This presentation premiered at WaterSmart Innovations

watersmartinnovations.com

Dealing with Consumption Data Outliers During Conservation Planning

John M. Clayton, Jack Kiefer, and Lisa Krentz – Hazen and Sawyer
Dave Bracciano, Nisai Wanakule, and Tirusew Asefa – Tampa Bay Water

WaterSmart Innovations Conference and Exposition
October 4th, 2017
South Point Hotel and Conference Center, Las Vegas, NV

Agency Background

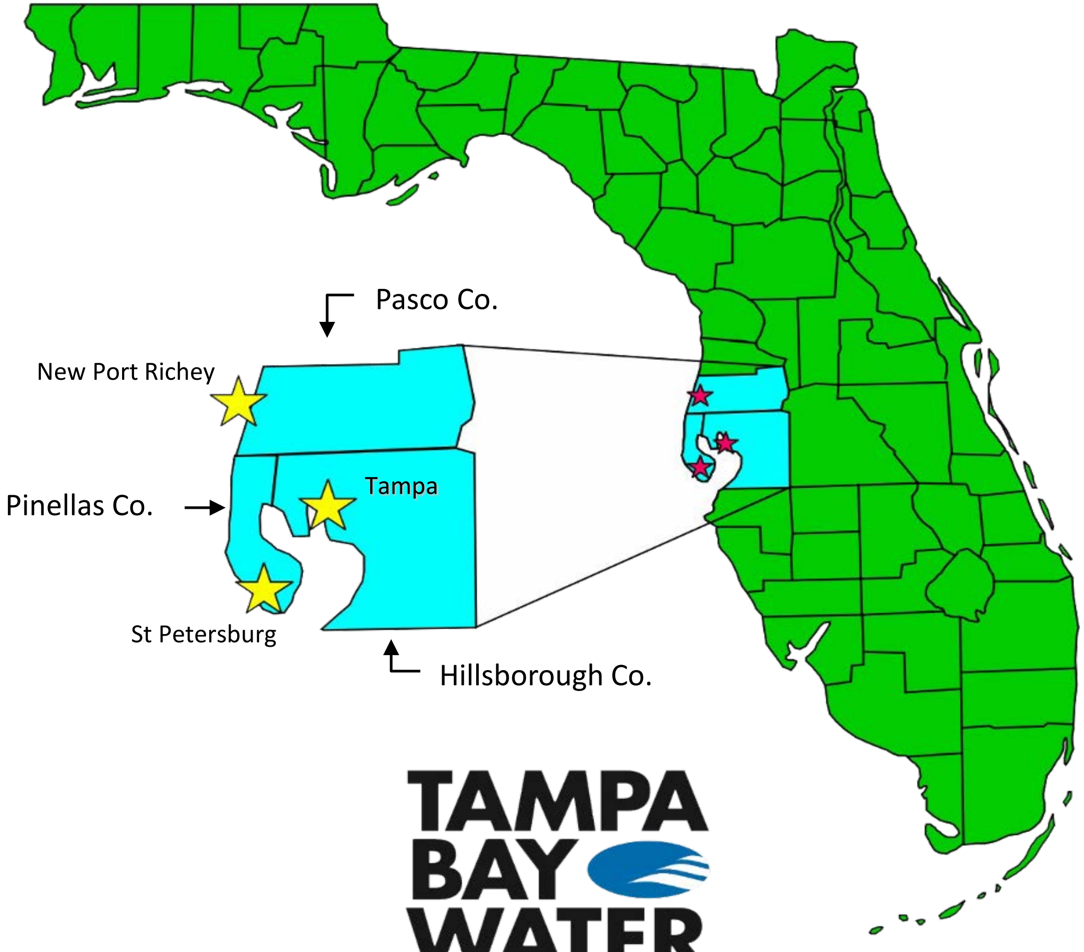
Regional water supply authority serving over 2.4 million customers

Six member governments, across three counties

Member demands:

2015: 227 MGD

2035: 281 MGD (baseline)

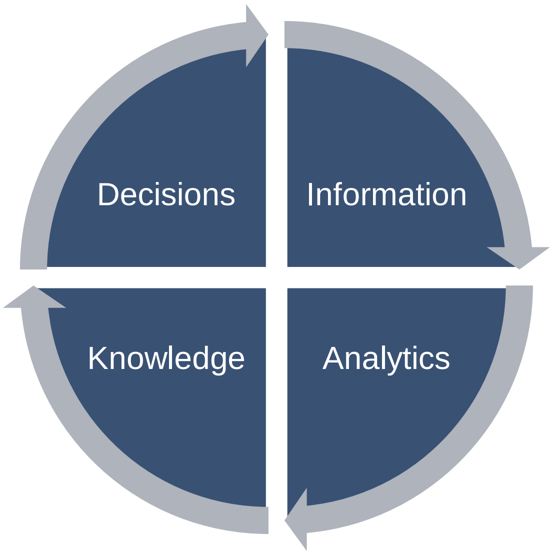


TAMPA BAY WATER

Supplying Water To The Region

Long-Term Demand Forecasting System (LTDFS)

- LTDFS designed to:
  - Track water consumption, socioeconomic, economic and policy conditions
  - Provide inputs for demand forecasting models (updated periodically)
  - Prepare forecasts through implementation of models (annually)
  - Inform regional and member specific demand management efforts
  - Support water supply reliability ("just-in-time" supply development) efforts



Database objectives

Extensive LTDFS database effort to:

1. Provide water use data and property characteristics for all individual customers (locations) with ability to aggregate to larger geographies
2. Ensure acquired information can be maintained through time to support future evaluations
3. Standardize design so queries and analytical routines can be replicated and updated efficiently through time

Information developed for each location

- Water use class
  - Retail/billing
  - Property use code
- Historical sales of potable water
  - Monthly (1998-2016)
  - Domestic meter(s)
  - Irrigation meter(s)
- Access to reclaimed water
- Property characteristics
  - Dwelling units (residential)
  - Year built
  - Lot size/area
  - Building/heated area
  - Other

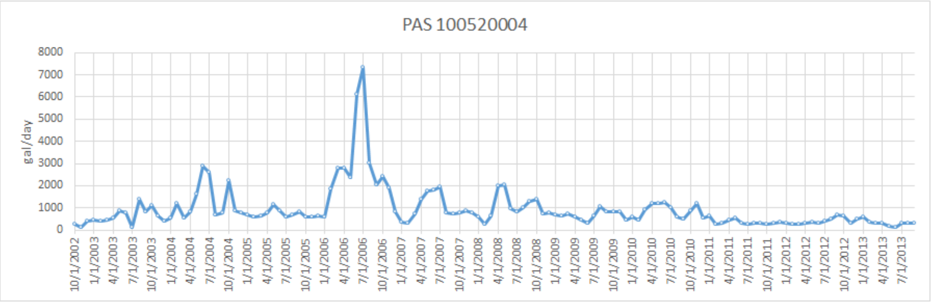
Locations and Small-Scale Geographies = More Noise in Consumption Data!

