

Compiling a Baseline Water Quality Database from Heterogeneous Data Sources: Lessons Learned

***Nenad Iricanin, Brian Turcotte
and R. Scott Huebner***

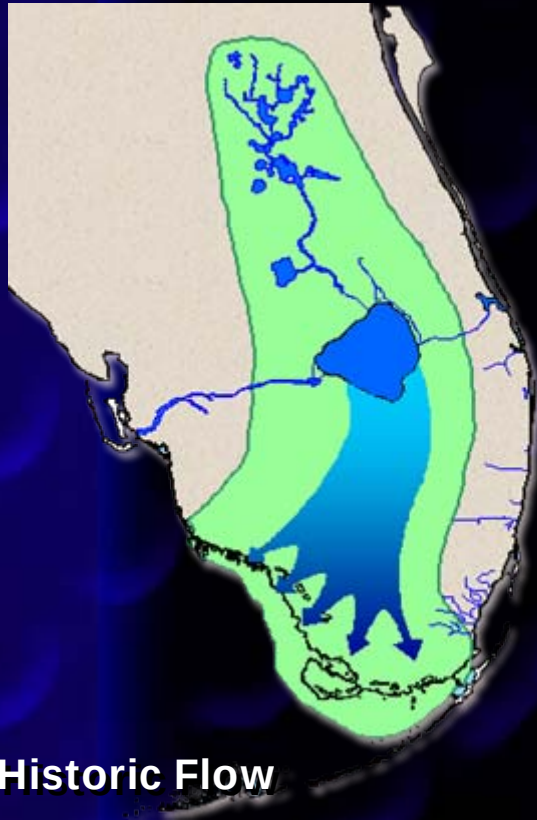
**Environmental Resource Assessment Department
South Florida Water Management District**

Goals of Project

- Establish a comprehensive water quality and hydrologic database for South Florida to characterize baseline conditions
- Consider a 10-year baseline period from May 1, 1991 through April 30, 2001
- Use baseline data to support Comprehensive Everglades Restoration Plan (CERP)

Comprehensive Everglades Restoration Plan (CERP)

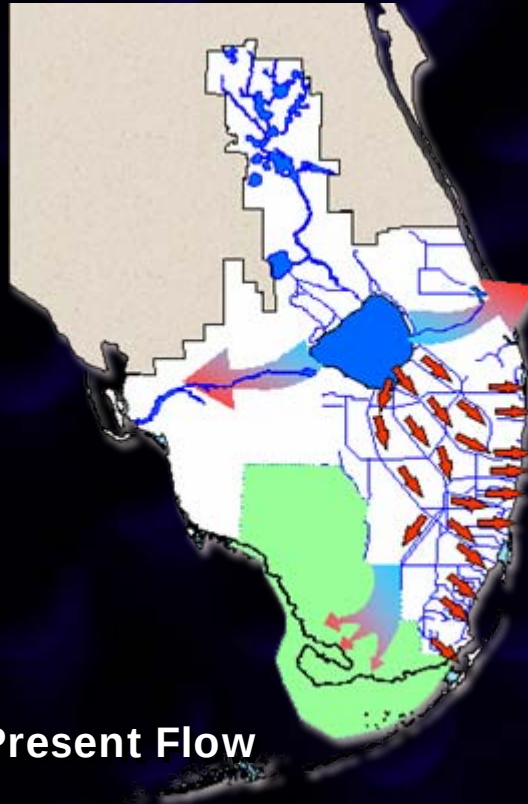
- Provides framework and guide to restore, protect and preserve the water resources of central and southern Florida (including the Everglades)
- Covers 16 counties over an 18,000 mile² area
- Plan was approved in the Water Resources Development Act (WRDA) of 2000
- The goal of CERP is to capture fresh water that is released to tide and restore the flow to the Everglades
- Redirect water to areas that need it most



Historic Flow

Sheet flow from Lake Okeechobee through the Everglades to Florida Bay

Most of the water is released from Lake Okeechobee goes to tide



Present Flow

Smaller amounts of fresh water released through Everglades to Florida Bay

Restore the fresh water flow from Lake Okeechobee through the Everglades to Florida Bay



CERP Planned Flow

Comprehensive Everglades Restoration Plan (CERP)

- One of the key documents is the CERP Monitoring and Assessment Plan (MAP)
- MAP is a product of an interagency interdisciplinary team known as Restoration, Coordination, and Verification (RECOVER)
- MAP identifies and describes *performance measures* and parameters of the natural and man-made systems in South Florida that will be used to assess restoration success

Comprehensive Everglades Restoration Plan (CERP) Water Quality Performance Measures

