

Mind Your Metadata

Exploiting Semantics for Configuration, Adaptation, and Provenance in Scientific Workflows

Yolanda Gil

Pedro Szekely

Craig Knoblock

Varun Ratnakar

Shubham Gupta

Maria Muslea

Fabio Silva

USC/ISI

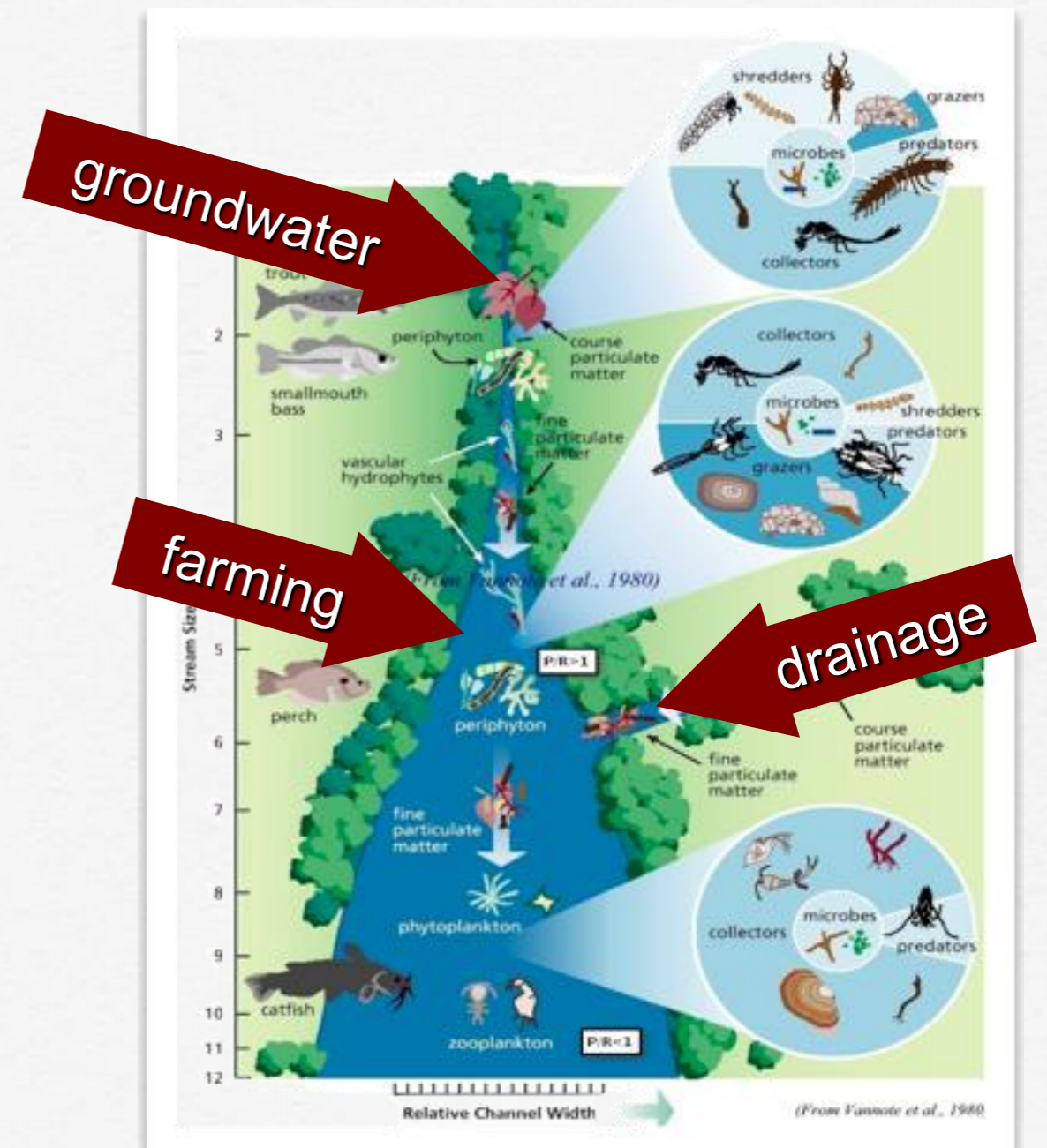
Tom Harmon

Sandra Villamizar

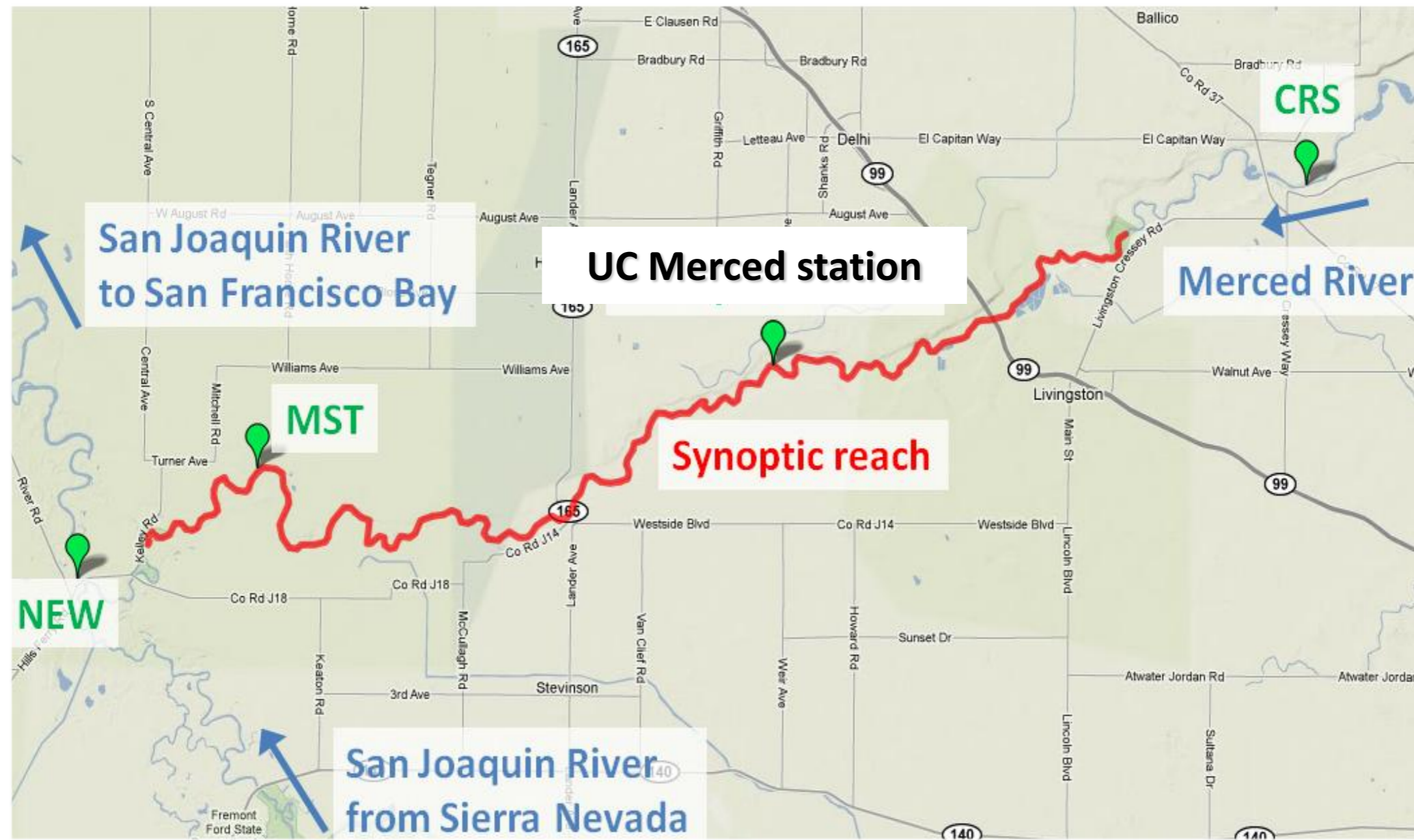
UC Merced

River Continuum vs Human Activities

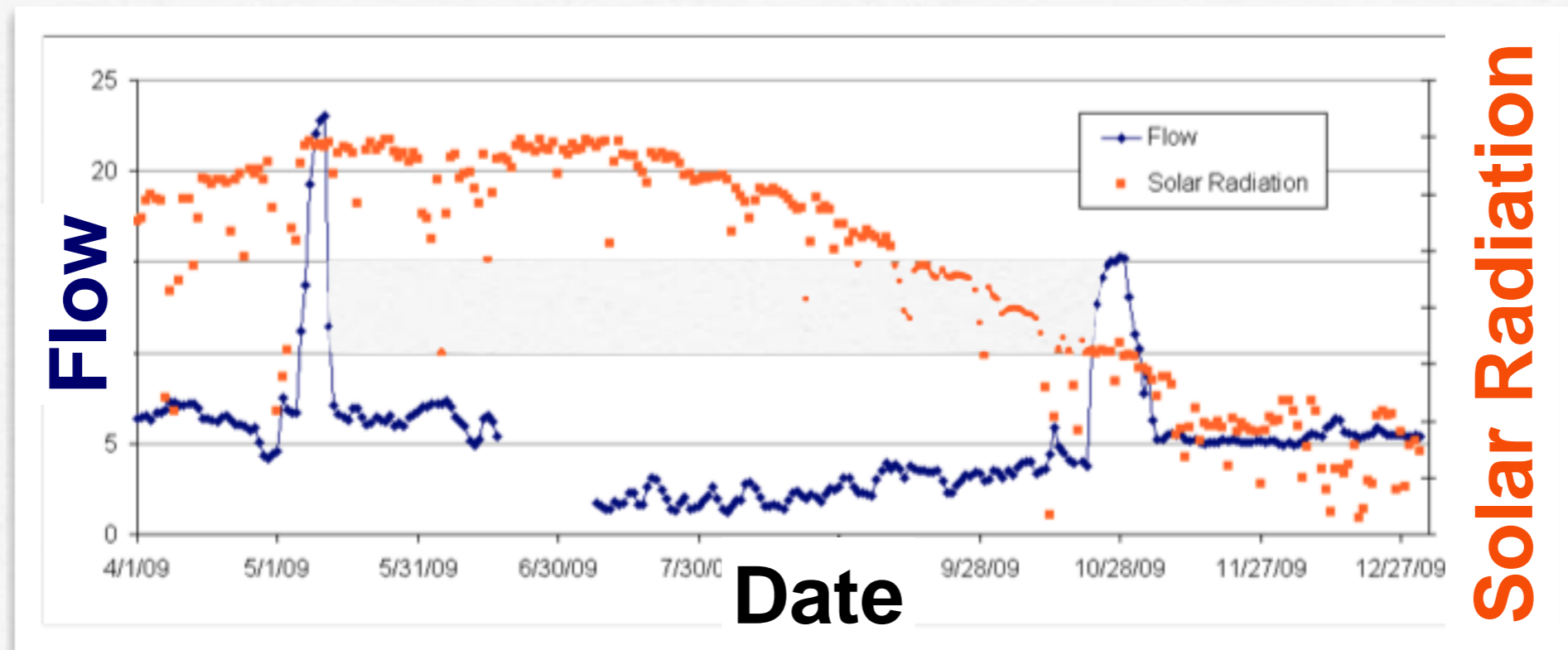
- River continuum: natural inputs, reactive transport
- Human intervention: Agricultural, industrial, municipal
- What management practices help/hurt?
- Can we restore natural behavior?



Case Study



Stream Metabolism Response to Human Disturbances

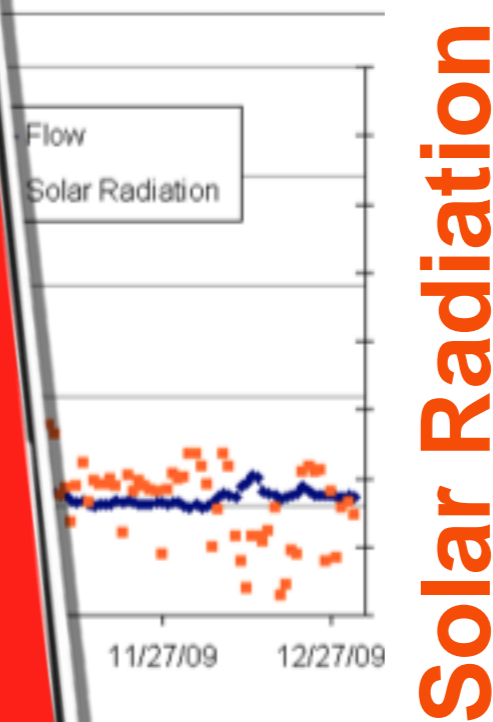


Pulse releases in the spring and fall to help the salmon runs

Stream Metabolism Response to Human Disturbances

... but how does this affect the ecology of the river?

... how about the effect of farmers?

