Predicted Results and Implications for Round 2

Michael J Messner, Ph.D. , USEPA-OGWDW
Song Qian, Ph.D., Univ. of Toledo, Dept. of Env. Sciences
Bentley Coffey, Ph.D., The Cadmus Group

November 15, 2012

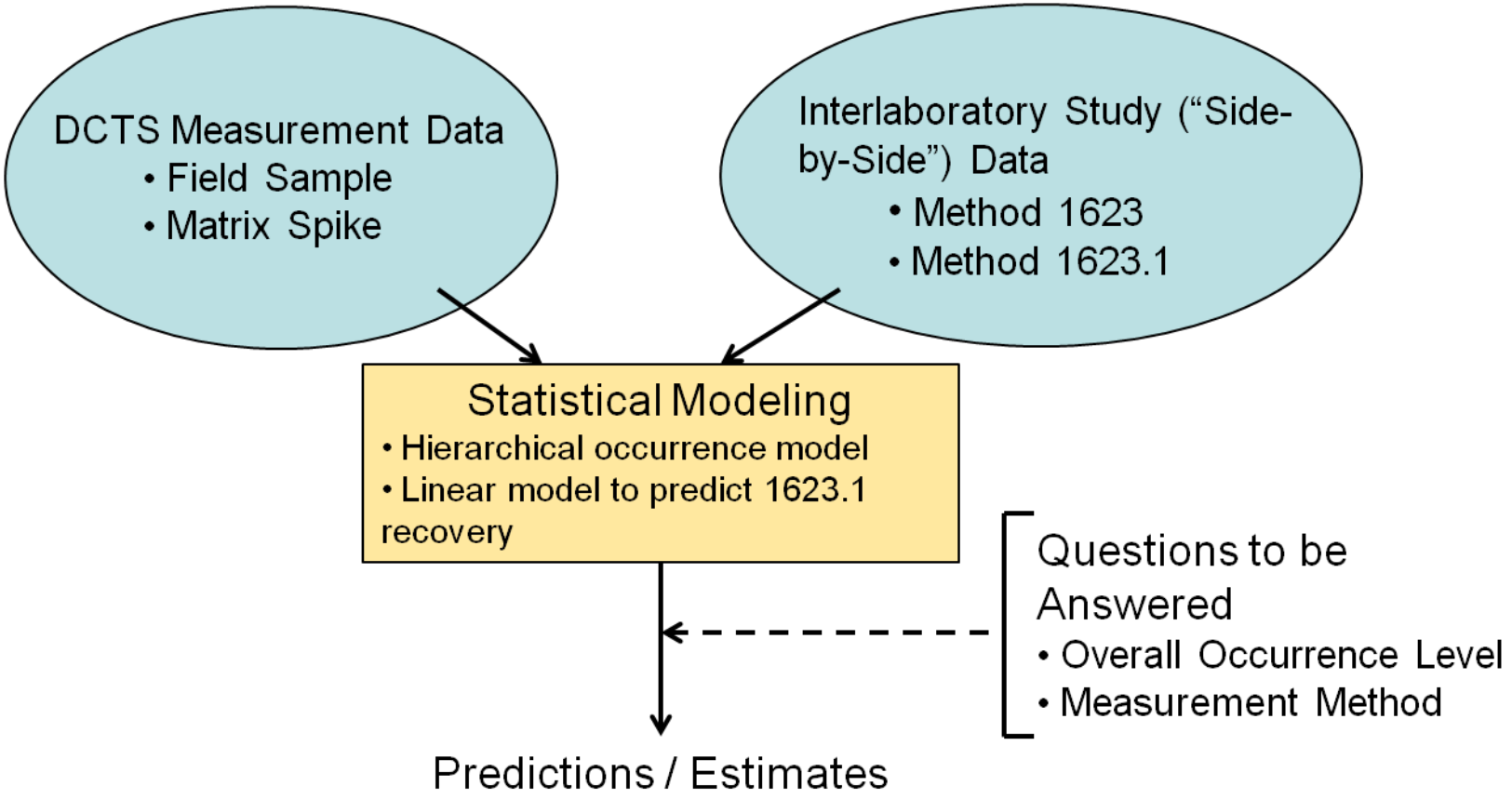
Outline

- Background
  - Questions to be addressed
  - Key Components
    - Data
    - Statistical Model
    - Assumptions
- Results addressing each question
- Summary / Recap

Questions

1. Can the model explain or reproduce the Round 1 data and outcomes?
2. What Round 1 outcomes would have been predicted using Method 1623.1?
3. For facilities placed in bin 1 during Round 1, what Round 2 outcomes are predicted (distribution of facilities across bins 1-4), assuming no change in occurrence levels
   - using Method 1623?
   - using Method 1623.1?
4. For facilities placed in bin 1 during Round 1, what would the outcomes look like if *Cryptosporidium* occurrence were to systematically increase or decrease?
   - Using Method 1623?
   - Using Method 1623.1?