- Major ions
- Biological
- Nutrients
- Physical parameters
- Trace metals/elements
- Solids
- Organic carbon
- Mercury
- PAHs, PCBs, pesticides / herbicides / fungicides
- Sulfur compounds
- Light attenuation

Comprehensive Everglades Restoration Plan (CERP)

- Performance measures are organized by distinct geographic areas called modules or *eco-regions*
- Six (6) eco-regions identified in MAP

Map of South Florida

Showing four (4) distinct geographic areas or eco-regions (or modules)

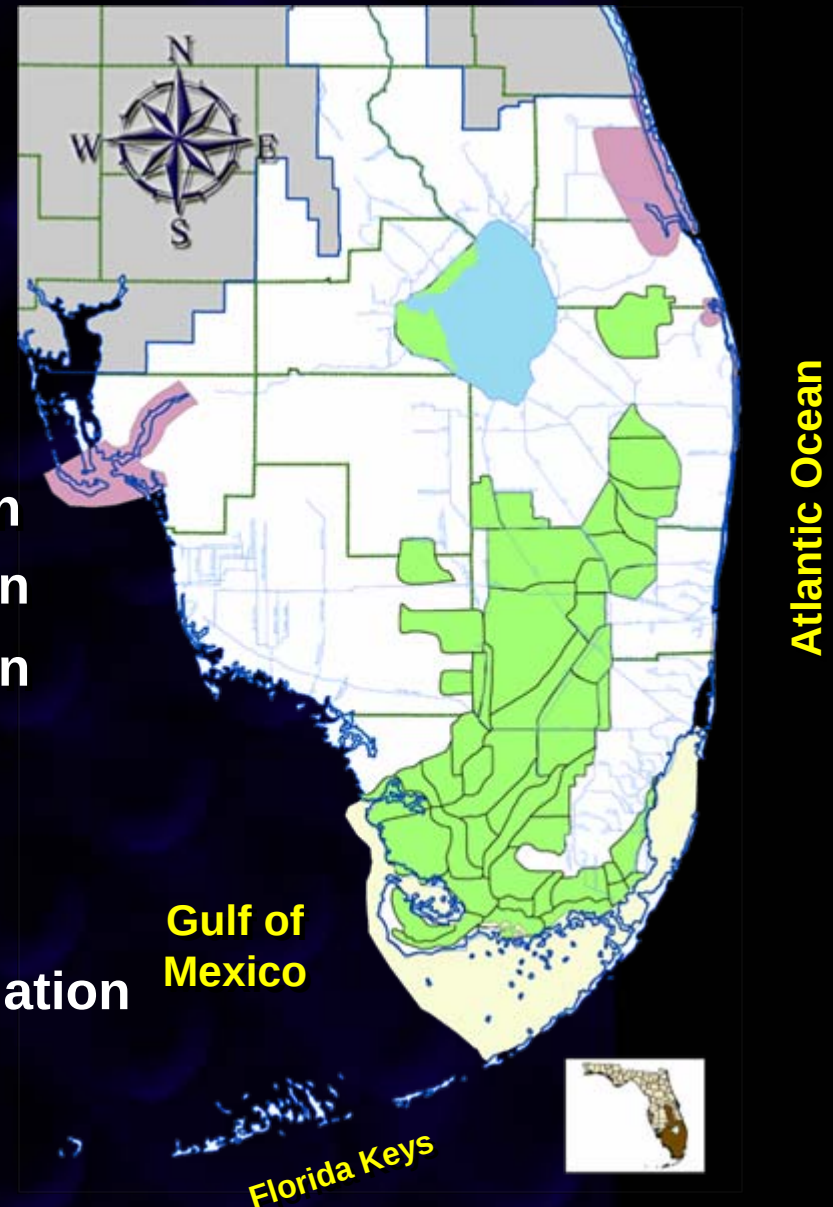
-  Lake Okeechobee Eco-region
-  Northern Estuaries Eco-region
-  Greater Everglades Eco-region
-  Southern Estuaries Eco-region

Two additional eco-regions:

- South Florida Hydrologic Network
- South Florida Mercury Bioaccumulation Monitoring Network