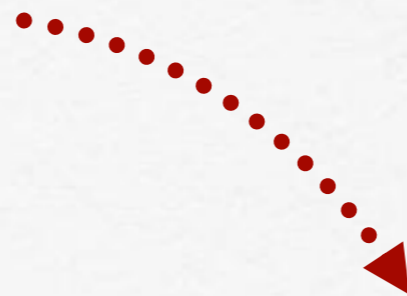
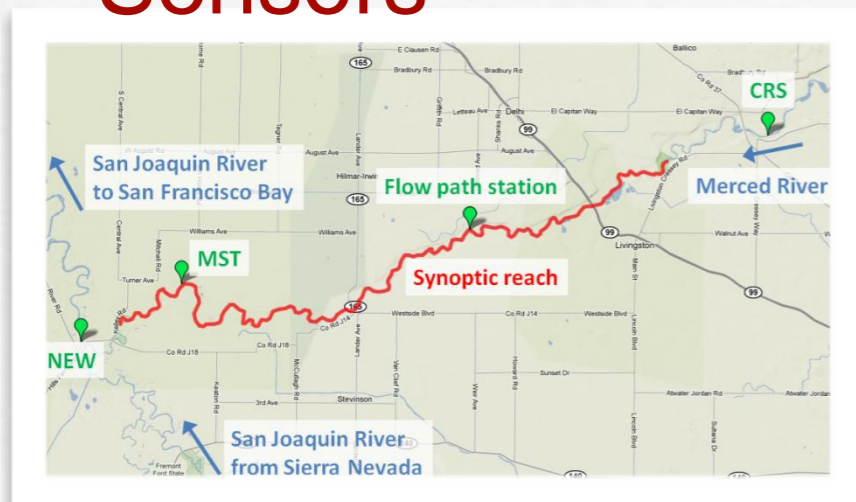


Pulse releases in the spring and fall to help the salmon runs

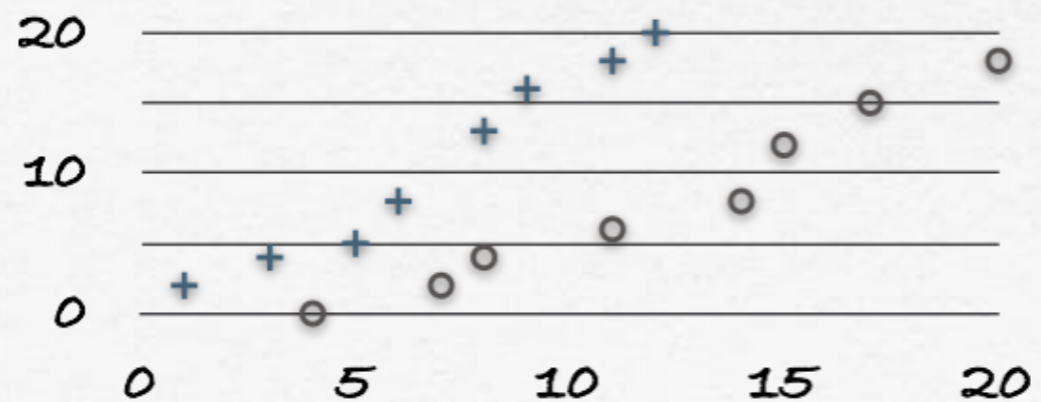
Aquatic Photosynthesis

Models of gross primary production (GPP),
community respiration (CR24)

Sensors



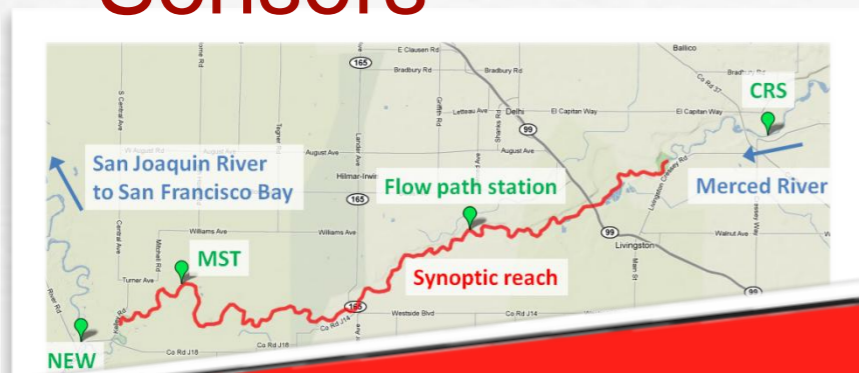
Analysis



Aquatic Photosynthesis

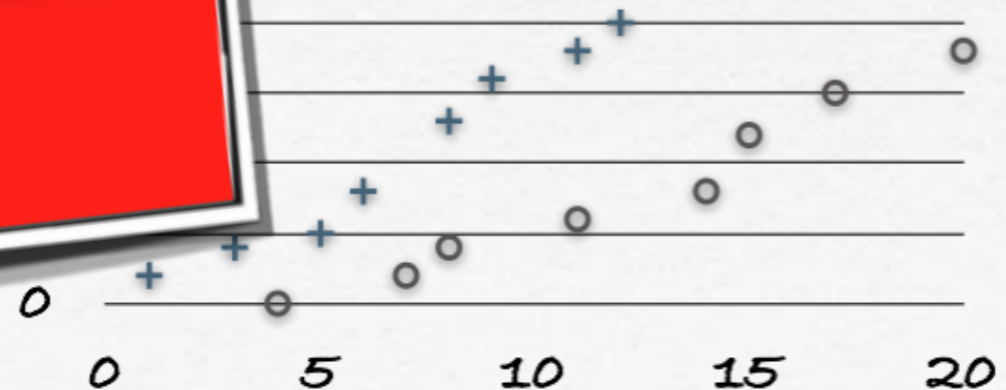
Models of gross primary production (GPP),
community respiration (CR24)

Sensors

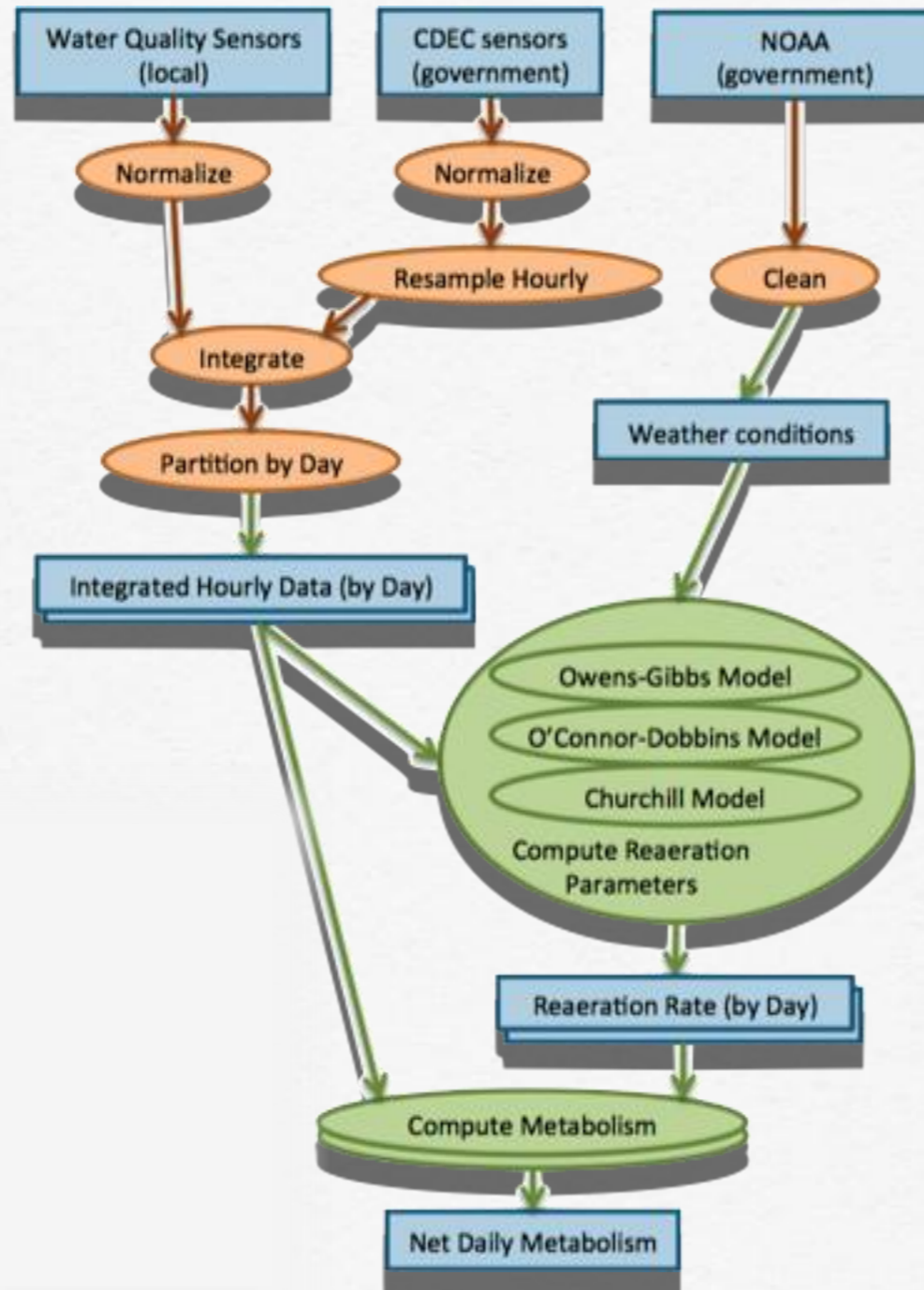


Analyses must be fast to produce
actionable information

Analysis



Workflow



Tom Harmon
environmental systems

Vision: Automated & Fast

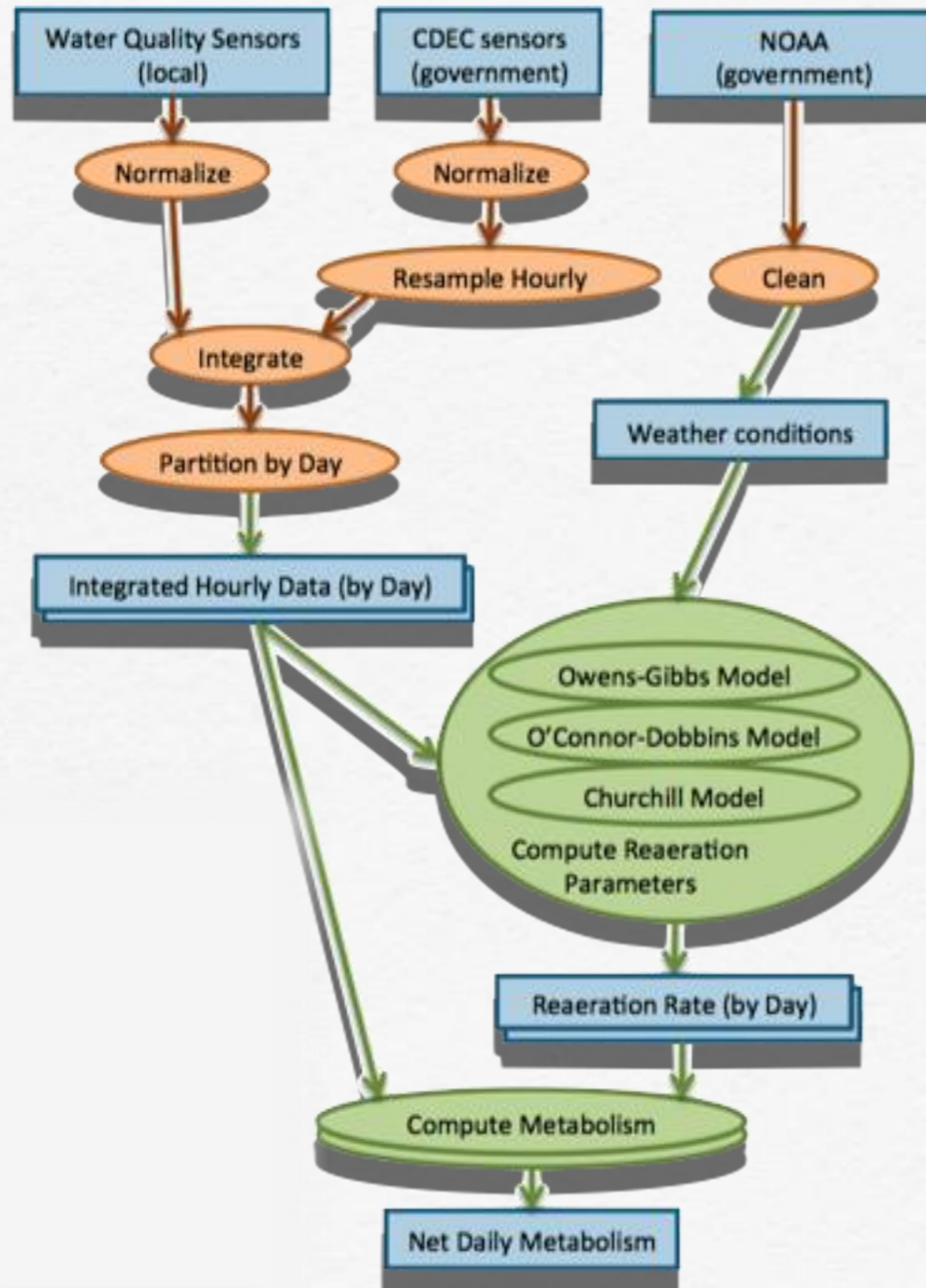


Reality: Difficult & Time Consuming



A large, bold, black question mark is centered on a white page from a spiral-bound notebook. The spiral binding is visible at the top of the page. The question mark is the sole focus of the image.

