The effects of posterior sampling design on management procedure performance in MSE

Samuel D. N. Johnson
SAFS Quantitative Seminar
January 25th, 2019
samuelj@sfu.ca

Simon Fraser University
School of Resource and Environmental Management

Quantitative Fisheries Research Group

SFU

Acknowledgements

Fisheries and Oceans, Canada

Rob Kronlund
Jaclyn Cleary

Landmark Fisheries Research

Beau Doherty

Supervisory Committee

Sean Cox
Ashleen Benson

Management Strategy Evaluation

Modern fisheries have entered a management-oriented paradigm

Reviews in Fish Biology and Fisheries **8**, 349–356 (1998)

POINTS OF VIEW

Tidier fisheries management requires a new MOP (management-oriented paradigm)

WILLIAM K. DE LA MARE

Australian Antarctic Division, Channel Highway, Kingston, Tasmania, 7050, Australia. E-mail: bill_de@autdiv.gov.au

Achieving management outcomes	>	Fitting an unbiased stock assessment in any 1 year

"Pretty good management will do." - De La Mare 2006

"Minimum sustainable whinge." - Pope 1983

Modern fisheries have entered a management-oriented paradigm

MANAGEMENT STRATEGY EVALUATION - THE LIGHT ON THE HILL

A.D.M. Smith

CSIRO Division of Fisheries
GPO Box 1538
Hobart TAS 7001

Experiences in the evaluation and implementation of management procedures

D. S. Butterworth, and A. E. Punt

Butterworth, D. S., and Punt, A. E. 1999. Experiences in the evaluation and implementation of management procedures. – ICES Journal of Marine Science, 56: 985-998.

Design of operational management strategies for achieving fishery ecosystem objectives

Keith J. Sainsbury, André E. Punt, and
Anthony D. M. Smith

Sainsbury, K. J., Punt, A. E., and Smith, A. D. M. 2000. Design of operational management strategies for achieving fishery ecosystem objectives. – ICES Journal of Marine Science, 57: 731–741.

MSE helps us choose management procedures under uncertainty

- At its heart, MSE is a risk analysis, asking “What are the consequences of a certain decision, given the current state of knowledge about a system
- Management Strategy Evaluation uses closed loop simulation to test candidate management procedures (decisions) under system uncertainty (operating models)
- Management procedures are combinations of data collection, assessment methods, and harvest control rules
- MPs are evaluated against quantitative management objectives