<http://www.everglades.org>

sfwmd.gov

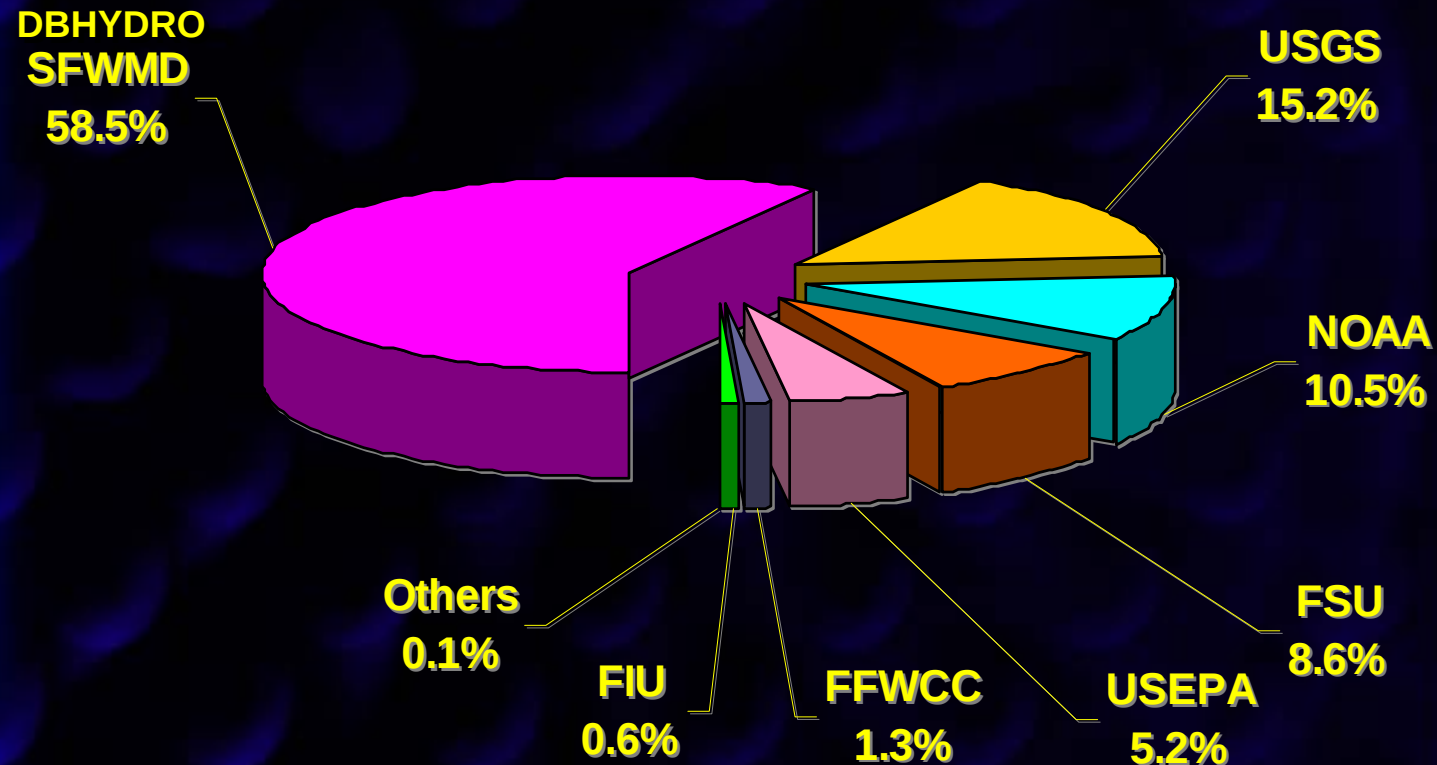


RECOVER Baseline Database

- Provide a historic database of pre-CERP (or baseline) conditions for water quality and hydrological performance measures as identified in the MAP
- 758 studies by 70 organizations identified as potential data sources
- Approximately 290 studies met the requirements for the baseline database
- A total of 79.8 million data records loaded by a contractor into baseline database
- Approximately 36 million records loaded from District's corporate database, DBHYDRO.

