Mind Your Metadata

ploiting Semantics for Configuration, Adaptation, a Provenance in Scientific Workflows

Yolanda Gil
Pedro Szekely
Craig Knoblock
Varun Ratnakar
Shubham Gupta
Maria Muslea
Fabio Silva

Tom Harmon
Sandra Villamizar

UC Merced

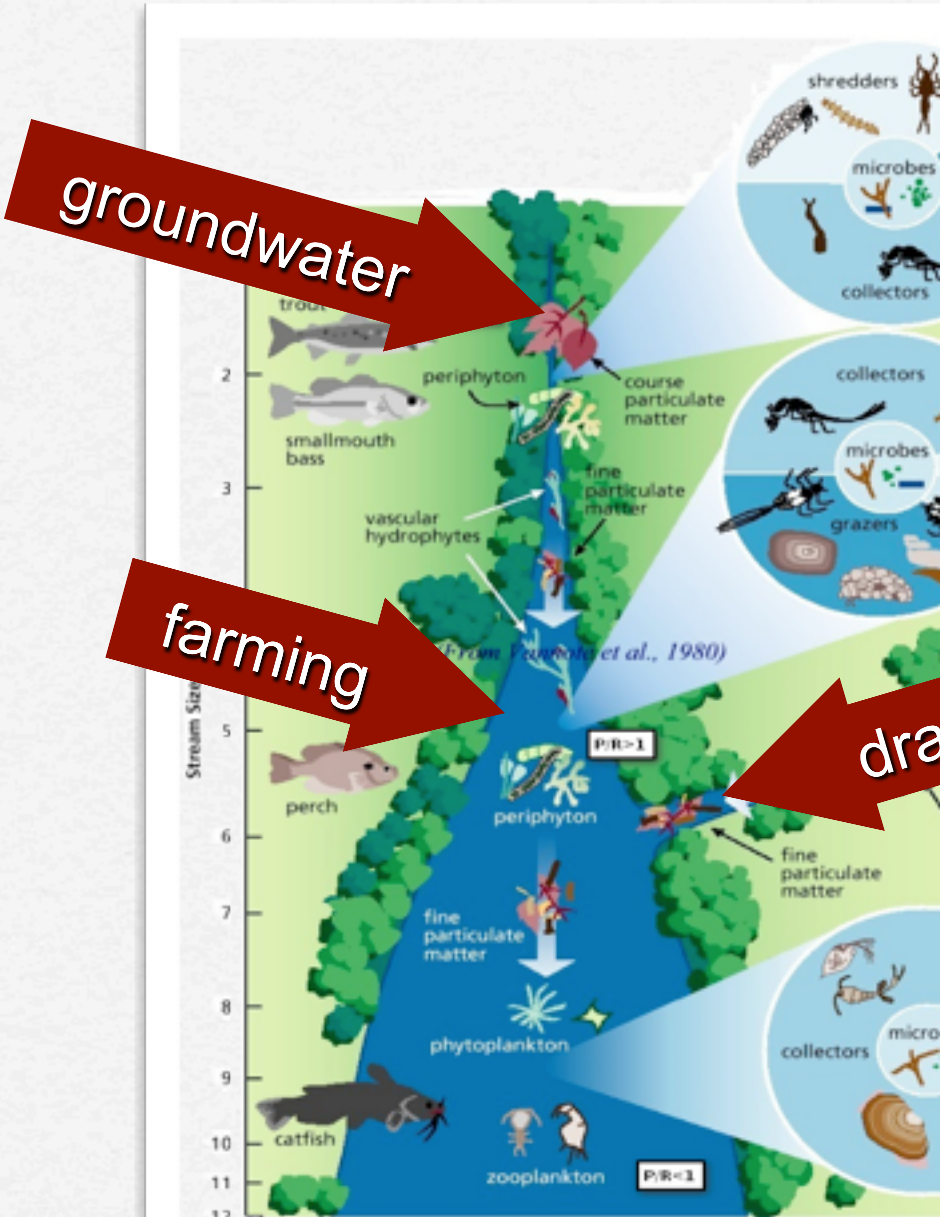
River Continuum vs Human Activities

River continuum: natural inputs, reactive transport

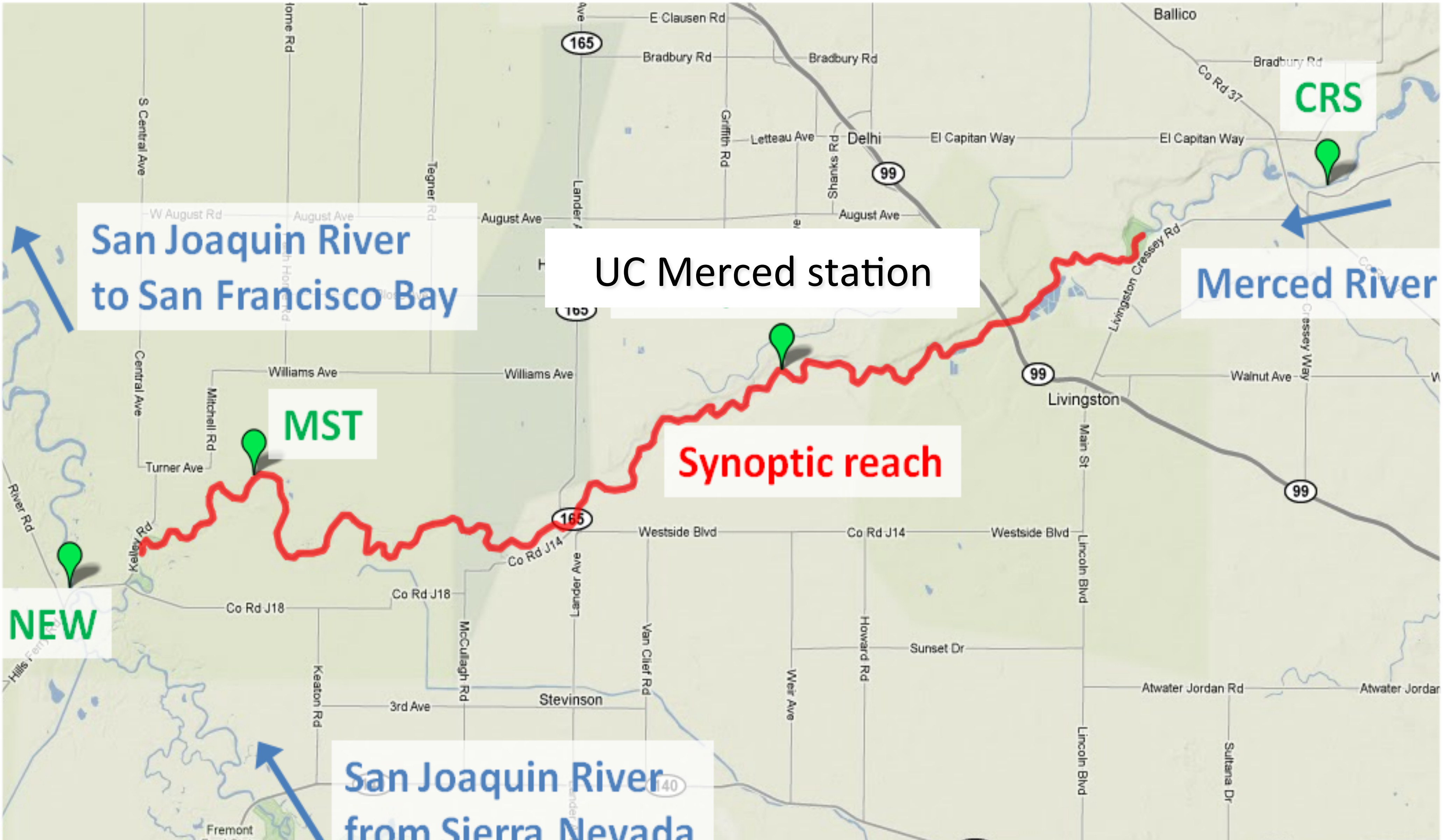
Human intervention: Agricultural, industrial, municipal

What management practices help/hurt?

Can we restore natural behavior?



Case Study



San Joaquin River to San Francisco Bay

UC Merced station

Merced River

CRS

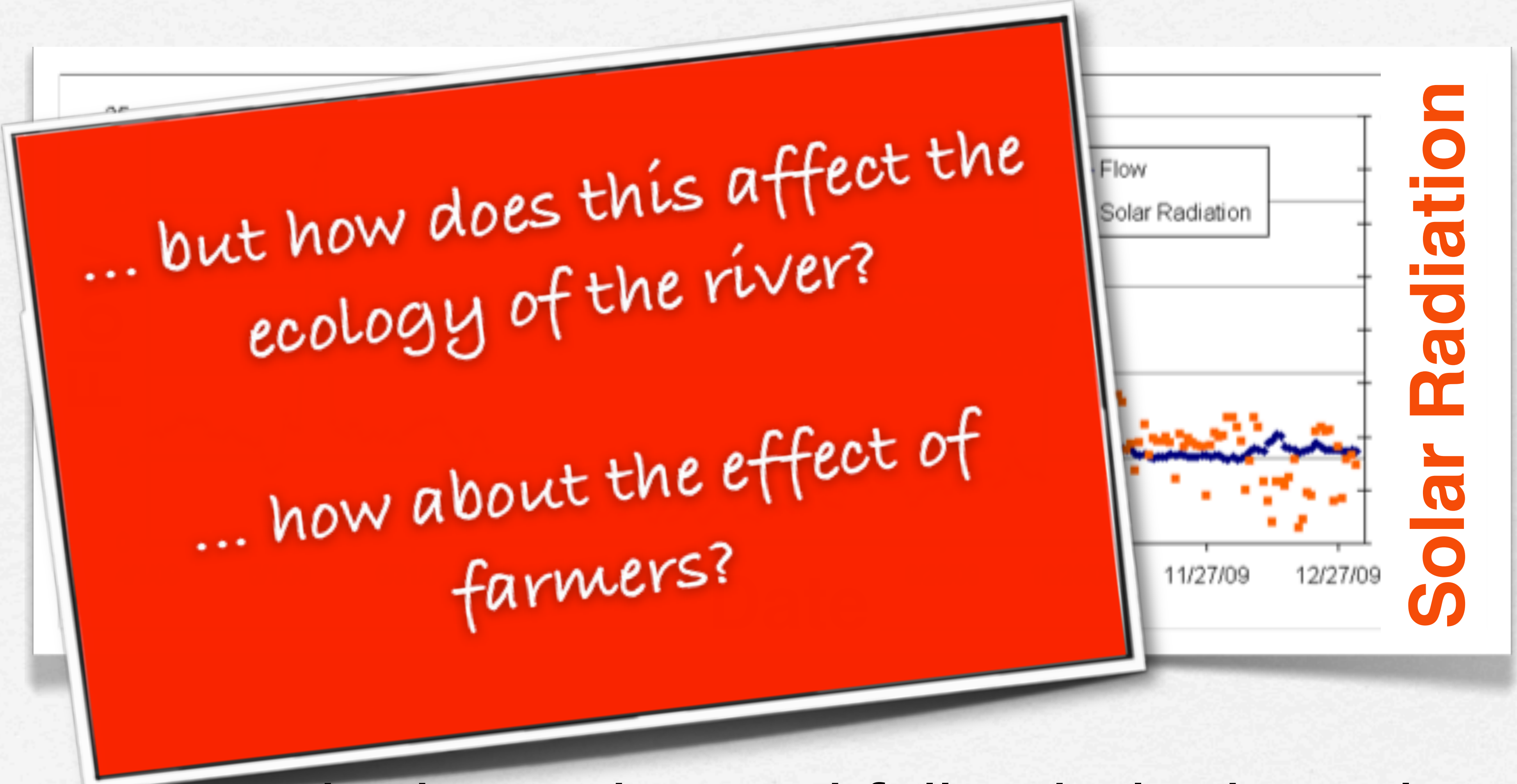
MST

NEW

Synoptic reach

San Joaquin River from Sierra Nevada

Stream Metabolism Response to Human Disturbances



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

... how about the effect of farmers?