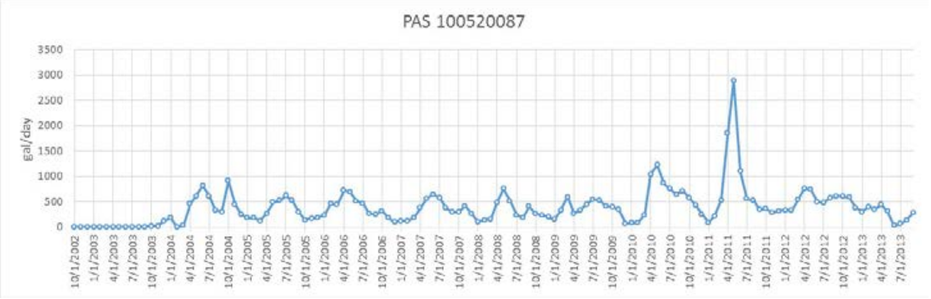
Hundreds of thousands of locations, tens of millions of monthly consumption points

Outliers can be anywhere

Potential to obfuscate or bias small-scale analyses

Can we manually spot and correct/flag them all?





Locations and Small-Scale Geographies = More Noise in Consumption Data!

Hundreds of thousands of locations, tens of millions of monthly consumption points

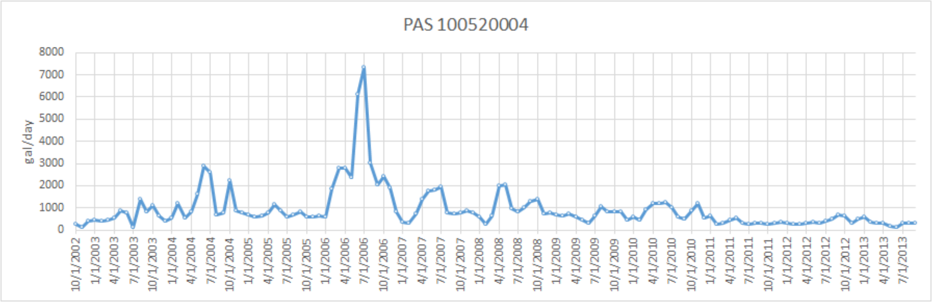
Outliers can be anywhere

Potential to obfuscate or bias small-scale analyses

Can we manually spot and correct/flag them all?

NO WAY!

We need automated screening procedures



SF Consumption in Tampa Bay

What is typical and what is not?

Single-family non-irrigator in Tampa area: 100-200 gal/day (gpd) average across a month

One irrigation cycle might dispense 2500 gallons

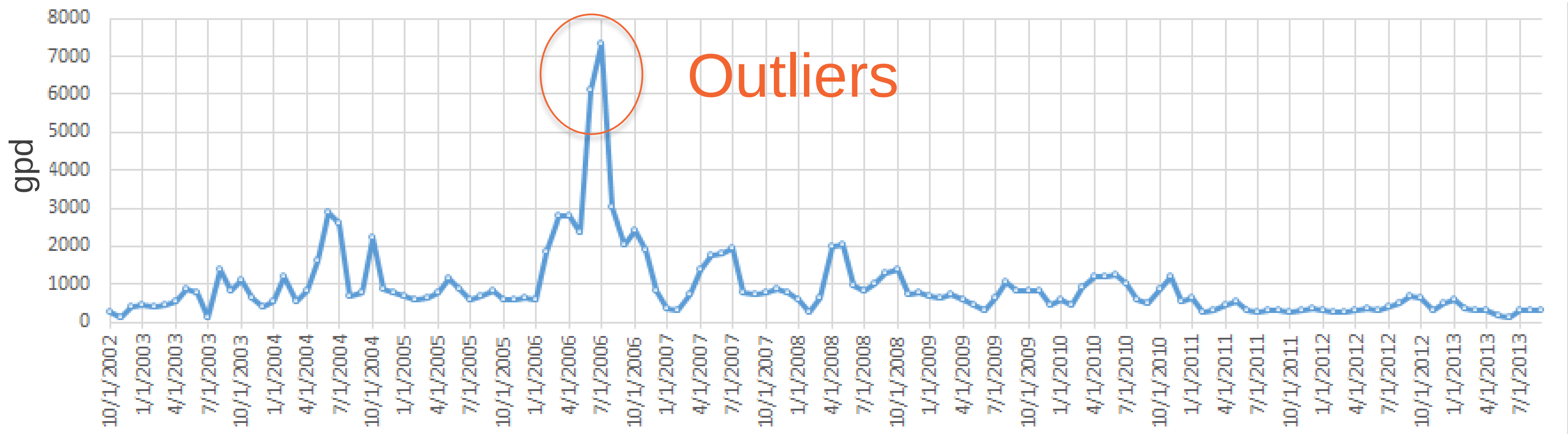
- 1 irrigation/week: 450-550 gpud in a month
- 2 irrigations/week: 750-850 gpud in a month
- 3 irrigations/week: 1100-1200 gpud in a month
- Daily irrigation: 2600-2700 gpud in a month

What Is An Outlier?

Physically speaking…

SF HH consumption becomes more physically unreasonable as it increases beyond about 2000 gpud

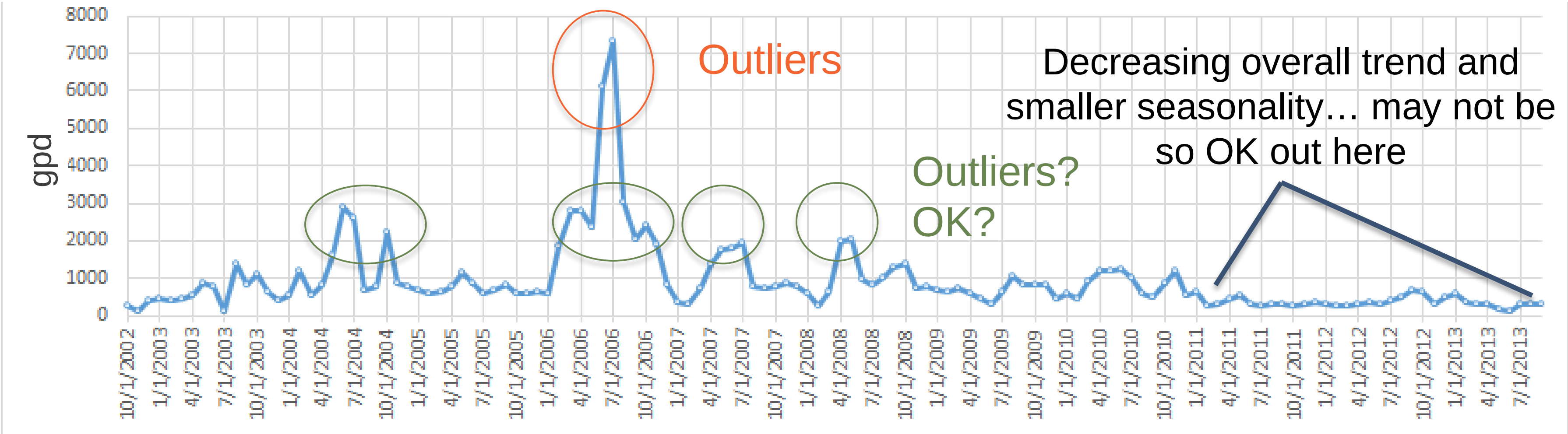
Leaks? Billing corrections/irregularities not related to actual use?



What Is An Outlier?

