 stations were in operation in 1990. Of the 106 stations in operation, only 23 stations had complete datasets (no missing data). To create an input dataset for the EDEN water-surface model, 168 data records were hindcasted and 46 data records had missing data gaps filled.

The water-level data records were hindcasted using empirical water-level estimation models. Of the 214 water-level estimation models used to hindcast and fill data records, 189 are linear regression models and 25 are artificial neural network models. Ninety percent of the models have percent model error (root mean square error divided by the range of the measured data) of 10 percent or less. The median coefficient of determination for the models is 0.91 and the median standard error is 0.19 foot. In addition to evaluating performance of each water-level estimation model, 2-centimeter (cm) vertical resolution water-level maps were generated using the hindcasted data and evaluated for irregular contours indicating a potential problem with the hindcasted data. May 1990 had the lowest water levels in the Everglades of the 21-year dataset used for the hindcasting study. To estimate these extreme low-water levels, the models needed to extrapolate beyond the range of the measured data. Model extrapolation assumes that the relation between input and output variables remains constant and limited model extrapolation is often appropriate. For the extreme low water in 1990, however, it was decided that large water-level estimation model extrapolations were not appropriate and the hindcasted dataset was begun on January 1, 1991, and ended on December 31, 1999. The hindcasted and gap-filled data provide reasonable estimates of station-specific water-level data for an extended historical period to inform research and natural resource management in the Everglades.

References Cited

Conrads, P.A., and Roehl, E.A., Jr., 2007, Hydrologic record extension of water-level data in the Everglades Depth Estimation Network (EDEN) using artificial neural network models, 2000–2006: U.S. Geological Survey Open-File Report 2007–1350, 56 p., accessed October 23, 2014, at *http://pubs.water.usgs.gov/ofr2007–1350*.

Conrads, P.A., and Petkewich, M.D., 2009, Estimation of missing water-level data for the Everglades Depth Estimation Network (EDEN): U.S. Geological Survey Open-File Report 2009–1120, 53 p.

Daamen, R.C., Roehl, E.A., Jr., and Conrads, P.A., 2010, Development of inferential sensors for real-time quality control of water-level data for the Everglades Depth Estimation Network, in Proceedings of the 2010 South Carolina Water Resources Conference, Columbia, S.C., 2010, 4 p.

Davis, J.H., Jr., 1943, Vegetation map of Southern Florida: Florida Geological Survey, Bulletin 25, figure 71, 1 sheet, 1:400,000 scale.

German, E.R., 2000, Regional evaluation of evapotranspiration in the Everglades: U.S. Geological Survey Water-Resources Investigations Report 00–4217, 48 p.

Helsel, R.M., and Hirsch, D.R., 1995, Studies in environmental science 49—Statistical methods in water resources: Amsterdam, Elsevier Science, 529 p.

Hinton, G.E., 1992, How neural networks learn from experience: Scientific American, v. 267, no. 3, p. 145–151.

Jensen, B.A., 1994, Expert systems—Neural networks, chap. 1.7 of Lipták, B.G., ed., Instrument engineers' handbook (3d ed.): Radnor, Pa., CRC Press, p. 48–54.

Jones, J.W., Desmond, G.B., Henkle, Charles, and Glover, Robert, 2012, An approach to regional wetland digital elevation model development using a differential global positioning system and a custom-built helicopter-based surveying system: International Journal of Remote Sensing, v. 33, p. 450–465.

Jones, J.W., and Price, S.D., 2007a, Conceptual design of the Everglades Depth Estimation Network (EDEN) grid: U.S. Geological Survey Open-File Report 2007–1200, 20 p.

Jones, J.W., and Price, S.D., 2007b, Initial Everglades Depth Estimation Network (EDEN) digital elevation model research and development: U.S. Geological Survey Open-File Report 2007–1034, 29 p.

Jones, J.W., Desmond, G.B., Henkle, Charles, and Glover, Robert, 2012, An approach to regional wetland digital elevation model development using a differential global positioning system and a custom-built helicopter-based surveying system: International Journal of Remote Sensing, v. 33, p. 450–465.

Lodge, T.E., 1994, The Everglades handbook, Understanding the ecosystem: Delray Beach, Fla., St. Lucie Press, 228 p.

McPherson, B.F., Higer, A.L., Gerould, S., and Kantrowitz, I.H., 1995, South Florida Ecosystem Program of the U.S. Geological Survey: U.S. Geological Survey Fact Sheet FS-134-95, accessed June 19, 2012, at *http://sofia.usgs.gov/publications/fs/134-95/index.html*.

Pearlstine, L., Higer, A., Palaseanu, M., Fujisaki, I., and Mazzotti, F., 2007, Spatially continuous interpolation of water stage and water depths using the Everglades Depth Estimation Network (EDEN): Gainesville, Fla., Institute of Food and Agricultural, University of Florida, CIR1521, 18 p., 2 apps.

Petkewich, M.D., and Conrads, P.A., 2013, Estimation of missing water-level data for the Everglades Depth Estimation Network (EDEN), 2013 update: U.S. Geological Survey Open-File Report 2013–1251, 45 p.