Current Method

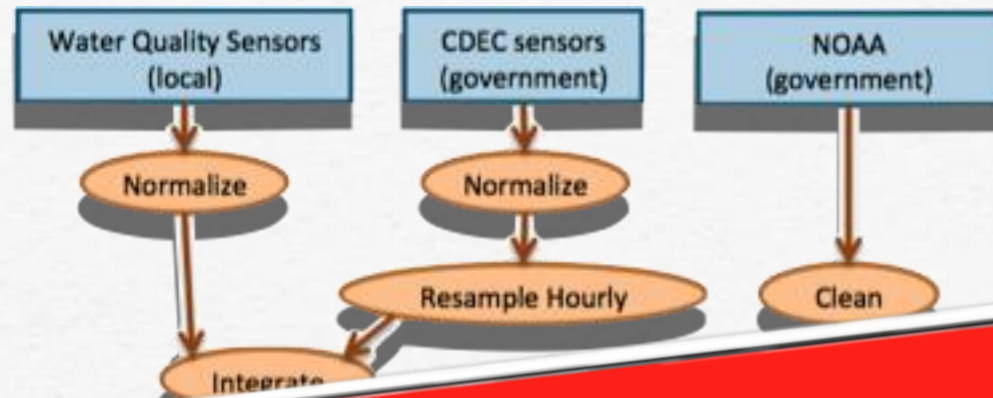


Manual
Data
Preparation



Custom
Scripts

Current Method



Manual

n

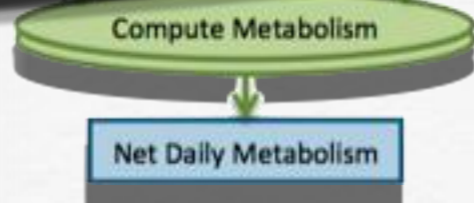
Multiple, separate tools

High learning costs

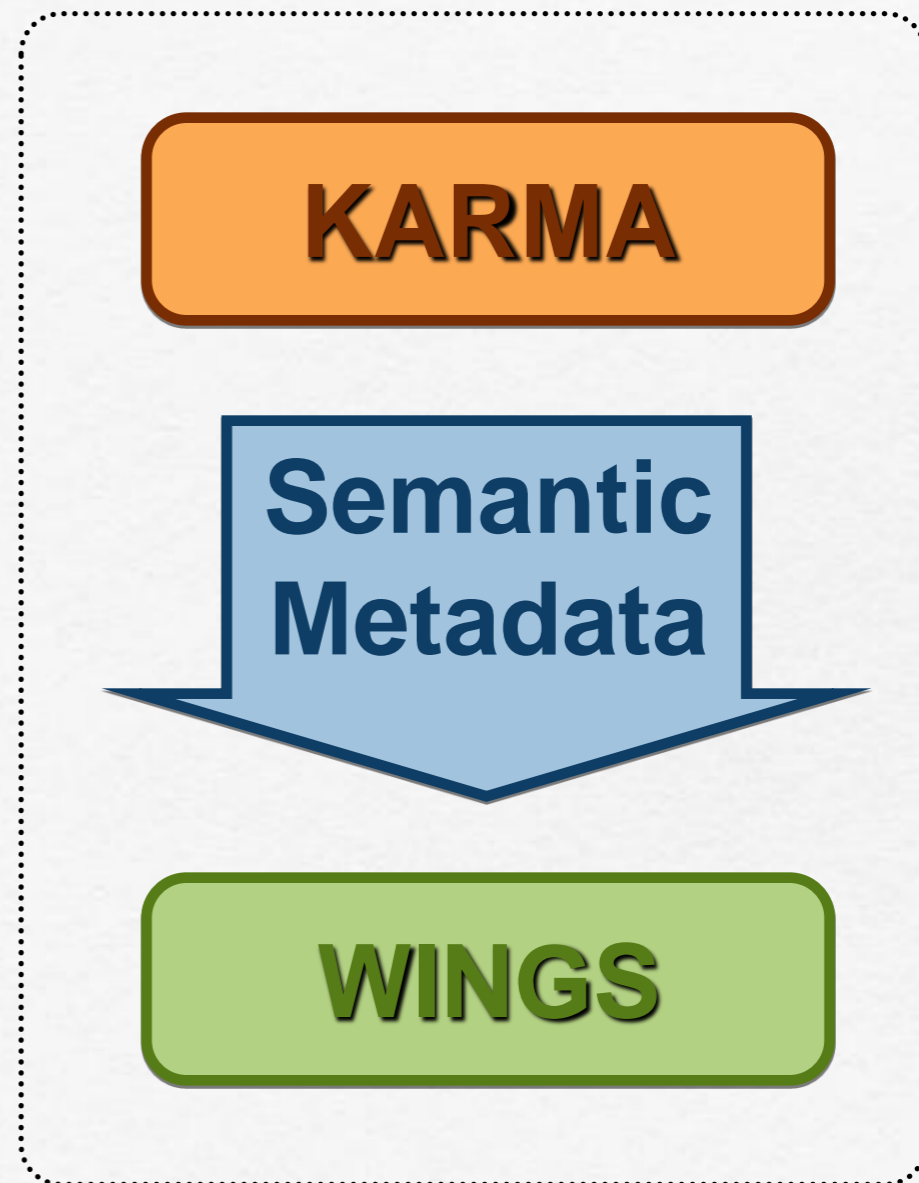
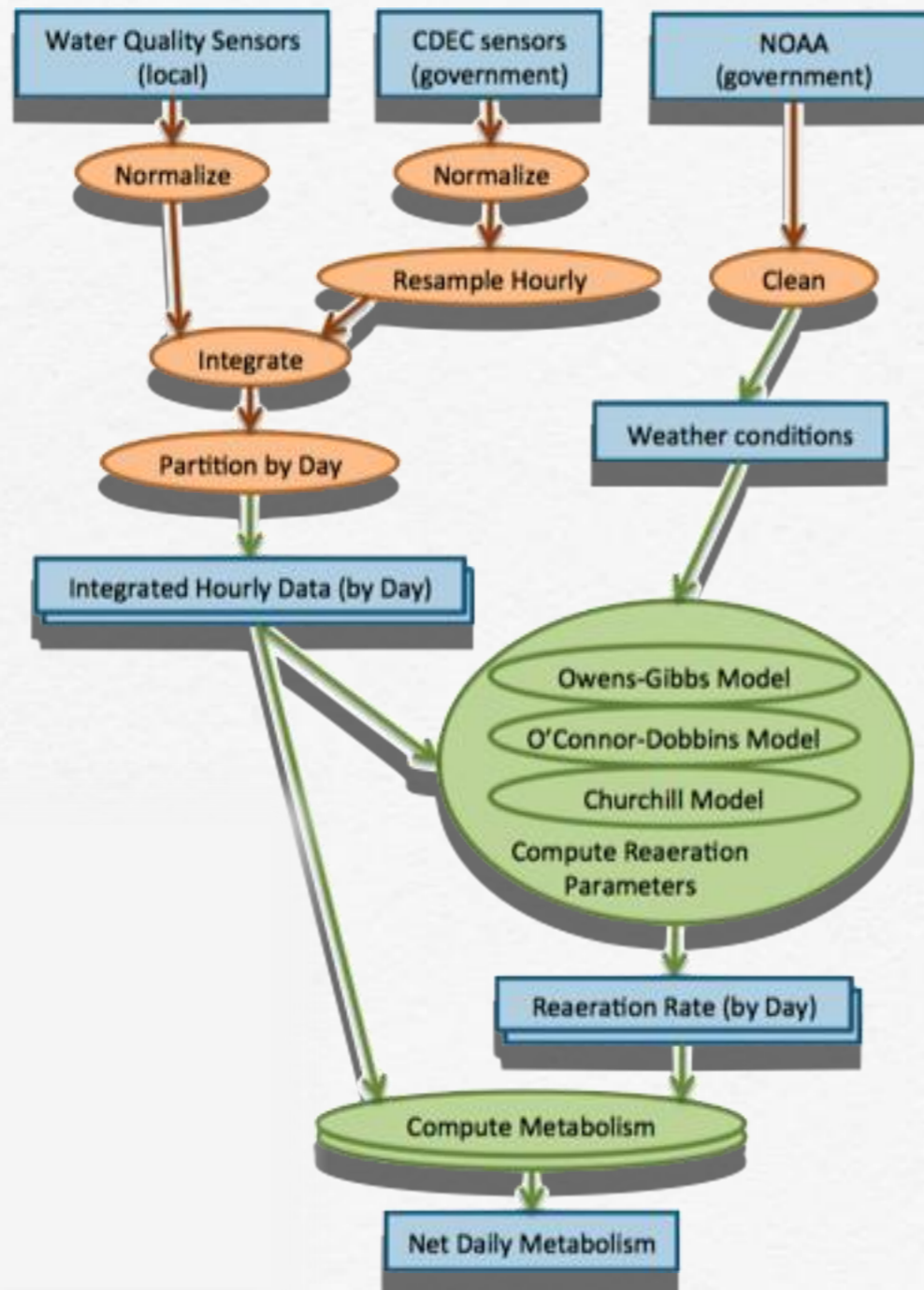
Ad hoc, by-hand movement of data & tool invocation

Data does not "flow" across tools

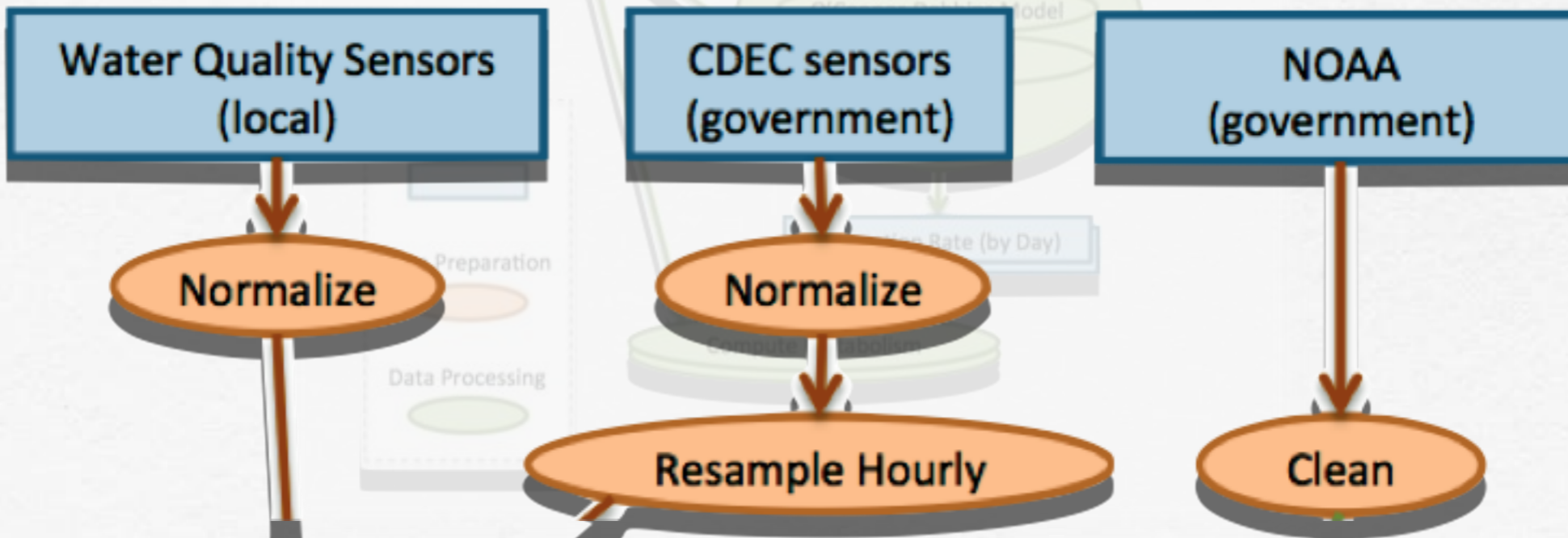
Scripts



Our Approach



Data Sources



[Tuchinda et al TWEB'11; Tuchinda et al IUI'08, IUI'07]

KARMA

The screenshot displays the KARMA v0.4 application window. At the top, there is a menu bar with options: Table, Script, Alignment, and Column. Below the menu bar, there are two tabs labeled 'Source1' and 'Source2'. A table is visible with four columns, each with a header 'Data Type' and 'Column Name'. Below the table, there are three main panels: 'Web Services', 'Inputs', and 'Outputs'. The 'Web Services' panel contains a list of service names, with 'CDEC - Event Data' selected. The 'Inputs' panel has two dropdown menus: 'Station ID' (set to 'SMN') and 'Sensor' (set to 'Choose Value'). The 'Outputs' panel lists '-Date', '-Time', and '-Value'. An 'Execute' button is located at the bottom left of the interface.

