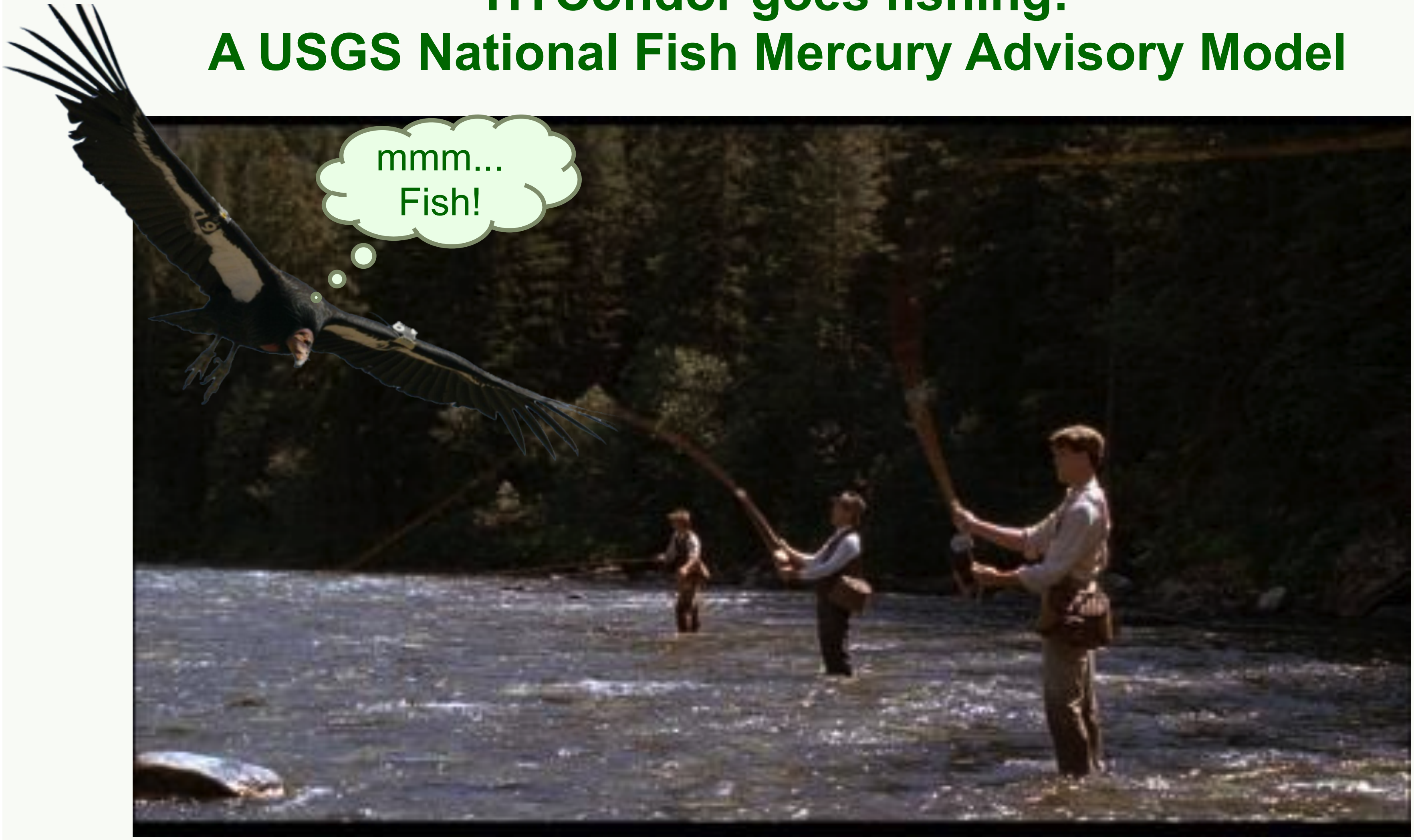
HTCondor goes fishing:
A USGS National Fish Mercury Advisory Model



Mike Fienen, Research Hydrologist
US Geological Survey, Wisconsin Water Science Center
Mark Brigham, David Donato, David Lorenz

Validation of the National Descriptive Model of Mercury in Fish

Mercury, fish, and human health

A descriptive statistical model

Validation efficiency

The role of HTCondor

Monument to Minamata
Mercury Poisoning Victims

What’s the deal with Mercury?

Powerful neurotoxin in humans

Bioaccumulates in fish

Generally, mercury concentration increases with background concentration and fish size

Important social justice implications

USGS as a national science agency is responsible for looking at these issues on a national scale

What's the deal with Mercury?