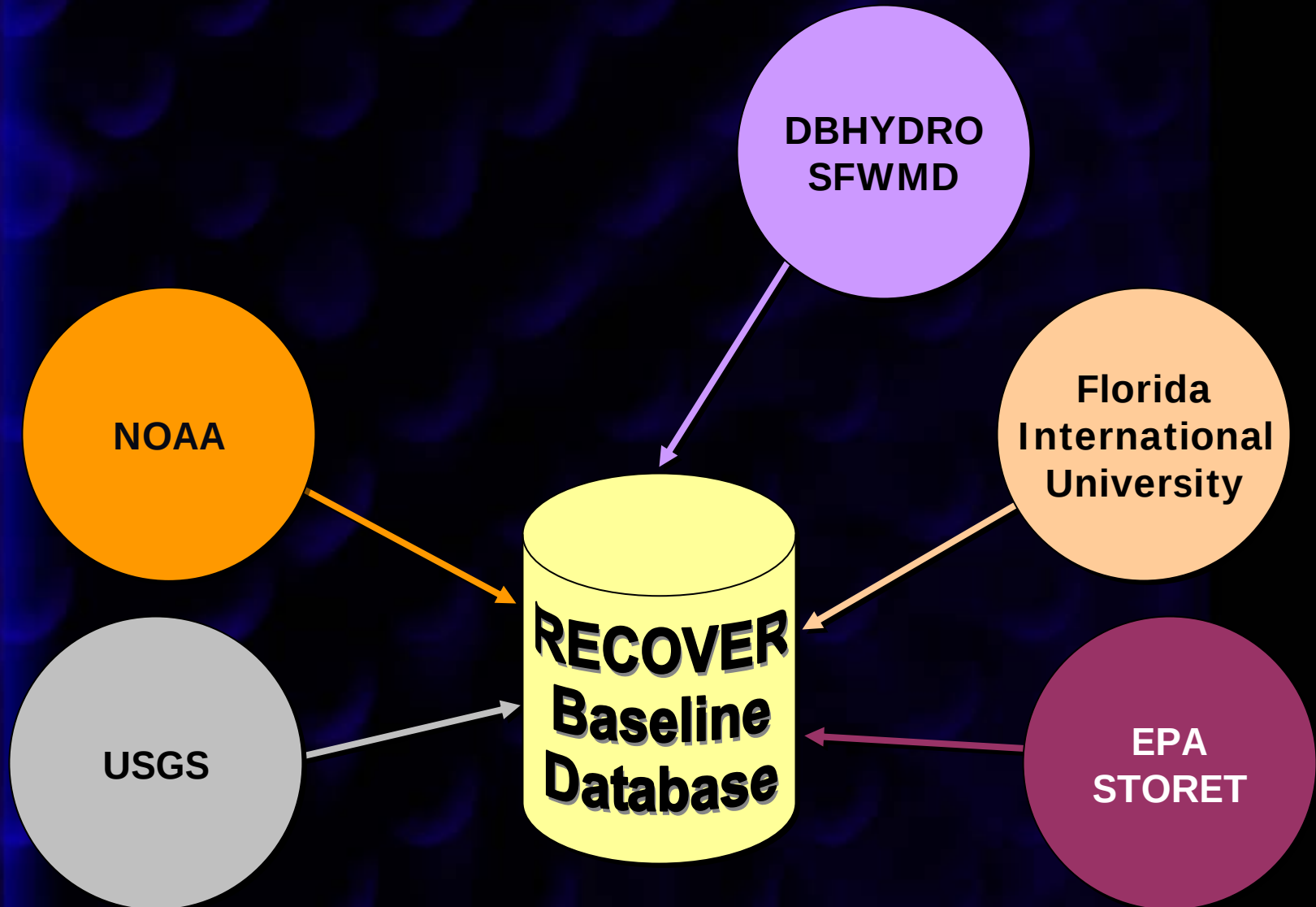
Contributing Agencies

79,819,210 records – 17,557,991 unusable records = 62,261,219 records
(Both Hydrologic and Water Quality)