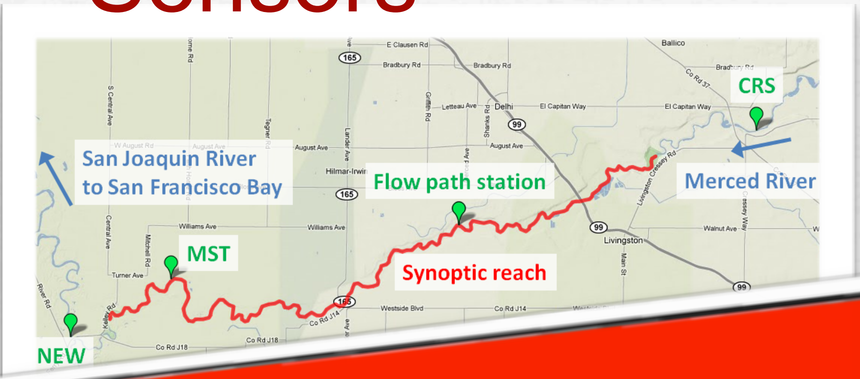
Solar Radiation

se releases in the spring and fall to help the salmon r

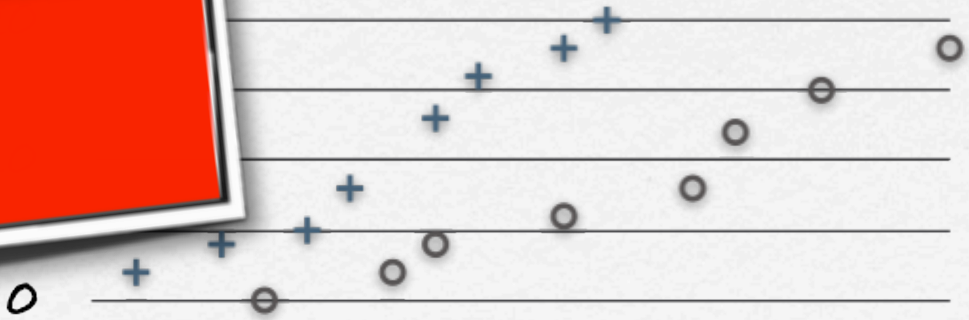
Aquatic Photosynthesis

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

Sensors

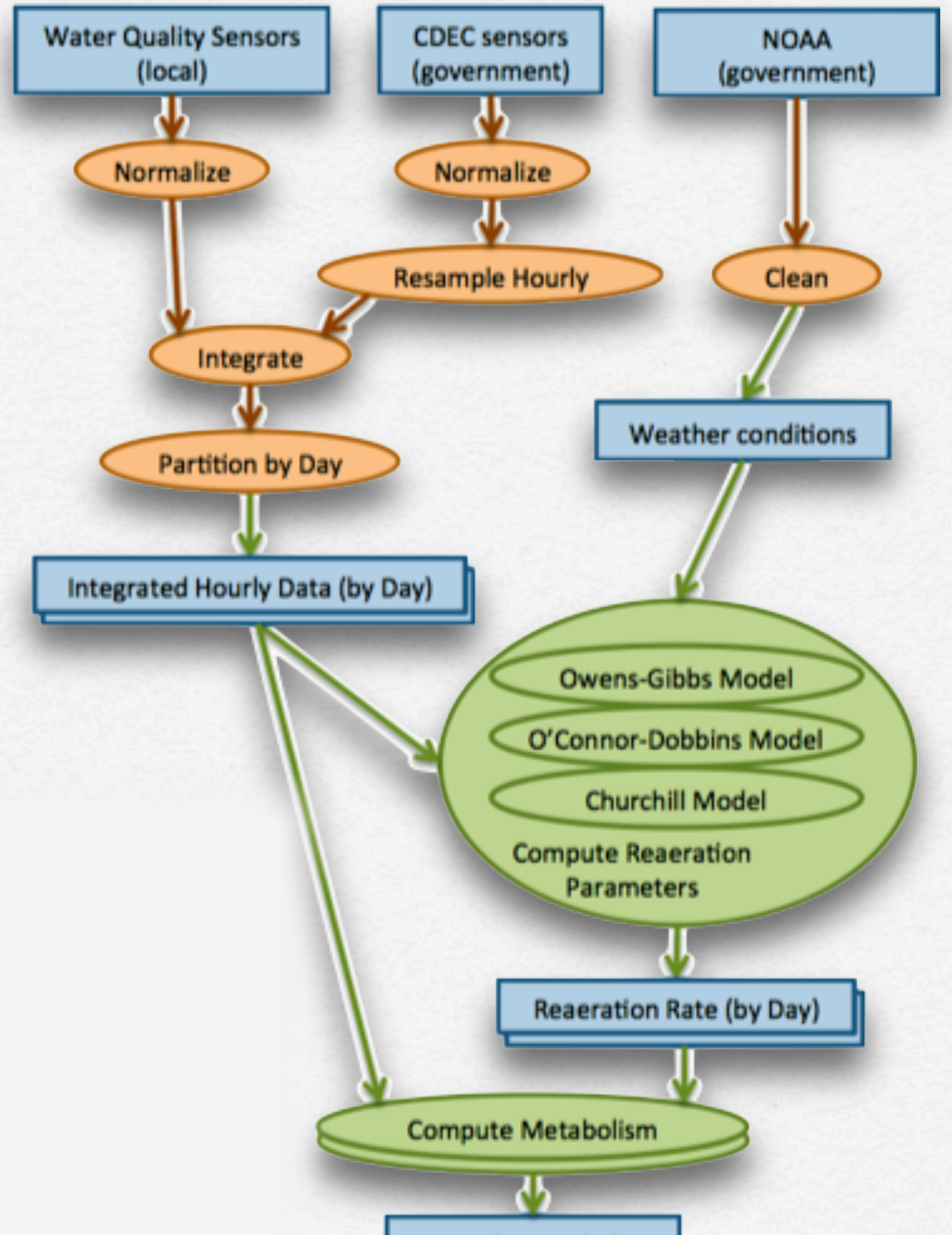


Analysis



alyses must be fast to produce actionable information

Workflow



Tom Harmon

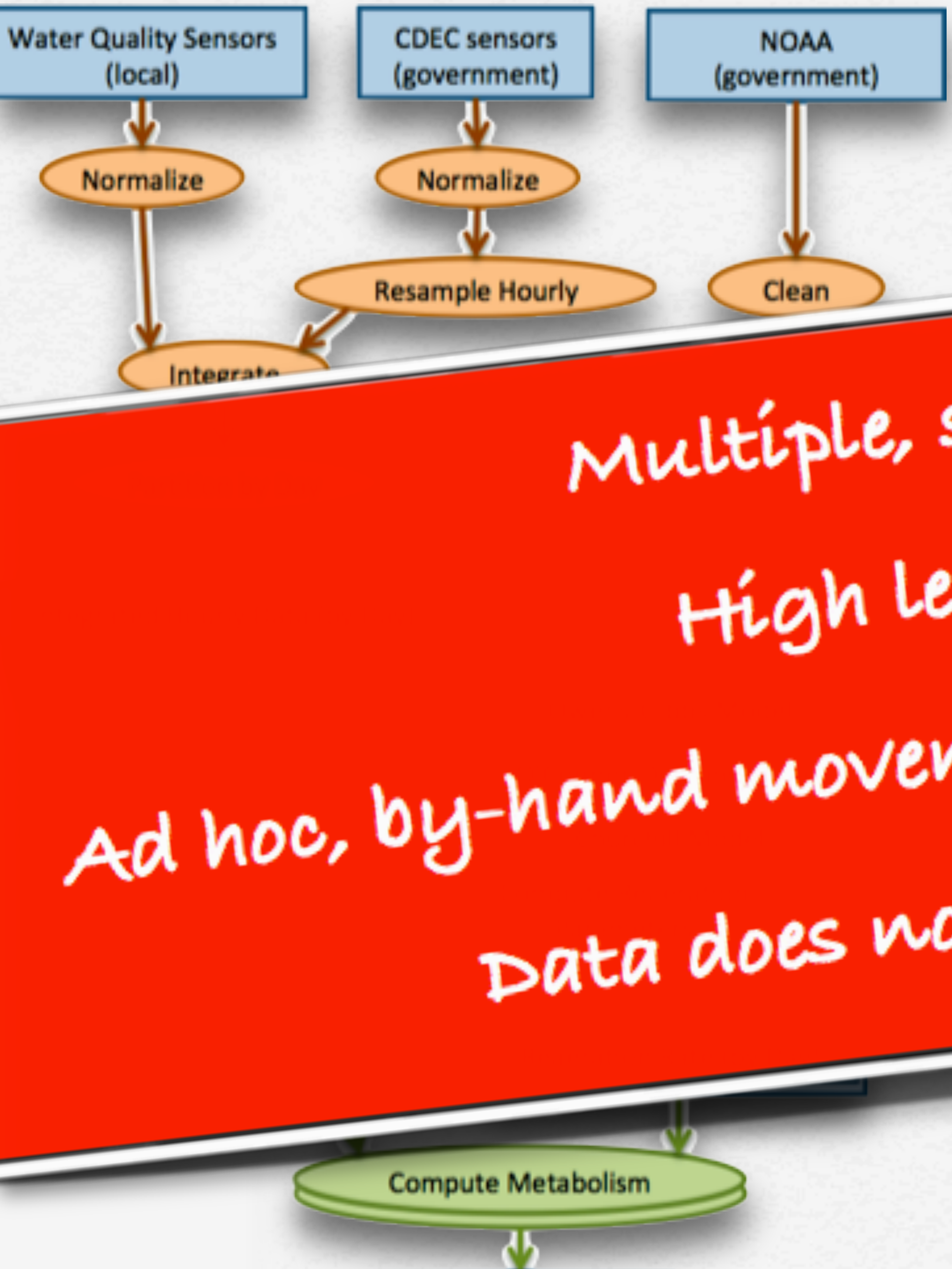
environmental system

Vision: Automated & Fast

Reality: Difficult & Time Consuming

?

Current Method



Manua

Multiple, separate tools

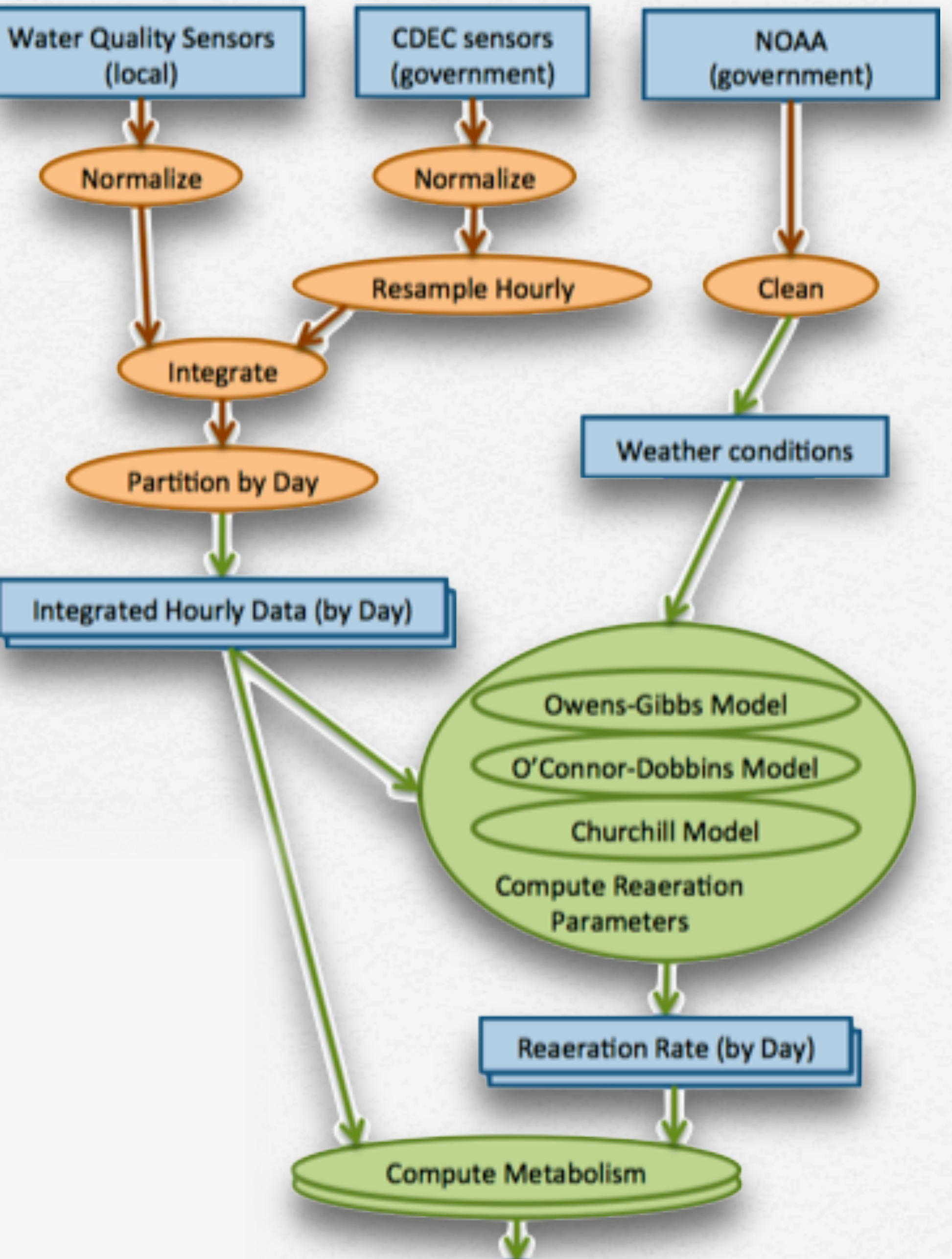
High learning costs

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

Data does not "flow" across tools

Scripts

Our Approach



Water Quality Sensors (local)

CDEC sensors (government)

NOAA (government)

Normalize

Normalize

Resample Hourly

Clean

Integrate

Weather conditions

Partition by Day

Integrated Hourly Data (by Day)

Owens-Gibbs Model

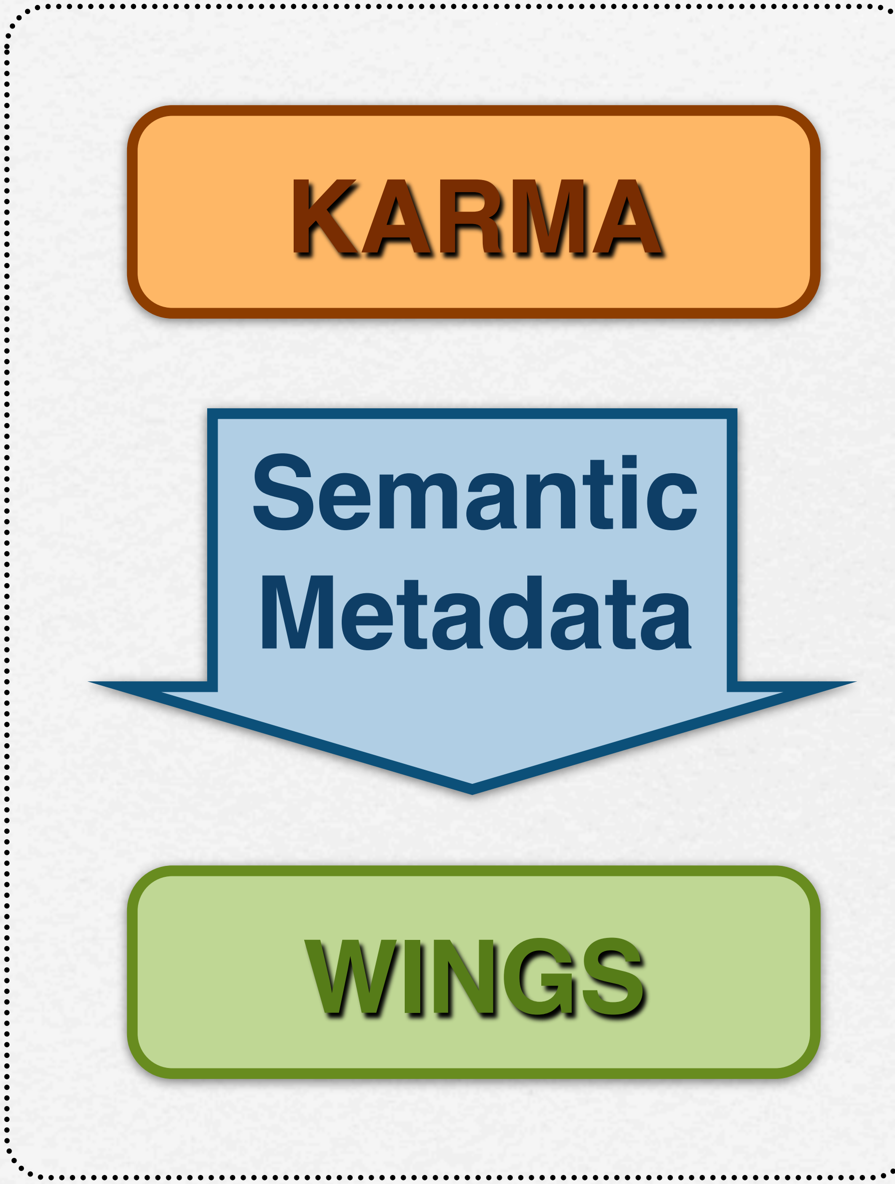
O'Connor-Dobbins Model

Churchill Model

Compute Reaeration Parameters

Reaeration Rate (by Day)