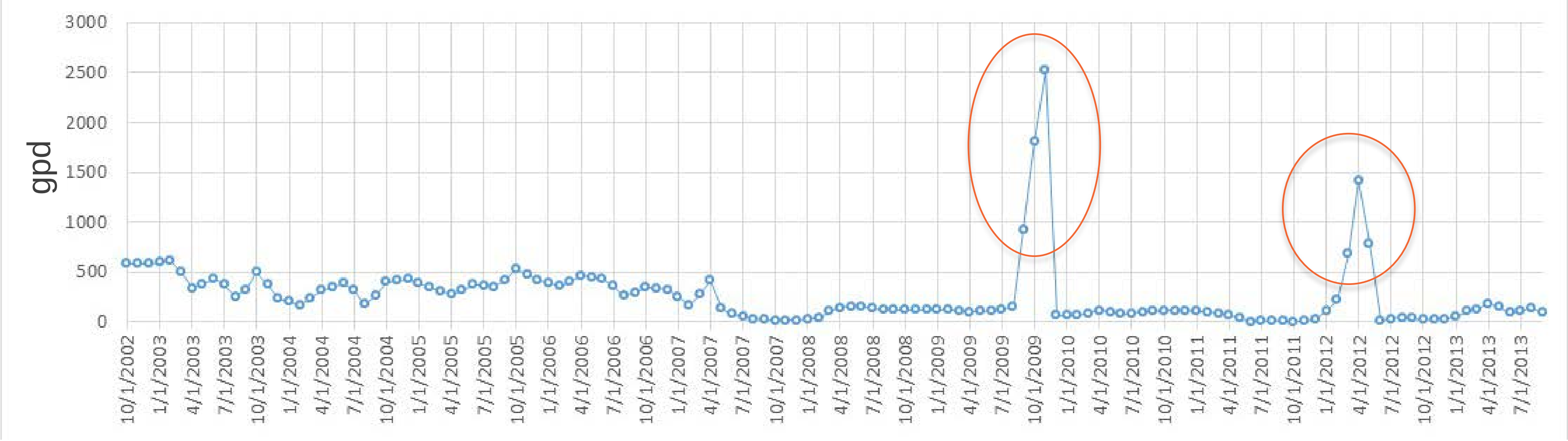
Also depends on how individual records relate to overall trend, seasonality at each SF household

- Both can change over time (changing customers at same household, changing fixtures and efficiency)



What Is An Outlier?

Outliers can be physically reasonable but way out of character for a given household

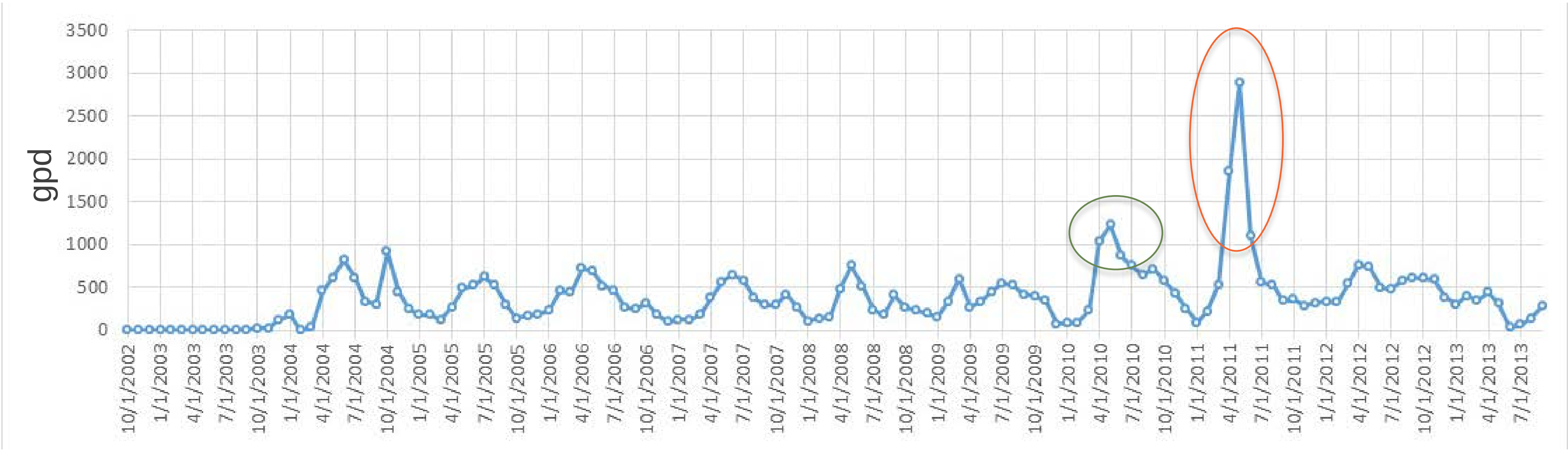


What Is An Outlier?

Sometimes outlier status is not obvious

Somewhat out of character, but these are in Spring (hot/dry) season

Also, previous year had high Spring consumption



Several Common Screening Methods

- Global gpd threshold
  - One threshold does not fit all
- Individual gpd thresholds (e.g. top n% for each household)
  - Not all households really have outliers
  - Strong seasonality and changing patterns over time – could discard real and critical data for our analyses
- Neither approach has literature-based statistical guidance on outlier detection

New Screening Method for Tampa Bay Water

1) Bulk-screen monthly SF consumption records

→ Peak gpd > some physically-based threshold

2) Detrend and deseasonalize monthly gpd series for each household

→ Provides series of normalized residuals

3) Analyze residual time series to detect outliers

→ Data points that stand out in their own time environment, even after accounting for trend and seasonality

→ Statistical method for normalized data: Cook's D

1) SF Monthly Consumption Screening

Many Options

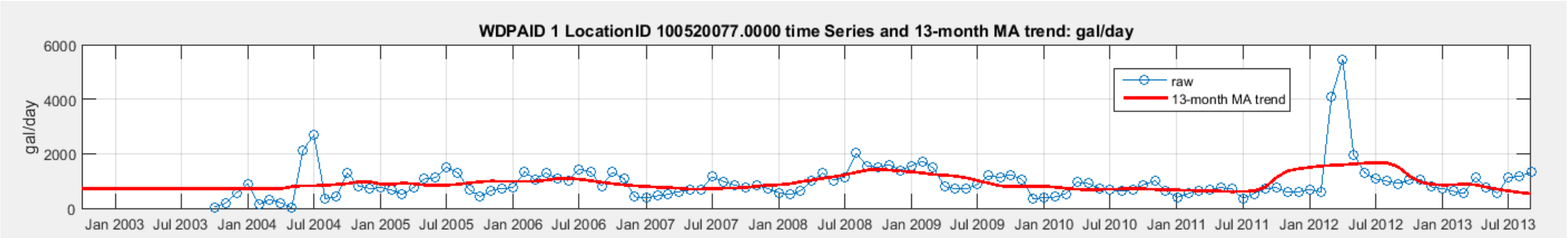
peak gpd threshold	Total Households	% of all Households	total Household/ months	% of all Household/ months
2000	30893	6.3%	3476231	5.9%
2500	17936	3.6%	2003792	3.4%
3000	11143	2.3%	1237017	2.1%
4000	5171	1.0%	571162	1.0%
5000	2819	0.6%	310777	0.5%

total SF Households	492823
total SF Household/months	59173132

2a) Detrend Each Household's Gpd

Calculate trend

- 13-month centered weighted moving average of gpd (1/24 on months 1 and 13, 1/12 on months 2-12)