Web Services

WebService Name
Buildings2StreetNames
CDEC Simple
CDEC - Event Data
CDEC - FLOW, Daily Mean

Inputs

Station ID: SMN

Sensor: Choose Value

- 1 - River Stage(feet)
- 14 - Battery Voltage(volts)
- 20 - Flow(cfs)
- 146 - Temperature, Water(C)
- 100 - Electrical Cond.(us/cm)
- 61 - Dissolved Oxygen(mg/l)

Outputs

- Date
- Time
- Value

Execute

Data Import

The screenshot displays the Karma_v0.4 application window. At the top, there is a menu bar with 'Table', 'Script', 'Alignment', and 'Column'. Below the menu bar, a red dashed box highlights a table with the following data:

String	String	String	String	String
Station ID	Start Date	Date	Time	RIVER STAGE (feet)
SMN	03/10/2010	20100309	2300	52.68
SMN	03/10/2010	20100309	2315	52.68
SMN	03/10/2010	20100309	2330	52.68
SMN	03/10/2010	20100309	2345	52.66
SMN	03/10/2010	20100310	0000	52.69
SMN	03/10/2010	20100310	0015	52.67
SMN	03/10/2010	20100310	0030	52.66
SMN	03/10/2010	20100310	0045	52.66
SMN	03/10/2010	20100310	0100	52.67
SMN	03/10/2010	20100310	0115	52.64
SMN	03/10/2010	20100310	0130	52.65

Below the table, there are buttons for 'Import', 'Clean', 'Integrate', and 'Publish'. A red dashed box also highlights a configuration panel with the following sections:

- Wrapper:** Database, Excel, CSV, KML, **WebService**
- Web Services:** WebService Name list including 'CDEC - Event Data' (selected).
- Inputs:** Station ID (SMN), Sensor (1 - River Stage(feet)), Start Date.
- Outputs:** -Date, -Time, -Value.
- Execute** button.

Need to Clean Data

Date	Time	RIVER STAGE (feet)
20100309	2300	52.68
20100309	2315	52.68

CDEC

timestamp	WXT510P
2010-03-10 00:00:00	760
2010-03-10 00:15:00	760

HYDROLAB

Date	Time	Temp	Cond
03/09/2010	23:00	13.4	1181.00
03/09/2010	23:15	13.4	1179.00

Required
Format

Need to Clean Data



Date	Time	Depth (feet)
20100309	2300	
20100309		

CDEC

60 Files for
1 month!

HYDROLAB



Date	Time	Temp	Cond
03/09/2010	23:00	13.4	1181.00
03/09/2010	23:15	13.4	1179.00

Required
Format

Data Cleaning with KARMA

The screenshot displays the Karma_v0.4 application window. At the top, there are tabs for 'Table', 'Script', 'Alignment', and 'Column'. Below these, there are sub-tabs for 'CDEC - Event Data0', 'Source2', 'Source3', and 'Cleaning Table'. The main area contains a table with three columns: 'String', 'Data Type', and 'Data Type'. The first column contains dates, the second contains 'User Defined Values', and the third contains 'Final Values'. A red dashed circle highlights the 'Clean' button, the 'Data Cleaning for:' dropdown menu (set to 'Column 2'), and the radio button options for 'Use original extracted values' and 'Use user defined values'.

String	Data Type	Data Type
Date	User Defined Values	Final Values
20100309	03/09/2010	
20100309		
20100309		
20100309		
20100310		
20100310		
20100310		
20100310		
20100310		
20100310		
20100310		
20100310		
20100310		

Import Clean Integrate Publish

Data Cleaning for: Column 2

Final result:

- Use original extracted values
- Use user defined values

Data Cleaning with KARMA

Karma_v0.4

EDBC - Event Data	Source2	Source3	Cleaning Table
String			Data Type
Date			User Defined Values
20100309			03/09/2010
20100309			
20100309			
20100309			

20100310

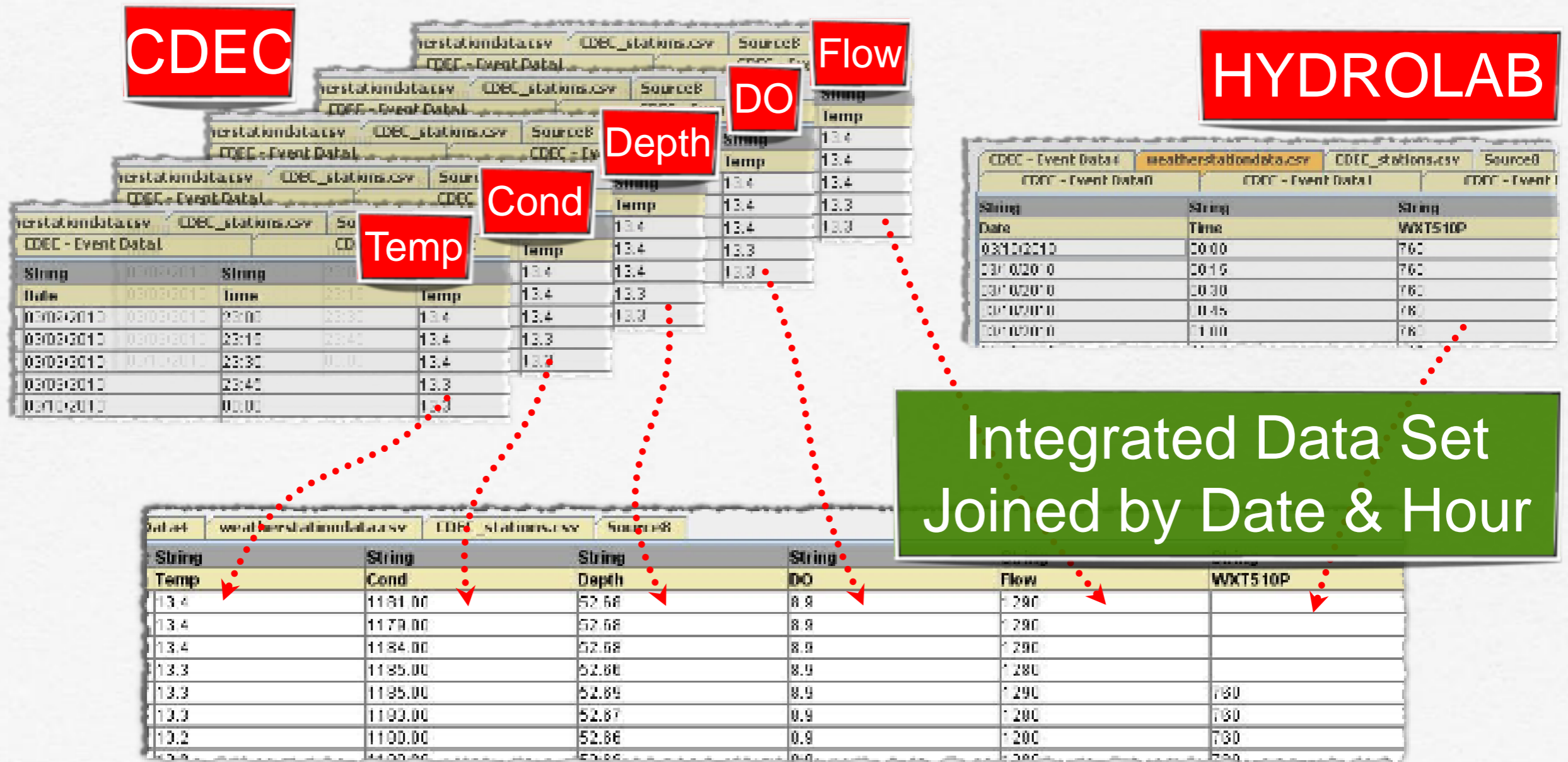
EDBC - Event Data	Source2	Source3	Cleaning Table
String			Data Type
Date			User Defined Values
20100309			03/09/2010
20100309			03/09/2010
20100309			03/09/2010
20100309			03/09/2010

Use user defined values

User provides example

KARMA generates cleaning rule

Need to Integrate All the Sources



Integrated Dataset


Karma_v0.4

Table Script Alignment Column

CDEC - Event Data0 CDEC - Event Data1 CDEC - Event Data2 CDEC - Event Data3 CDEC - Event Data4 weatherstationdata.csv CDEC_stations.csv

forSite	forDate	String	String	String	String	String
Station ID	Start Date	Date	Time	Temp	Cond	D
SMN	03/10/2010	03/09/2010	23:00	13.4	1181.00	5
SMN	03/10/2010	03/09/2010	23:15	13.4	1179.00	5
SMN	03/10/2010	03/09/2010	23:30	13.4	1184.00	5
SMN	03/10/2010	03/09/2010	23:45	13.3	1185.00	5
SMN	03/10/2010	03/10/2010	00:00	13.3	1185.00	5
SMN	03/10/2010	03/10/2010	00:15	13.2	1182.00	5

Import Clean Integrate Publish



KARMA Generates Data Processing Script

```
ImportWSSource("CDEC - Event Data", "SMN", "146", "$1", "$2"); SetColumnName("CDEC - Event Data0", "4", "Date"); ApplyCleanRule("CDEC - Event Data0", "Date", "20100309", "03/09/2010"); DeleteColumnCommand("Sensor"); DeleteColumnCommand("End Date"); SetColumnName("CDEC - Event Data0", "3", "Time"); SetColumnName("CDEC - Event Data0", "4", "Temp"); ApplyCleanRule("CDEC - Event Data0", "Time", "2300", "23:00"); SwitchToEmptySourceTab(1); ImportWSSource("CDEC - Event Data", "SMN", "100", "$1", "$2"); SetColumnName("CDEC - Event Data1", "4", "Date"); ApplyCleanRule("CDEC - Event Data1", "Date", "20100309", "03/09/2010"); DeleteColumnCommand("Sensor"); DeleteColumnCommand("Start Date"); DeleteColumnCommand("End Date"); SetColumnName("CDEC - Event Data1", "2", "Time"); SetColumnName("CDEC - Event Data1", "3", "Cond"); ApplyCleanRule("CDEC - Event Data1", "Time", "2300", "23:00"); SwitchToEmptySourceTab(2); ImportWSSource("CDEC - Event Data", "SMN", "1", "$1", "$2"); SetColumnName("CDEC - Event Data2", "4", "Date"); ApplyCleanRule("CDEC - Event Data2", "Date", "20100309", "03/09/2010"); DeleteColumnCommand("Sensor"); DeleteColumnCommand("Start Date"); DeleteColumnCommand("End Date"); SetColumnName("CDEC - Event Data2", "2", "Time"); SetColumnName("CDEC - Event Data2", "3", "Depth"); ApplyCleanRule("CDEC - Event Data2", "Time", "2300", "23:00"); SwitchToEmptySourceTab(3); ImportWSSource("CDEC - Event Data", "SMN", "61", "$1", "$2"); SetColumnName("CDEC - Event Data3", "4", "Date"); ApplyCleanRule("CDEC - Event Data3", "Date", "20100309", "03/09/2010"); DeleteColumnCommand("Sensor"); DeleteColumnCommand("Start Date"); DeleteColumnCommand("End Date"); SetColumnName("CDEC - Event Data3", "2", "Time"); SetColumnName("CDEC - Event Data3", "3", "DO"); ApplyCleanRule("CDEC - Event Data3", "Time", "2300", "23:00"); SwitchToEmptySourceTab(4); ImportWSSource("CDEC - Event Data", "SMN", "20", "$1", "$2"); SetColumnName("CDEC - Event Data4", "4", "Date"); ApplyCleanRule("CDEC - Event Data4", "Date", "20100309", "03/09/2010"); DeleteColumnCommand("Sensor"); DeleteColumnCommand("Start Date"); DeleteColumnCommand("End Date"); SetColumnName("CDEC - Event Data4", "2", "Time"); SetColumnName("CDEC - Event Data4", "3", "Flow"); ApplyCleanRule("CDEC - Event Data4", "Time", "2300", "23:00"); SwitchToSourceTab(0); join("CDEC - Event Data0", "CDEC - Event Data1", "Cond"); join("CDEC - Event Data0", "CDEC - Event Data2", "Depth"); join("CDEC - Event Data0", "CDEC - Event Data3", "DO"); join("CDEC - Event Data0", "CDEC - Event Data4", "Flow"); SwitchToEmptySourceTab(5); ImportCSVSource("..\data\CDEC_stations.csv"); ImportColumnFromCSV("Station ID", "0", "true"); ImportColumnFromCSV("Metadata", "1", "true"); ImportColumnFromCSV("Name", "2", "true"); ImportColumnFromCSV("Elevation", "3", "true"); ImportColumnFromCSV("Latitude", "4", "true"); ImportColumnFromCSV("Longitude", "5", "true"); SwitchToSourceTab(0); join("CDEC - Event Data0", "CDEC_stations.csv", "Latitude"); join("CDEC - Event Data0", "CDEC_stations.csv", "Longitude"); PublishToWS("WINGS Portal", "TEST_CDEC_WEATHER_$3", "CDEC - Event Data0");
```