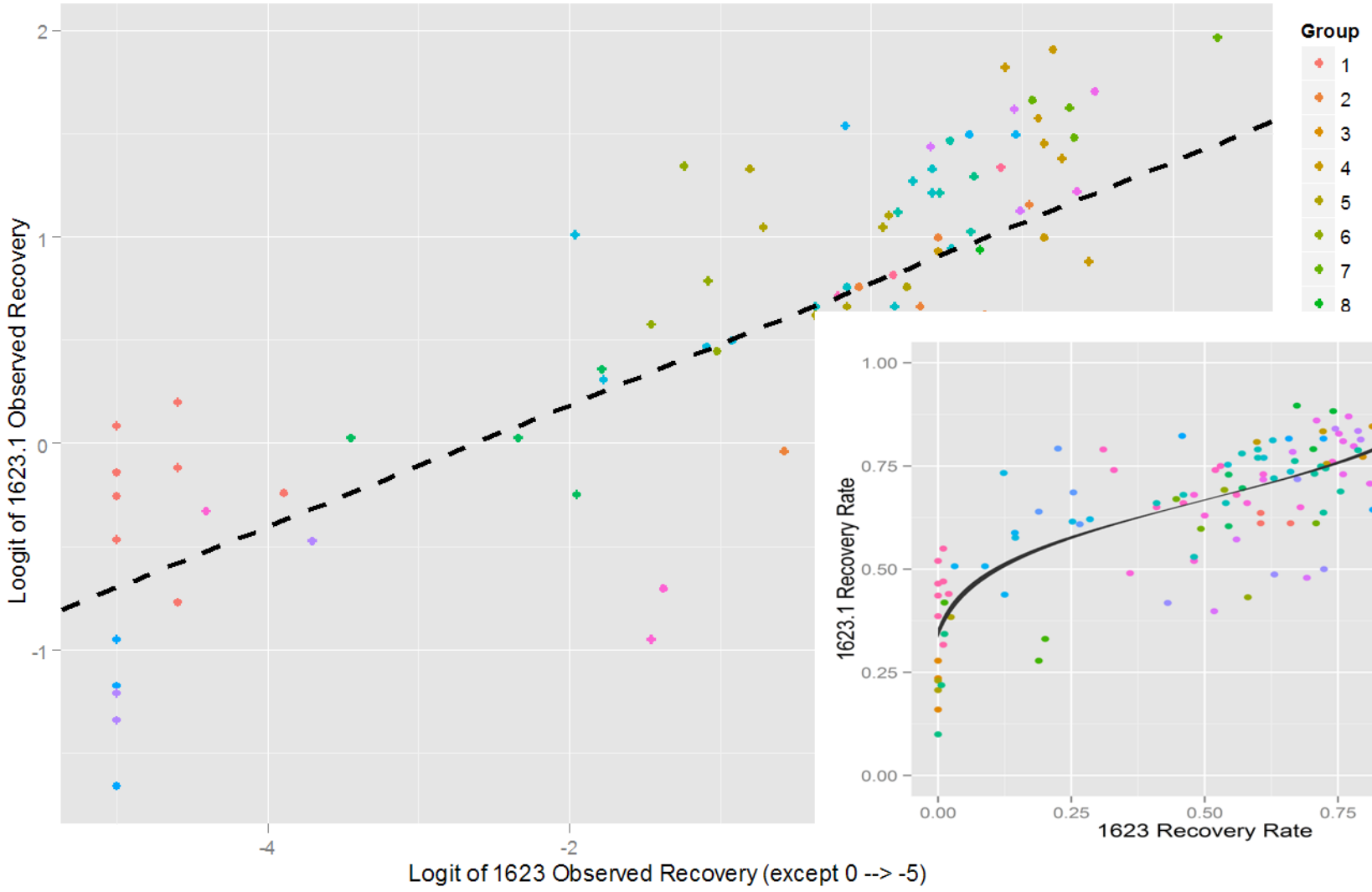
Key Analytical Components



DCTS Data Used

- "Cleaned-up" DCTS monitoring data (available online at http://water.epa.gov/lawsregs/rulesregs/sdwa/lt2/upload/cryptodatacleaned.csv)
  - Source water type info is taken from *E coli* dataset
  - Data from unfiltered source water, facilities with fewer than 20 *Cryptosporidium* field measurements and facilities with blended sources (reporting other than actual counts and volumes) and Schedule 4 systems are excluded.
  - Missing Schedule numbers were inferred from reported populations served.
  - These include no grandfathered data and no data from facilities that committed to 5.5 log overall treatment.

Methods 1623 & 1623.1 Interlaboratory Data Used



Statistical Models

- Previous slide shows a linear model for predicting Method 1623.1 recovery, as a function of Method 1623 recovery.
- The *Cryptosporidium* occurrence model is more complex:
  - Similar to the model described in the LT2 EA
    - Lognormal distribution of facility means.
    - At each facility, concentration varies lognormally over time.
  - Log-odds of recovery varies normally.

Cryptosporidium Occurrence Model, continued

- Occurrence “effects” are included for individual facilities, and also for water type and Schedule (system size).
- Recovery “effects” are included for individual laboratories.
- The recovery model also includes probabilities of zero recovery that vary from lab to lab for Method 1623.

- *Appended slides include mathematical notation that more succinctly define the model.*

Assumptions for Predictive Modeling

Assumptions:

- Facilities contributing data represent the population of facilities that are required to monitor.
- Matrix spike recoveries represent recoveries in field samples.
- Between- and within-location variances (defined on log-scale) are stable over time, and are used to predict Round 2 outcomes, given multiplicative shifts in *Cryptosporidium* concentrations.
- For Round 2 simulations, every facility samples monthly and uses its maximum running annual average to determine bin placement.

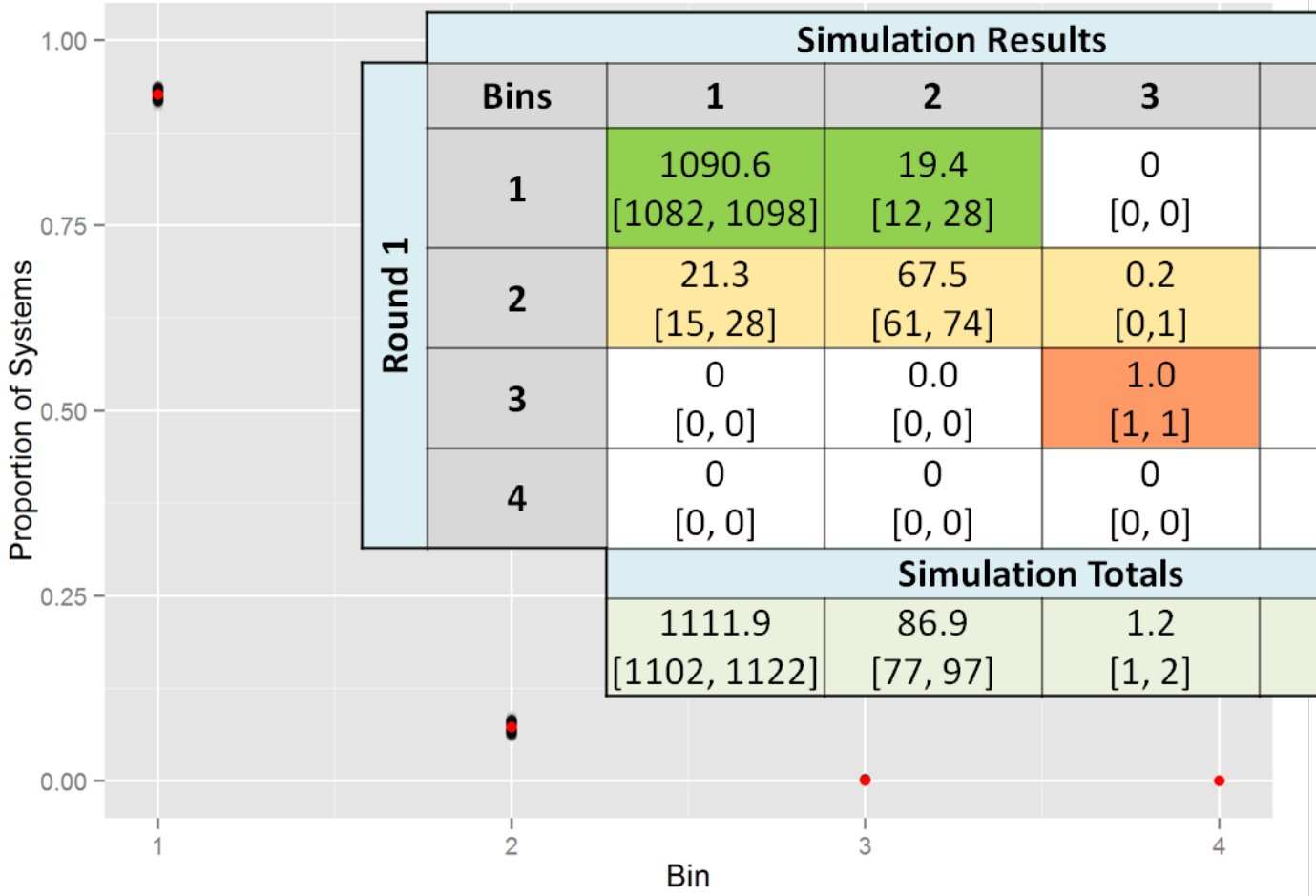
Uncertainty

- Bayesian Markov Chain Monte Carlo samples are used to convey uncertainty about model parameters
  - *Cryptosporidium* occurrence and Method 1623 recovery
  - Logistic model for predicting Method 1623.1 recovery from Method 1623 recovery

1. Can the model explain or reproduce the Round 1 data and outcomes?

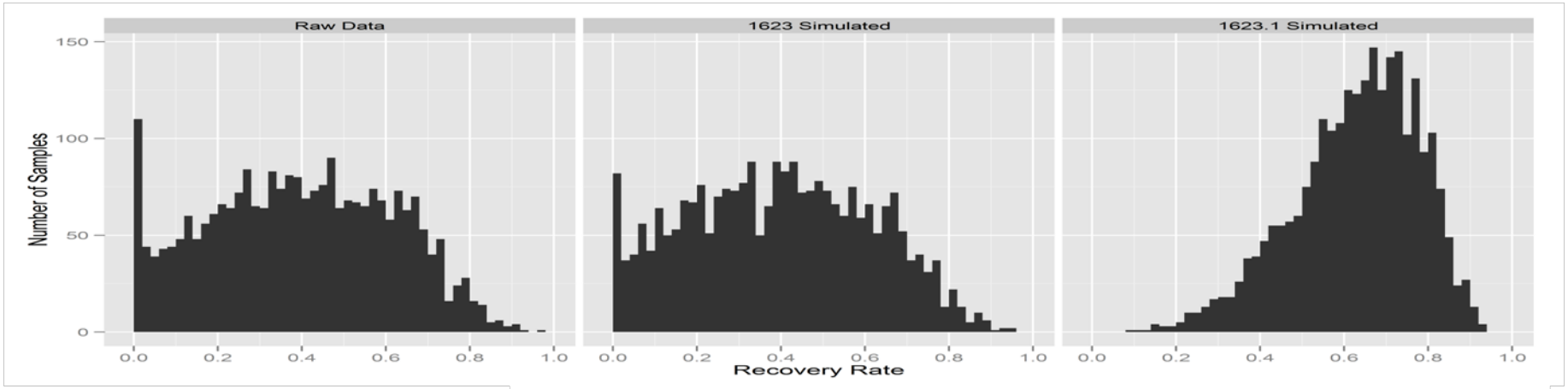
- Performance is indicated qualitatively by watching replicate chains, looking for convergence, & autocorrelation. All checks indicate the model is performing properly.
- Next slides shows that the model predicts outcomes that are like actual.