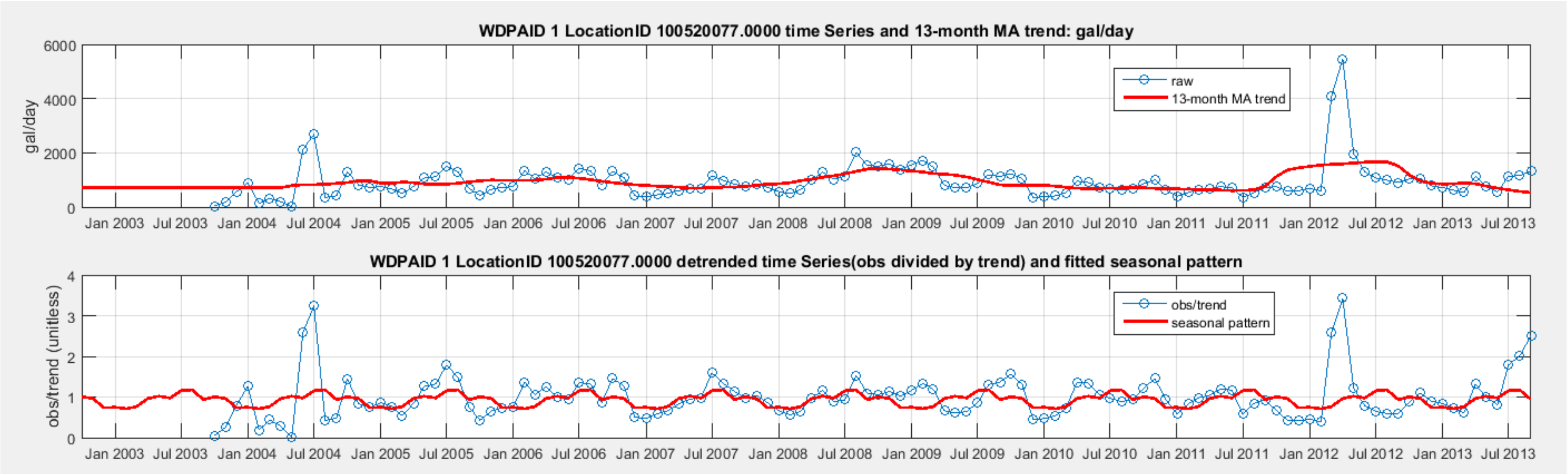
Hazen

2a) Detrend Each Household's Gpd

Calculate trend

13-month centered weighted moving average of gpd (1/24 on months 1 and 13, 1/12 on months 2-12)