KARMA Generates Data Processing Script

```
ImportWSSource("CDEC - Event Data", "SMN", "146", "$1", "$2");SetColumnName("CDEC  
- Event Data0", "4", "Date");ApplyCleanRule("CDEC - Event
```

```
ImportWSSource("CDEC - Event Data", "SMN", "146", "$1", "$2");  
SetColumnName("CDEC - Event Data0", "4", "Date");  
ApplyCleanRule("CDEC - Event Data0", "Date", "20100309", "03/09/2010");  
DeleteColumnCommand("Sensor");  
DeleteColumnCommand("End Date");  
SetColumnName("CDEC - Event Data0", "3", "Time");  
SetColumnName("CDEC - Event Data0", "4", "Temp");  
ApplyCleanRule("CDEC - Event Data0", "Time", "2300", "23:00");  
SwitchToEmptySourceTab(1);  
ImportWSSource("CDEC - Event Data", "SMN", "100", "$1", "$2");  
SetColumnName("CDEC - Event Data1", "4", "Date");
```

```
Data0", "CDEC_stations.csv", "Longitude");PublishToWS("WINGS  
Portal", "TEST_CDEC_WEATHER_03", "CDEC - Event Data0");
```


Publishing Processed Data to WINGS

SMN	03/10/2010	03/10/2010	00:30	13.2
SMN	03/10/2010	03/10/2010	00:45	13.2
SMN	03/10/2010	03/10/2010	01:00	13.2
SMN	03/10/2010	03/10/2010	01:15	13.2
SMN	03/10/2010	03/10/2010	01:30	13.1

Import Clean Integrate **Publish**

HTML KML XML CSV Text File Database RDF **WebService**

Web Services

WebService Name
WINGS Portal

Inputs

File Name
WEATHER_2010_03_10

File Content
CDEC - Event Data0

Semantic Metadata for Input Files

```
<?xml version="1.0" encoding="UTF-8" ?>
```

```
<rdf:RDF
```

```
  xml:base="http://www.isi.edu/dc/Water/library.owl#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:dc="http://www.isi.edu/dc/ontology.owl#"
  xmlns:dcdom="http://www.isi.edu/dc/Water/ontology.owl#"
  xmlns="http://www.isi.edu/dc/Water/library.owl#">
```

```
<dcdom:Daily_Sensor_Data rdf:ID="FILENAME">
```

```
  <dcdom:forDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">DATE</dcdom:forDate>
  <dcdom:forSite rdf:datatype="http://www.w3.org/2001/XMLSchema#string">STATIONID</dcdom:forSite>
  <dcdom:siteLatitude rdf:datatype="http://www.w3.org/2001/XMLSchema#float">LATITUDE</dcdom:siteLatitude>
  <dcdom:siteLongitude rdf:datatype="http://www.w3.org/2001/XMLSchema#float">LONGITUDE</dcdom:siteLongitude>
  <dcdom:slope rdf:datatype="http://www.w3.org/2001/XMLSchema#float">SLOPE</dcdom:slope>
  <dcdom:velocity rdf:datatype="http://www.w3.org/2001/XMLSchema#float">VELOCITY</dcdom:velocity>
  <dcdom:depth rdf:datatype="http://www.w3.org/2001/XMLSchema#float">DEPTH</dcdom:depth>
  <dcdom:flow rdf:datatype="http://www.w3.org/2001/XMLSchema#float">FLOW</dcdom:flow>
  <dcdom:barpress rdf:datatype="http://www.w3.org/2001/XMLSchema#float">760</dcdom:barpress>
```

```
</dcdom:Daily_Sensor_Data>
```

```
</rdf:RDF>
```


Semantic Metadata for Input Files

```
<?xml version="1.0" encoding="UTF-8" ?>
```

```
<rdf:RDF
```

```
  xml:base="http://www.isi.edu/dc/Water/library"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:dc="http://www.isi.edu/dc/ontology.owl#"
  xmlns:dcdom="http://www.isi.edu/dc/Water/ontology.owl#"
  xmlns="http://www.isi.edu/dc/Water/library.owl#"
```

```
<dcdom:Daily_Sensor_Data rdf:ID="FILENAME">
```

```
<dcdom:forDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">DATE</dcdom:forDate>
```

```
<dcdom:forSite rdf:datatype="http://www.w3.org/2001/XMLSchema#string">SITE</dcdom:forSite>
```

```
<dcdom:siteLatitude rdf:datatype="http://www.w3.org/2001/XMLSchema#float">LATITUDE</dcdom:siteLatitude>
```

```
<dcdom:siteLongitude rdf:datatype="http://www.w3.org/2001/XMLSchema#float">LONGITUDE</dcdom:siteLongitude>
```

```
<dcdom:slope rdf:datatype="http://www.w3.org/2001/XMLSchema#float">SLOPE</dcdom:slope>
```

```
<dcdom:velocity rdf:datatype="http://www.w3.org/2001/XMLSchema#float">VELOCITY</dcdom:velocity>
```

```
<dcdom:depth rdf:datatype="http://www.w3.org/2001/XMLSchema#float">DEPTH</dcdom:depth>
```

```
<dcdom:flow rdf:datatype="http://www.w3.org/2001/XMLSchema#float">FLOW</dcdom:flow>
```

```
<dcdom:barpress rdf:datatype="http://www.w3.org/2001/XMLSchema#float">760</dcdom:barpress>
```

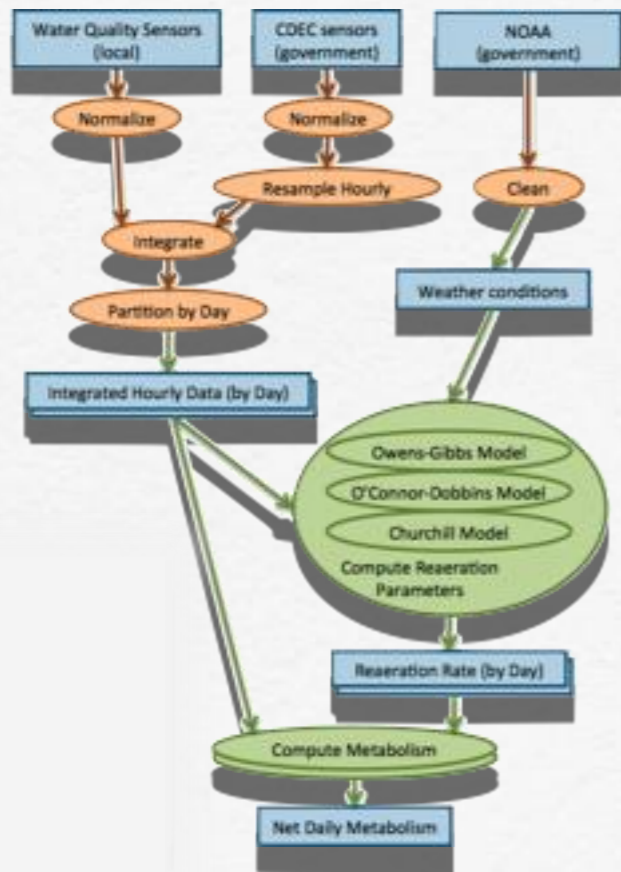
```
</dcdom:Daily_Sensor_Data>
```

```
</rdf:RDF>
```

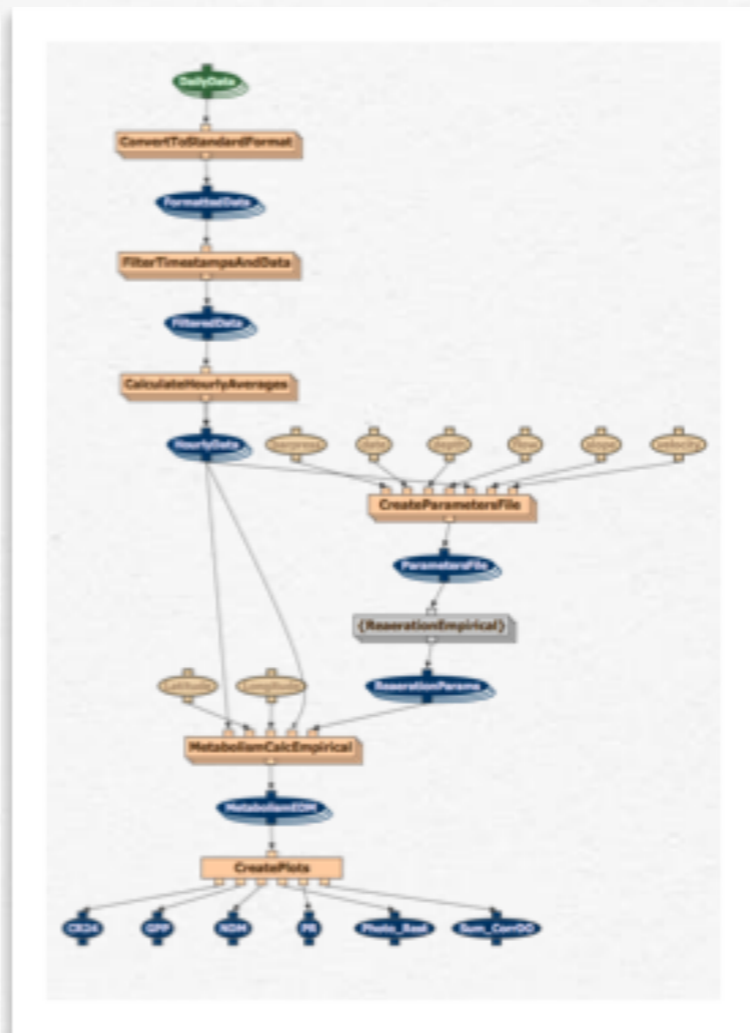
Automatically
Generated by
KARMA

[Gil et al JETAI'11; Gil et al IEEE-IS'11; Gil et al e-Science'09; Kim et al JCC'08]

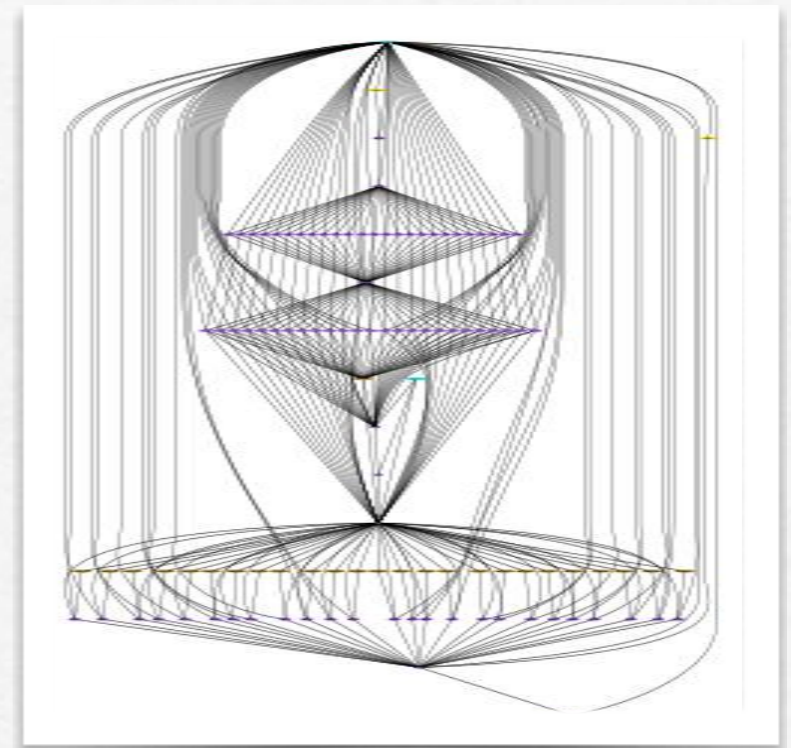
Workflows with WINGS



Conceptual workflow



WINGS Workflow



Workflow execution

WINGS Received Metadata from KARMA

The screenshot displays the WINGS DataBrowser interface. On the left, a file tree shows the following structure:

- DataObject
 - AlgoFile
 - PlotImage
 - Sensor_Data
 - Daily_Data
 - Daily_Parameters
 - NTM_Parameters
 - Daily_Sensor_Data
 - CDEC_WEATHER_2010_03_02
 - CDEC_WEATHER_2010_03_03
 - CDEC_WEATHER_2010_03_04
 - CDEC_WEATHER_2010_03_05
 - CDEC_WEATHER_2010_03_06
 - CDEC_WEATHER_2010_03_07
 - CDEC_WEATHER_2010_03_08
 - CDEC_WEATHER_2010_03_09
 - CDEC_WEATHER_2010_03_10
 - CDEC_WEATHER_2010_03_11

The right pane, titled 'DataBrowser', shows the selected file 'CDEC_WEATHER_2010_03_10'. It includes a 'View File' button and a 'Delete File' button. Below these is the title 'Metadata for CDEC_WEATHER_2010_03_10' and a 'Save Metadata' button. A table displays the metadata for this file:

Name	Value
barpress	760
depth	1.6564940150390628
flow	1213.7113
forDate	2010-03-10
forSite	SMN
siteLatitu...	37.347214
siteLongi...	-120.976181
slope	0.0001
usedAlg...	
velocity	0.5606918448339844