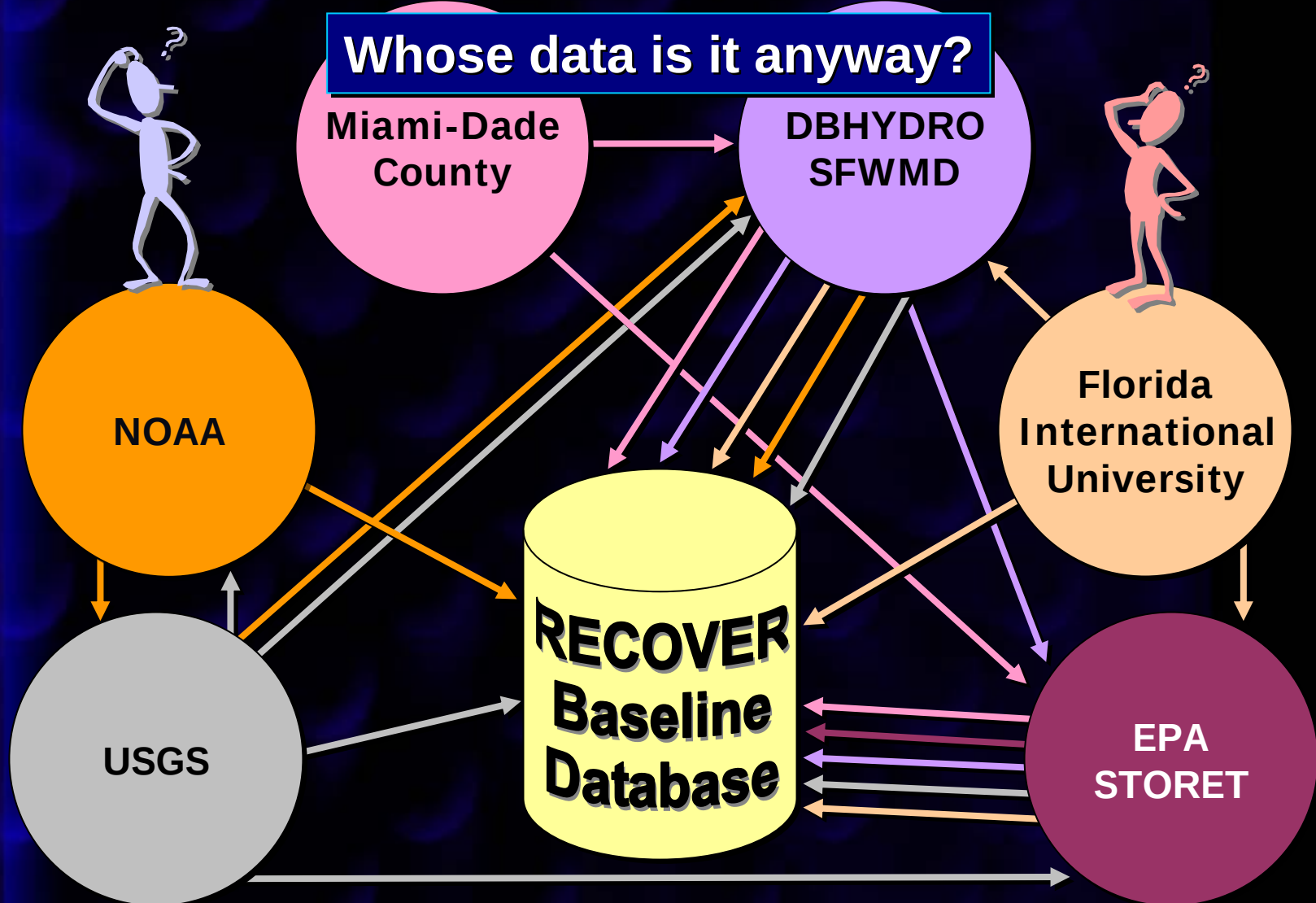


Water Quality Records = 35.4 million

Data Sources for RECOVER Database



Data Sources for RECOVER Database



Database Design Re-Think or “Digital Epiphany”

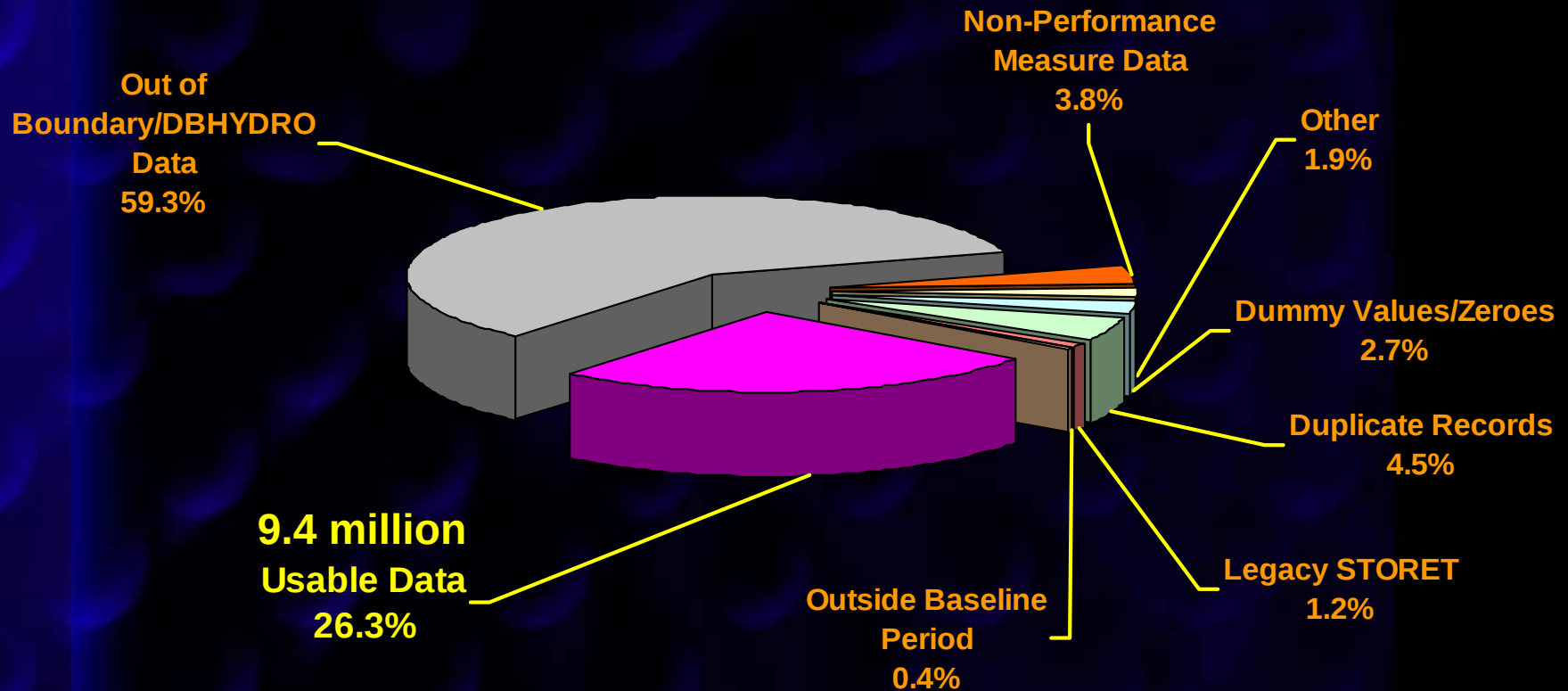
- **Separate baseline database for water quality and hydrologic data was scrapped**
- **Data in baseline database that did not reside in DBHYDRO (or “found data”) would be loaded into DBHYDRO**
 - **More than 50 % of water quality and hydrologic data for South Florida was already in DBHYDRO**
 - **DBHYDRO is a dynamic database**
- **Data records to be loaded into DBHYDRO would have to adhere to DBHYDRO standards**