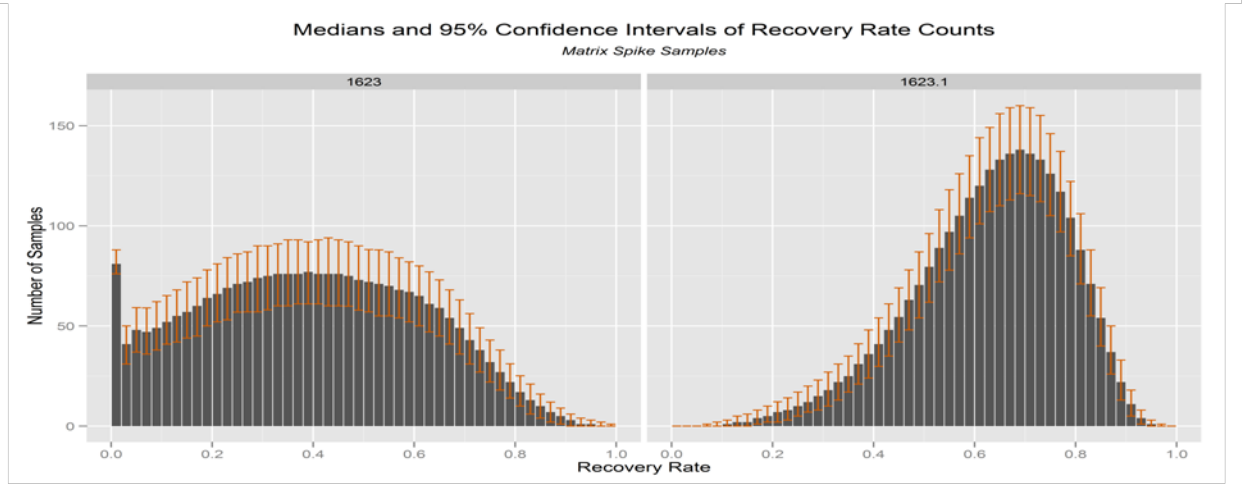
Model's Simulation of Round 1 Binning with 1623 vs. Actual Round 1 Binning



Round 1 Bins	Simulation Results: 1	2	3	4	Round 1 Totals
1	1090.6 [1082, 1098]	19.4 [12, 28]	0 [0, 0]	0 [0, 0]	1,110
2	21.3 [15, 28]	67.5 [61, 74]	0.2 [0,1]	0 [0, 0]	89
3	0 [0, 0]	0.0 [0, 0]	1.0 [1, 1]	0 [0, 0]	1
4	0 [0, 0]	0 [0, 0]	0 [0, 0]	0 [0, 0]	0
Simulation Totals	1111.9 [1102, 1122]	86.9 [77, 97]	1.2 [1, 2]	0 [0, 0]	

Model Simulated Recovery Rates vs. Recovery Ratio from Matrix Spikes

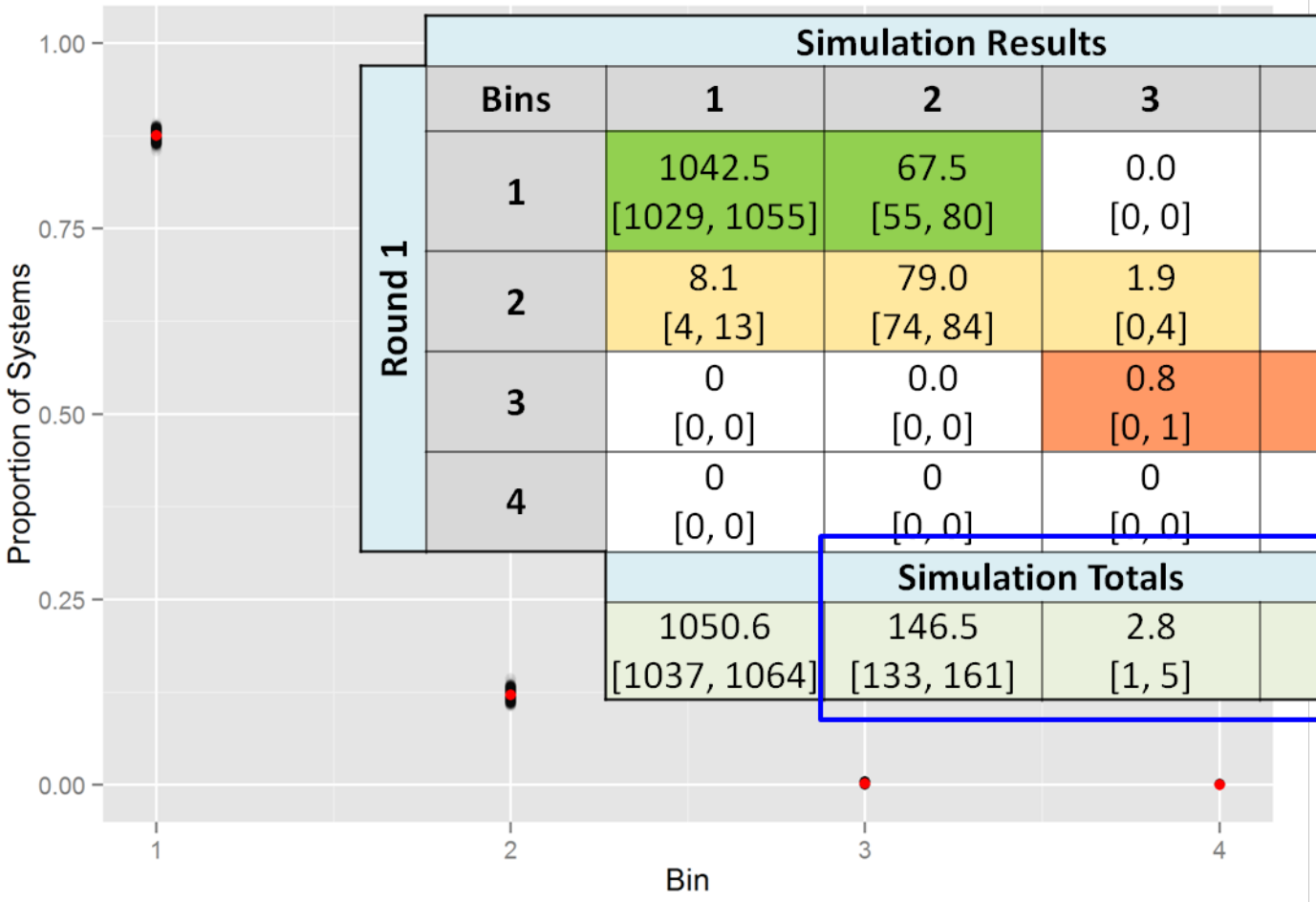




2. What Round 1 outcomes would have been predicted using Method 1623.1?

- Start with same concentrations and 1623 recoveries as used to address Question 1.
- For each record, randomly sample a 1623.1 recovery, conditional on the 1623 recovery.
- Randomly draw a new count for each field sample and determine the new bin placement for each facility.

Model's Simulation of Round 1 Binning with 1623.1 vs. Actual Round 1 Binning



Round 1	Simulation Results				Round 1 Totals
Bins	1	2	3	4	
1	1042.5 [1029, 1055]	67.5 [55, 80]	0.0 [0, 0]	0 [0, 0]	1,110
2	8.1 [4, 13]	79.0 [74, 84]	1.9 [0,4]	0.0 [0, 0]	89
3	0 [0, 0]	0.0 [0, 0]	0.8 [0, 1]	0.2 [0, 1]	1
4	0 [0, 0]	0 [0, 0]	0 [0, 0]	0 [0, 0]	0
	Simulation Totals				
	1050.6 [1037, 1064]	146.5 [133, 161]	2.8 [1, 5]	0.2 [0, 1]	

Total 150 in "action bins" (≈70% increase).