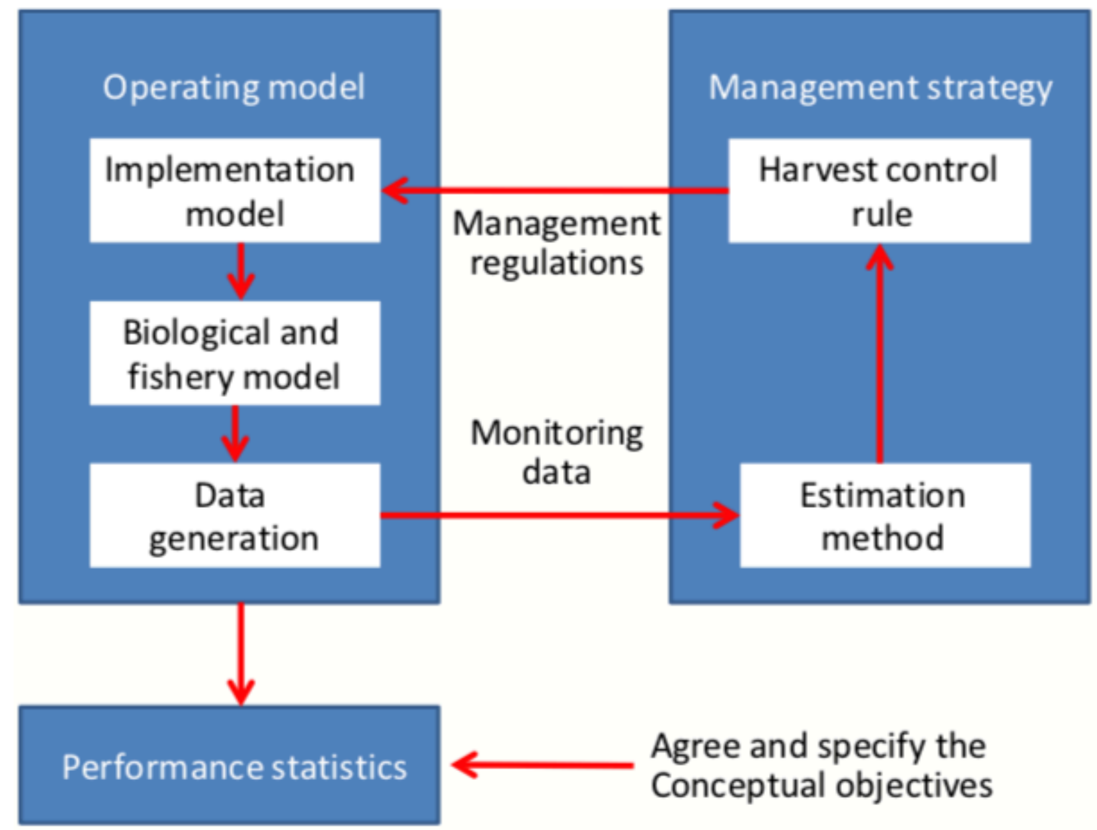
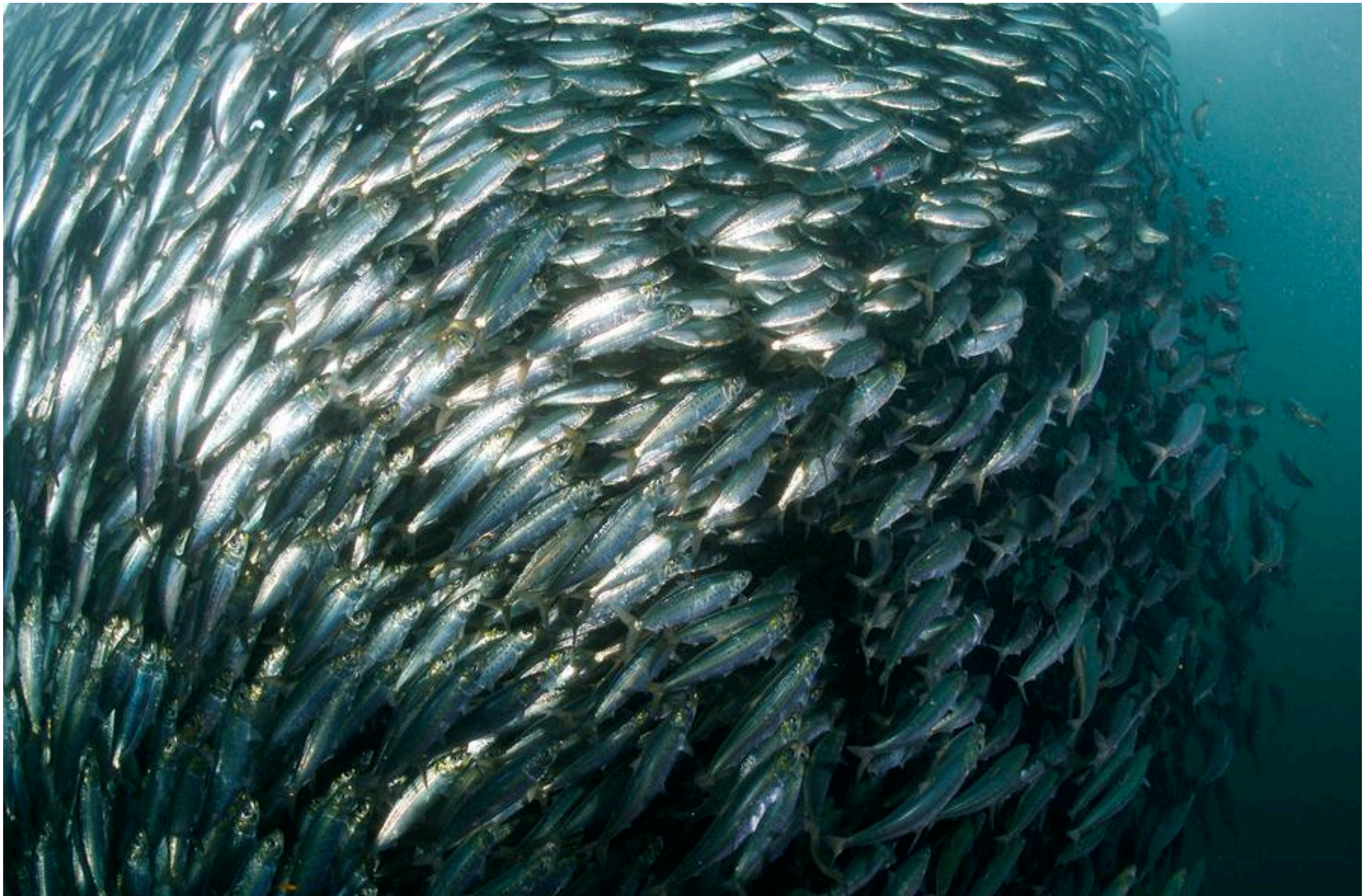