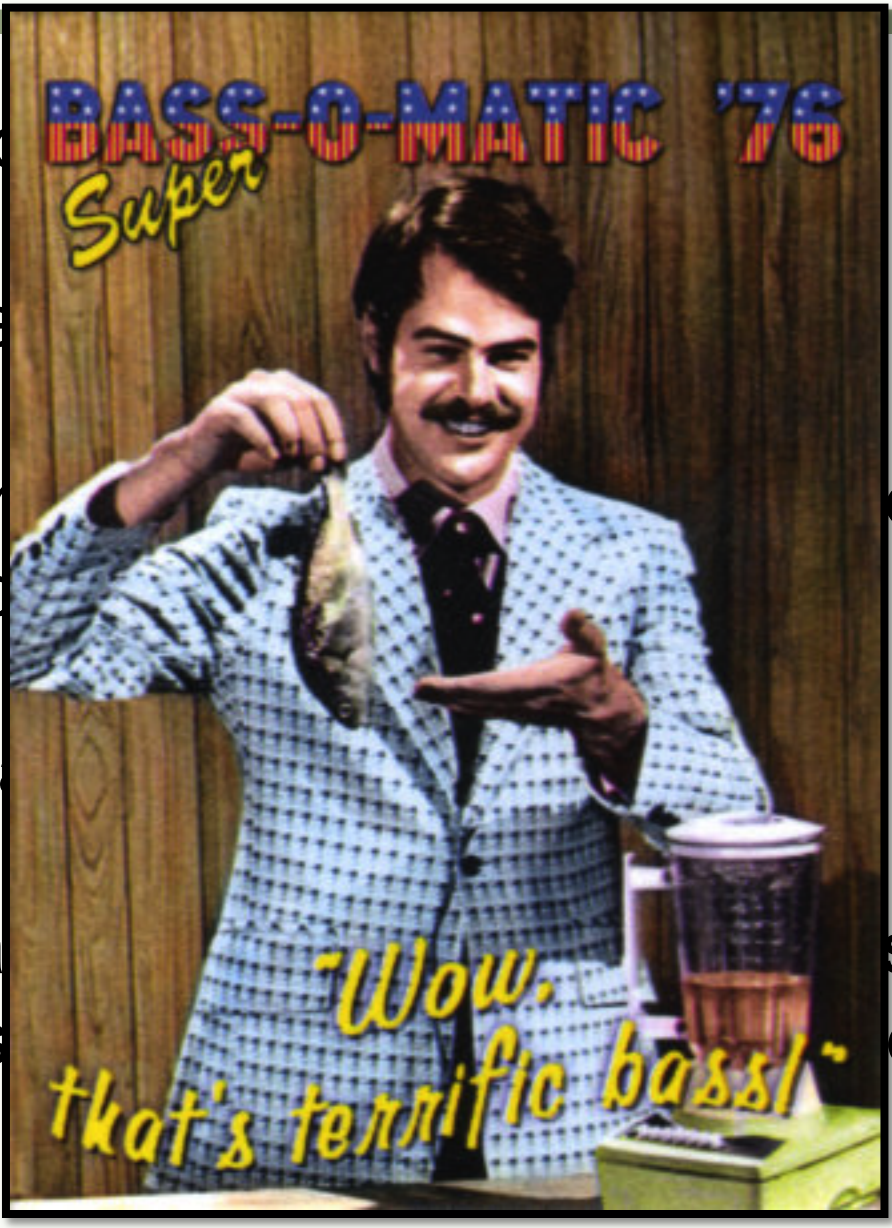
Powerful neuro

Bioaccumulate

Generally, mer es with background co

Important soci

USGS as a na sponsible for looking at thes e



The National Descriptive Model of Mercury in Fish

Analysis of Covariance (general linear model) applied to a large, national fish-mercury data set.

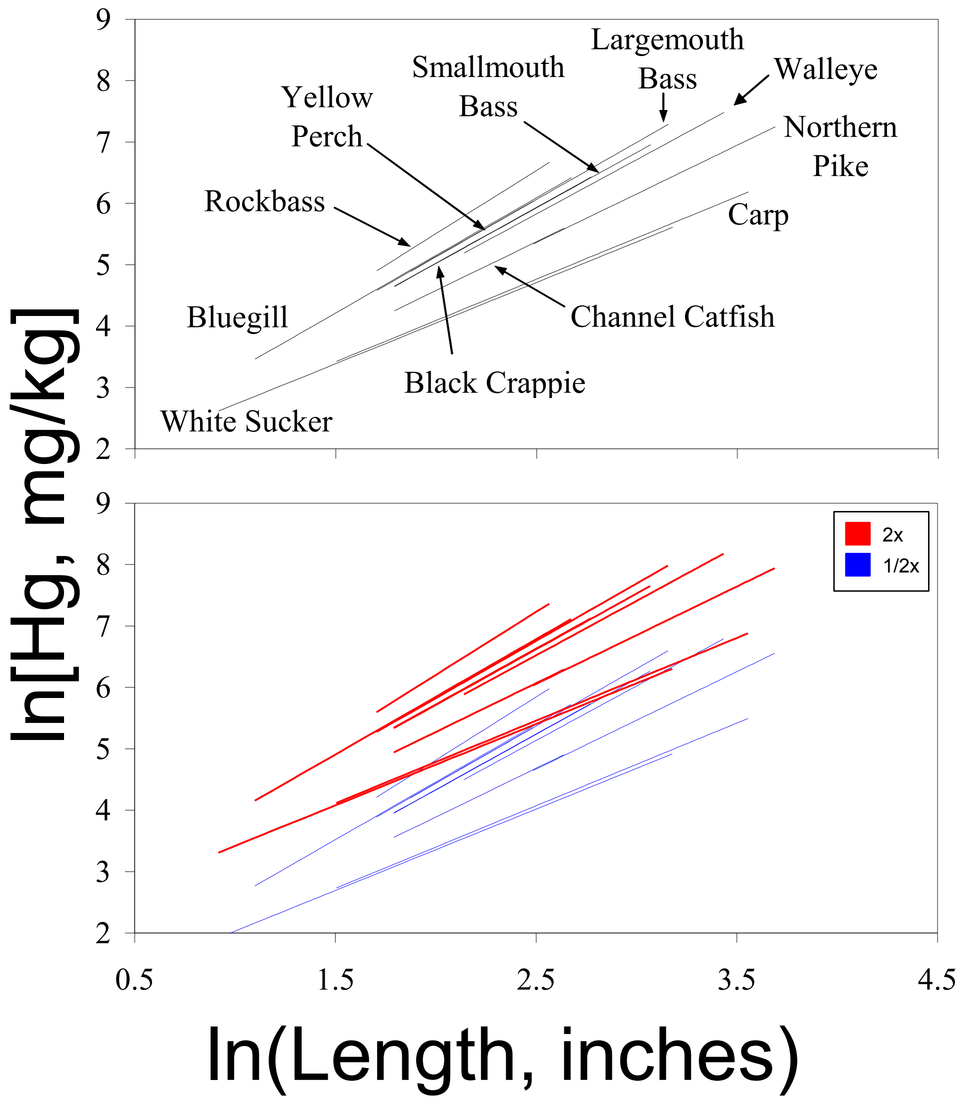
Relates fish [*Hg*] to fish length for many **events** and **species & cuts**, simultaneously.