3. For the 1110 facilities placed in Bin 1 during Round 1, what Round 2 outcomes are predicted (distribution of facilities across bins 1-4), assuming no change in the facility-specific occurrence distributions
   - using Method 1623?
   - using Method 1623.1?

Round 2 Predictions by Source Type

Round 2 Simulations of Binning 1110 Plants in Bin 1 in Round 1	Method 1623		Method 1623.1	
	Bin 1	Bins 2-4	Bin 1	Bins 2-4
Reservoirs/Lakes	577.6 [571, 584]	11.3 [5, 18]	567.6 [559, 576]	21.4 [13, 30]
Flowing Streams	374.6 [364, 384]	30.4 [21, 42]	347.0 [335, 360]	58.0 [45, 71]
All	1065.2 [1053, 1076]	44.8 [34, 58]	1025.2 [1010, 1041]	85.0 [68, 101]

- This table focuses on the 1110 facilities that were placed in Bin 1, based on Round 1 data. *(Not all were classified as Reservoirs/Lakes or Flowing Streams)*
- Overall occurrence levels are assumed to be the same as in Round 1.
- Method 1623.1's higher recovery rate increases the numbers of facilities in Bins 2-4.
- Even with Method 1623.1, few reservoir/lakes are assigned to Bins 2-4.

4. For facilities placed in Bin 1 during Round 1, what would the outcomes look like if *Cryptosporidium* occurrence were to systematically increase or decrease
   - using Method 1623?
   - Using Method 1623.1?

"Increase or decrease" is by factor-of-three. Each facility mean concentration is multiplied or divided by three.

Modeled Round 2 Outcomes: Method 1623 and 1623.1

Plants in Bins 2-4 under Alternative Scenarios on Occurrence Distribution

Of 1110 Plants in Bin 1 in Round 1	Method 1623			Method 1623.1		
	×(3)	×(1)	×(1/3)	×(3)	×(1)	×(1/3)
Reservoirs/Lakes	7.62%	1.92%	0.29%	12.58%	3.63%	0.59%
Flowing Streams	22.57%	7.51%	1.39%	35.26%	14.32%	3.11%
All (includes GWUDI & both)	13.18%	4.04%	0.71%	20.93%	7.64%	1.54%

Summary of Model-Based Estimates

- The modeling does a good job of reproducing the Round 1 data and outcomes.
- Modeled estimates for Round 2 (assuming same average concentrations, Method 1623, 10 L samples and 24 samples for all facilities) are similar to the observed Round 1 results, as would be expected.
- Modeled estimates for Round 2 as above, but with Method 1623.1, show more facilities placed in the higher bins, due to improved recovery for the new method.
- Assumed changes in overall occurrence levels (3x or 1/3 of Round 1 observations) result in modeled estimates with expected (increased or decreased) number of facilities in the higher bins, and again with more occurrence in higher bins using 1623.1 versus 1623.

If you need additional information on the models and model-based predictions, please contact:

Mike Messner at:
messner.michael@epa.gov

Or Ken Rotert at:
rotert.kenneth@epa.gov

Appended Slides

- Occurrence Model Details: Slides 22 - 35
- Additional Results Tables: Slides 36 - 37

Statistical Models

Distribution of oocysts Recovered from Spiked Sample:

$$y_{sj} \sim Bin(r_{sj}, Crypto_j^{spiked})$$

$$Count \sim Pois\left(\lambda = Crypto^{spiked} \times Rec \times e^{\varepsilon}\right)$$

Statistical Models

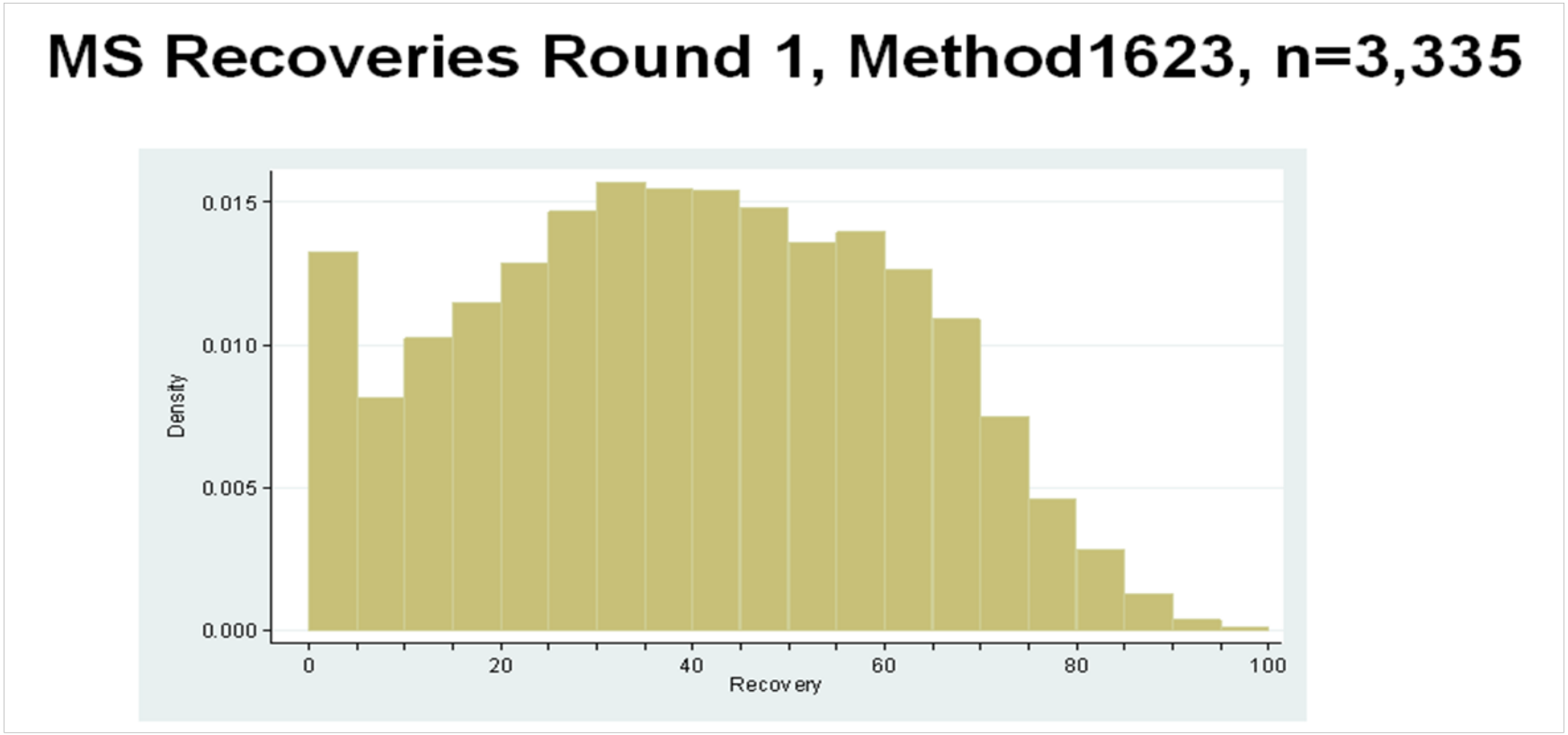


Figure : Extra 0s (more than we bargained for)

Statistical Models

Account for extra 0s and over-dispersion:

$$Rec = Failed \times 0 + (1 - Failed) \times Rate$$

$$\begin{aligned} \lambda_{sj} &= (1 - z_j) r_{k[j]} Crypto_j^{spiked} \\ z_j &\sim Bern(p_{0k[j]}) \\ \text{logit}(r_k) &\sim N(\theta_R, \sigma_5^2) \\ \text{logit}(p_{0k}) &\sim N(\theta_{R0}, \sigma_6^2) \end{aligned}$$