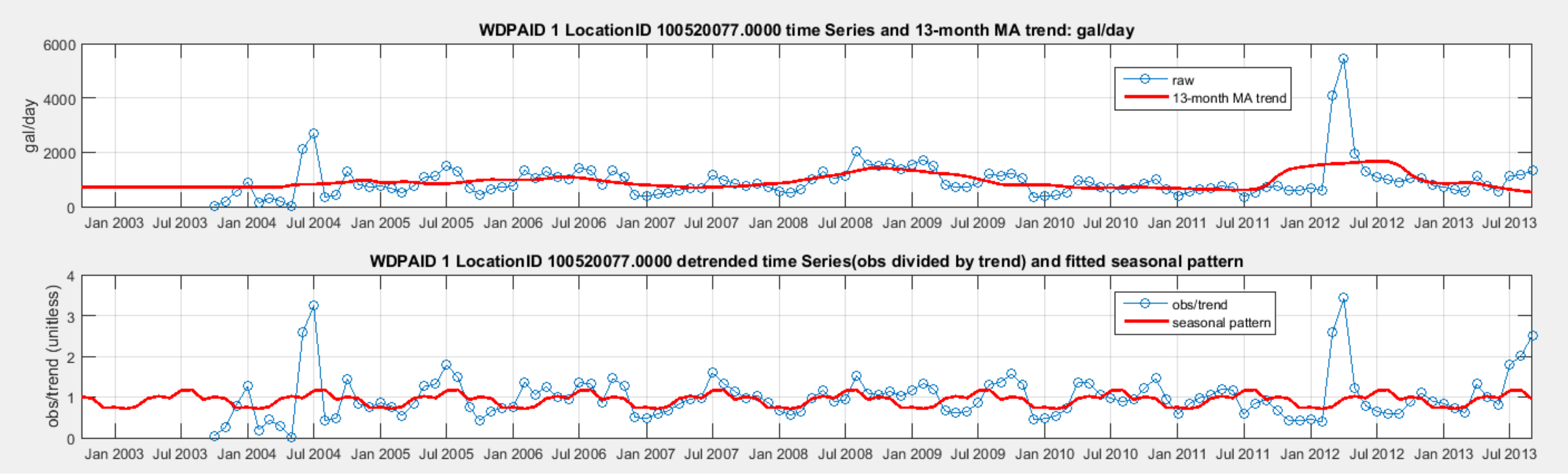
Detrended = gpd / trend



2b) Deseasonalize Each Household's Gpd

Calculate seasonal pattern

Regress detrended series using monthly fixed effects

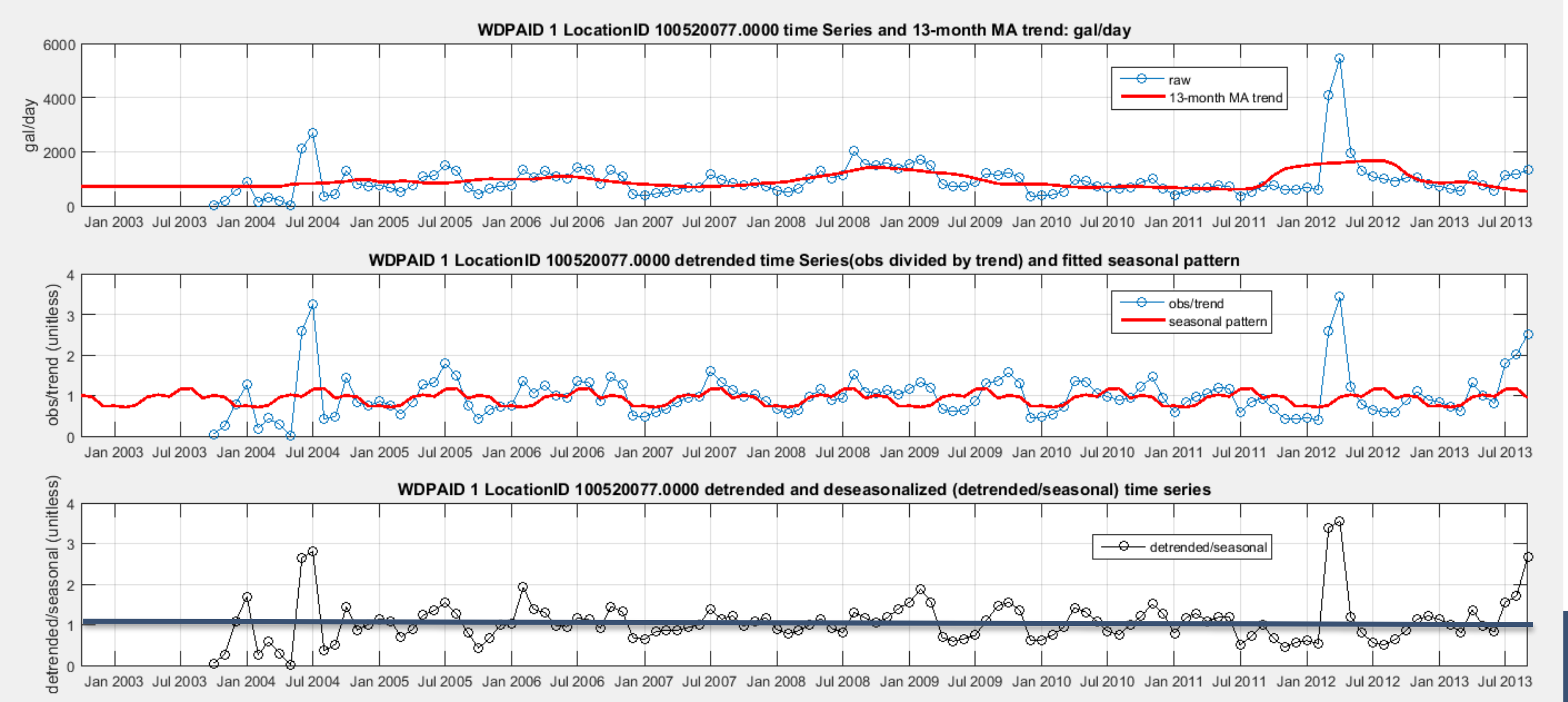


2b) Deseasonalize Each Household's Gpd

Calculate seasonal pattern

- Regress detrended series using monthly fixed effects

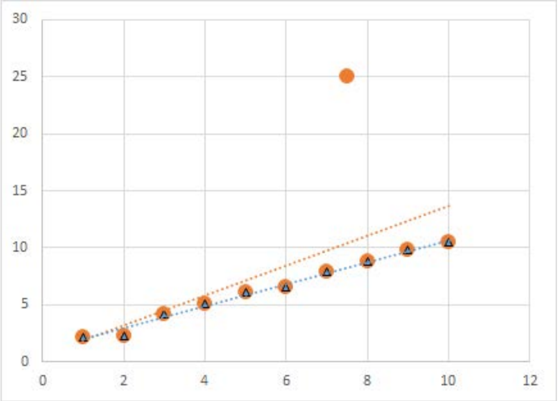
Residual = detrended / seasonal



3) Analyze residual time series to detect outliers

Linear regression method to detect outliers: Cook's D statistic

- Common output from regression packages
- A statistical value for each data point produced by fitting
- Indicates how much each data point influences regression coefficients
- Common guidance: if Cook's D > 4/n then point is outlier

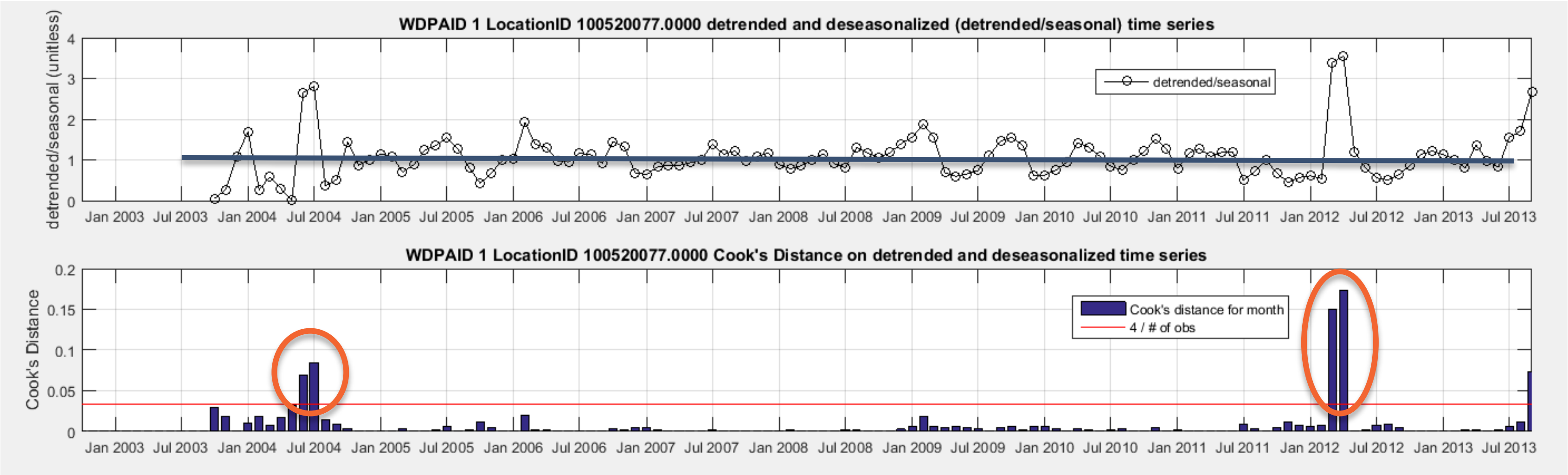


$$D_i = \frac{(y_i - \hat{y}_i)^2}{(p+1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right]$$

D_i = Cook's distance measure for observation i
$y_i - \hat{y}_i$ = the residual for observation i
h_i = the leverage for observation i
p = the number of independent variables
s = the standard error of the estimate