Compute Metabolism

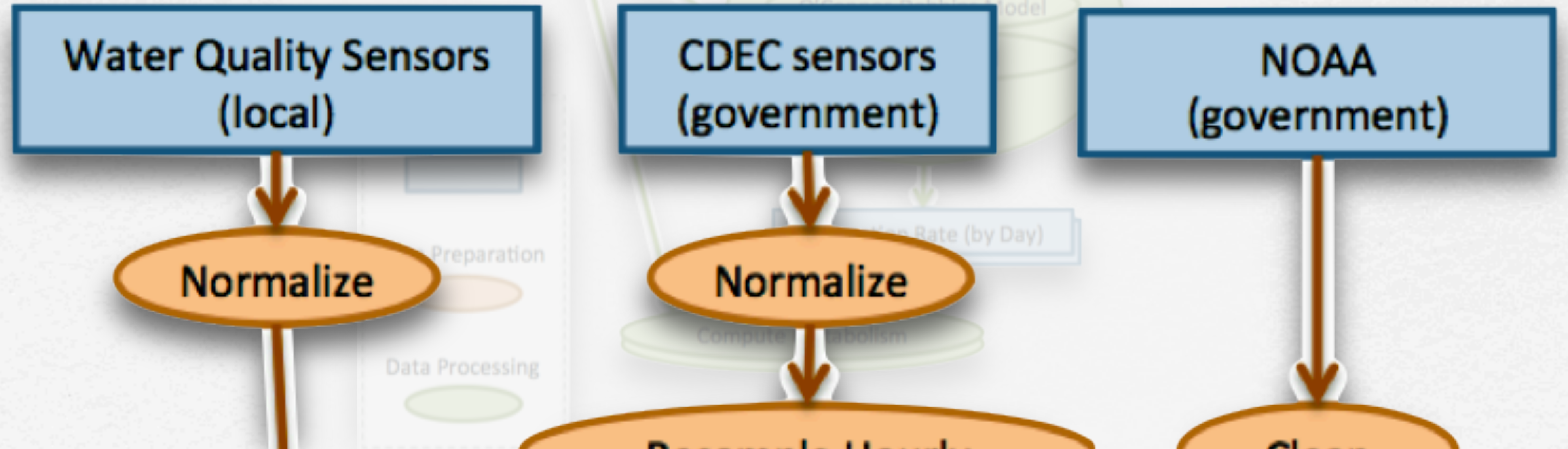


KARMA

Semantic Metadata

WINGS

Data Sources



Water Quality Sensors (local)

CDEC sensors (government)

NOAA (government)

Normalize

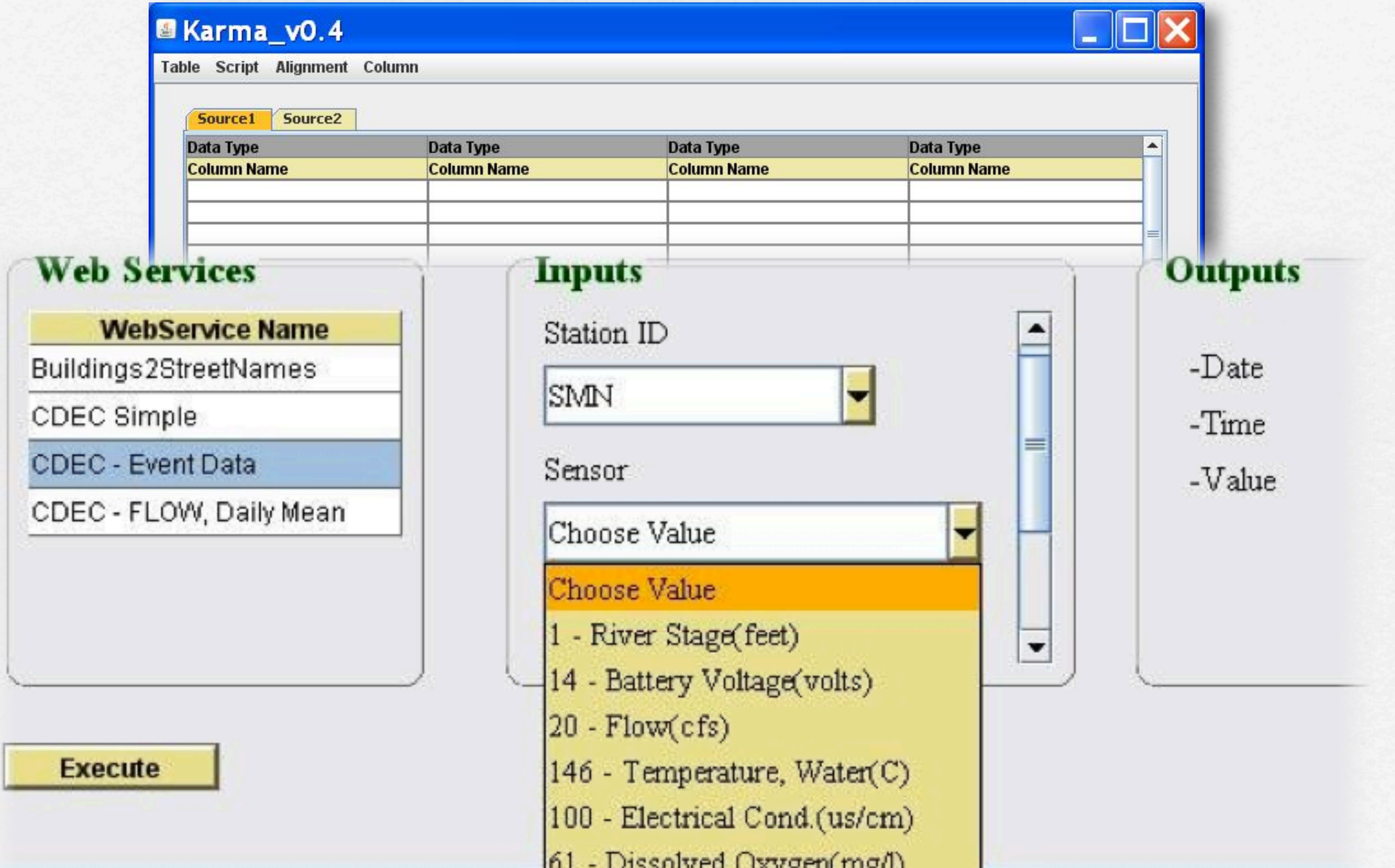
Normalize

Resample Hourly

Clean

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

KARMA



Data Type	Data Type	Data Type	Data Type
Column Name	Column Name	Column Name	Column Name

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

Station ID

SMN

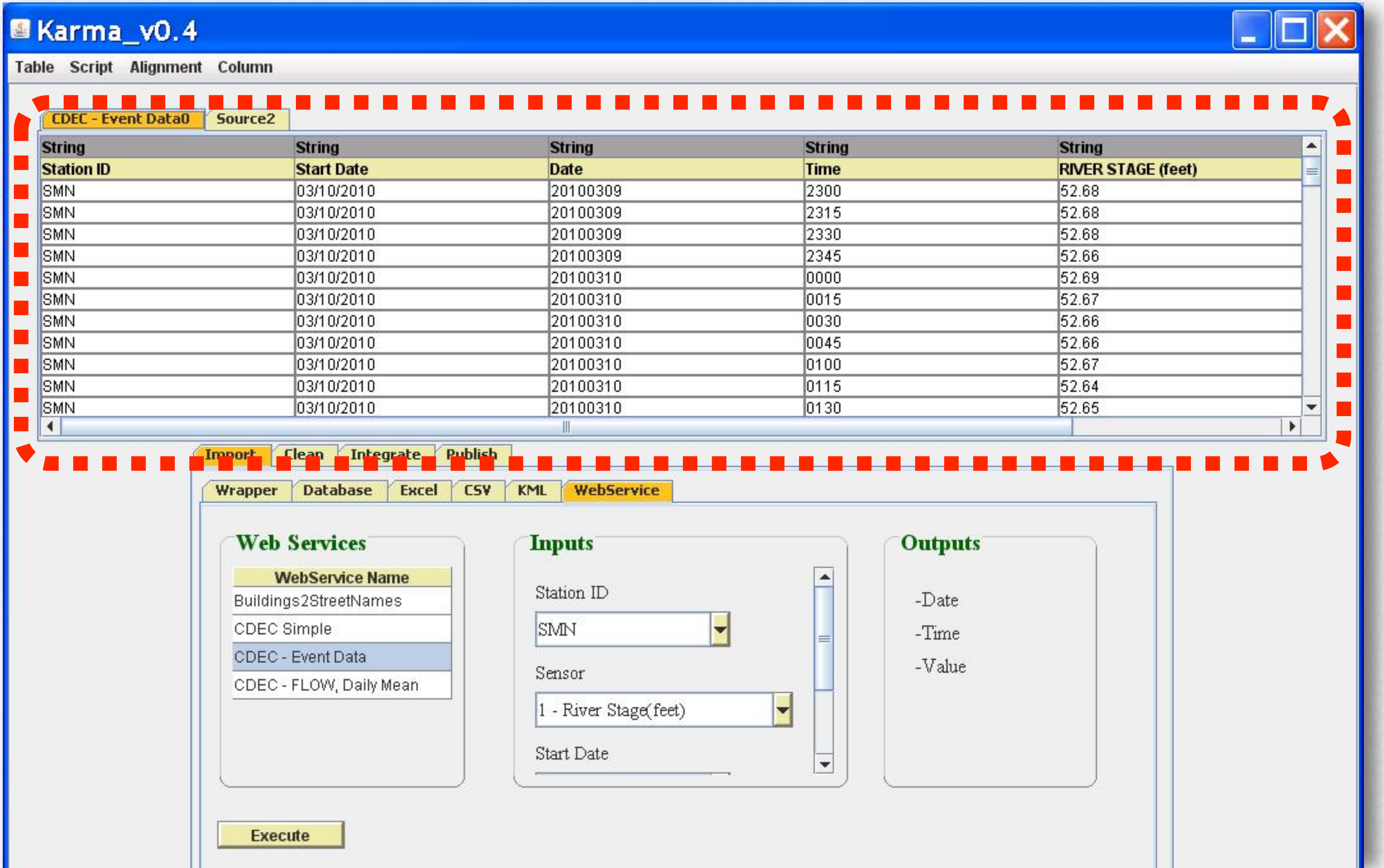
Sensor

Choose Value

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)

-Date
-Time
-Value

Data Import



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

Web Services

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

Inputs

Station ID: SMN

Sensor: 1 - River Stage(feet)