Optional amount of additional detail:

$$logit(Rate) \sim N(\rho_k, \sigma_5^2)$$

$$Failed \sim Bern(Prob(0|Lab = k))$$

$$logit(Prob(0|Lab = k)) \sim N(\theta_k, \sigma_6^2)$$

$$logit(q) = \log\left(\frac{q}{1-q}\right)$$

Statistical Models

Occurrence in Field Monitoring Data:

$$
\begin{aligned}
y_i &\sim Pois(\lambda_i) \\
\lambda_i &= (1 - z_i) r_{k[i]} V_i C_i e^{\epsilon_i} \\
\log(C_i) &\sim N(\mu_0 + \mu_{1l[i]} + \mu_{2m[i]} + \mu_{3n[i]}, \sigma_1^2) \\
\mu_{1l} &\sim N(0, \sigma_2^2) \\
\mu_{2m} &\sim N(0, \sigma_3^2) \\
\mu_{3n} &\sim N(0, \sigma_4^2)
\end{aligned}
$$

$$Count \sim Pois\left(\lambda = \left[Vol \times Conc^{field}\right] \times Rec \times e^{\varepsilon}\right)$$

$$Rec = (1 - Failed) \times Rate$$

$$\log(Conc) \sim N(\mu_0 + \mu_1 + \mu_2 + \mu_3, \sigma_1^2)$$

Location Effects: $\mu_1 \sim N(0, \sigma_2^2)$

Size Effects: $\mu_2 \sim N(0, \sigma_3^2)$

Source Effects: $\mu_3 \sim N(0, \sigma_4^2)$

Statistical Models

Putting them together:

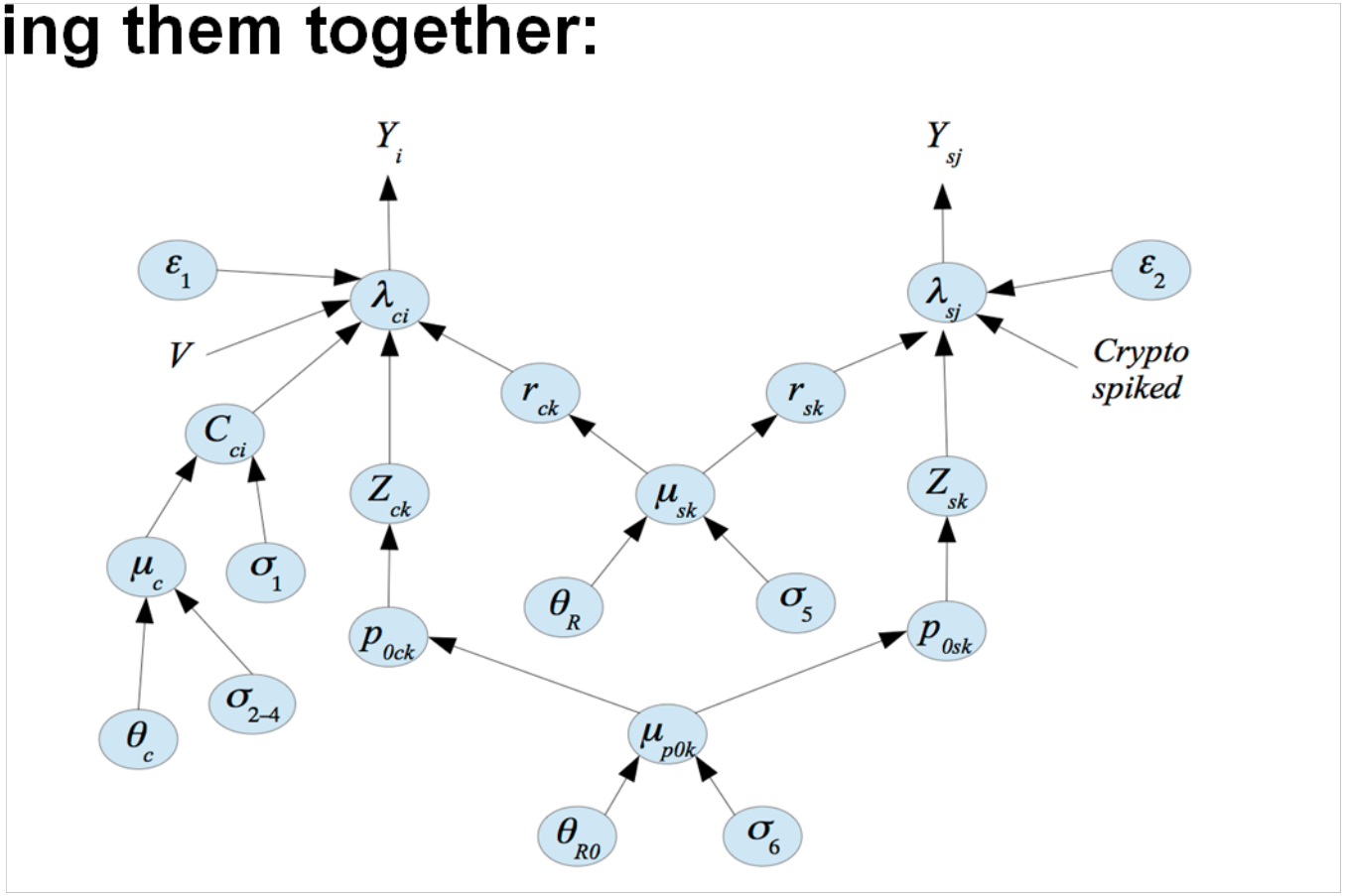
Model of oocysts Recovered from Spiked Sample

+

Model of Occurrence in Field Monitoring Data

Statistical Models

Putting them together:



Computation

Markov Chain Monte Carlo (MCMC)

Implemented initially in JAGS and later in Stan – using both packages reassures reliability of results