Figure 1 Conceptual overview of the management strategy evaluation modelling process.

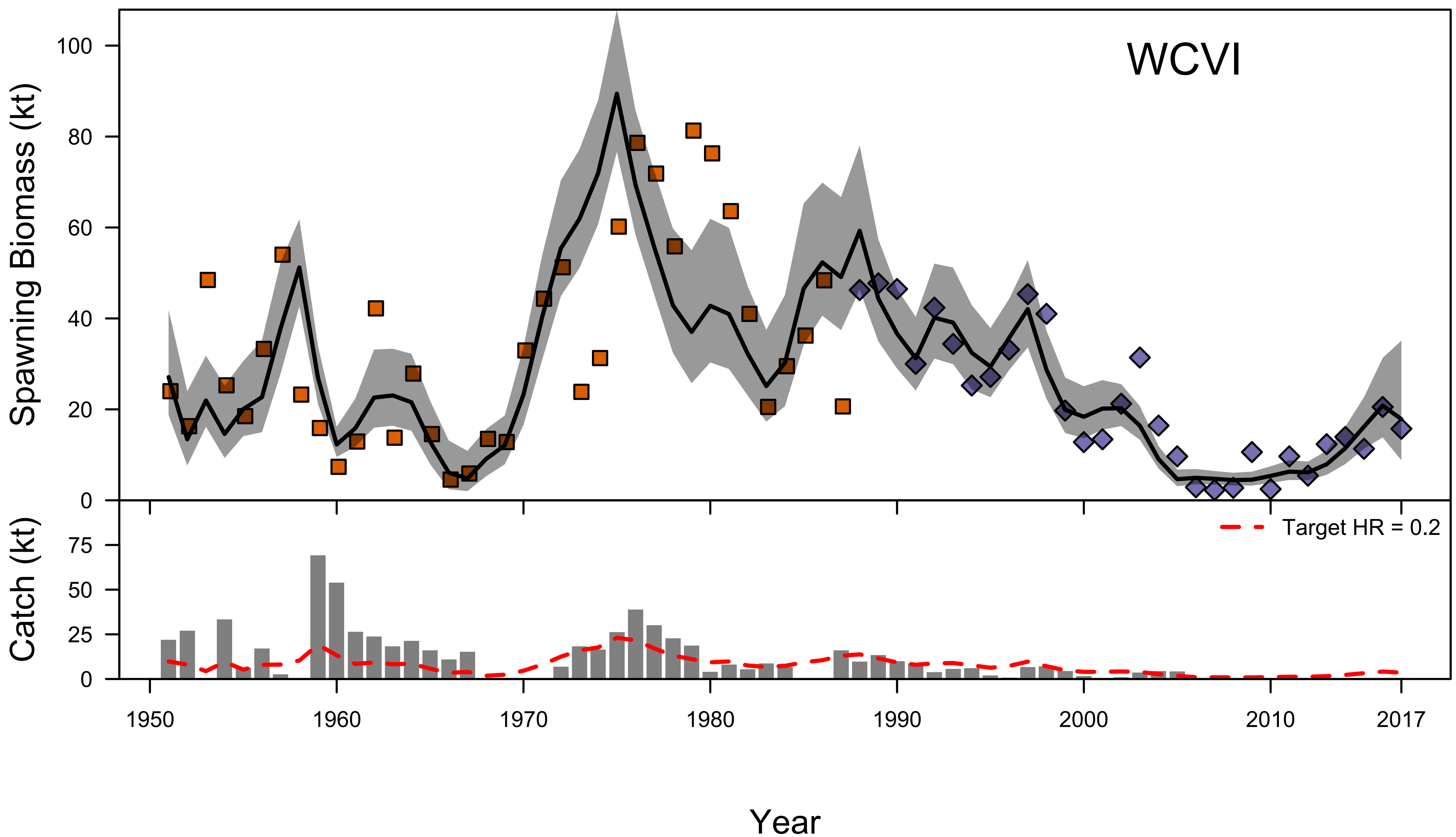
Punt et al 2016