Start Date

Outputs

-Date
-Time
-Value

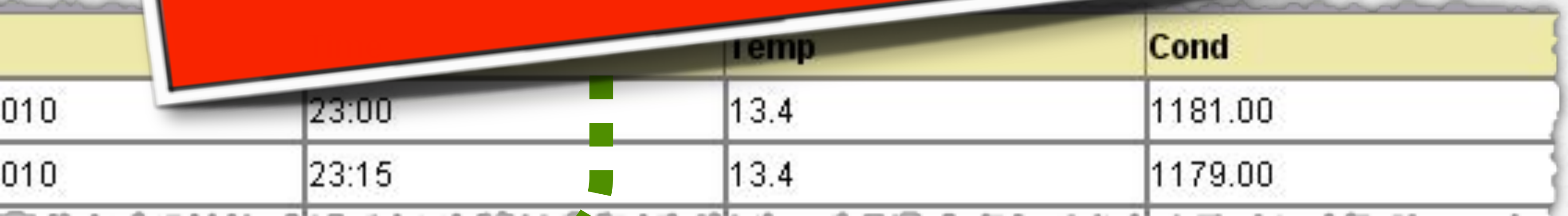
Need to Clean Data

	Time	GE (feet)
0309	2300	
0309		

60 Files for 1 month!

Cl

HYDRO

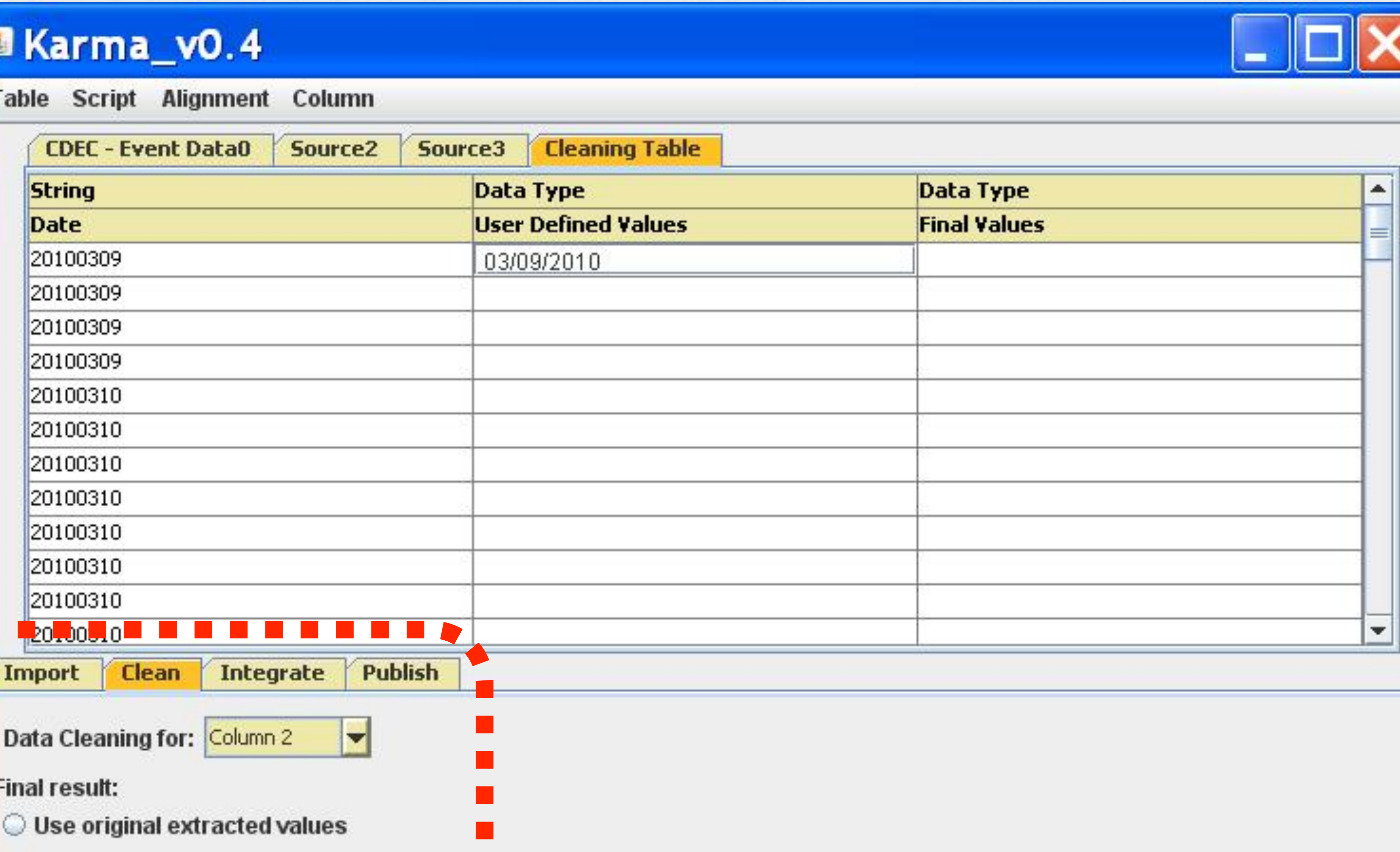


		Temp	Cond
010	23:00	13.4	1181.00
010	23:15	13.4	1179.00

Requ

Fo

Data Cleaning with KARMA



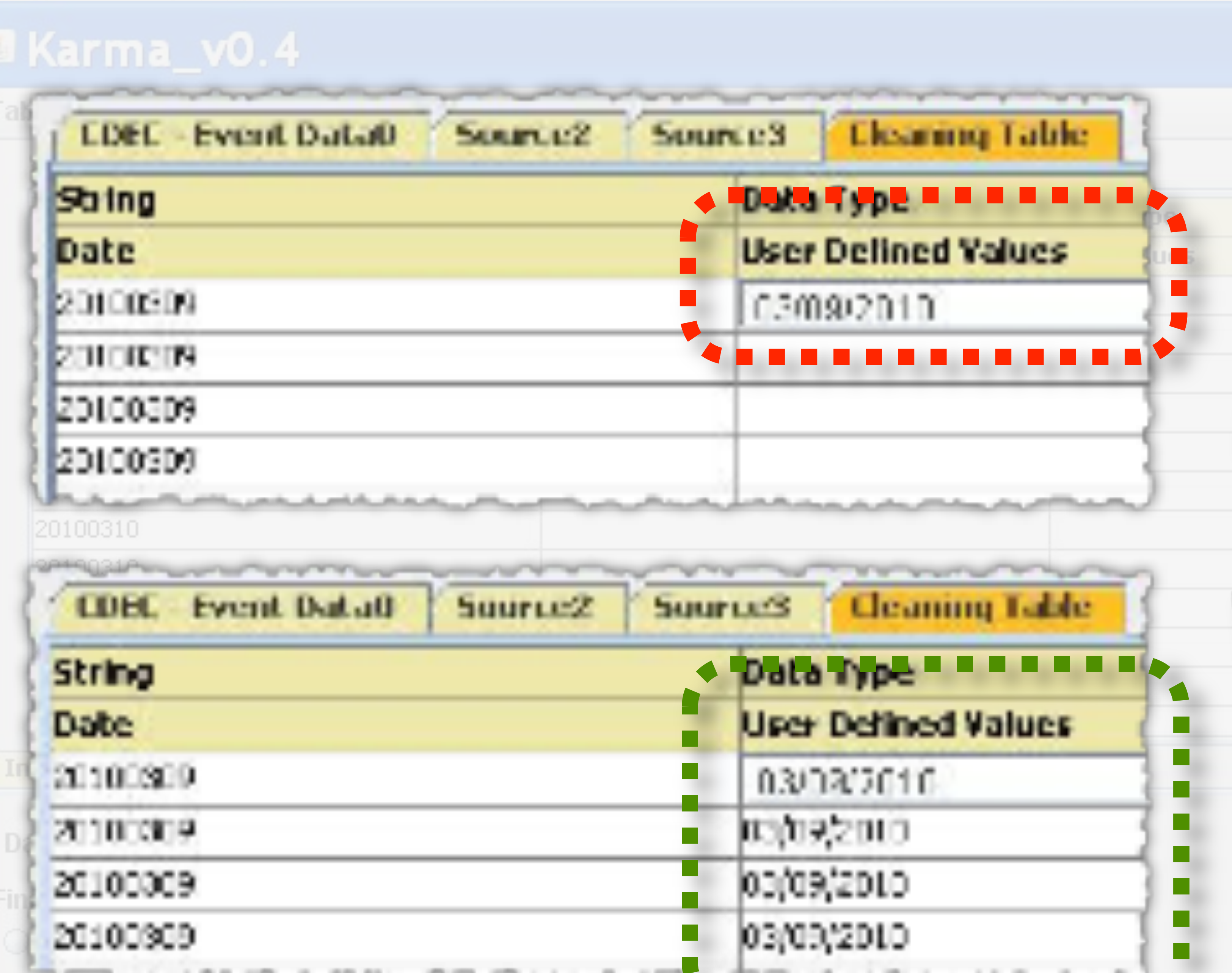
String	Data Type	Data Type
Date	User Defined Values	Final Values
20100309	03/09/2010	
20100309		
20100309		
20100309		
20100310		
20100310		
20100310		
20100310		
20100310		
20100310		
20100310		
20100310		

Data Cleaning for: Column 2

Final result:

○ Use original extracted values

Data Cleaning with KARMA



Karma_v0.4

CDEC - Event Data0 | Source2 | Source3 | Cleaning Table

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

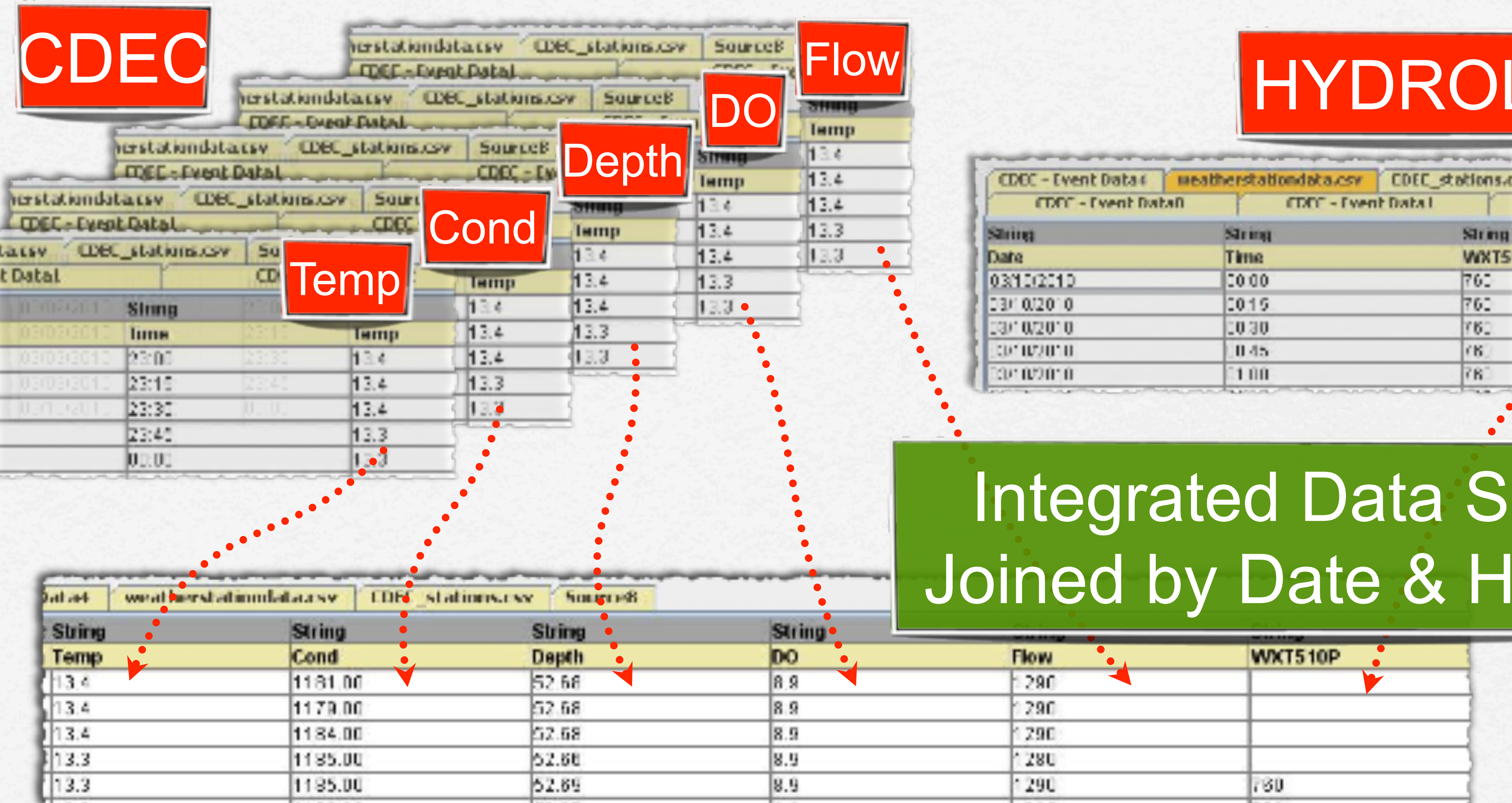
User provides example

CDEC - Event Data0 | 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

KARMA generates cleaning rule

eed to Integrate All the Source



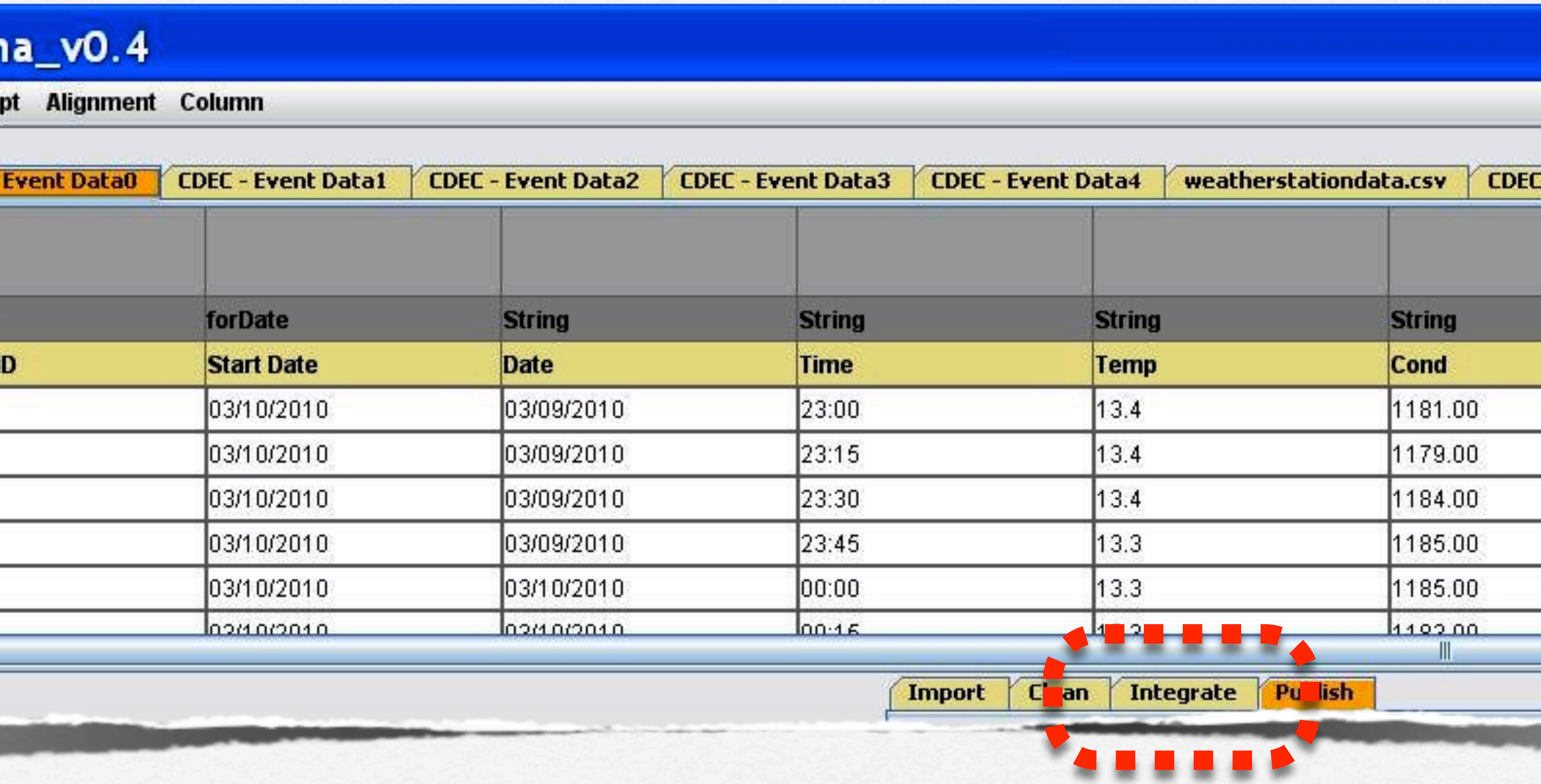
Integrated Data S
Joined by Date & H

CDEC - Event Data4	weatherstationdata.csv	CDEC_stations.c
CDEC - Event Data0	CDEC - Event Data1	
String	String	String
Date	Time	WXT5
03/10/2010	00:00	760
03/10/2010	00:15	760
03/10/2010	00:30	760
03/10/2010	00:45	760
03/10/2010	01:00	760