3) Analyze residual time series to detect outliers

Fit a regression: gpd = constant



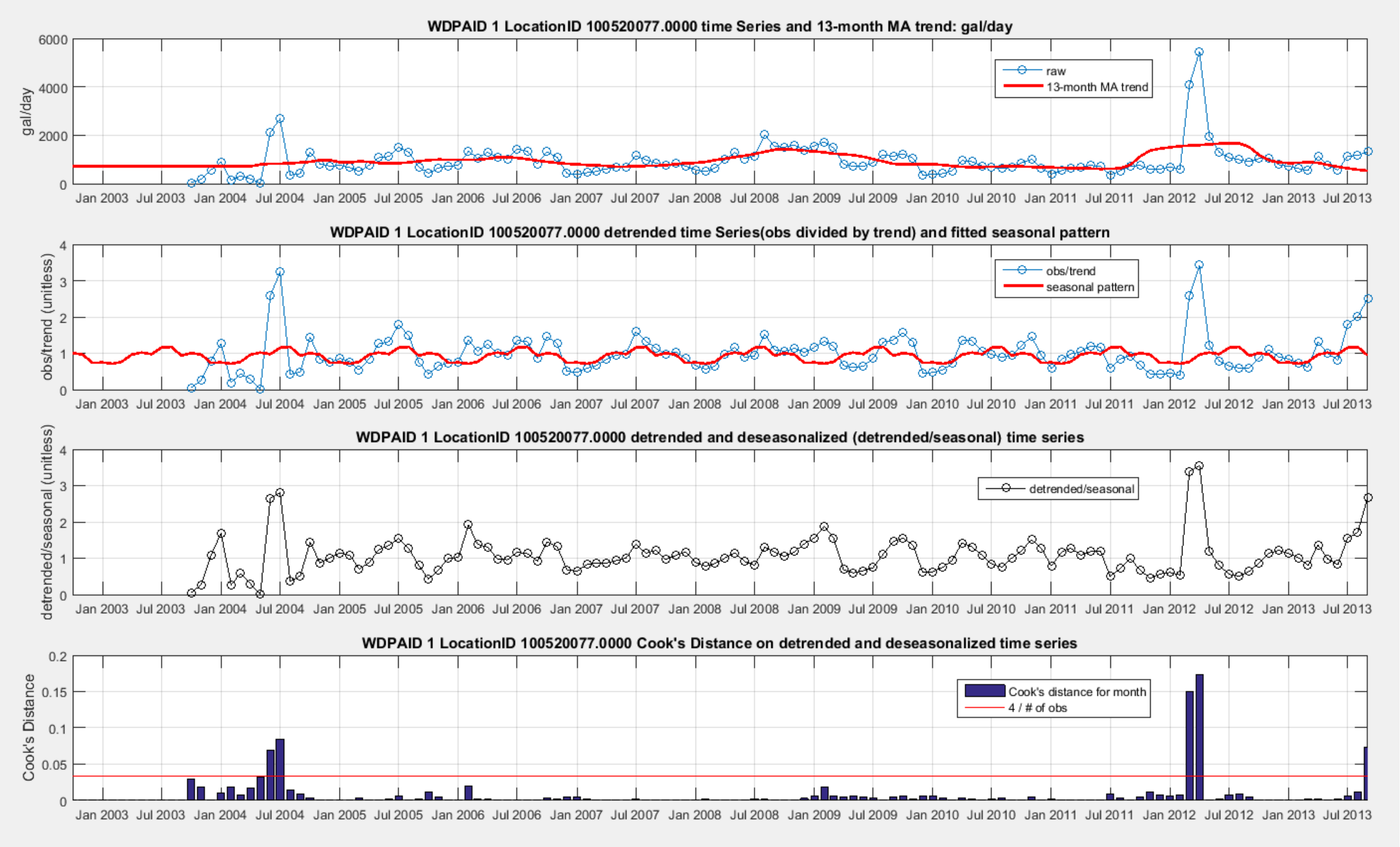
SF Screening Method w/ Cook's Distance

Many options now…

- Gpd > Fixed threshold (2000, 2500, 3000…)
- Cook's D > threshold (4/# obs, 8/# obs…)
- Gpd > threshold AND Cook's D > threshold
- Gpd > threshold OR Cook's D > threshold