Example Decision Context: West Coast of Vancouver Island Herring

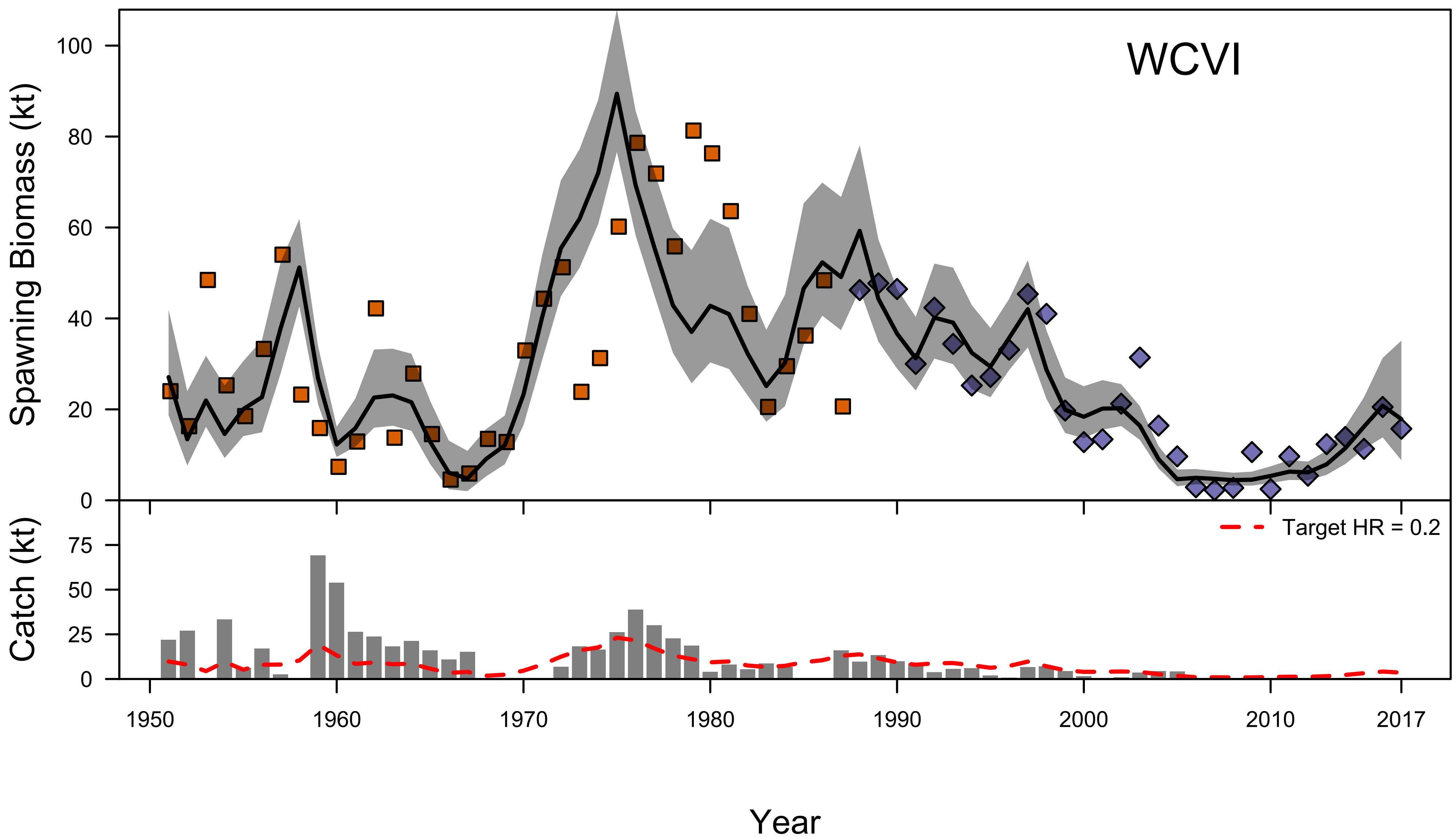


Example Decision Context: West Coast of Vancouver Island Herring