Rantz, S.E., and others, 1982, Measurement and computation of streamflow: U.S. Geological Survey Water-Supply Paper 2175, 631 p.

Richardson, J.R., Bryant, W.L., Kitchens, W.M., Mattson, J.E., and Pope, K.R., 1990, An evaluation of the refuge habitats and relationships to water quality, quantity, and hydroperiods—A synthesis report: Gainesville, Fla., Florida Cooperative Fish and Wildlife Research Unit, University of Florida,166 p.

Risley, J.C., Roehl, E.A., Jr., and Conrads, P.A., 2003, Estimating water temperatures in small streams in western Oregon using neural network models: U.S. Geological Survey Water-Resources Investigations Report 02–4218, 60 p.

Roehl, Ed, Jr., Conrads, P.A., and Cook, J.B., 2003, Discussion of "Using complex permittivity and artificial neural networks for contaminant prediction": Journal of Environmental Engineering, v. 129, p. 1069–1071.

Roehl, E., Risley, J., Stewart, J., and Mitro, M., 2006, Numerically optimized empirical modeling of highly dynamic, spatially expansive, and behaviorally heterogeneous hydrologic systems—Part 1, in Proceedings of the International Environmental Modelling and Software Society 2006 Summit on Environmental Modelling and Software, Burlington, Vt., 6 p, accessed July 8, 2014, at *http://www.iemss.org/iemss2006/*.

Rosenblatt, F., 1958, The perceptron, a probabilistic model for information storage and organization in the brain: Psychological Review, v. 65, p. 386–408.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J., 1986, Learning internal representations by error propagation, in Rumelhart, D.E., McClelland, J.L., and the PDP Research Group, Parallel distributed processing—Explorations in the microstructure of cognition, volume 1: Cambridge, Mass., MIT Press, p. 318–362.

South Florida Natural Resource Center, 2005, An assessment of the Interim Operational Plan: Homestead, Fla., South Florida Natural Resources Center, Everglades National Park, Project Evaluation Report, SFNRC Technical Series 2005:2, 47 p., accessed DATE, at *http://www.nps.gov/ever/naturescience/upload/SecureIOPReportSmall.pdf*.

Stewart, J., Mitro, M., Roehl, E., and Risley, J., 2006, Optimized empirical modeling of highly dynamic, spatially expansive, and behaviorally heterogeneous hydrologic systems—Part 2, in Gourbesville, Philippe, Cunge, Jean, Guinot, Vincent, and Liong, Shie-Yui, eds., Proceedings of the 7th International Conference on Hydroinformatics, August 2006, Nice, France, v. 2, p. 831–838.

Telis, P.A., 2005, Project description, South Florida Surface Water Monitoring Network for the Support of MAP Projects (known as EDEN, Everglades Depth Estimation Network): accessed October 2, 2007, at *http://sofia.usgs.gov/projects/eden/*.

Telis, P.A., 2006, The Everglades depth estimation network (EDEN) for support of ecological and biological assessments: U.S. Geological Survey Fact Sheet 2006–3087, 4 p.

Telis, P.A., and Henkel, Heather, 2009, Everglades Depth Estimation Network (EDEN) applications—Tools to view, extract, plot, and manipulate EDEN data: U.S. Geological Survey Fact Sheet 2009–3052, 4 p.

Telis, P.A., Xie, Zhixiao, Lui, Zhongwei, Li, Yingru, and Conrads, P.A., 2014, The Everglades Depth Estimation Network (EDEN) surface-water model Version 2: U.S. Geological Survey Scientific Investigations Report 2014–5209

U.S. Army Corps of Engineers, 1999, Central and Southern Florida Project comprehensive review study: Jacksonville, Fla., Final Integrated Feasibility Report and Programmatic Environmental Impact Statement, [variously paged], 4 annexes, 15 apps.

U.S. Army Corps of Engineers, 2011, Everglades Restoration Transition Plan final environmental impact statement—Volume 1: U.S. Army Corps of Engineers, Jacksonville District, [variously paged].

U.S. Fish and Wildlife Service, 2000, Arthur R. Marshall Loxahatchee National Wildlife Refuge Comprehensive Conservation Plan: U.S. Fish and Wildlife Service, 362 p.

U.S. Geological Survey, 2007, Facing tomorrow's challenges—U.S. Geological Survey science in the decade 2007–2017: U.S. Geological Survey Circular 1309, x + 70 p.

Weiss, S.M., and Indurkhya, Nitin, 1998, Predictive data mining—A practical guide: San Francisco, Morgan Kaufmann Publishers, Inc., 228 p.

Xie, Zhixiao, Liu, Zhongwei, Jones, J.W., Higer, A.L., and Telis, P.A., 2011, Landscape unit based digital elevation model development for the freshwater wetlands within the Arthur C. Marshall Loxahatchee National Wildlife Refuge, Southeastern Florida: Applied Geography, v. 31, no. 2, p. 401–412.