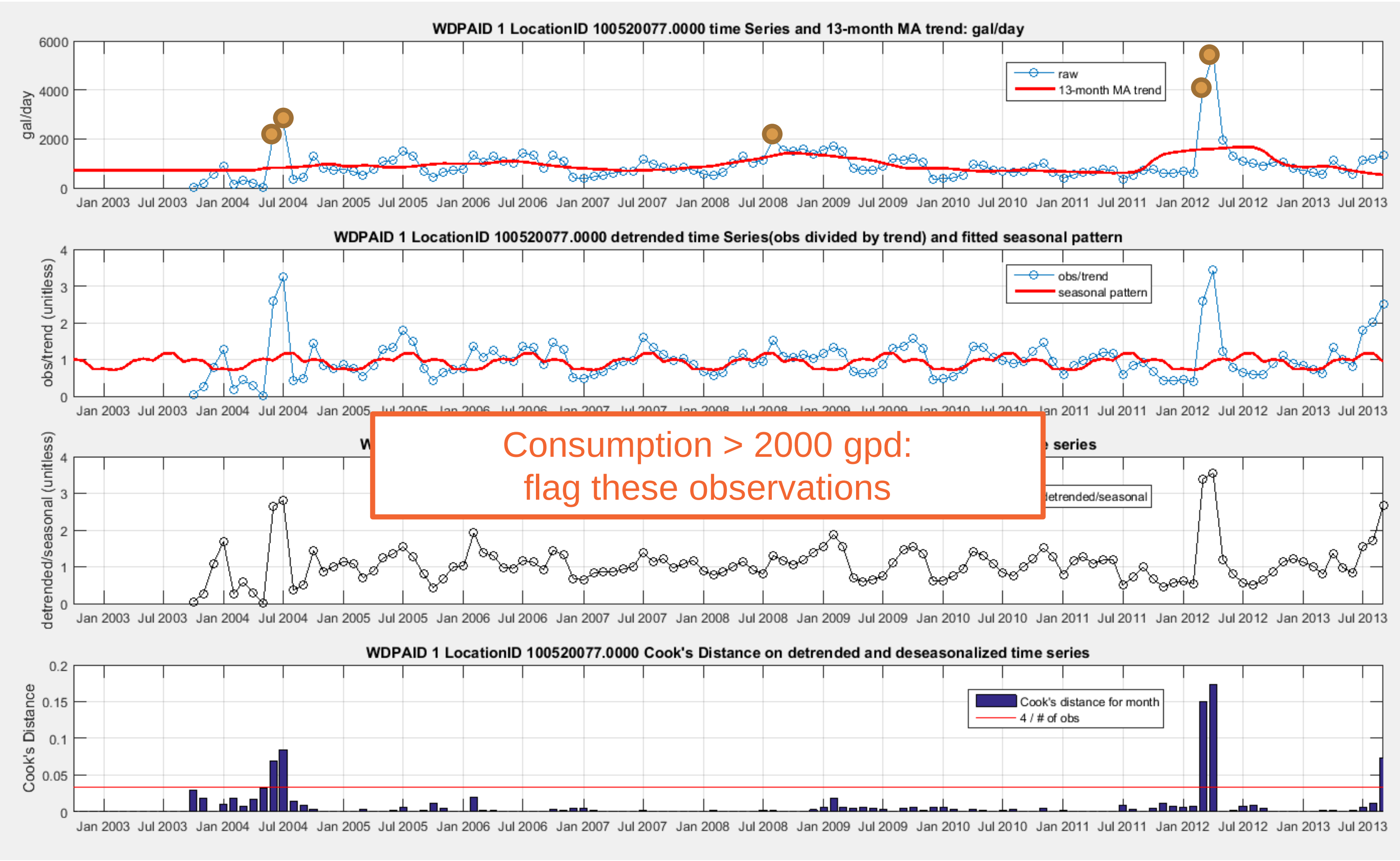
Example

Combining GPD and Cook's D Thresholds



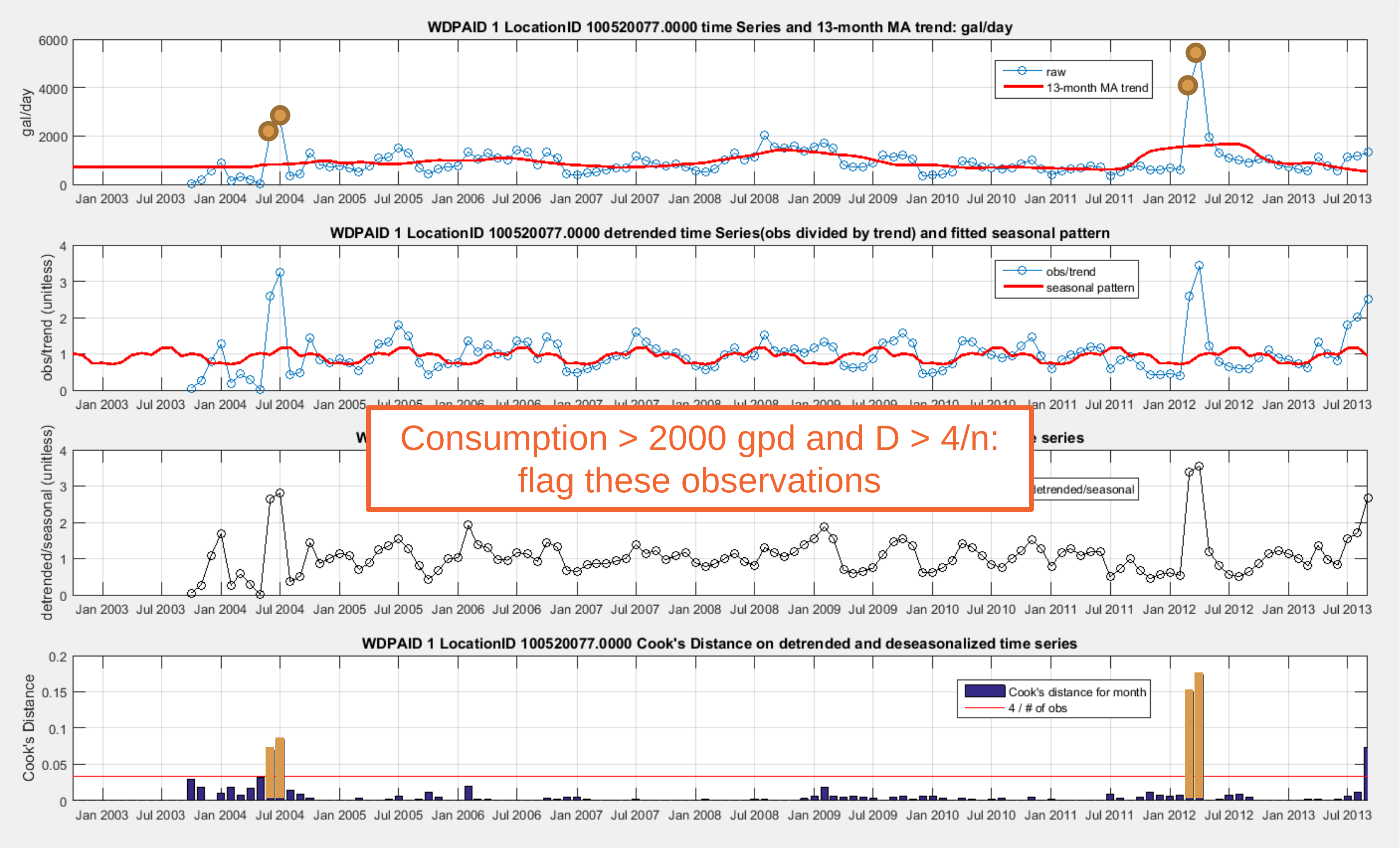
Example

Combining GPD and Cook's D Thresholds



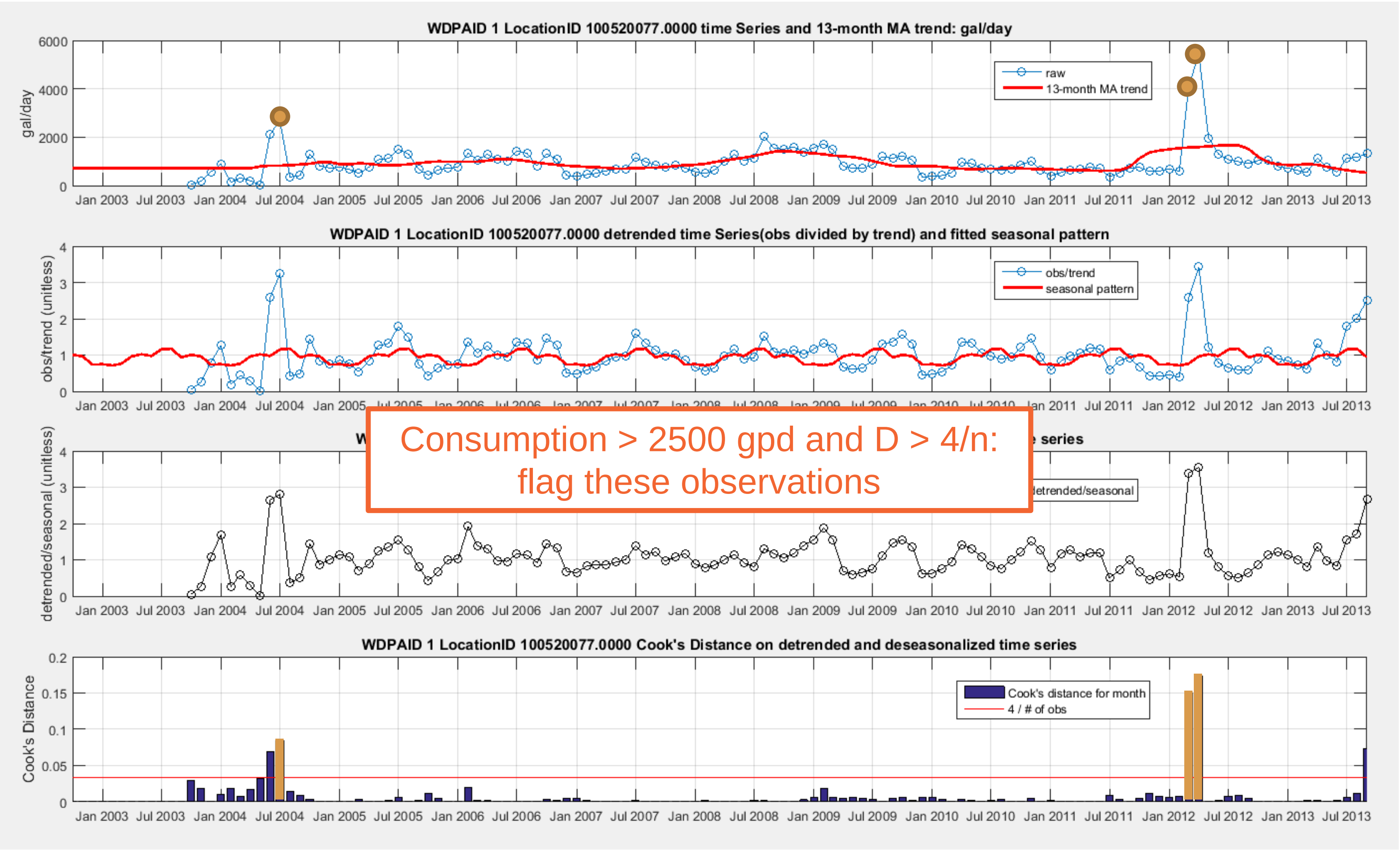
Example

Combining GPD and Cook's D Thresholds



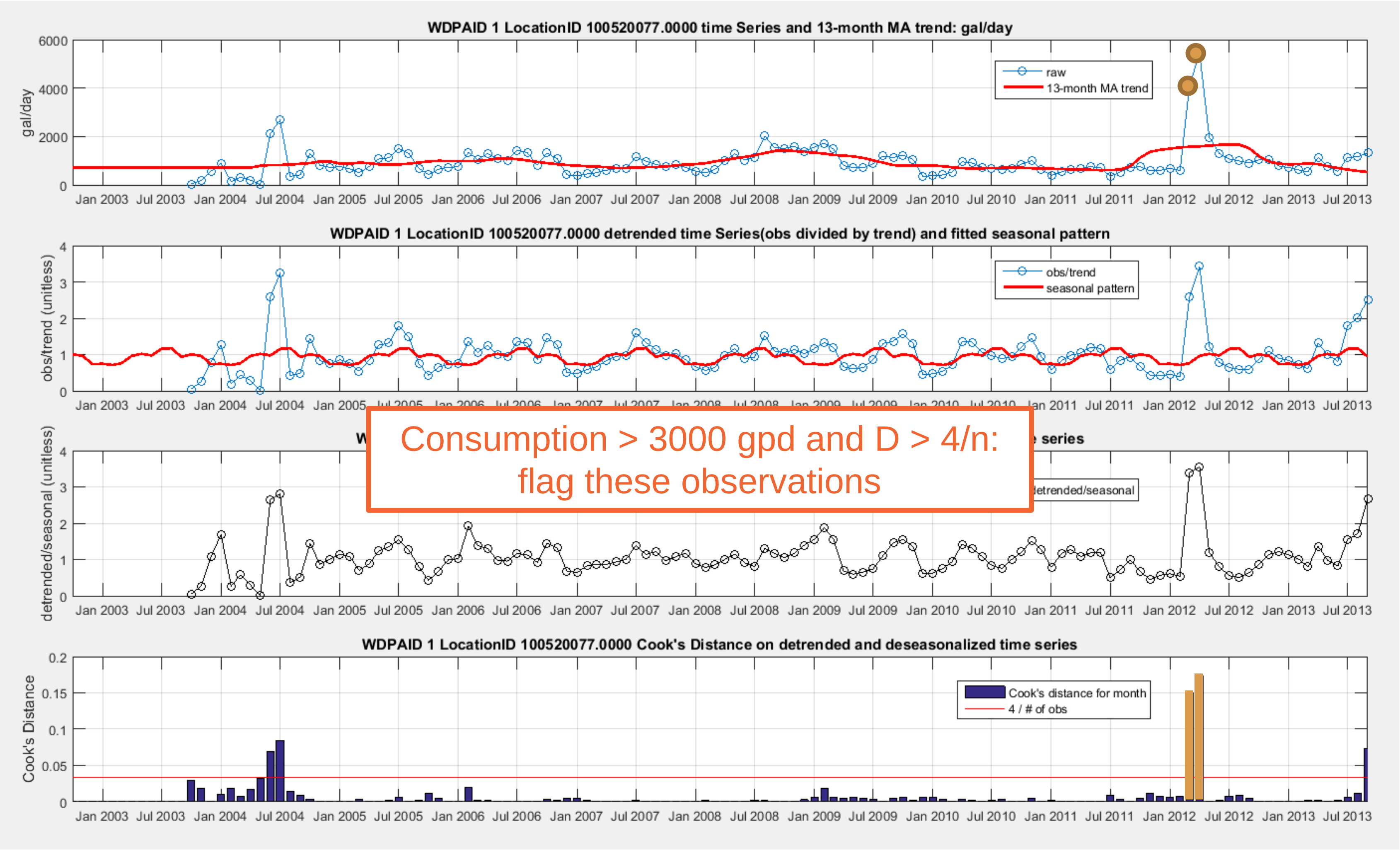
Example

Combining GPD and Cook's D Thresholds



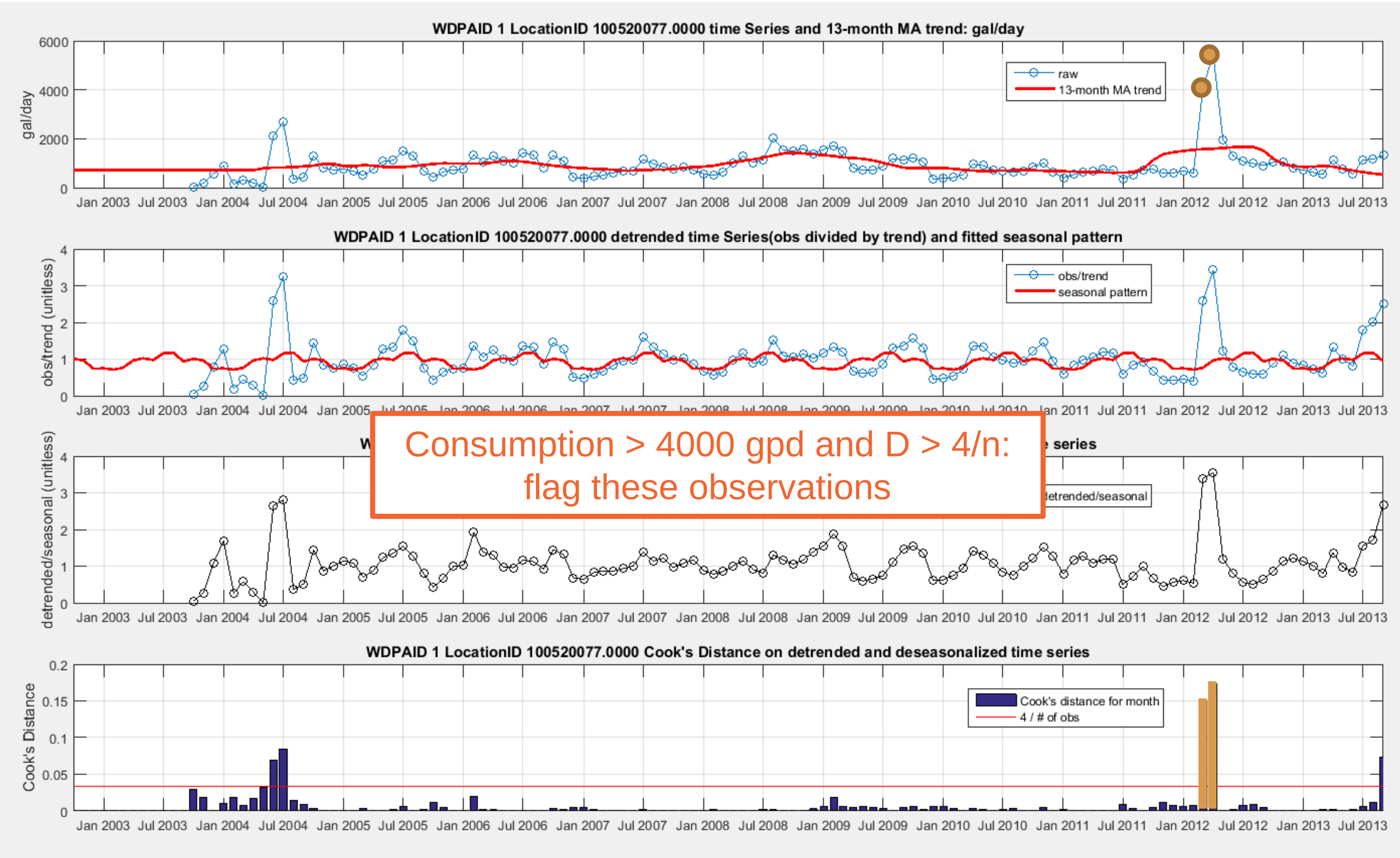
Example

Combining GPD and Cook's D Thresholds



Example

Combining GPD and Cook's D Thresholds



All Methods Require Judgement

Although outlier detection methods are quantitative, still requires qualitative decisions

- Threshold selections
- What do about identified outliers

Conclusion

Detrend + Deseasonalize + Analyze Residuals using Cook's D statistic

Better confidence as an automated method for mass-screening of outliers

- Provides more info for outlier judgement than gpd thresholds: Intel on time-environment of consumption data
- Statistical characterization of departures by established means
- Individual visual assessments still possible
- In absence of visual assessment, analyst still knows there is a rational mechanical basis for identifying the outliers

Thank You!

Cooks4 vs Cooks8

Locations and Observations with Flags

gpd threshold	Qualifying Locations	Cook4: 1+ month > Cook4		Cook4: 1+ month > Cook4 & >gpd thresh	
		Locations	%	Locations	%
2000	32725	32457	99.2%	30314	92.6%
2500	21992	21790	99.1%	20758	94.4%
3000	17927	17758	99.1%	16685	93.1%
4000	15241	15089	99.0%	13588	89.2%
5000	12024	11903	99.0%	10680	88.8%

gpd threshold	Obs in Qualifying Locations	Cook4: months > Cook4		Cook4: months > Cook4 & >gpd thresh	
		Obs	%	Obs	%
2000	3648048	142269	3.9%	55468	1.5%
2500	2438028	93843	3.8%	41582	1.7%
3000	1991471	77491	3.9%	34755	1.7%
4000	1701716	68380	4.0%	27564	1.6%
5000	1335844	53452	4.0%	19350	1.4%

tot SF locations in WYs 2003-2013	555019
tot location/months in WYs 2003-2013	64945036

Frequency of flags per location

2000 gpd threshold