$$\ln[Hg+1] = \alpha j + \beta k \ln[length+1] + \varepsilon$$

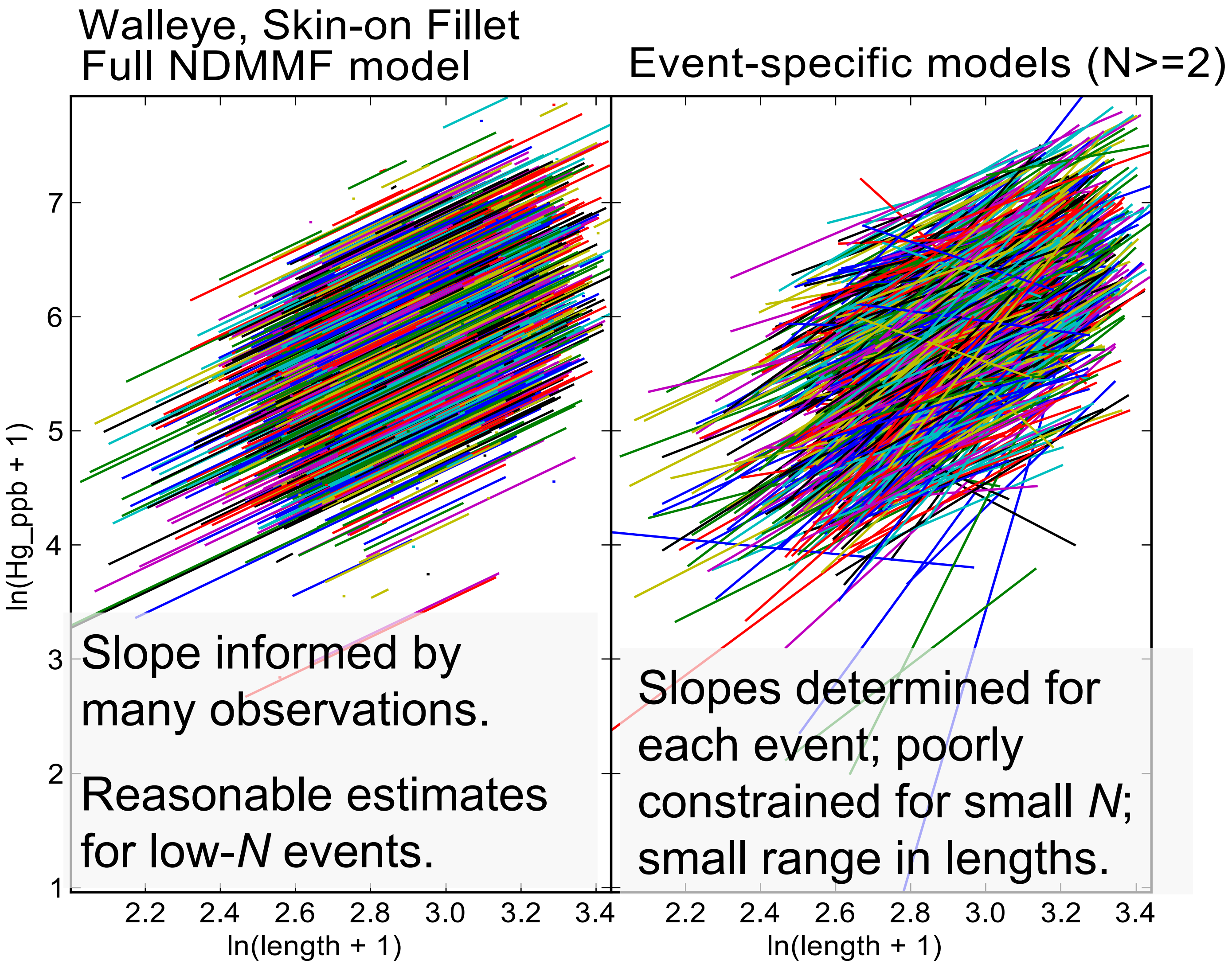
Slope (length parameter) for k^{th} species-cut combination.

Intercept (event parameter) for j^{th} event. All fish samples collected at a site in a calendar year comprise an event.