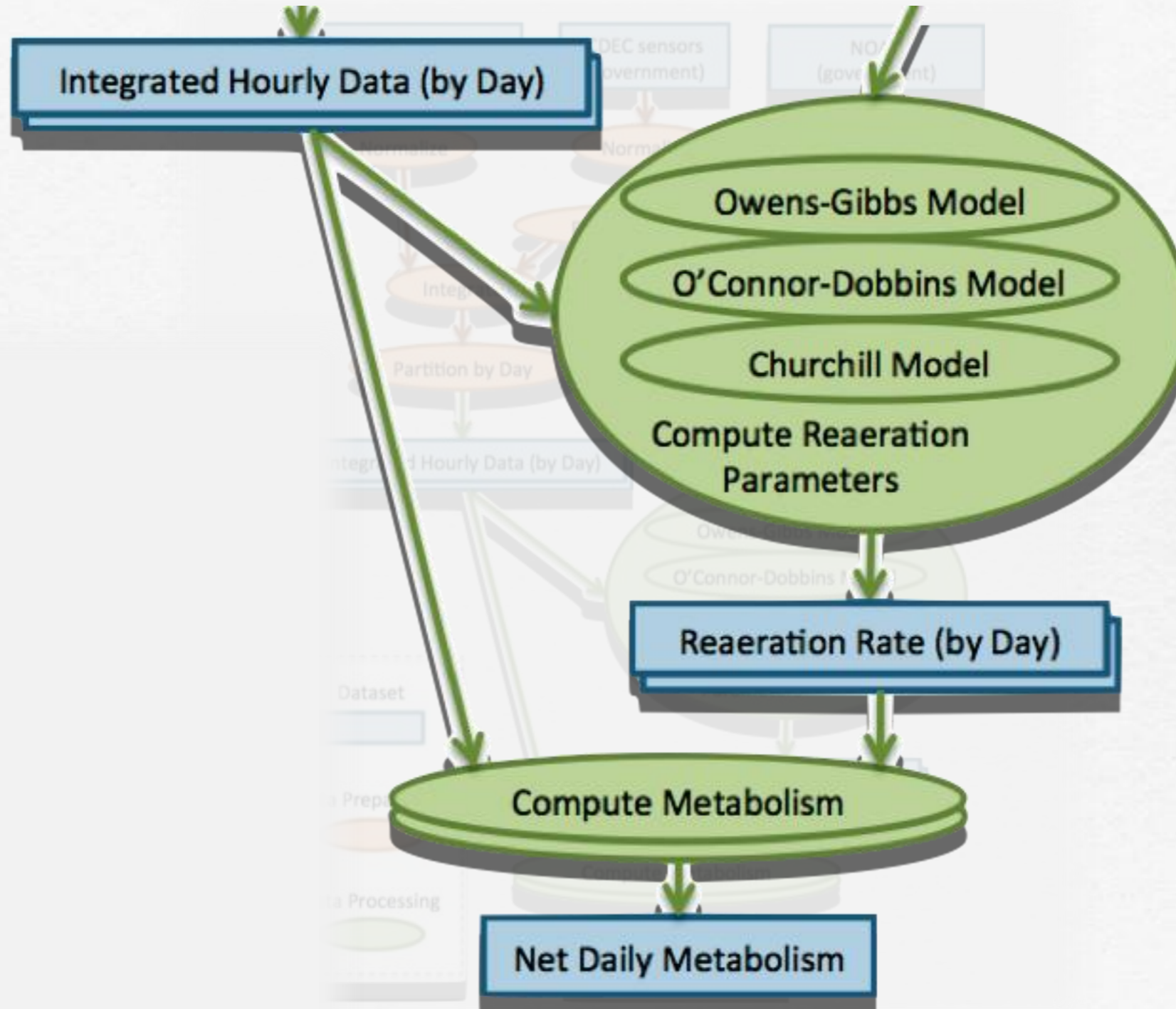
WINGS Received Metadata from KARMA

Metadata automatically associated with each input file

The screenshot shows a file browser window with a directory tree on the left and a metadata table on the right. The directory tree includes folders for 'Daily_Parameters' and 'Daily_Sensor_Data', with the latter containing a series of files named 'CDEC_WEATHER_2010_03_02' through 'CDEC_WEATHER_2010_03_14'. The file 'CDEC_WEATHER_2010_03_10' is selected. The metadata table on the right lists various parameters and their values for this file.

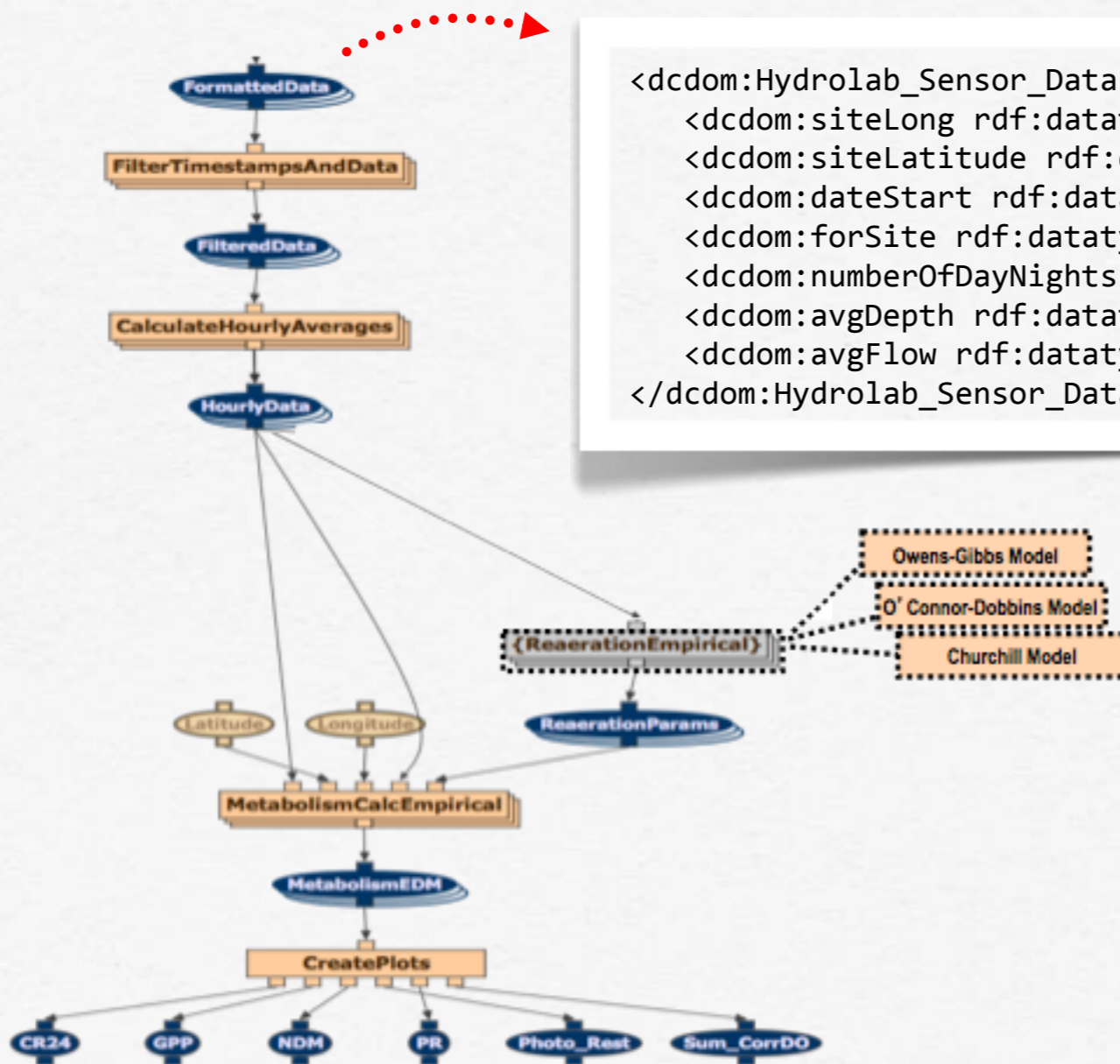
Name	Value
barpress	760
depth	1.6564940150390628
flow	1213.7113
forDate	2010-03-10
forSite	SMN
siteLatitu...	37.347214
siteLongi...	-120.976181
slope	0.0001
usedAlg...	
velocity	0.5606918448339844

Workflow



Using Metadata in Workflow Execution

Metadata



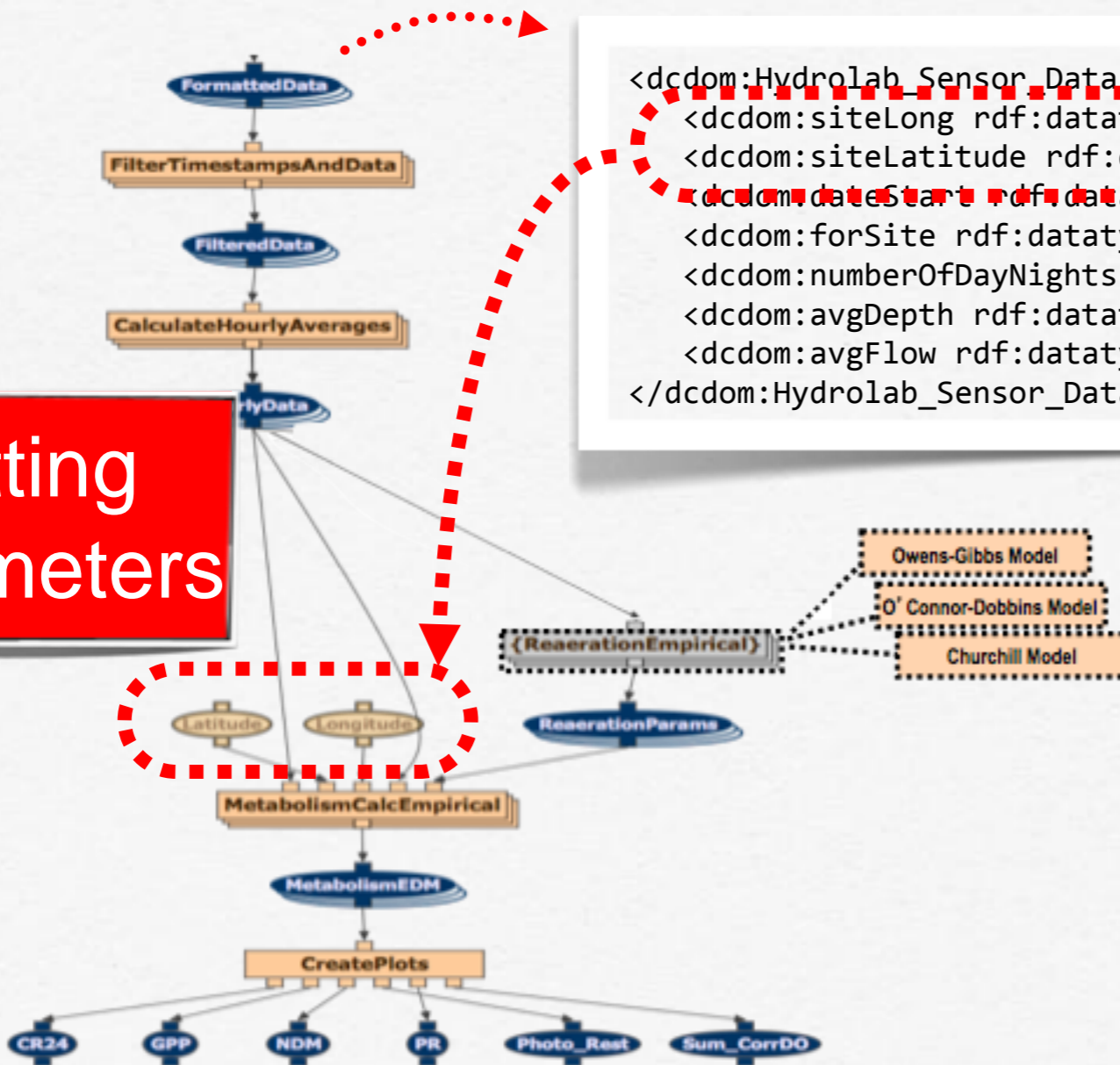
```
<dcdom:HydroLab_Sensor_Data rdf:ID="HydroLab-CDEC-04272011">
  <dcdom:siteLong rdf:datatype="float">-120.931</dcdom:siteLongitude>
  <dcdom:siteLatitude rdf:datatype="float">37.371</dcdom:siteLatitude>
  <dcdom:dateStart rdf:datatype="date">2011-04-27</dcdom:dateStart>
  <dcdom:forSite rdf:datatype="string">MST</dcdom:forSite>
  <dcdom:numberOfDayNights rdf:datatype="int">1</dcdom:numberOfDayNights>
  <dcdom:avgDepth rdf:datatype="float">4.523957</dcdom:avgDepth>
  <dcdom:avgFlow rdf:datatype="float">2399</dcdom:avgFlow>
</dcdom:HydroLab_Sensor_Data>
```