Initial RECOVER Baseline Database Water Quality Data Scrub

- Remove records for stations outside District boundaries using GIS
- Remove records with dummy variables or zeroes
- Remove duplicate records (initial screening)
- Remove records outside the baseline period and non-performance measure data
- Remove records loaded from DBHYDRO and Legacy STORET

Initial RECOVER Baseline Database Water Quality Data Scrub

Water Quality Records = 35.4 million

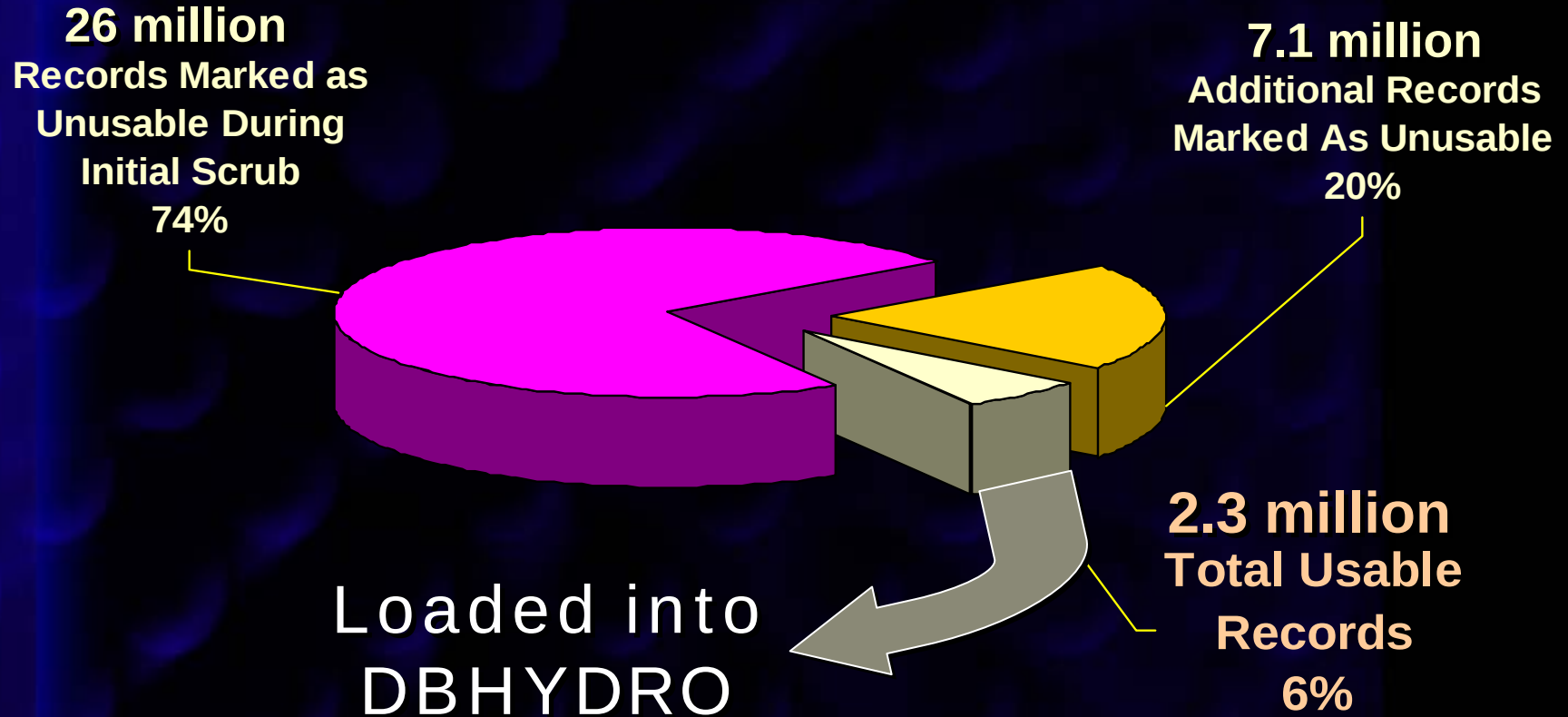


Final RECOVER Baseline Database Water Quality Data Scrub

Affecting 9.4 Million Usable Records

- **Remove records for stations outside eco-regions using GIS**
- **Remove duplicate records (final screening)**
- **Consolidate 15-minute instantaneous data for temperature, pH, specific conductivity and dissolved oxygen into daily averages (reduced records from 1.9 million to 9,600 records)**

Final RECOVER Baseline Database Water Quality Data Scrub



Original Number of Water Quality Records = 35.4 million

Database Format Issues

- Date and Time stored in the database:

- YYYYMMDD, MM/DD/YYYY, MM/DD/YYYY
HH:MM:SS

- Latitude and Longitude:

- 999.999°, DEGMINSEC, etc.

- Values were stored as numeric or text

- Reference sheet from contributing agencies was required to understand field names in the database

Parameter Naming Convention

Each agency has its own naming convention, parameter names need to be standardized

PARAM_NAME	NEW_PARAM_NAME
NH3+ORGT	Total Dissolved Nitrogen (TDN) as N
Nitrogen, mixed forms (NH3)+(NH4)+organic+(NO2)+(NO3)	Total Dissolved Nitrogen (TDN) as N
TN	Total Nitrogen (TN) as N
TNSWF	Total Nitrogen (TN) as N
TNW	Total Nitrogen (TN) as N
TNWDUP	Total Nitrogen (TN) as N
TNWW	Total Nitrogen (TN) as N
Chl-A	chlorophyll a, spectrophotometer, corrected for pheophytin
Chla	chlorophyll a, spectrophotometer, corrected for pheophytin
Chlorophyll a (Corrected)	chlorophyll a, spectrophotometer, corrected for pheophytin
chla	chlorophyll a, spectrophotometer, corrected for pheophytin
TOTAL PHOSPHORUS	Phosphorus as P
Total Phosphorus in Porewater	Phosphorus as P
PHOSPHORUS, HYDROLYZABLE + ORTHO, TOT, AUTOANALYZER	Phosphorus as P
TPO4(mg/L)	Phosphorus as P

Units

Most agencies use uniform parameter units; however,....

For example:

Nitrate = 0.021 mg/L

Is it

0.021 mg/L NO₃ as N

Or

0.021 mg/L as NO₃?

This can result in a 343 %
difference in concentrations



PARAM_NAME	NEW_PARAM_NAME
"Nitrogen, Kjeldahl"	Kjeldahl Nitrogen as N
	Total Kjeldahl Nitrogen (TKN) as N
Nitrogen, Kjeldahl	Kjeldahl Nitrogen as N
	Total Kjeldahl Nitrogen (TKN) as N
BAP	Benzo[a]pyrene
	Biologically available phosphorus
CHLA	Chlorophyll a
	chlorophyll a, fluorometer
N+N	Nitrite (NO ₂) + nitrate (NO ₃)
	Nitrite (NO ₂) + nitrate (NO ₃) as N
NH ₃ -N	Ammonia (NH ₃) + ammonium (NH ₄) as N
	Ammonia (NH ₃) as N
NH ₄	Ammonium (NH ₄)
	Ammonium (NH ₄) as N
NO ₃	Nitrate (NO ₃) as N
	Nitrate (NO ₃) as NO ₃
NO ₃ -	Nitrate (NO ₃)
	Nitrate (NO ₃) as N
O-PO ₄	Orthophosphate as P
	Orthophosphate as PO ₄