Conceptualized model for 11 species, 3 events



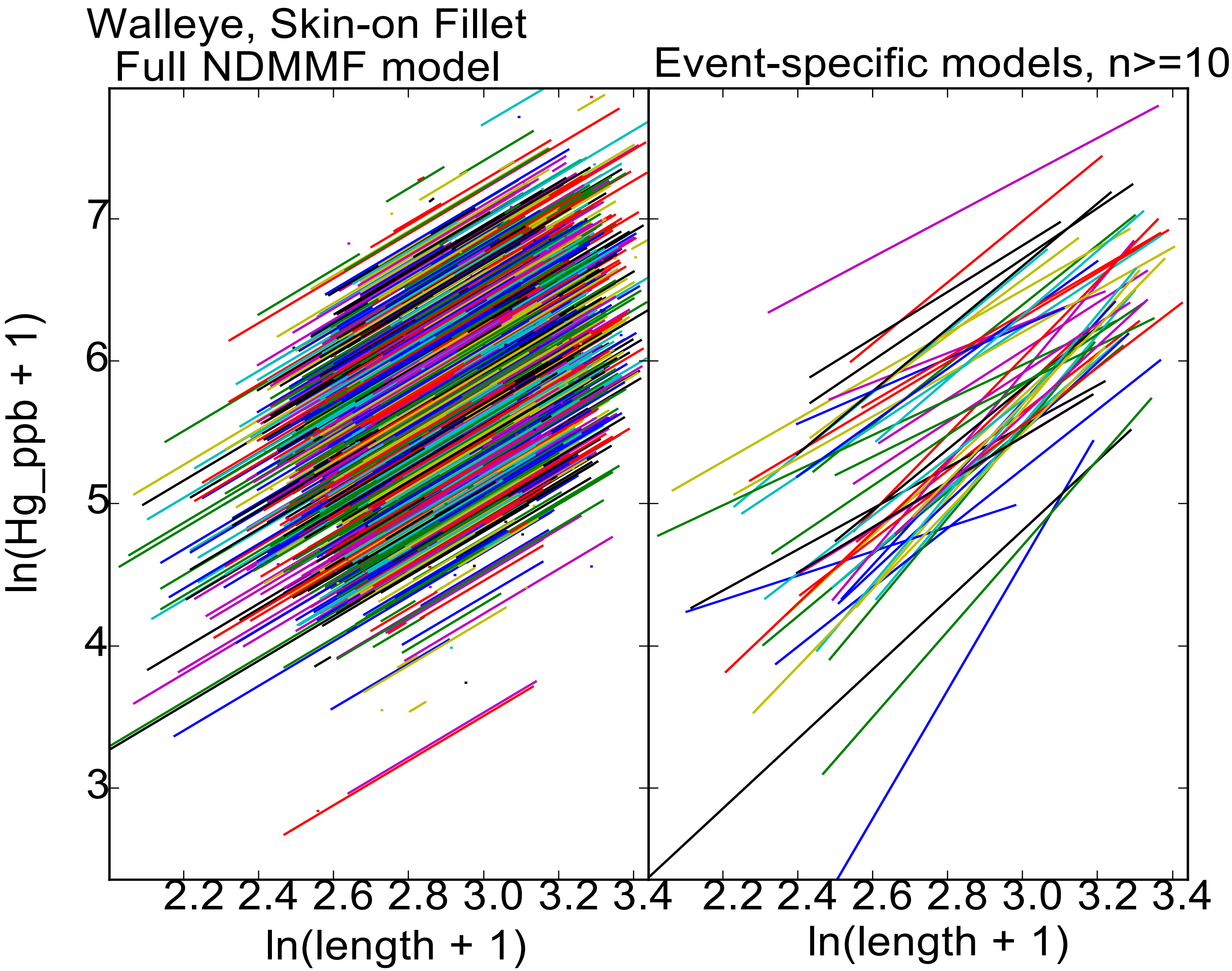
Model assumes constant slope for species-cut



Slope informed by many observations.

Reasonable estimates for low-*N* events.

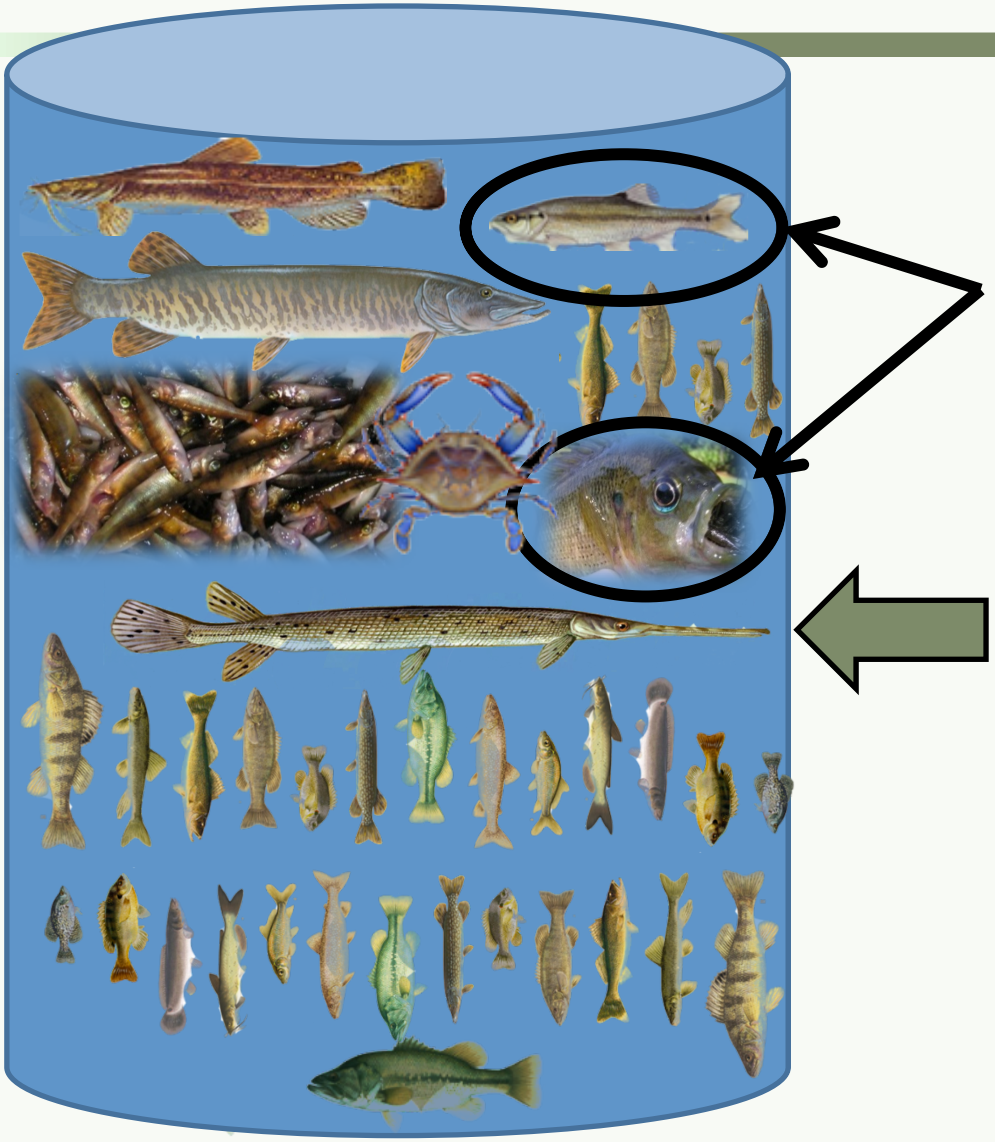
Slopes determined for each event; poorly constrained for small *N*; small range in lengths.