Using Metadata in Workflow Execution

Metadata

```
<dcdom:Hydrolab_Sensor_Data rdf:ID="Hydrolab-CDEC-04272011">  
<dcdom:siteLong rdf:datatype="float">-120.931</dcdom:siteLongitude>  
<dcdom:siteLatitude rdf:datatype="float">37.371</dcdom:siteLatitude>  
<dcdom:dateStart rdf:datatype="date">2011-04-27</dcdom:dateStart>  
<dcdom:forSite rdf:datatype="string">MST</dcdom:forSite>  
<dcdom:numberOfDayNights rdf:datatype="int">1</dcdom:numberOfDayNights>  
<dcdom:avgDepth rdf:datatype="float">4.523957</dcdom:avgDepth>  
<dcdom:avgFlow rdf:datatype="float">2399</dcdom:avgFlow>  
</dcdom:Hydrolab_Sensor_Data>
```

Setting Parameters



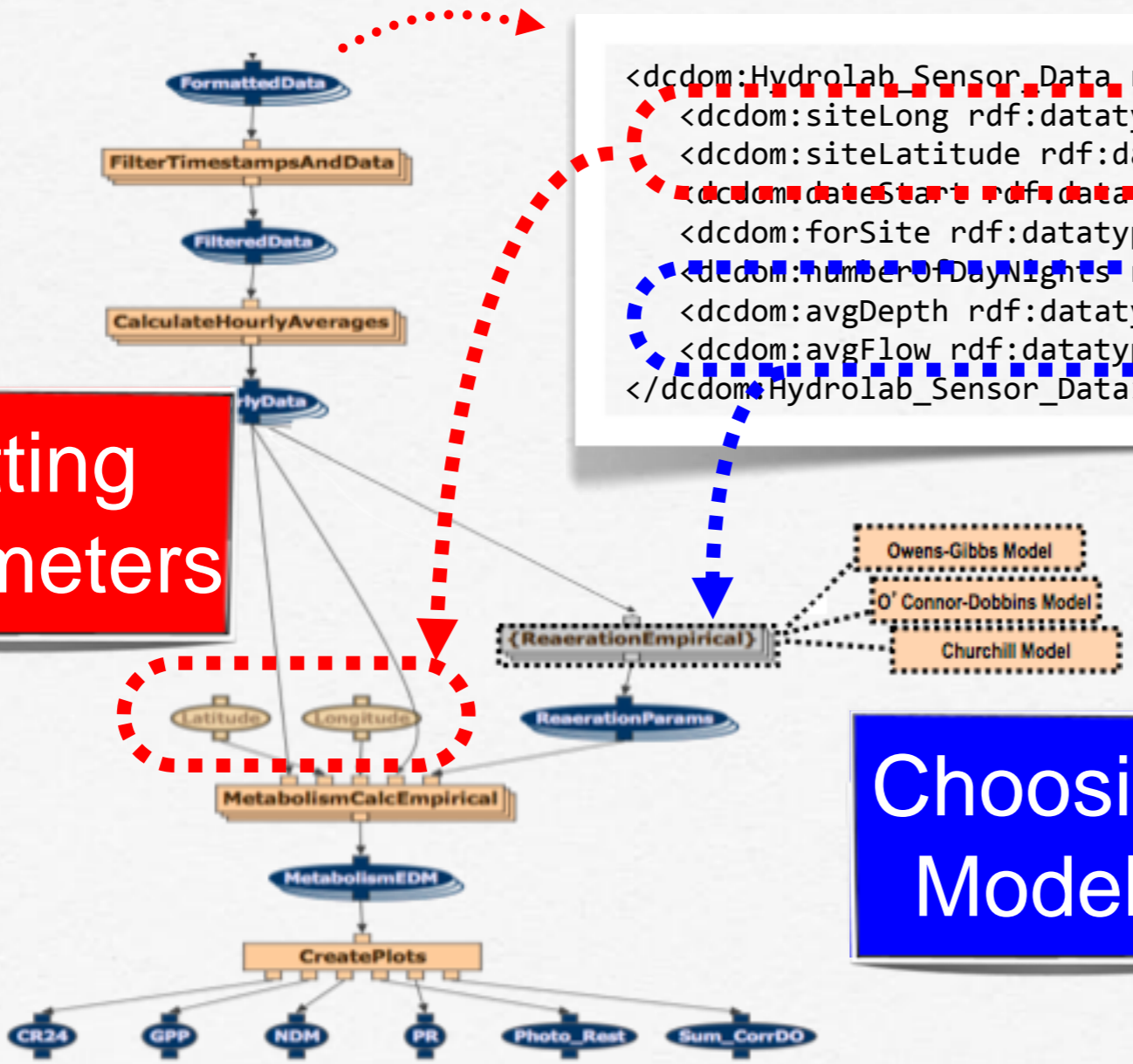
Using Metadata in Workflow Execution

Metadata

```
<dcdom:Hydrolab_Sensor_Data rdf:ID="Hydrolab-CDEC-04272011">  
<dcdom:siteLong rdf:datatype="float">-120.931</dcdom:siteLongitude>  
<dcdom:siteLatitude rdf:datatype="float">37.371</dcdom:siteLatitude>  
<dcdom:dateStart rdf:datatype="date">2011-04-27</dcdom:dateStart>  
<dcdom:forSite rdf:datatype="string">MST</dcdom:forSite>  
<dcdom:numberOfDayNights rdf:datatype="int">1</dcdom:numberOfDayNights>  
<dcdom:avgDepth rdf:datatype="float">4.523957</dcdom:avgDepth>  
<dcdom:avgFlow rdf:datatype="float">2399</dcdom:avgFlow>  
</dcdom:Hydrolab_Sensor_Data>
```

Setting Parameters

Choosing Models



Workflow Results

http://seagull.isi.edu/marbles/wpaccessresults.html

Most Visited Getting Started Yolanda's Frequent... Latest Headlines CSD Dict/Thes ISI ISD IKC YG YG Pointers News Google Maps Wikipedia Popular NYT Home Page

Run Workflows Access Results (PNG Image, 744x1180 pixels) NSF FastLane Proposal Status

My Runs

Delete Reload

Template	Progress	Start Time	End Time
AquaFlow-1 Run ID:64, Request ID:AquaFlow-1_Run_4d7a562851efe	Finished!	9:04:50 am, Mar 11, 2011	9:07:11 am, Mar 11, 2011

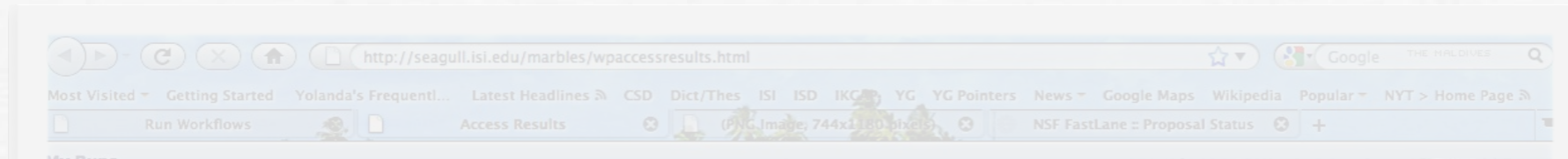
ResultBrowser **AquaFlow-1_Run_4d7a**

Data Run Log Workflow Documentation

Get HTML Get RDF

Variable	Bindings
Data: Input (4)	
1 Hydrolab_Data	Hydrolab_Jan_2011_NEXSENS_txt (168 KB)
2 InputDepthFile	{ depth_Jan_txt (57 B) }
3 Latitude	{ 37.37 }
4 Longitude	{ -120.93 }
Data: Intermediate (7)	
5 MeanNightDO	{ DO_MST_2011-01-01_0 (3 KB, Save), DO_MST_2011-01-01_1 (3 KB, Save), DO_MST_2011-01-01_2 (3 KB, Save), DO_MST_2011-01-01_3 (3 KB, Save), DO_MST_2011-01-01_4 (3 KB, Save), DO_MST_2011-01-01_5 (3 KB, Save), DO_MST_2011-01-01_6 (3 KB, Save), DO_MST_2011-01-01_7 (3 KB, Save), DO_MST_2011-01-01_8 (3 KB, Save), DO_MST_2011-01-01_9 (2 KB, Save), DO_MST_2011-01-01_10 (2 KB, Save), DO_MST_2011-01-01_11 (2 KB, Save), DO_MST_2011-01-01_12 (2 KB, Save), DO_MST_2011-01-01_13 (3 KB, Save), DO_MST_2011-01-01_14 (3 KB, Save), DO_MST_2011-01-01_15 (3 KB, Save), DO_MST_2011-01-01_16 (3 KB, Save), DO_MST_2011-01-01_17 (3 KB, Save), DO_MST_2011-01-01_18 (3 KB, Save), DO_MST_2011-01-01_19 (3 KB, Save), DO_MST_2011-01-01_20 (3 KB, Save), DO_MST_2011-01-01_21 (3 KB, Save), DO_MST_2011-01-01_22 (3 KB, Save), DO_MST_2011-01-01_23 (3 KB, Save), DO_MST_2011-01-01_24 (3 KB, Save), DO_MST_2011-01-01_25 (3 KB, Save), DO_MST_2011-01-01_26 (3 KB, Save), DO_MST_2011-01-01_27 (3 KB, Save), DO_MST_2011-01-01_28 (3 KB, Save), DO_MST_2011-01-01_29 (3 KB, Save) }
6 OutputDailyParams	{ Params_MST_2011-01-01_0 (59 B, Save), Params_MST_2011-01-01_1 (68 B, Save), Params_MST_2011-01-01_2 (66 B, Save), Params_MST_2011-01-01_3 (68 B, Save), Params_MST_2011-01-01_4 (66 B, Save), Params_MST_2011-01-01_5 (66 B, Save), Params_MST_2011-01-01_6 (66 B, Save), Params_MST_2011-01-01_7 (66 B, Save), Params_MST_2011-01-01_8 (66 B, Save), Params_MST_2011-01-01_9 (56 B, Save), Params_MST_2011-01-01_10 (46 B, Save), Params_MST_2011-01-01_11 (56 B, Save), Params_MST_2011-01-01_12 (66 B, Save), Params_MST_2011-01-01_13 (57 B, Save), Params_MST_2011-01-01_14 (68 B, Save), Params_MST_2011-01-01_15 (66 B, Save), Params_MST_2011-01-01_16 (66 B, Save), Params_MST_2011-01-01_17 (68 B, Save), Params_MST_2011-01-01_18 (68 B, Save), Params_MST_2011-01-01_19 (68 B, Save), Params_MST_2011-01-01_20 (68 B, Save), Params_MST_2011-01-01_21 (68 B, Save), Params_MST_2011-01-01_22 (66 B, Save), Params_MST_2011-01-01_23 (68 B, Save), Params_MST_2011-01-01_24 (66 B, Save), Params_MST_2011-01-01_25 (66 B, Save), Params_MST_2011-01-01_26 (57 B, Save), Params_MST_2011-01-01_27 (68 B, Save), Params_MST_2011-01-01_28 (66 B, Save), Params_MST_2011-01-01_29 (66 B, Save) }
7 OutputHourlyAvgedData	{ AvgHourly_MST_2011-01-01_0 (882 B, Save), AvgHourly_MST_2011-01-01_1 (871 B, Save), AvgHourly_MST_2011-01-01_2 (887 B, Save), AvgHourly_MST_2011-01-01_3 (905 B, Save), AvgHourly_MST_2011-01-01_4 (886 B, Save), AvgHourly_MST_2011-01-01_5 (858 B, Save), AvgHourly_MST_2011-01-01_6 (863 B, Save), AvgHourly_MST_2011-01-01_7 (854 B, Save),

Workflow Results



{ DO_MST_2011-01-01_0 (3 KB, Save), DO_MST_2011-01-01_1 (3 KB, Save), DO_MST_2011-01-01_2 (3 KB, Save), DO_MST_2011-01-01_3 (3 KB, Save), DO_MST_2011-01-01_4 (3 KB, Save), DO_MST_2011-01-01_5 (3 KB, Save), DO_MST_2011-01-01_6 (3 KB, Save), DO_MST_2011-01-01_7 (3 KB, Save), DO_MST_2011-01-01_8 (3 KB, Save), DO_MST_2011-01-01_9 (2 KB, Save), DO_MST_2011-01-01_10 (2 KB, Save), DO_MST_2011-01-01_11 (2 KB, Save), DO_MST_2011-01-01_12 (2 KB, Save), DO_MST_2011-01-01_13 (3 KB, Save), DO_MST_2011-01-01_14 (3 KB, Save), DO_MST_2011-01-01_15 (3 KB, Save), DO_MST_2011-01-01_16 (3 KB, Save), DO_MST_2011-01-01_17 (3 KB, Save), DO_MST_2011-01-01_18 (3 KB, Save), DO_MST_2011-01-01_19 (3 KB, Save), DO_MST_2011-01-01_20 (3 KB, Save), DO_MST_2011-01-01_21 (3 KB, Save), DO_MST_2011-01-01_22 (3 KB, Save), DO_MST_2011-01-01_23 (3 KB, Save), DO_MST_2011-01-01_24 (3 KB, Save), DO_MST_2011-01-01_25 (3 KB, Save), DO_MST_2011-01-01_26 (3 KB, Save), DO_MST_2011-01-01_27 (3 KB, Save), DO_MST_2011-01-01_28 (3 KB, Save), DO_MST_2011-01-01_29 (3 KB, Save) }