Data4 | weatherstationdata.csv | CDEC_stations.csv | SourceB

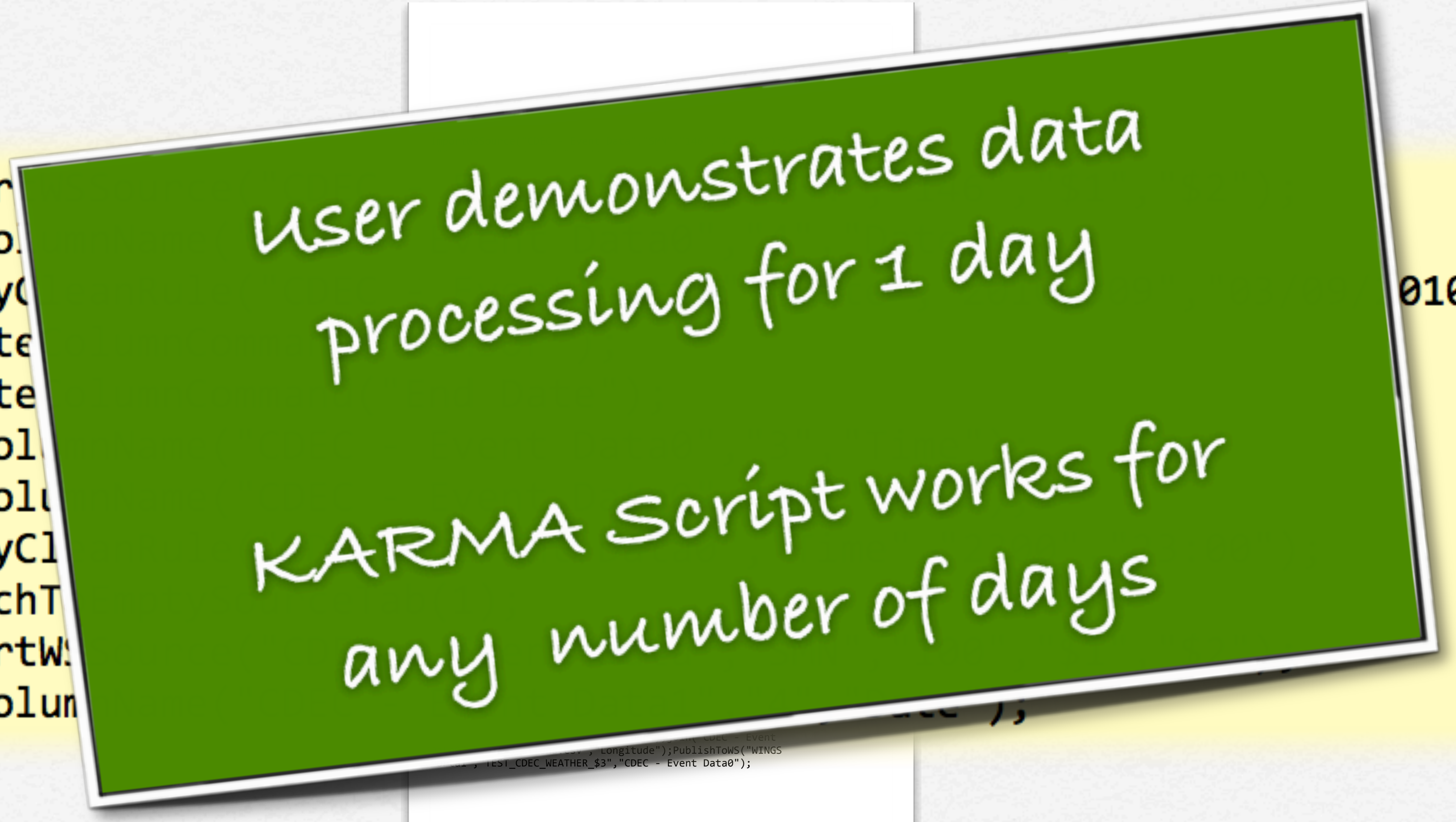
String	String	String	String	String	String
Temp	Cond	Depth	DO	Flow	WXT510P
13.4	1181.00	52.68	8.9	1290	
13.4	1179.00	52.68	8.9	1290	
13.4	1184.00	52.68	8.9	1290	
13.3	1185.00	52.68	8.9	1280	
13.3	1185.00	52.69	8.9	1290	760

Integrated Dataset



	forDate	String	String	String	String
D	Start Date	Date	Time	Temp	Cond
	03/10/2010	03/09/2010	23:00	13.4	1181.00
	03/10/2010	03/09/2010	23:15	13.4	1179.00
	03/10/2010	03/09/2010	23:30	13.4	1184.00
	03/10/2010	03/09/2010	23:45	13.3	1185.00
	03/10/2010	03/10/2010	00:00	13.3	1185.00
	03/10/2010	03/10/2010	00:15	13.3	1183.00

RMA Generates Data Processing Sc



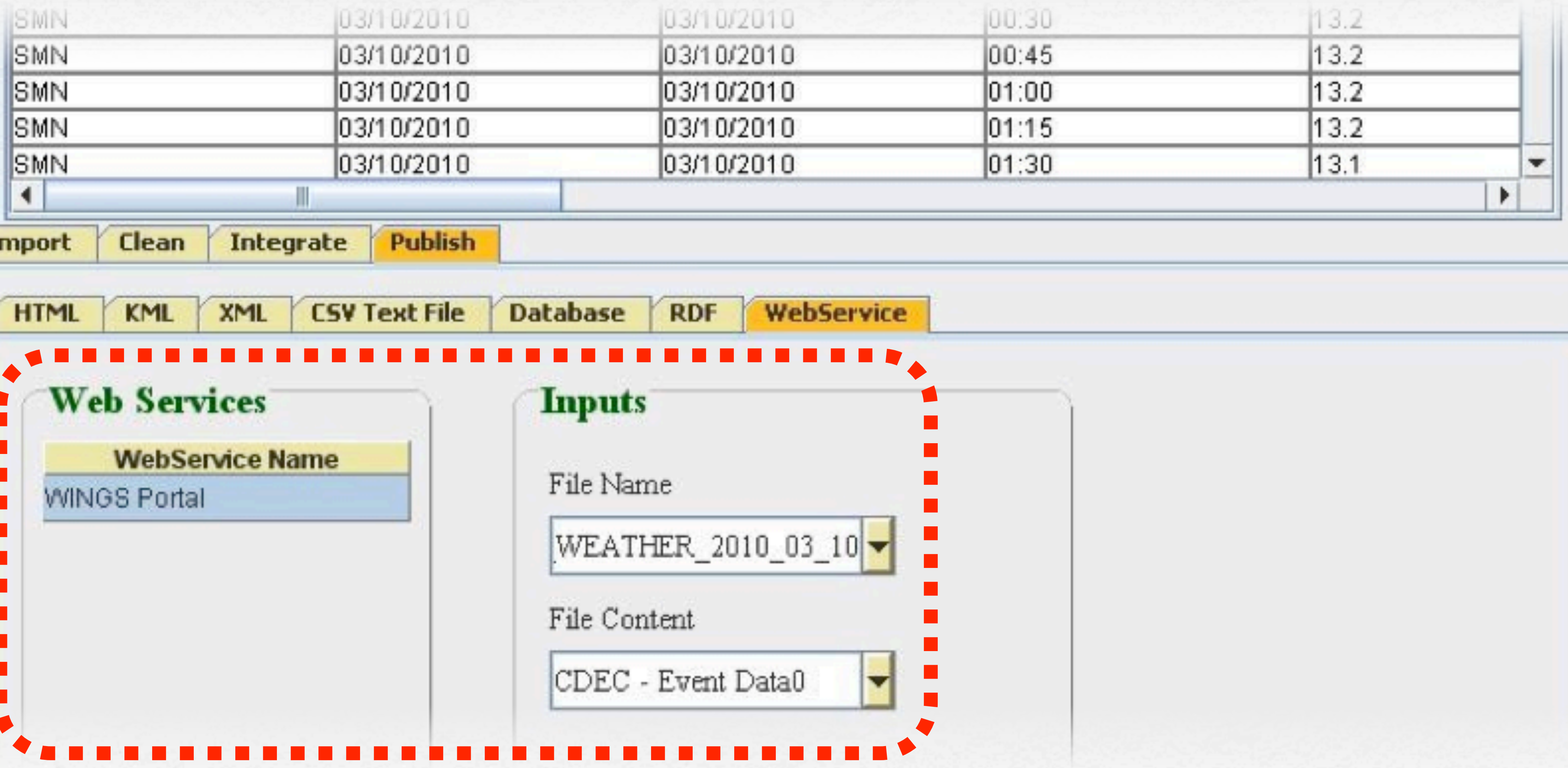
```
r
o
yC                                                   016
te
te
ol
olu
yCl
chT
rtWS
olum                                          ate");
```

```
CDEC - Event
, Longitude");PublishToWS("WINGS
, TEST_CDEC_WEATHER_$3","CDEC - Event Data0");
```