Walleye, Skin-on Fillet
Full NDMMF model

Event-specific models, n>=10

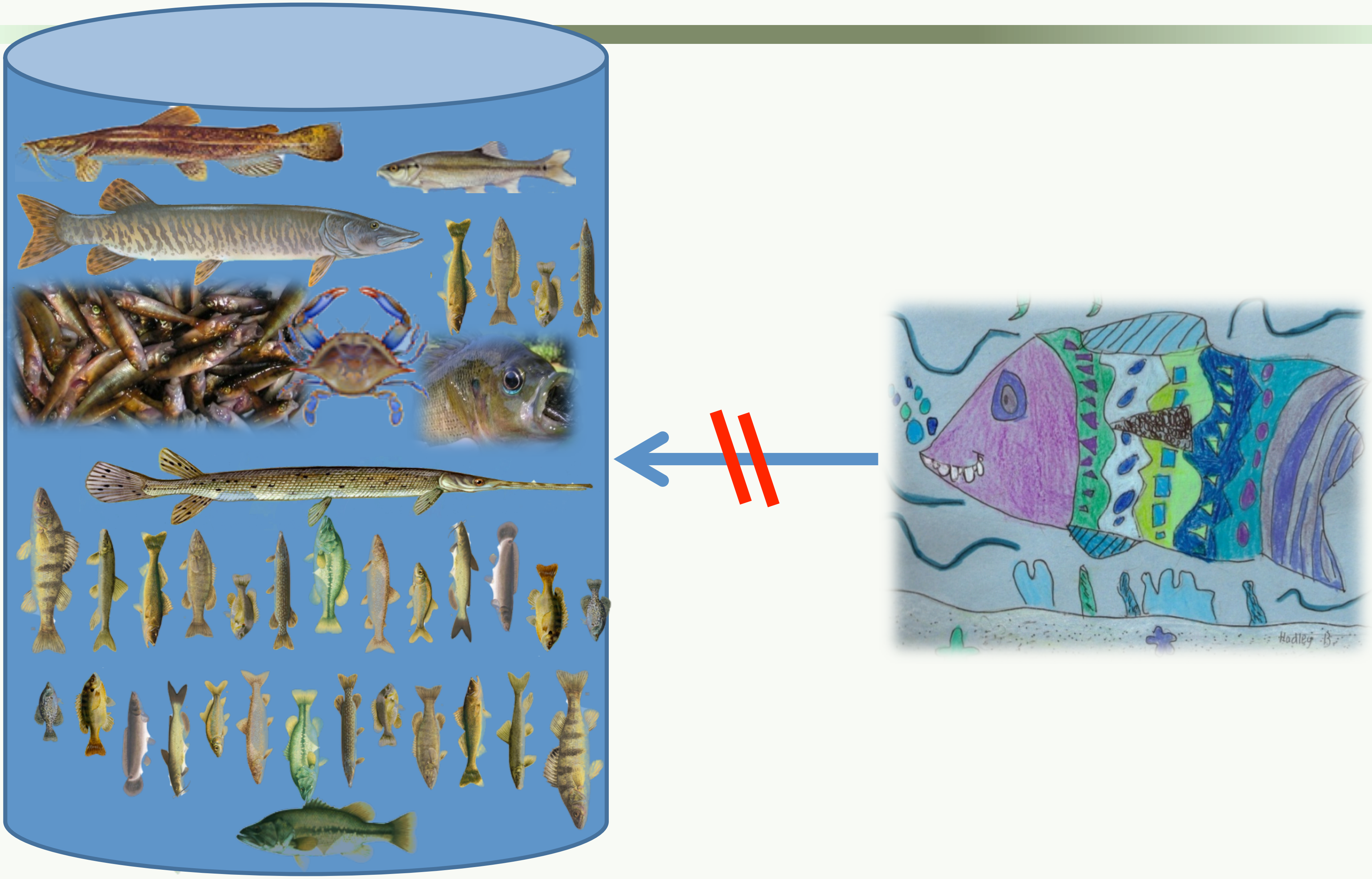
Strength in numbers! Large *inference space*



Sampled species at an event

Can estimate (predict) [Hg] for same event, for *all* species in data set.