{ Params_MST_2011-01-01_0 (59 B, Save), Params_MST_2011-01-01_1 (68 B, Save), Params_MST_2011-01-01_2 (66 B, Save), Params_MST_2011-01-01_3 (68 B, Save), Params_MST_2011-01-01_4 (66 B, Save), Params_MST_2011-01-01_5 (66 B, Save), Params_MST_2011-01-01_6 (66 B, Save), Params_MST_2011-01-01_7 (66 B, Save), Params_MST_2011-01-01_8 (66 B, Save), Params_MST_2011-01-01_9 (56 B, Save), Params_MST_2011-01-01_10 (46 B, Save), Params_MST_2011-01-01_11 (56 B, Save), Params_MST_2011-01-01_12 (66 B, Save), Params_MST_2011-01-01_13 (57 B, Save), Params_MST_2011-01-01_14 (68 B, Save), Params_MST_2011-01-01_15 (66 B, Save), Params_MST_2011-01-01_16 (66 B, Save), Params_MST_2011-01-01_17 (68 B, Save), Params_MST_2011-01-01_18 (68 B, Save), Params_MST_2011-01-01_19 (68 B, Save), Params_MST_2011-01-01_20 (68 B, Save), Params_MST_2011-01-01_21 (68 B, Save), Params_MST_2011-01-01_22 (66 B, Save), Params_MST_2011-01-01_23 (68 B, Save), Params_MST_2011-01-01_24 (66 B, Save), Params_MST_2011-01-01_25 (66 B, Save), Params_MST_2011-01-01_26 (57 B, Save), Params_MST_2011-01-01_27 (68 B, Save), Params_MST_2011-01-01_28 (66 B, Save), Params_MST_2011-01-01_29 (66 B, Save) }

Params_MST_2011-01-01_6 (66 B, Save), Params_MST_2011-01-01_7 (66 B, Save), Params_MST_2011-01-01_8 (66 B, Save), Params_MST_2011-01-01_9 (56 B, Save), Params_MST_2011-01-01_10 (46 B, Save), Params_MST_2011-01-01_11 (56 B, Save), Params_MST_2011-01-01_12 (66 B, Save), Params_MST_2011-01-01_13 (57 B, Save), Params_MST_2011-01-01_14 (68 B, Save), Params_MST_2011-01-01_15 (66 B, Save), Params_MST_2011-01-01_16 (66 B, Save), Params_MST_2011-01-01_17 (68 B, Save), Params_MST_2011-01-01_18 (68 B, Save), Params_MST_2011-01-01_19 (68 B, Save), Params_MST_2011-01-01_20 (68 B, Save), Params_MST_2011-01-01_21 (68 B, Save), Params_MST_2011-01-01_22 (66 B, Save), Params_MST_2011-01-01_23 (68 B, Save), Params_MST_2011-01-01_24 (66 B, Save), Params_MST_2011-01-01_25 (66 B, Save), Params_MST_2011-01-01_26 (57 B, Save), Params_MST_2011-01-01_27 (68 B, Save), Params_MST_2011-01-01_28 (66 B, Save), Params_MST_2011-01-01_29 (66 B, Save) }

7 OutputHourlyAvgedData

{ AvgHourly_MST_2011-01-01_0 (882 B, Save), AvgHourly_MST_2011-01-01_1 (871 B, Save), AvgHourly_MST_2011-01-01_2 (887 B, Save), AvgHourly_MST_2011-01-01_3 (905 B, Save), AvgHourly_MST_2011-01-01_4 (886 B, Save), AvgHourly_MST_2011-01-01_5 (858 B, Save), AvgHourly_MST_2011-01-01_6 (863 B, Save), AvgHourly_MST_2011-01-01_7 (854 B, Save),

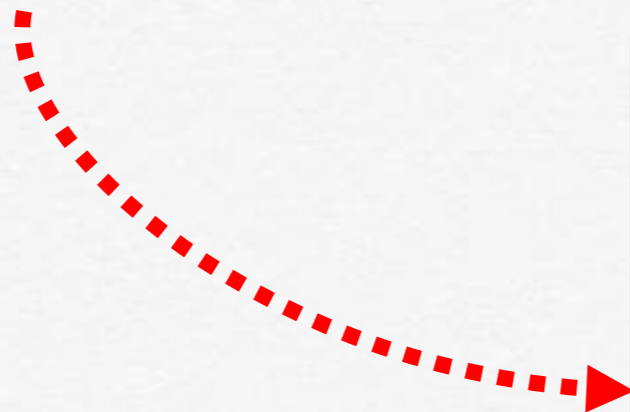
Workflow Results Have Metadata

WINGS automatically generates metadata for each output file

```
<dcdom:Metabolism_Results rdf:ID="Metabolism_Results-CDEC-04272011">  
  <dcdom:siteLong rdf:datatype="float">-120.931</dcdom:siteLongitude>  
  <dcdom:siteLatitude rdf:datatype="float">37.371</dcdom:siteLatitude>  
  <dcdom:dateStart rdf:datatype="date">2011-04-27</dcdom:dateStart>  
  <dcdom:forSite rdf:datatype="string">MST</dcdom:forSite>  
  <dcdom:numberOfDayNights rdf:datatype="int">1</dcdom:numberOfDayNights>  
  <dcdom:avgDepth rdf:datatype="float">4.523957</dcdom:avgDepth>  
  <dcdom:avgFlow rdf:datatype="float">2399</dcdom:avgFlow>  
</dcdom: Metabolism_Results>
```


WINGS Generates Provenance Metadata

```
SELECT ?url WHERE {  
  ?data dcdom:usedAlgorithm dcdom:ODM .  
  ?data rdf:type dcdom:Metabolism_Estimates .  
  ?data wflow:hasLocation ?url  
}
```



Metadata for Metabolism_SMN_2010_03_03Z_ODM	
Save Metadata	
Name	Value
velocity	0.66163415
usedAlgorithm	dcdom:ODM
slope	1.0E-4
siteLongitude	-120.97618
siteLatitude	37.347214
forSite	SMN
forDate	2010-03-03Z
flow	1581.6842
depth	1.0403947
temperature	750.0

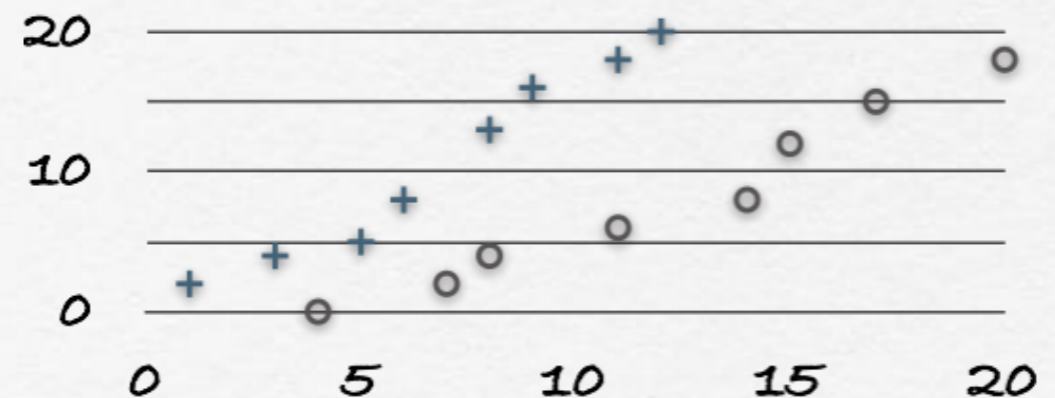
Aquatic Photosynthesis

Models of gross primary production (GPP),
community respiration (CR24)

Sensors



Analysis



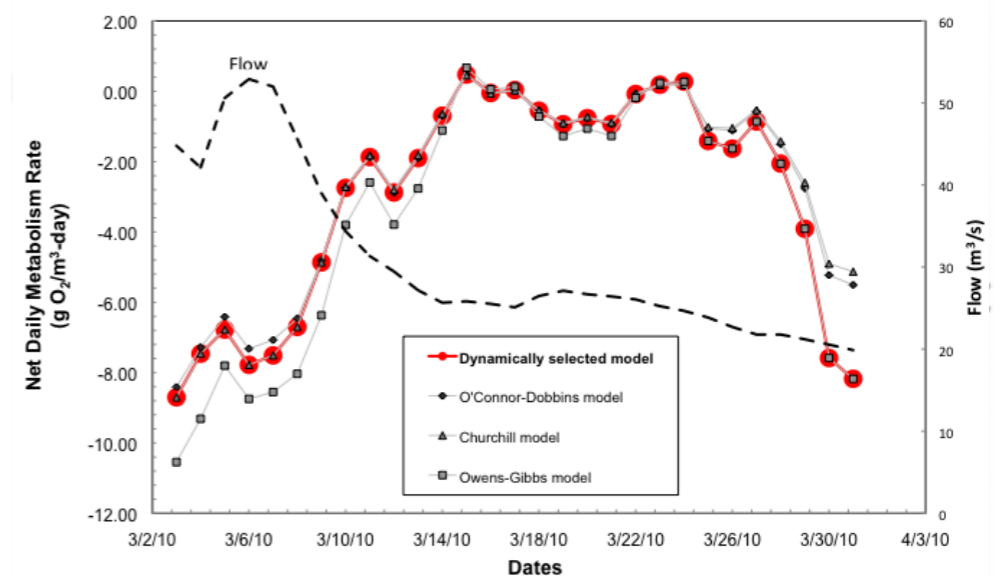
Aquatic Photosynthesis

Models of gross primary production (GPP),
community respiration (CR24)

Sensors



Workflow Results



Aquatic Photosynthesis

Models of gross primary production (GPP),
community respiration (CR24)

Sensors



Workflow Results



Summary

- ❑ Tools for end-users
- ❑ End to end support
- ❑ Data import, cleaning, integration
- ❑ Automated workflow execution
- ❑ Captures metadata provenance



Related Work

- Data integration:
 - Data Wrangler [Kandel et al 2011]
 - Google Refine [Huynh et al]
- Workflow systems:
 - VisTrails [Howe et al 2008],
 - Kepler [Barseghian et al 2010]
- Many tools generate provenance metadata, often in RDF
 - None generate other kinds of metadata
 - None use metadata to configure models



California's Central Valley Water Project

Complex network of rivers, reservoirs, canals, groundwater basins



Grand/Long-range vision

–Observe, model, manage water resources to optimize stream ecology while sustaining society's water needs

California Data Exchange Center

The screenshot shows the California Data Exchange Center website. The browser address bar displays the URL: `cdec.water.ca.gov/cgi-progs/stationInfo?station_id=MST`. The page header includes the CA.gov logo and the text "DEPARTMENT OF WATER RESOURCES California Data Exchange Center". A navigation menu contains links for Home, Query Tools, Precipitation, River Forecast, River Stages/Flow, Reservoirs, Snow, Stations, and Weather. Below the menu, there are links for "Lookup Station Metadata", "Real-time Data Stations", and "Daily Data Stations".

The main content area is titled "MERCED RIVER NEAR STEVINSON". It includes a "Map of surrounding area" link and a table of station metadata:

Station ID	MST	Elevation	82' ft
River Basin	MERCED R	County	MERCED
Hydrologic Area	SAN JOAQUIN RIVER	Nearby City	STEVINSON
Latitude	37.371000°N	Longitude	120.931000°W
Operator	CA Dept of Water Resources	Data Collection	

Below the metadata table is a "River Stage Definitions" table:

Datum 0	0.00' NGVD	Peak of Record	12/05/1950 73.80'
Monitor Stage	67.0'	Flood Stage	71.0'

The following data types are available online. Select one of the links below to retrieve recent data.

Sensor Description	Duration	Plot	Data Collection	Data A
ELECTRICAL CONDUCTIVITY MICRO S, us/cm	(daily)	(COND)	COMPUTED	07/01/200
FLOW, MEAN DAILY, cfs	(daily)	(FLOW)	COMPUTED	03/30/199
TEMPERATURE, WATER, deg f	(daily)	(TEMP W)	COMPUTED	07/01/200
BATTERY VOLTAGE, volts	(event)	(BAT VOL)	SATELLITE	02/08/200 07/04/200
FLOW, RIVER DISCHARGE, cfs	(event)	(FLOW)	COMPUTED	03/20/199
RIVER STAGE, feet	(event)	(RIV STG)	SATELLITE	03/20/199

On the left side of the page, there are two sections: "MOST POPULAR LINKS" and "RELATED LINKS".

MOST POPULAR LINKS

- Executive Summary
- Real-time Data
- Daily Data
- Monthly Data
- Historical Data
- Data Plotter
- Station Search
- Station Locator
- Daily Water Temperatures
- Reports
- Other Related Data Sources
- Contact CDEC Staff

RELATED LINKS

- California Cooperative Snow Surveys
- California Drought Information
- State Climatologist
- State Meteorologist
- Division of Flood Management
- Department of Water Resources

California Data Exchange Center

Date	Time	TempF
20091020	0	65.2
20091020	100	64.9
20091020	200	64.5
20091020	300	64.2
20091020	400	63.8
20091020	500	63.4
20091020	600	63.1
20091020	700	62.8
20091020	800	62.7
20091020	900	62.8
20091020	1000	63.2
20091020	1100	63.9
20091020	1200	64.6