Units

Other potential unit problems:

- Conductivity as: mS/cm or mmhos/cm rather than μ S/cm or μ mhos/cm
- Temperature as: °F rather than °C
- Depth in: feet rather than meters
- Concentrations in: mmolar or μ molar rather than mg/L or μ g/L
- Concentrations in: mmole/kg or μ mole/kg rather than mg/L or μ g/L

Data Loading Problems

■ Text files or space delimited:

- Columns may not line up correctly

ChemicalTbl_datamodified.dat											
Reference	CollectionDate	Test	Component	LabID	SampleLocation	StationName	Analysis	Date	Result	Units	Code
STR200023@08/30/00B	8/30/2000	S-BNA	Endosulfan	I	478930		STR200023	9/25/2000	150	ug/kg	U
STR200023@08/30/00B	8/30/2000	S-BNA	Endosulfan	II	478930		STR200023	9/25/2000	150	ug/kg	U
STR200023@08/30/00B	8/30/2000	S-BNA	Endrin	478930		STR200023	9/25/2000		150	ug/kg	U

■ Comma delimited or CSV files:

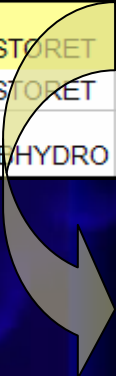
- Columns may not line up if a comma exists in any of the records

Data Loading Problems

EPA STORET

Mystery dissolved arsenic data in database (with dissolved value greater than total)

STUDY_ID	SAMPLE_DATE	STATION_ID	PARAM_NAME	NEW_PARAM_NAME	RESULT_VALUE	RESULT_UNIT
STORET	25-Jul-90	S5A_SFWMD	ARSENIC, INORGANIC DISS (UG/L AS AS)	Arsenic (As)	22.4	UG/L
STORET	25-Jul-90	S5A_SFWMD	ARSENIC, TOTAL (UG/L AS AS)	Arsenic (As)	5.17	UG/L
DBHYDRO	25-Jul-90	S5A	TOT. ARSENIC	Arsenic (As)	5.17	UG/L



This parameter appears in DBHYDRO as silica

Other Data Problems

- Some agencies have their own criteria for flagging data: ex. USGS uses “<” to flag data that is below the method detection limit (MDL) ; other agencies use “U”
- Some agencies use a negative number to denote the value is below the MDL
- Techniques used to consolidate data like global “search and replace” procedures can cause problems
 - K PAR → Potassium K
 - Total Dissolved Nitrogen → Fraction = Total

Lessons Learned

RECOVER Baseline Database

- **Team – important part of project: good balance of disciplines but weighted toward water quality expertise**
- **Scientific data incorrectly interpreted by database programmers**
- **Staging area very important when dealing with many data sources**
- **Some data were modified/calculated before being loaded into database, no original copy available for re-loading data**

Lessons Learned

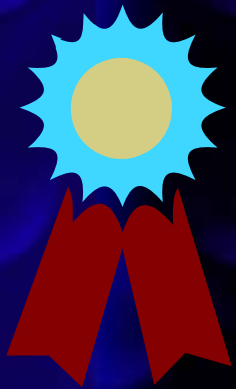
RECOVER Baseline Database

- Data were flattened/denormalized (no referential integrity)
- Lack of indices and/or no unique identifiers
- GIS Important tool: identification of incorrect coordinates or duplication of stations
- Station name problems – different names same location, different location same name
- Embedding one field's information in another

Lessons Learned

RECOVER Baseline Database

- **Elevation problems: NGVD29 vs NAVD88**
- **Technical data conversion problems: importing MS Access tables to Oracle caused numeric fields to drop values past the hundredths column**
- **All data are not stored or maintained equally (e.g., Legacy STORET)**
- **The majority of environmental (water quality and hydrologic) data for South Florida reside in the District's DBHYDRO database**



Acknowledgements

- Lisa Herlihy Sutron Corporation
- Julio Parades Laboratory Data Consultants, Inc.
- Steven Ziliak Laboratory Data Consultants, Inc.
- Andrew Bessio Sapphire Technologies



Final Thought QUESTIONS?



SOUTH FLORIDA WATER MANAGEMENT DISTRICT



Nenad Iricanin

Environmental Resource Assessment
(561) 682-2956
nirican@sfwmd.gov

no small mistakes!