…unless unconnected to at least one event



USGS CIDA Center for Integrated Data Analytics

This Model was Published but not Validated

2004 NDMMF solved by Maximum Likelihood regression, using PROC LIFEREG in SAS 9.3 on 64-bit Unix (35 hour forward run time)

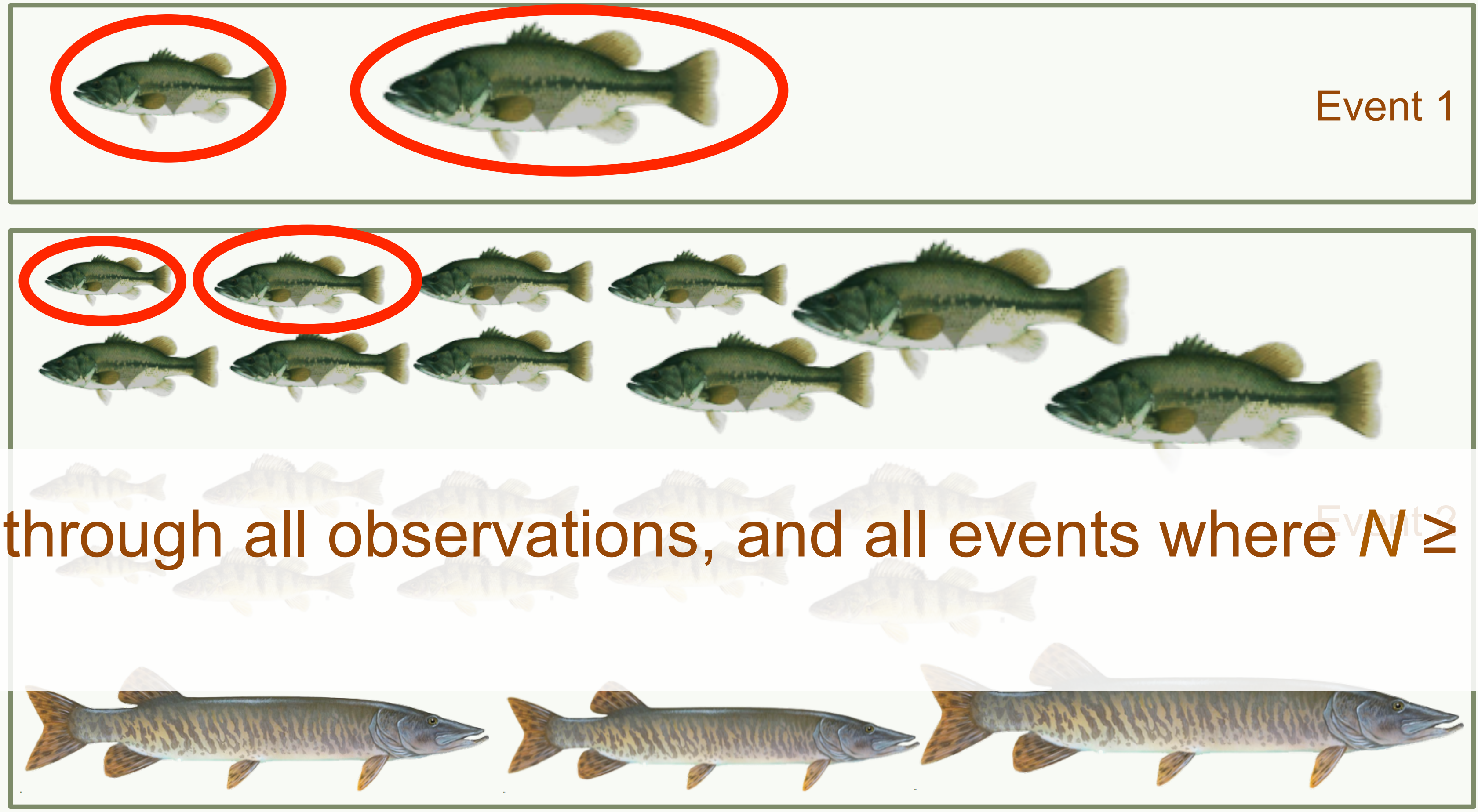
Data set updated in 2006:

- 102,000 observations; 10,000 sites across U.S.
- Used ~99,000 observation subset, where N>1, and all samples were “connected” to data base

Validation study in 2010-2012:

- Python code replaced manual spreadsheet fiddling to establish connectivity of events
- C-code facsimile model that converges in seconds
- Run in parallel on 76 CPUs with HTCondor

Leave-One-Out Cross-Validation



…through all observations, and all events where *N* ≥ 2.

The Role of HTCondor

High Throughput computing can increase efficiency…

…but more importantly, it can enable science that could not otherwise be done.