User demonstrates data processing for 1 day

KARMA Script works for any number of days

ublishing Processed Data to WING



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

Web Services

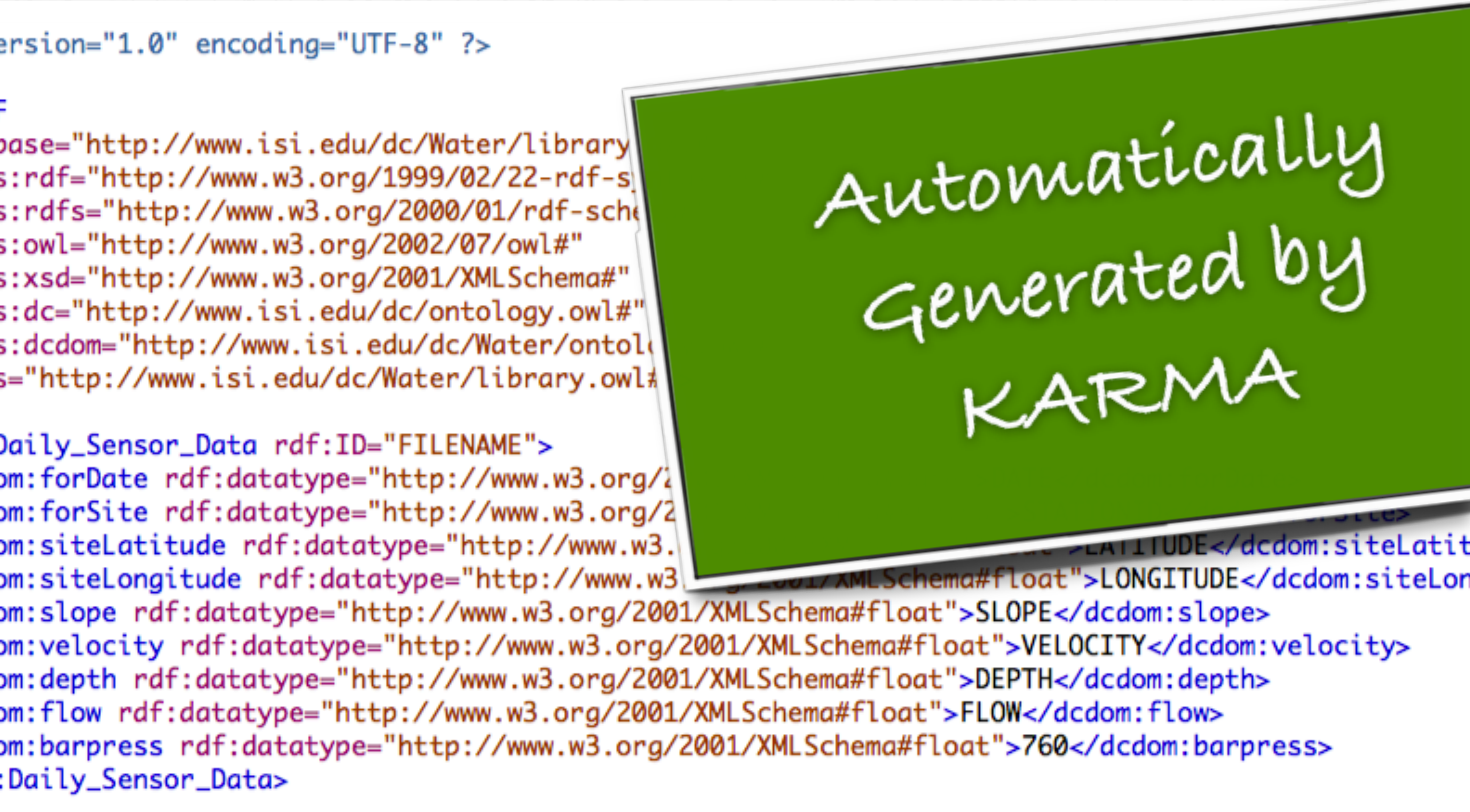
WebService Name
WINGS Portal

Inputs

File Name: WEATHER_2010_03_10

File Content: CDEC - Event Data0

Semantic Metadata for Input Files

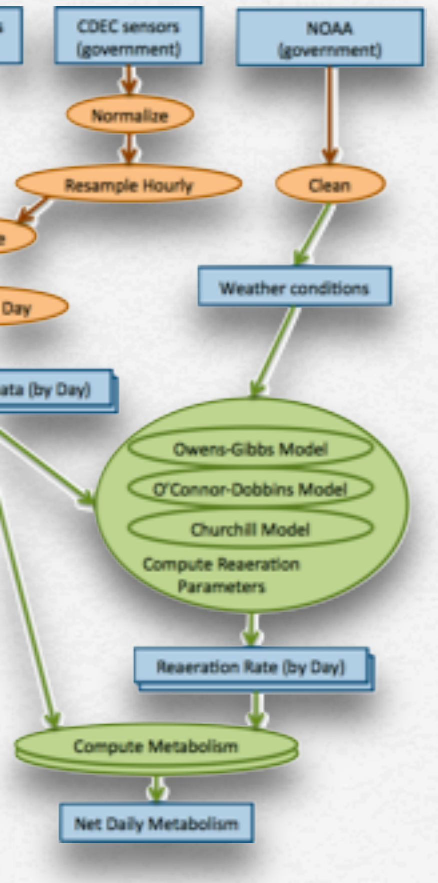


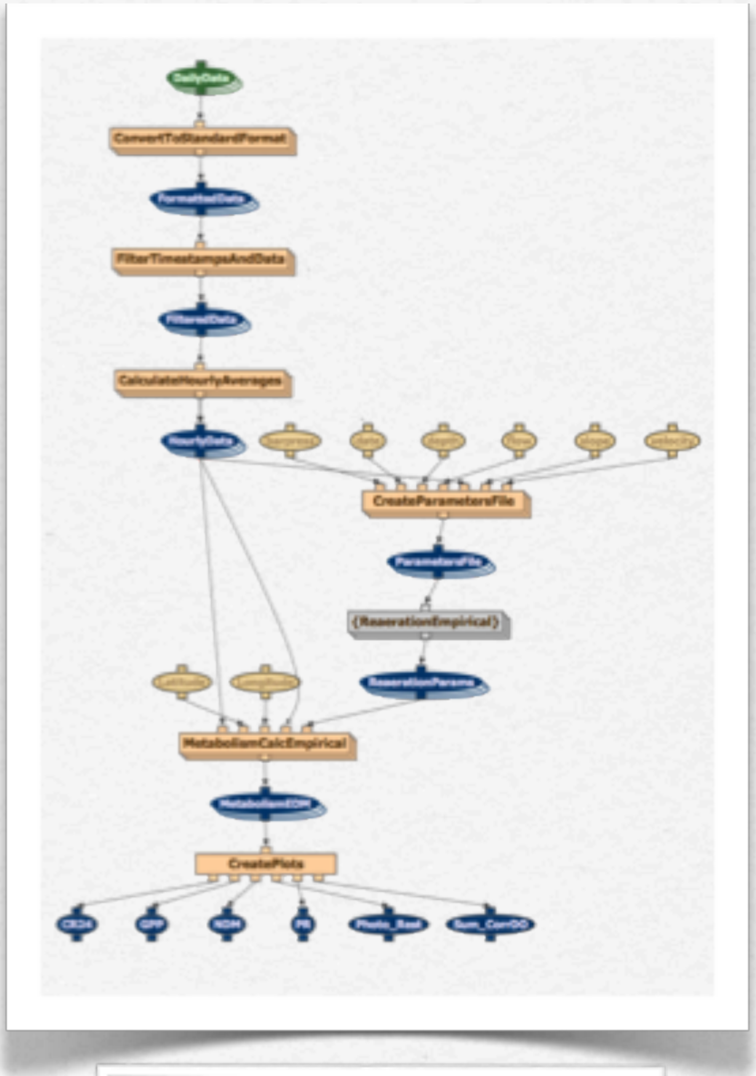
```
ersion="1.0" encoding="UTF-8" ?>

:
base="http://www.isi.edu/dc/Water/library
s:rdf="http://www.w3.org/1999/02/22-rdf-s
s:rdfs="http://www.w3.org/2000/01/rdf-sche
s:owl="http://www.w3.org/2002/07/owl#"
s:xsd="http://www.w3.org/2001/XMLSchema#"
s:dc="http://www.isi.edu/dc/ontology.owl#"
s:dcdom="http://www.isi.edu/dc/Water/ontol
s="http://www.isi.edu/dc/Water/library.owl#

Daily_Sensor_Data rdf:ID="FILENAME">
om:forDate rdf:datatype="http://www.w3.org/2
om:forSite rdf:datatype="http://www.w3.org/2
om:siteLatitude rdf:datatype="http://www.w3.                  LATITUDE</dcdom:siteLatit
om:siteLongitude rdf:datatype="http://www.w3                XMLSchema#float">LONGITUDE</dcdom:siteLon
om:slope rdf:datatype="http://www.w3.org/2001/XMLSchema#float">SLOPE</dcdom:slope>
om:velocity rdf:datatype="http://www.w3.org/2001/XMLSchema#float">VELOCITY</dcdom:velocity>
om:depth rdf:datatype="http://www.w3.org/2001/XMLSchema#float">DEPTH</dcdom:depth>
om:flow rdf:datatype="http://www.w3.org/2001/XMLSchema#float">FLOW</dcdom:flow>
om:barpress rdf:datatype="http://www.w3.org/2001/XMLSchema#float">760</dcdom:barpress>
:Daily_Sensor_Data>
```

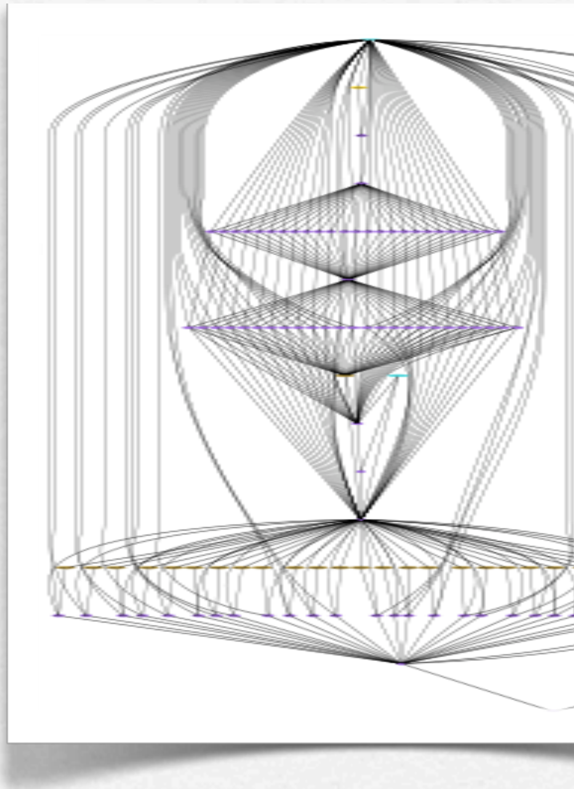
Automatically Generated by KARMA

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

Workflows with WINGS





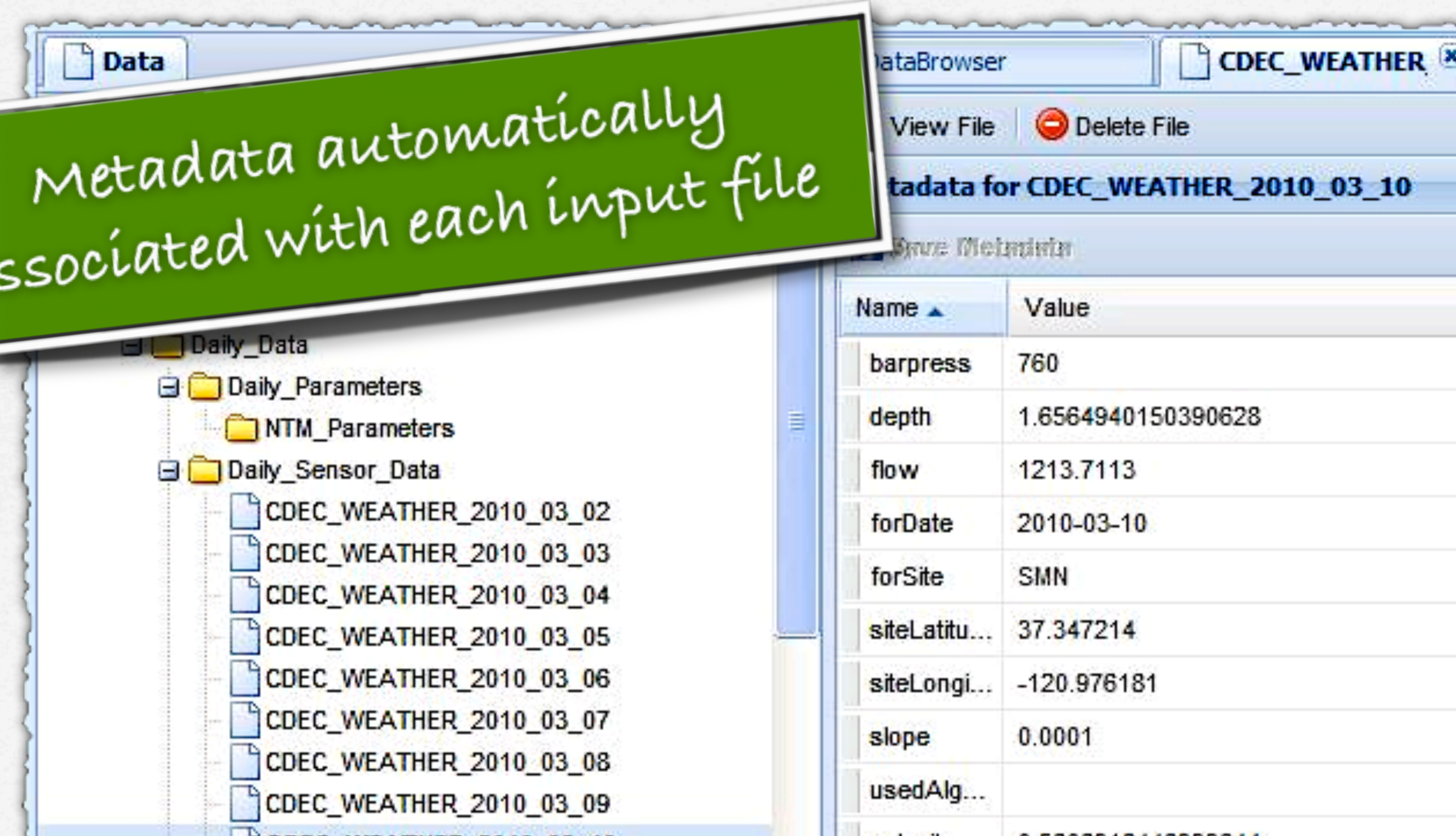


onceptual

WINGS

Workflow

VINGS Received Metadata from KARM



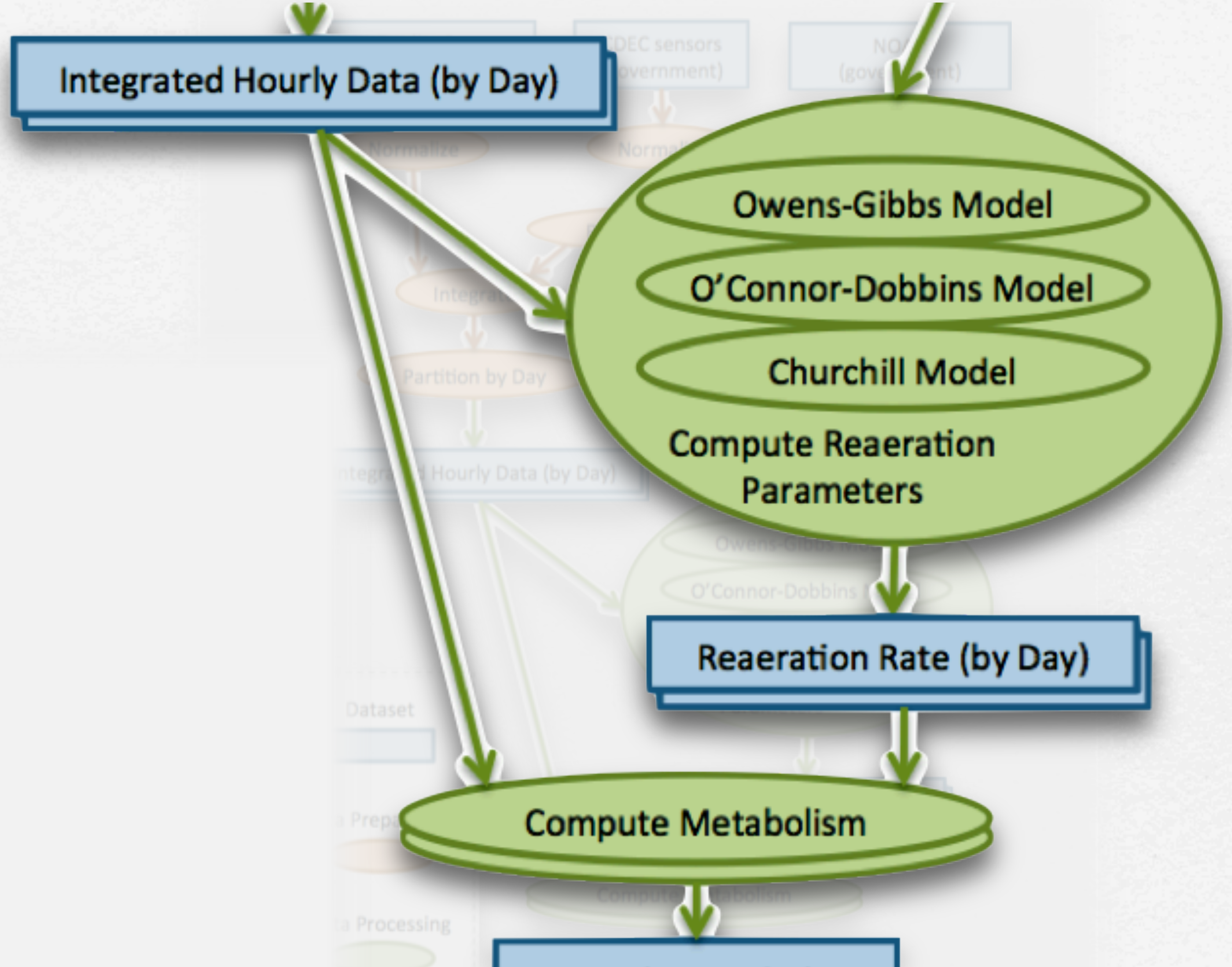
Metadata automatically
ssociated with each input file

tadata for CDEC_WEATHER_2010_03_10

- 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

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...	