Results

Lab performance assessment

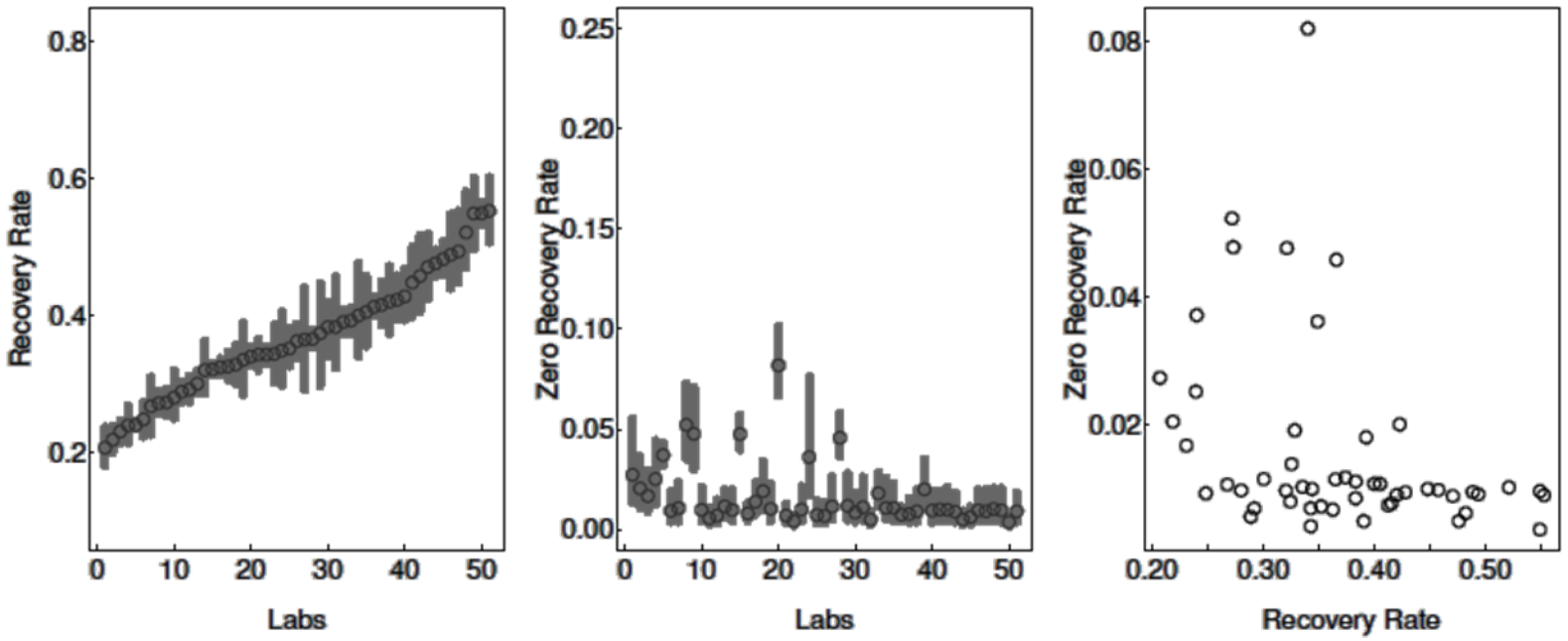


Figure : Large between lab variances

Results

Crypto occurrences

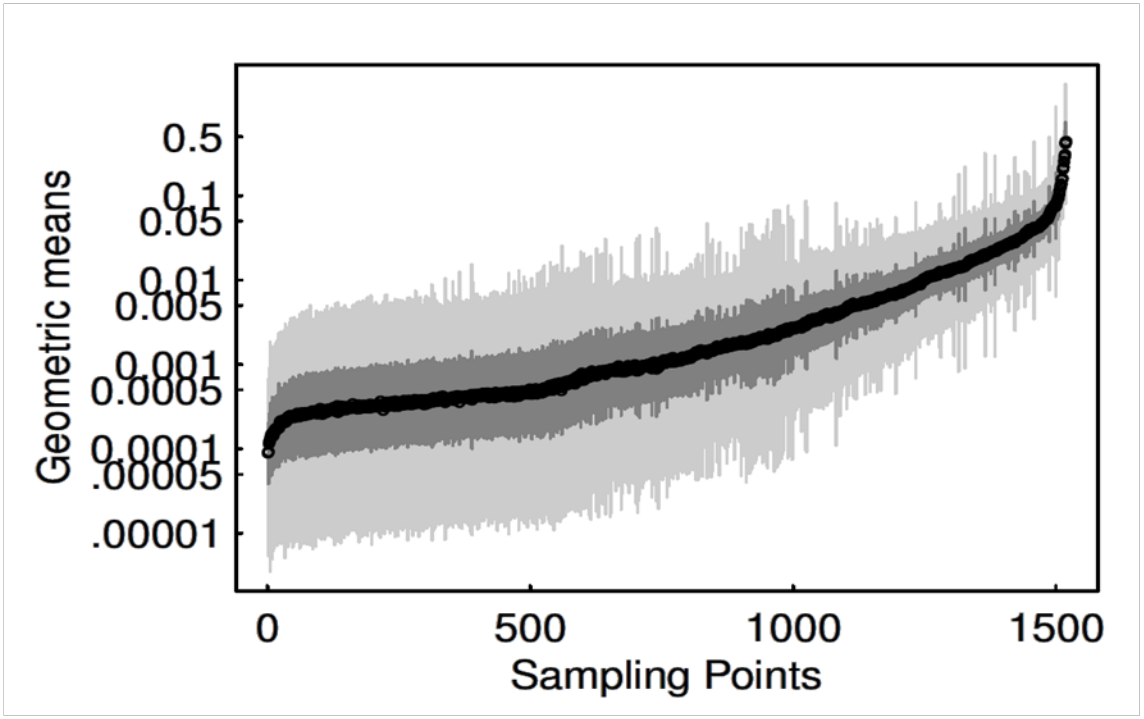
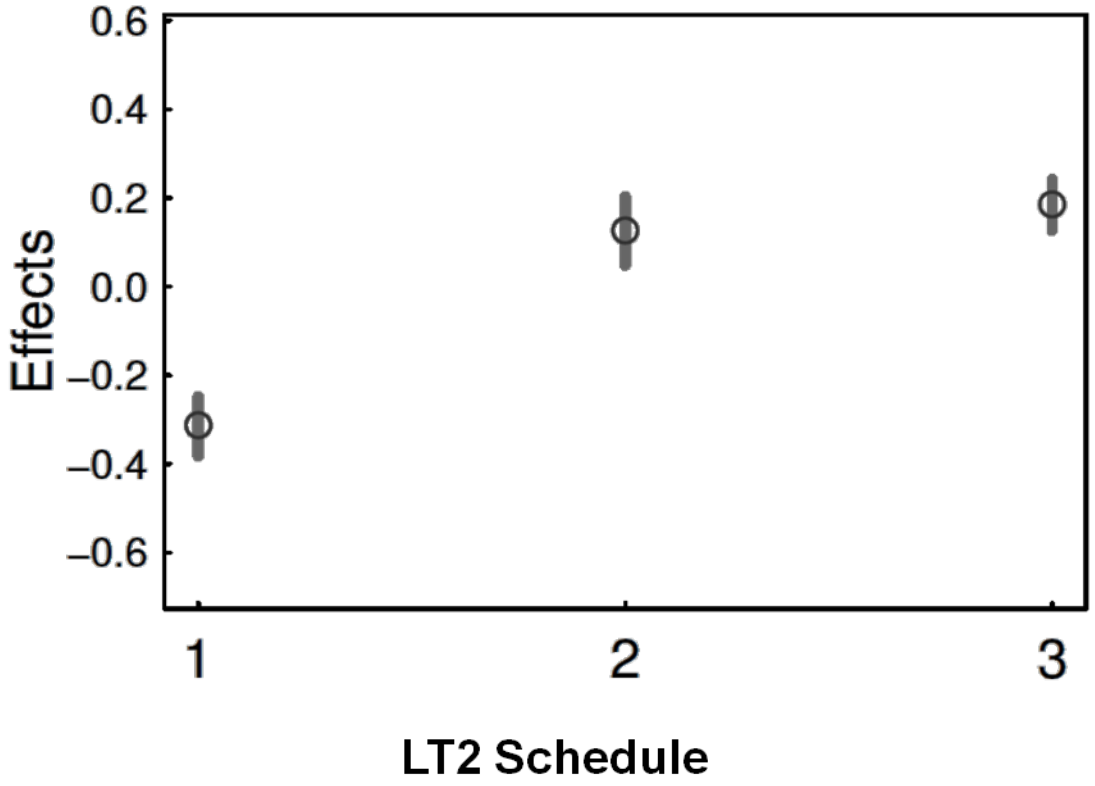