Leave-One-Out Cross-Validation

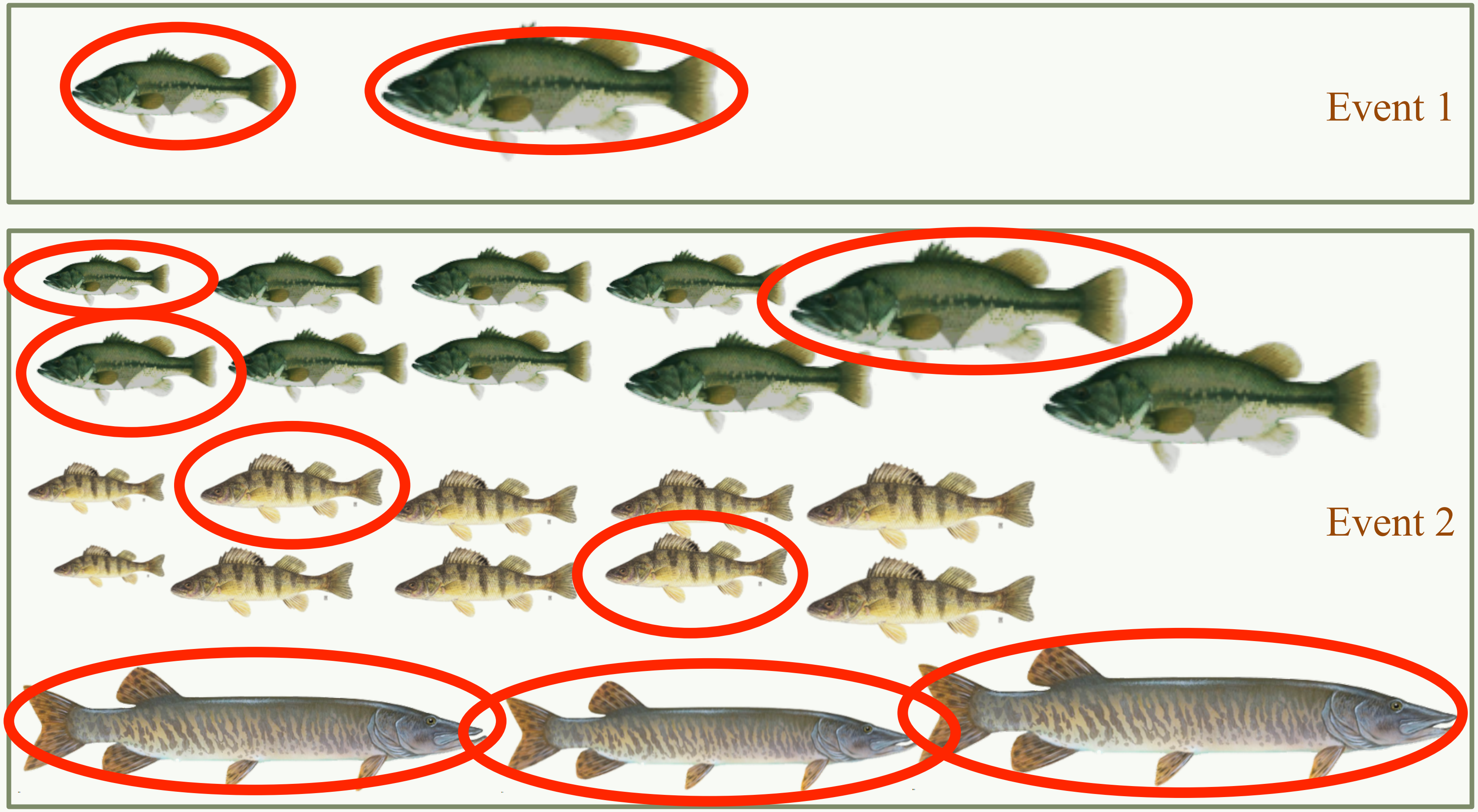
~100,000 runs x 35 hours = 400 CPU *years*

Reducing run times to ~3 minutes and distributing over 76 CPUs = 1-2 CPU *days*

Beware the notification setting!

```
notification = Never
universe = vanilla
log = log/worker_$(Cluster).log
output = log/worker_$(Cluster)_$(Process).out
error = log/worker_$(Cluster)_$(Process).err
requirements = ( (OpSys == "LINUX"))
executable = worker.sh ...
```

Repeated Random Sub-Sampling



Randomly drop 10% of samples. Remove unconnected species and *N*=1 events. Run model. Repeat 1,000 times (with replacement).

USGS CIDA Center for Integrated Data Analytics

Describe model performance by specific **cases**:

Species & Cut, State, Water body type, Site.

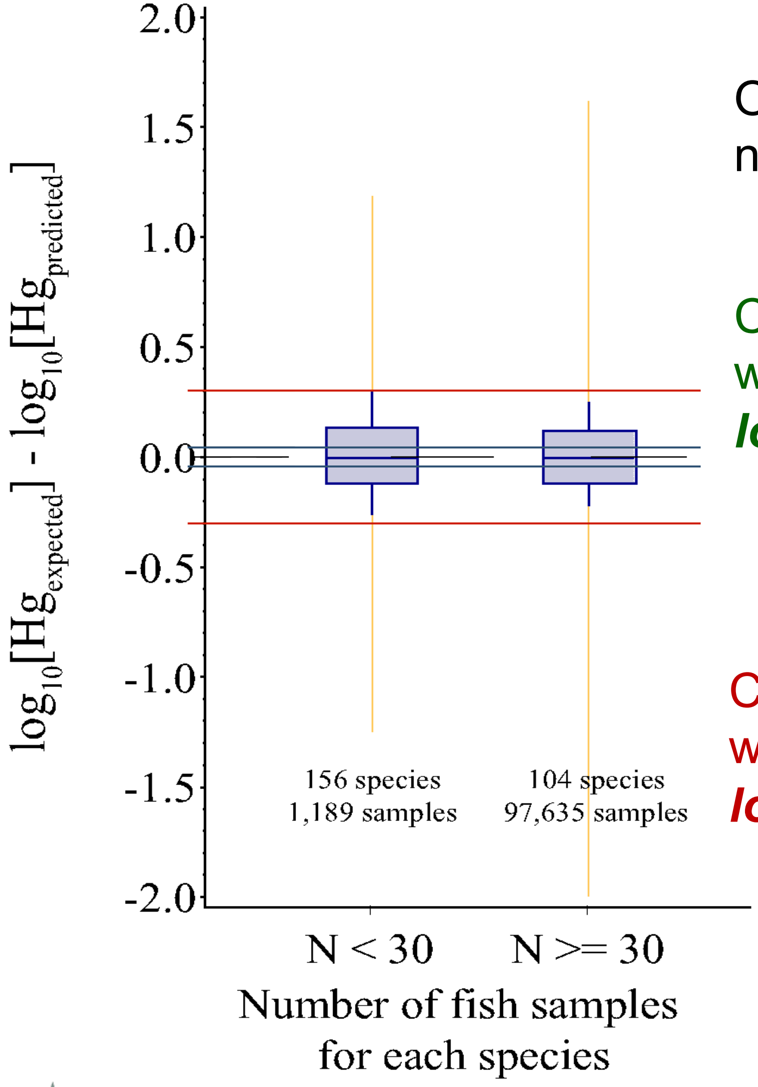
Prediction Errors:

$$PE = \log_{10}[Hg_{\mathit{expected}}] - \log_{10}[Hg_{\mathit{predicted}}]$$

Where:

$$[Hg_{\mathit{expected}}] = \left|\begin{array}{l} [Hg_{\mathit{observed}}] \text{ if detected} \\ \\ \text{inverse Mills ratio if} \\ [Hg_{\mathit{observed}}] < \text{reporting limit.} \end{array}\right.$$

Summary of Prediction errors: bias & variability



Overall, median PE ≈ 0, indicating no *bias*.

Cases where median [$Hg_{predicted}$] is within 10% of [$Hg_{expected}$]: ***low bias***

Cases where >80% of [$Hg_{predicted}$] are within 0.5x–2x of [$Hg_{expected}$]: ***low variability***

Conclusions

NDMMF incorporates greater inference space than traditional models

- Predicts for species & cuts that were not sampled in an event; traditional models cannot.

Validation shows low bias for all states and water-body types; and for most species & cuts and sites

- Some small-*N* cases showed larger bias. As more data are added, parameter estimates are better constrained.

HTCondor is essential to enable large-scale analysis

- It not only made our work more efficient, but enables science application that would otherwise not be possible

NDMMF Future Applications

More complete and efficient fish consumption advisories using a consistent national model

1×10^6+ samples from Environment Canada can be added to expand the range of applicability

Acknowledgements

Funding provided by

The USGS National Water Quality Assessment Program

Thanks to

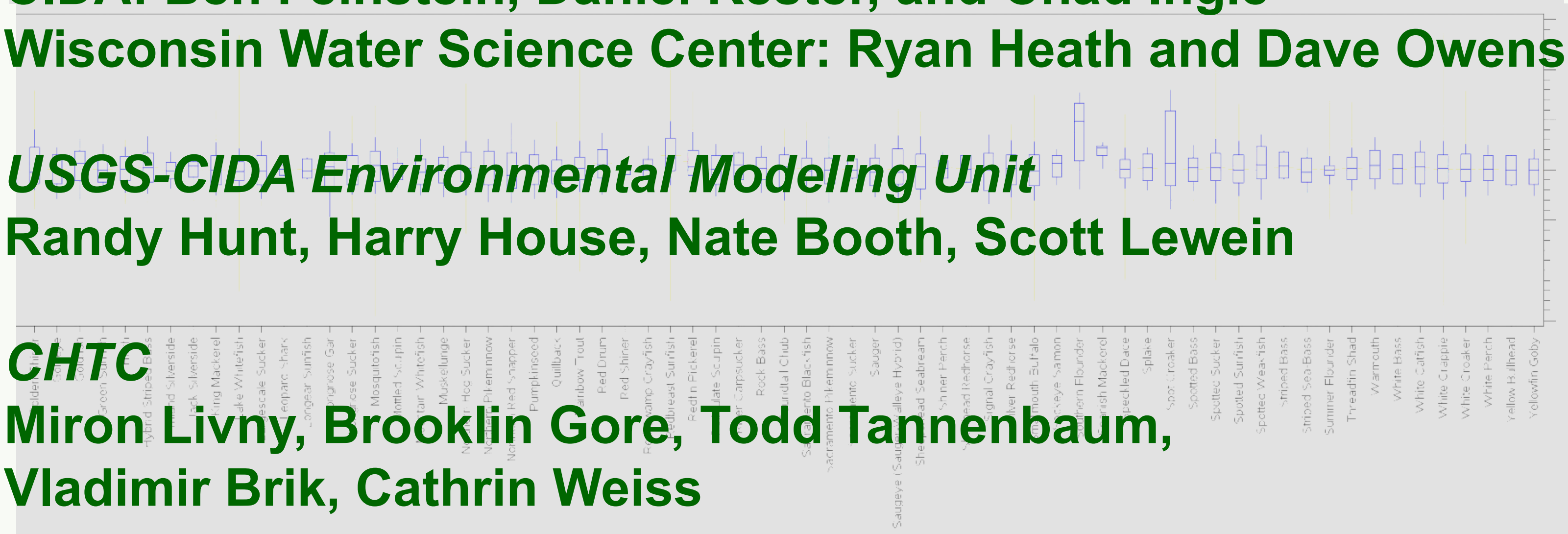
All the State and Local agencies for providing data

Steve Wente (USEPA) for writing the first version of NDMMF

IT support

CIDA: Ben Feinstein, Daniel Kester, and Chad Ingle

Wisconsin Water Science Center: Ryan Heath and Dave Owens



USGS-CIDA Environmental Modeling Unit

Randy Hunt, Harry House, Nate Booth, Scott Lewein

CHTC

Miron Livny, Brooklin Gore, Todd Tannenbaum,
Vladimir Brik, Cathrin Weiss