Workflow



Integrated Hourly Data (by Day)

Owens-Gibbs Model

O'Connor-Dobbins Model

Churchill Model

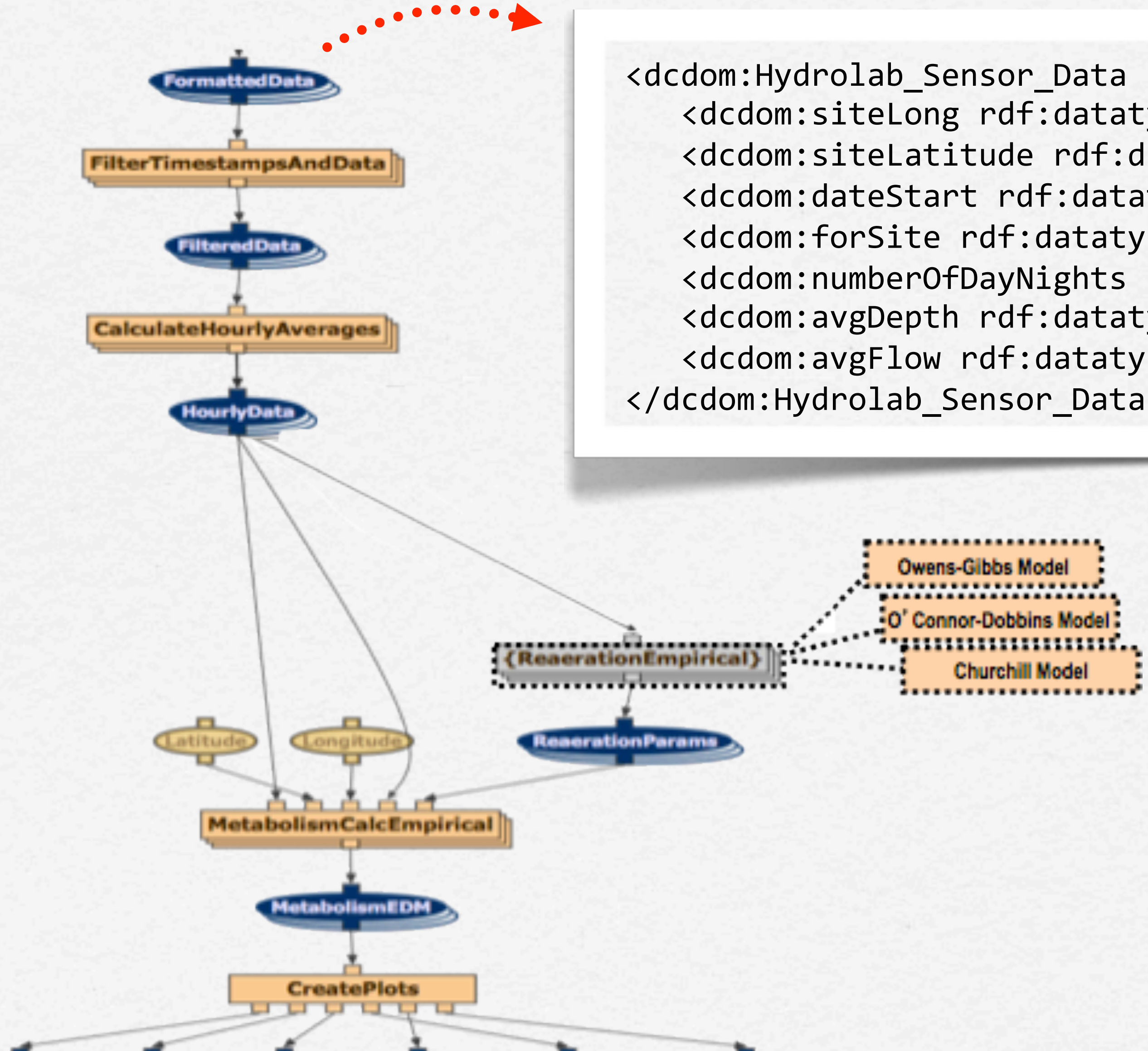
Compute Reaeration Parameters

Reaeration Rate (by Day)

Compute Metabolism

sing Metadata in Workflow Executio

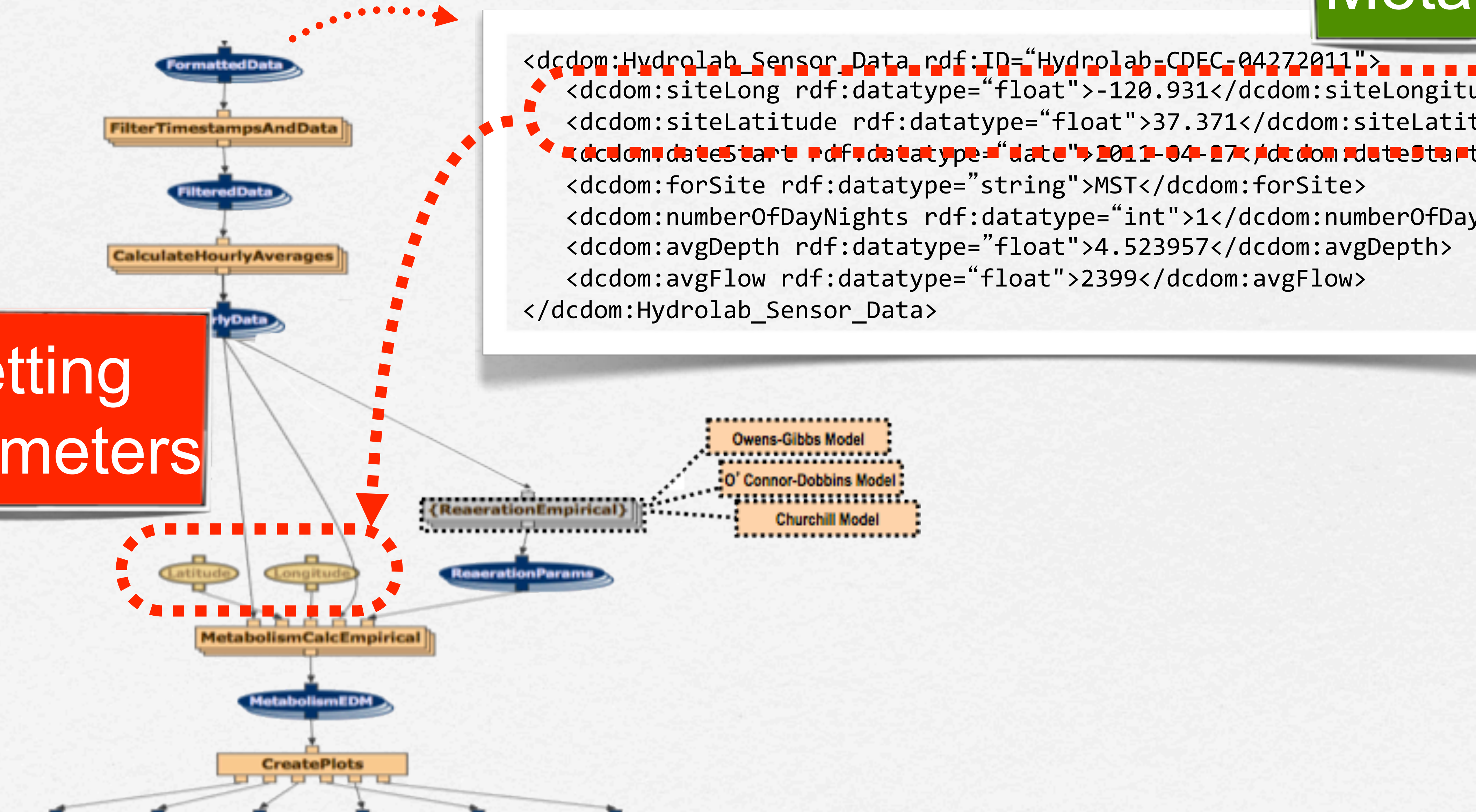
Meta



```
<dcdom:Hydrolab_Sensor_Data rdf:ID=“Hydrolab-CDEC-04272011">
   <dcdom:siteLong rdf:datatype=“float">-120.931</dcdom:siteLongitu
   <dcdom:siteLatitude rdf:datatype=“float">37.371</dcdom:siteLatit
   <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:numberOfDay
   <dcdom:avgDepth rdf:datatype=”float">4.523957</dcdom:avgDepth>
   <dcdom:avgFlow rdf:datatype=“float">2399</dcdom:avgFlow>
</dcdom:Hydrolab_Sensor_Data>
```

sing Metadata in Workflow Executio

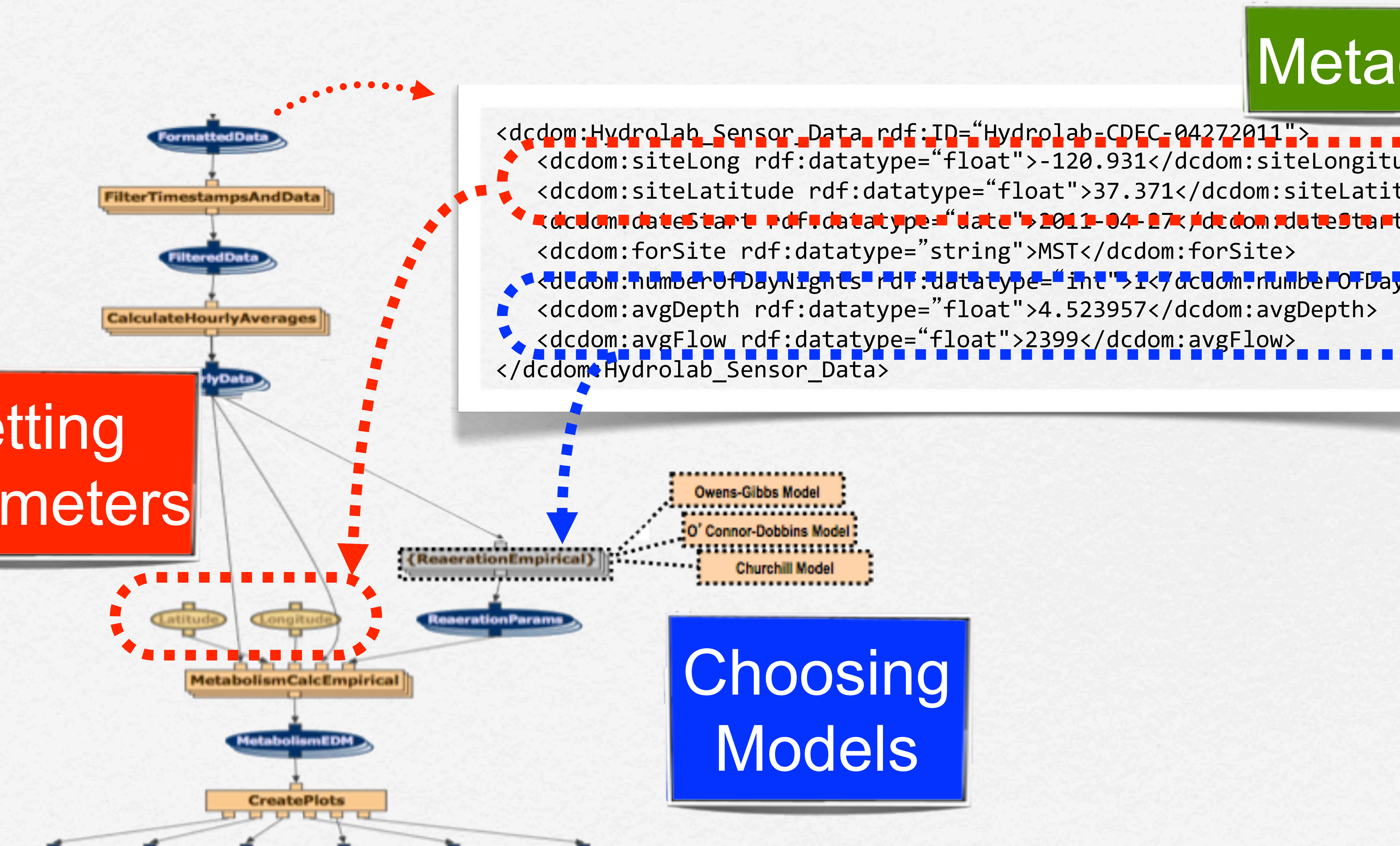
Meta



```
<dcdom:Hydrolab_Sensor_Data rdf:ID=“Hydrolab-CDEC-04272011">
   <dcdom:siteLong rdf:datatype=“float">-120.931</dcdom:siteLongitu
   <dcdom:siteLatitude rdf:datatype=“float">37.371</dcdom:siteLatit
   <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:numberOfDay
   <dcdom:avgDepth rdf:datatype=”float">4.523957</dcdom:avgDepth>
   <dcdom:avgFlow rdf:datatype=“float">2399</dcdom:avgFlow>
</dcdom:Hydrolab_Sensor_Data>
```