Figure : Geometric means for each sampling point

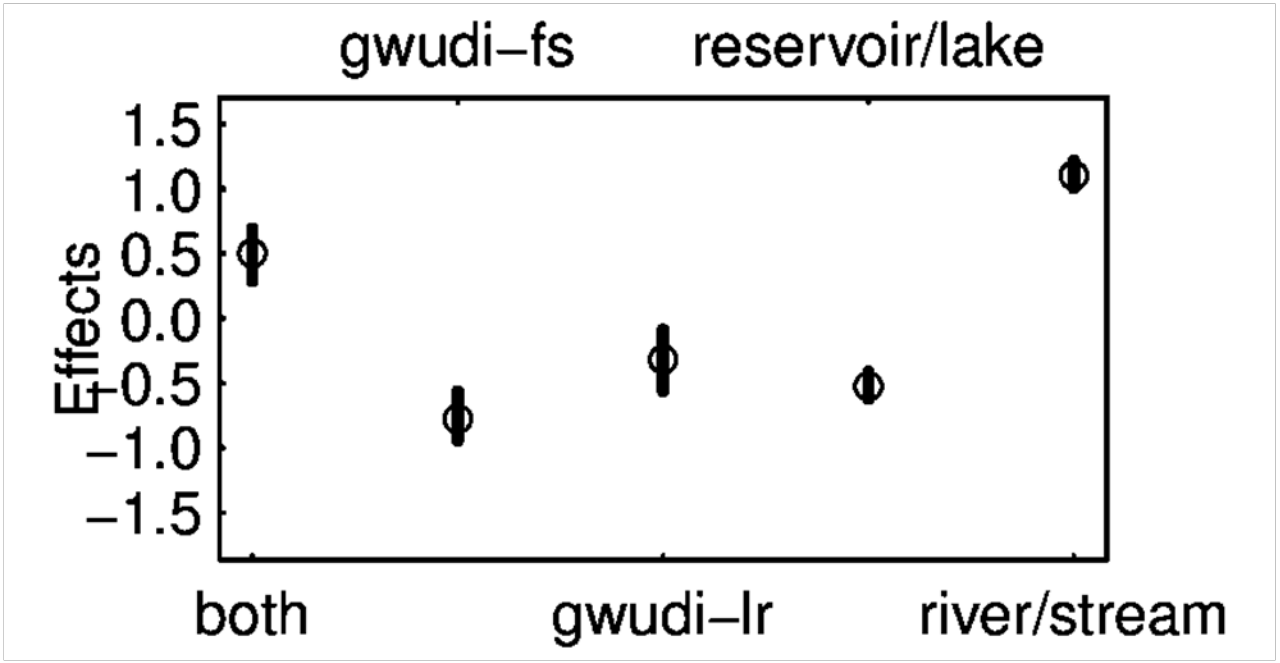
Results

The size effect



Results

The source water type effect



Results

Model Predicted versus Observed

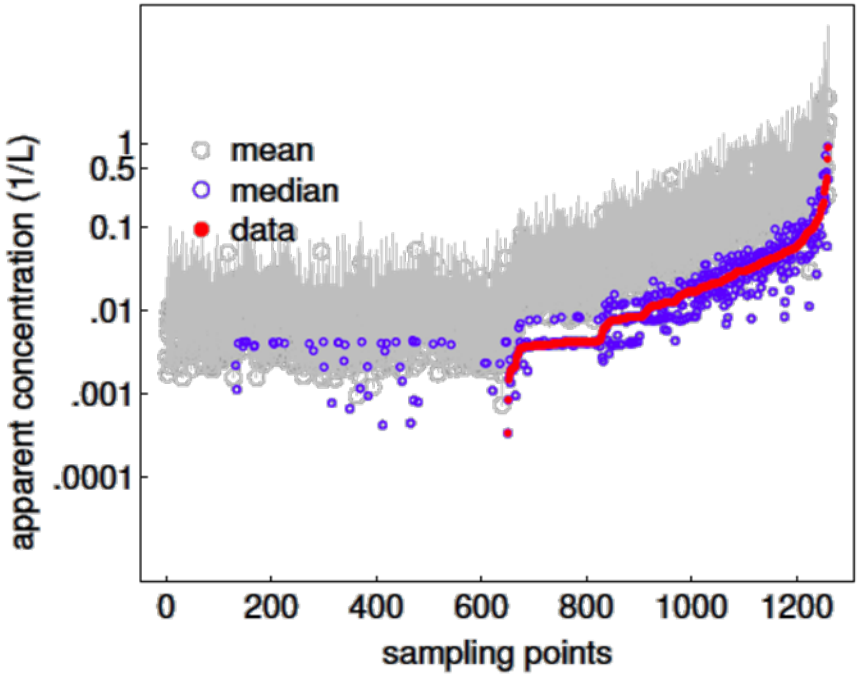
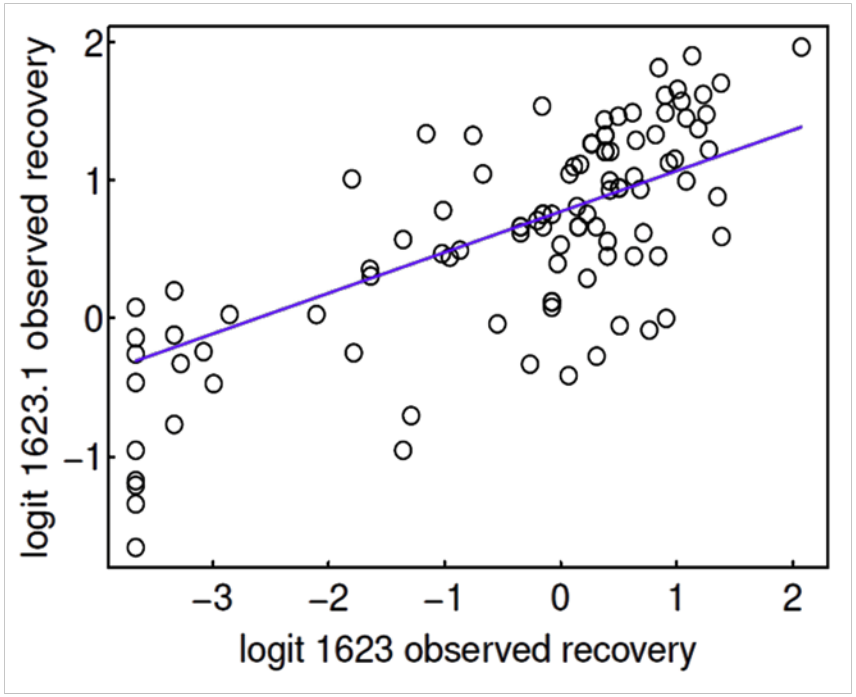
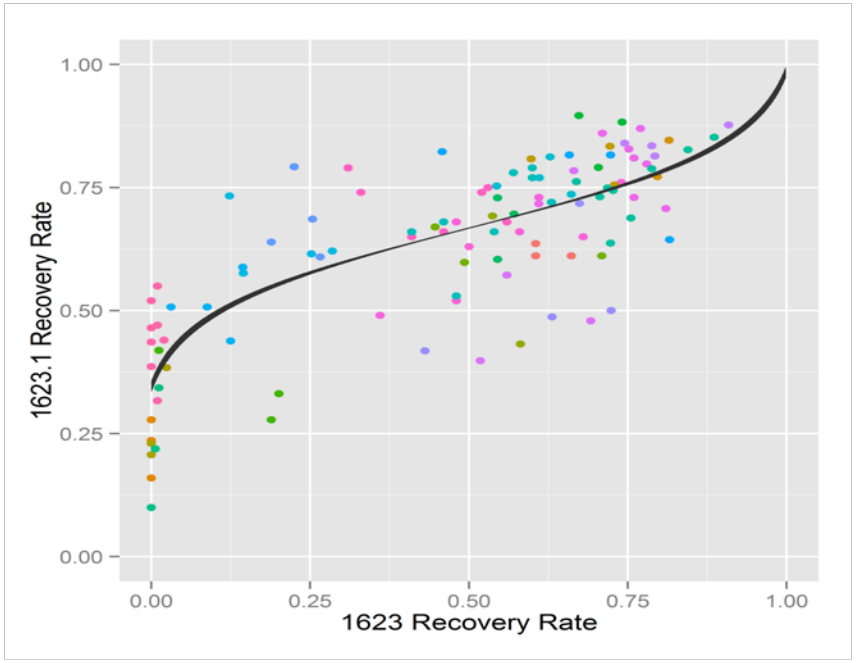


Figure : Predicted versus observed

Using the Model for Prediction

Enhanced Recovery with Method 1623.1





$$logit(Rec_{1623.1}) = \beta_0 + \beta_1 logit(Rec_{1623}) + \varepsilon$$

Discussion

Bayesian Hierarchical Models

- Hierarchical model (random effects model) is suited for EPA's needs – understanding the national distribution of source water crypto concentrations
- Bayesian approach allows an explicit assessment of uncertainty for a model that can capture these key complications in *Cryptosporidium* occurrence
- MCMC is flexible and parallel computation makes the process doable in a practical amount of time

Discussion

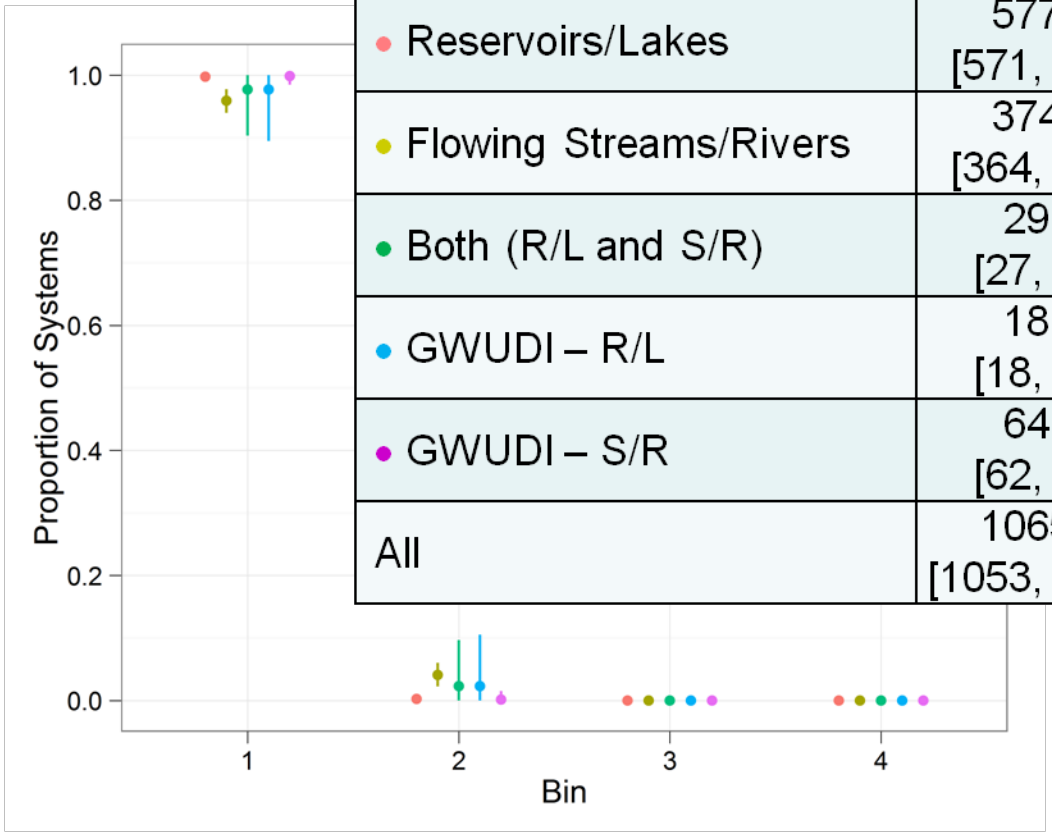
The model assumes that recovery of oocysts in field samples is similar to recovery of oocysts from matrix spike samples

- Environmental oocysts are present in much smaller numbers, but the probability of detection doesn't change with the number present.
- Environmental oocysts have aged in the environment, but are still as likely to be counted as the "fresh" oocysts that are spiked
- Analysts handle and assay field samples and spiked samples in the same manner. Their knowledge of the sample type (field or spike) doesn't influence how they treat the samples.
- Zero recovery in a fraction of spiked samples suggests that the same fraction of field samples may produce zeros in spite of any environmental oocysts that are present in the sample.

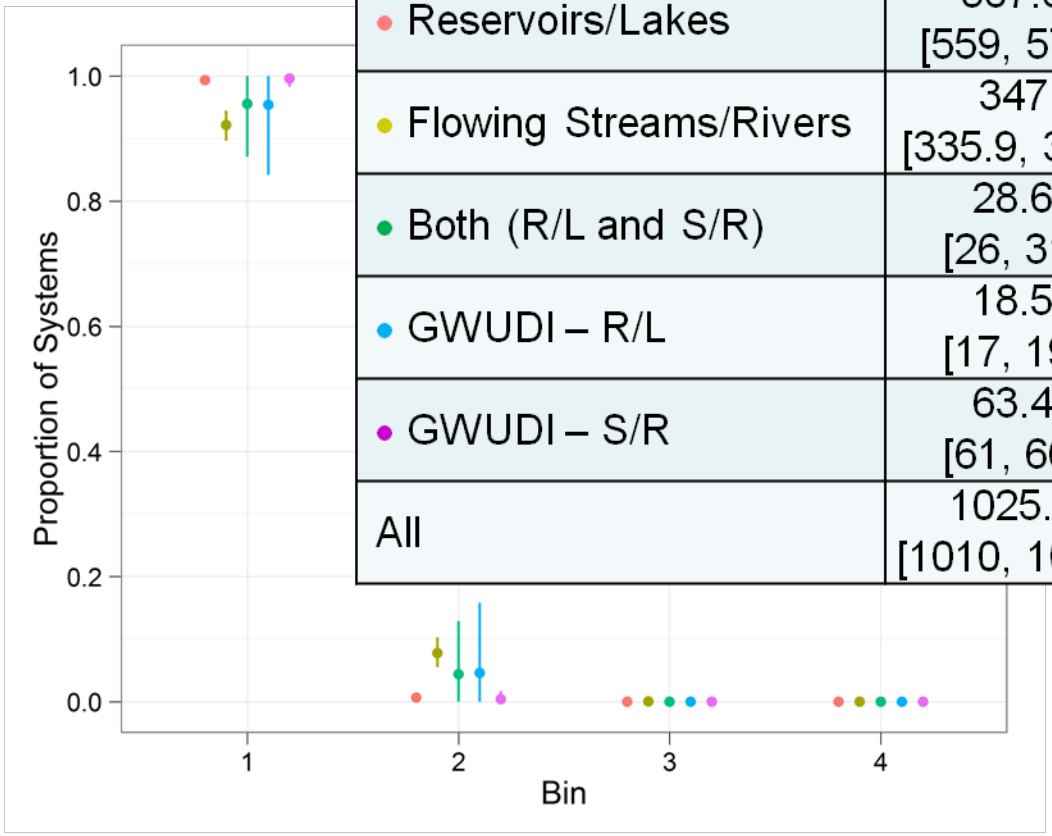
The model reveals that the majority of 0's in field samples are due to a) no oocysts in the volume assayed, b) undetected oocysts due to imperfect recovery, and c) undetected oocysts due to 0 recovery.

Modeled Round 2 Outcomes: Using Method 1623

Of 1110 Plants in Bin 1 in Round 1	Bin 1	Bin 2	Bin 3	Bin 4
Reservoirs/Lakes	577.6 [571, 584]	11.3 [5, 17.1]	0 [0, 1]	0 [0, 0]
Flowing Streams/Rivers	374.6 [364, 384]	30.3 [21, 41]	0.1 [0, 1]	0 [0, 0]
Both (R/L and S/R)	29.7 [27, 31]	1.3 [0, 4]	0 [0, 0]	0 [0, 0]
GWUDI – R/L	18.8 [18, 19]	0.2 [0, 1]	0 [0, 0]	0 [0, 0]
GWUDI – S/R	64.5 [62, 66]	1.5 [0, 4]	0 [0, 0]	0 [0, 0]
All	1065.2 [1053, 1076]	44.7 [34, 57]	0.2 [0, 1]	0 [0, 0]



Modeled Round 2 Outcomes: Using Method 1623.1



Of 1110 Plants in Bin 1 in Round 1	Bin 1	Bin 2	Bin 3	Bin 4
Reservoirs/Lakes	567.6 [559, 576]	21.3 [13, 29.1]	0.1 [0, 1]	0 [0, 0]
Flowing Streams/Rivers	347 [335.9, 360]	57.7 [45, 69]	0.3 [0, 2]	0 [0, 0]
Both (R/L and S/R)	28.6 [26, 31]	2.3 [0, 5]	0 [0, 0]	0 [0, 0]
GWUDI – R/L	18.5 [17, 19]	0.5 [0, 2]	0 [0, 0]	0 [0, 0]
GWUDI – S/R	63.4 [61, 66]	2.6 [0, 5]	0 [0, 0]	0 [0, 0]
All	1025.2 [1010, 1041]	84.4 [68, 99]	0.4 [0, 2]	0 [0, 0]

Crypto Counted	Frequency
0	38729
1	1721
2	548
3	238
4	130
5	86
6	46
7	26
8	26
9	19
10	11
11	6
12	9
13	2
14	4
15	6
16	0
17	2
18	0
19	1
20	1