tting
meters

sing Metadata in Workflow Executio



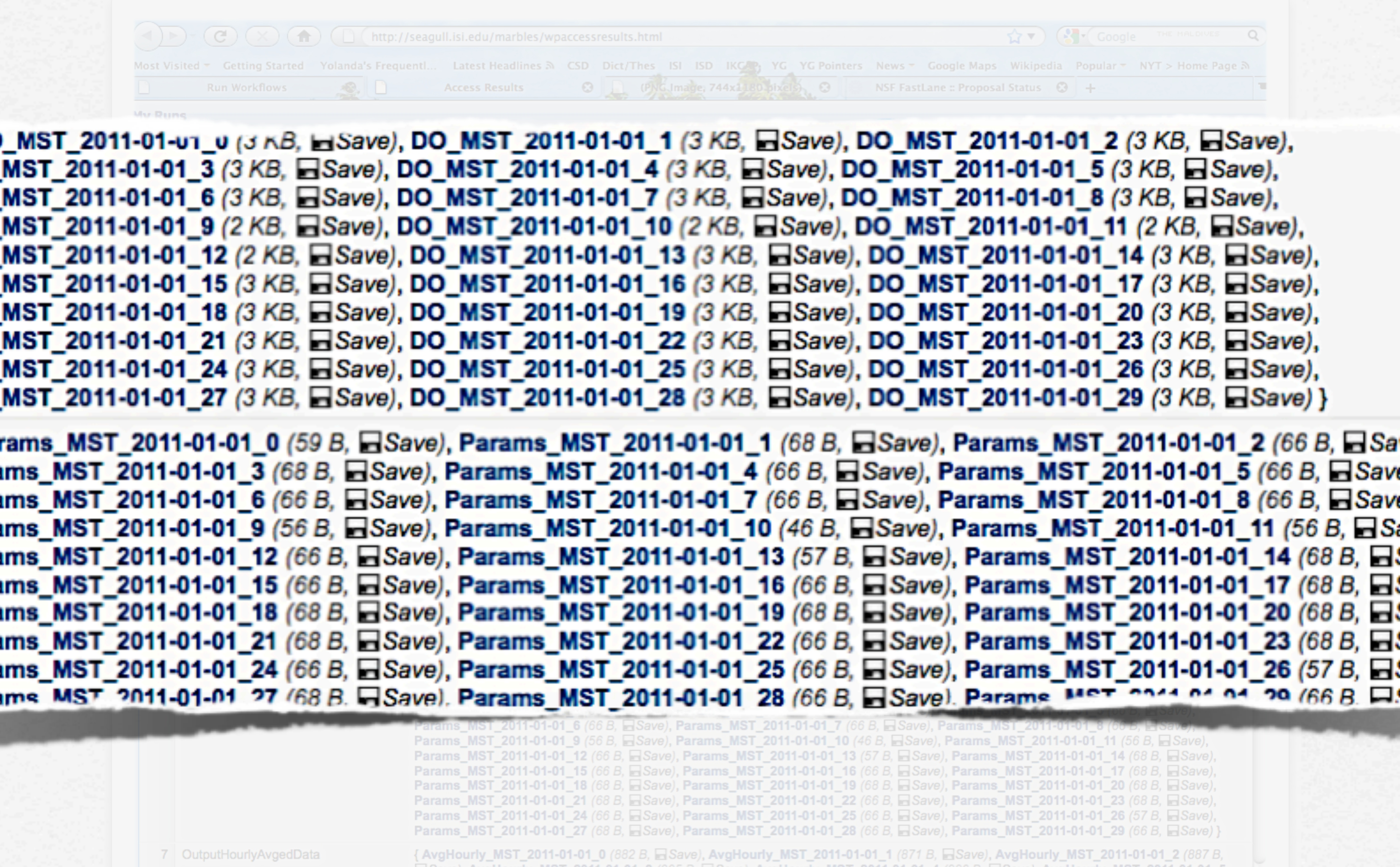
Meta

```
<dcdom:Hydrolab_Sensor_Data rdf:ID=“Hydrolab-CDEC-04272011">
   <dcdom:siteLong rdf:datatype=“float">-120.931</dcdom:siteLongitu
   <dcdom:siteLatitude rdf:datatype=“float">37.371</dcdom:siteLatit
   <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:numberOfDay
   <dcdom:avgDepth rdf:datatype=”float">4.523957</dcdom:avgDepth>
   <dcdom:avgFlow rdf:datatype=“float">2399</dcdom:avgFlow>
</dcdom:Hydrolab_Sensor_Data>
```

tting
meters

Choosing
Models

Workflow Results



_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),
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),
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),
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),
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),
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),
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),
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),
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),
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) }

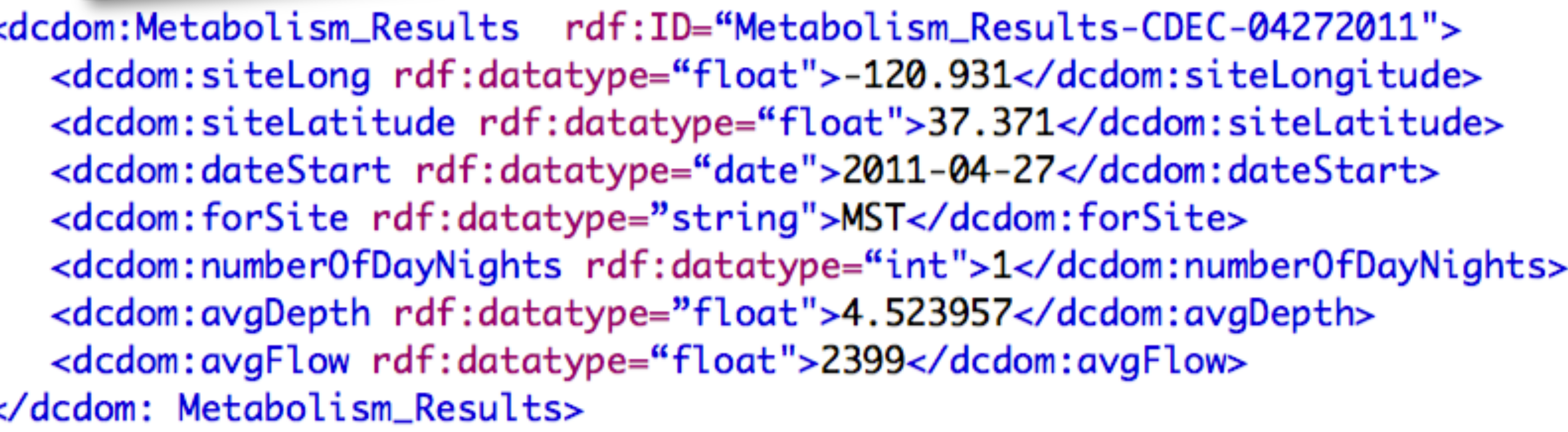
rams_MST_2011-01-01_0 (59 B, Save), Params_MST_2011-01-01_1 (68 B, Save), Params_MST_2011-01-01_2 (66 B, Sa
ams_MST_2011-01-01_3 (68 B, Save), Params_MST_2011-01-01_4 (66 B, Save), Params_MST_2011-01-01_5 (66 B, Sav
ams_MST_2011-01-01_6 (66 B, Save), Params_MST_2011-01-01_7 (66 B, Save), Params_MST_2011-01-01_8 (66 B, Sav
ams_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, S
ams_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,
ams_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,
ams_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,
ams_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,
ams_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,
ams_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,

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,

Workflow Results Have Metadata

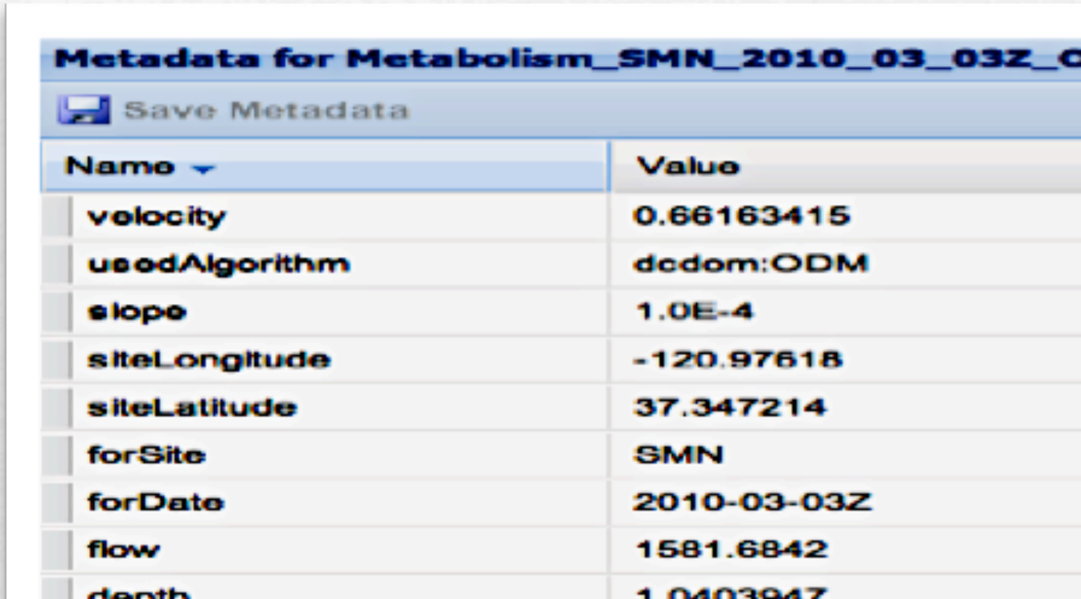
WINGS automatically generates metadata for each output file



```
<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>
```

VINGS Generates Provenance Metadat

```
LECT ?url WHERE {
ata dcdom:usedAlgorithm dcdom:ODM .
ata rdf:type dcdom:Metabolism_Estimates .
ata wflow:hasLocation ?url
```



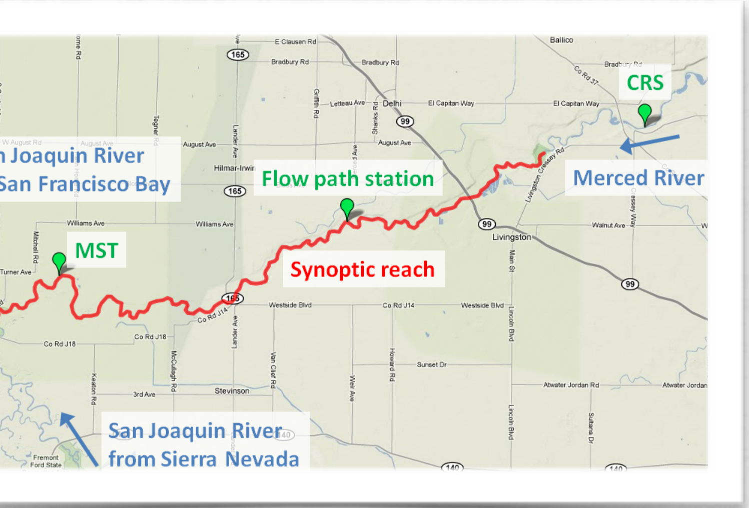
Metadata for Metabolism_SMN_2010_03_03Z_C

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

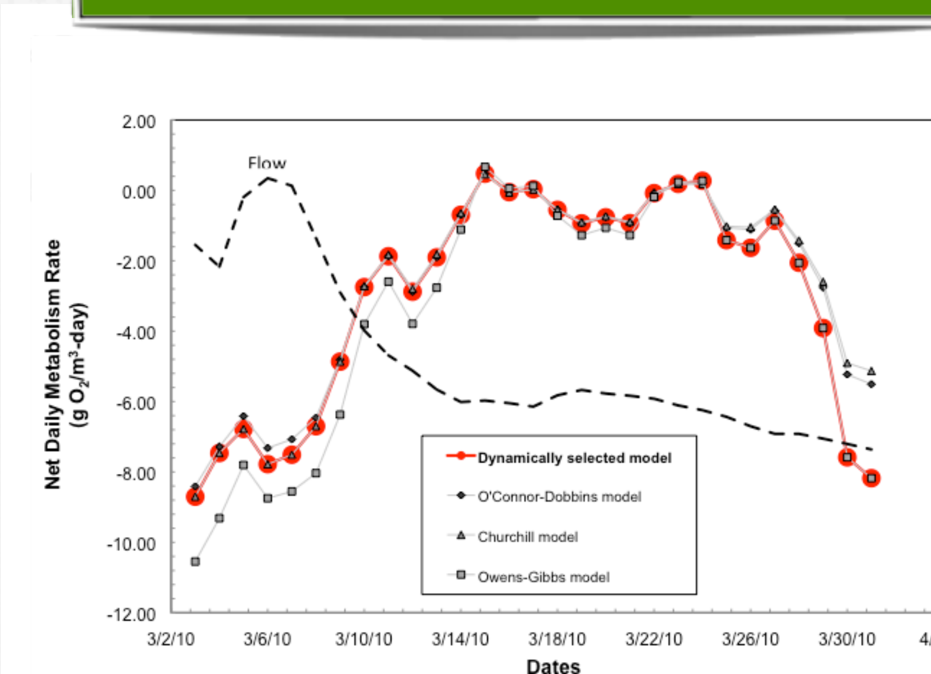
Aquatic Photosynthesis

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

ensors



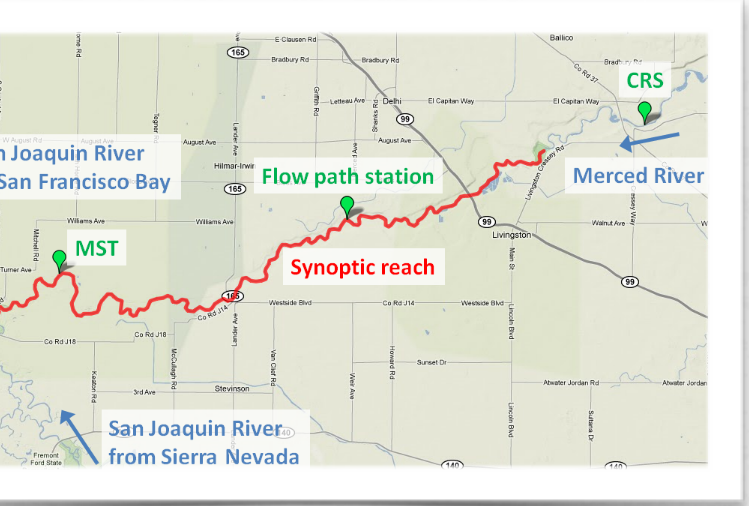
Workflow Results



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

THANK
YOU!