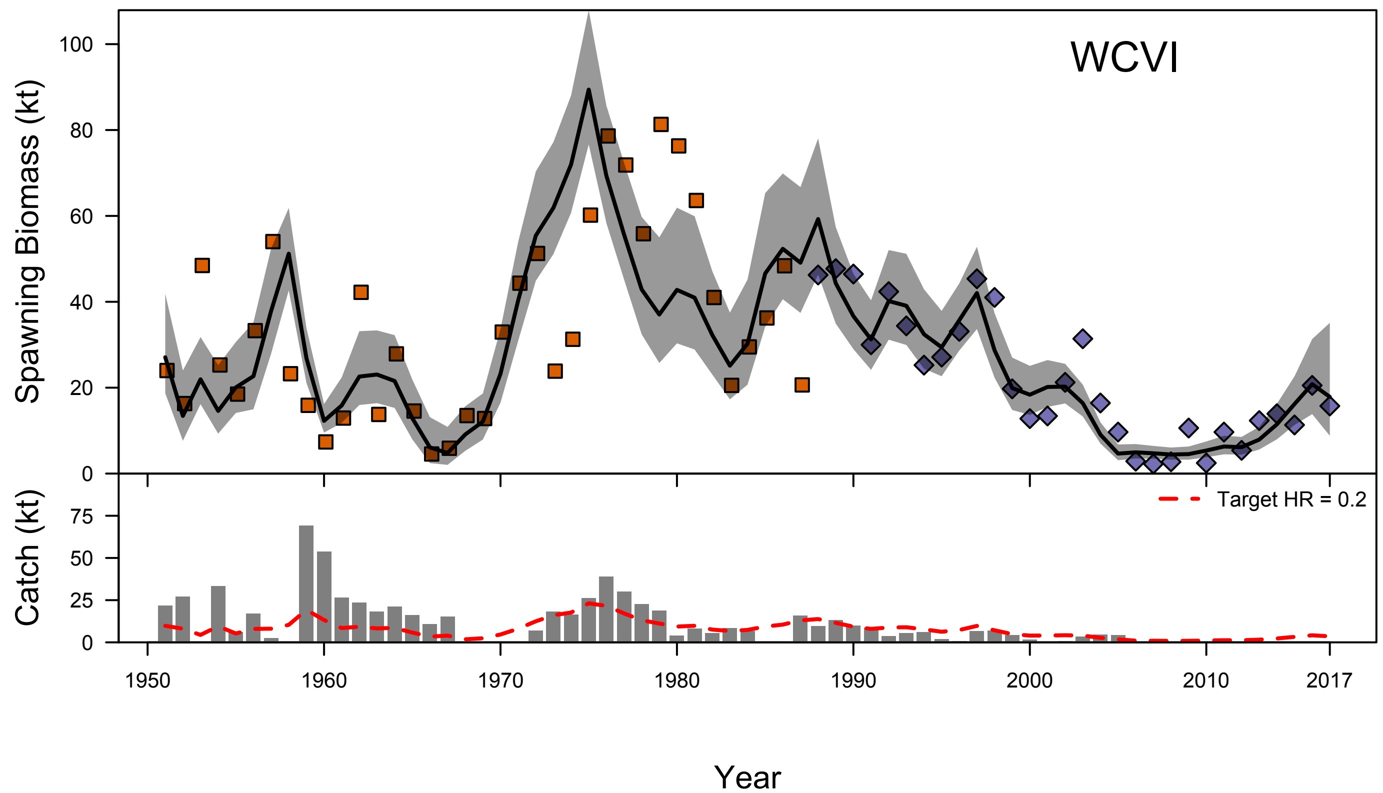
Example Decision Context: West Coast of Vancouver Island Herring



Example Decision Context: West Coast of Vancouver Island Herring

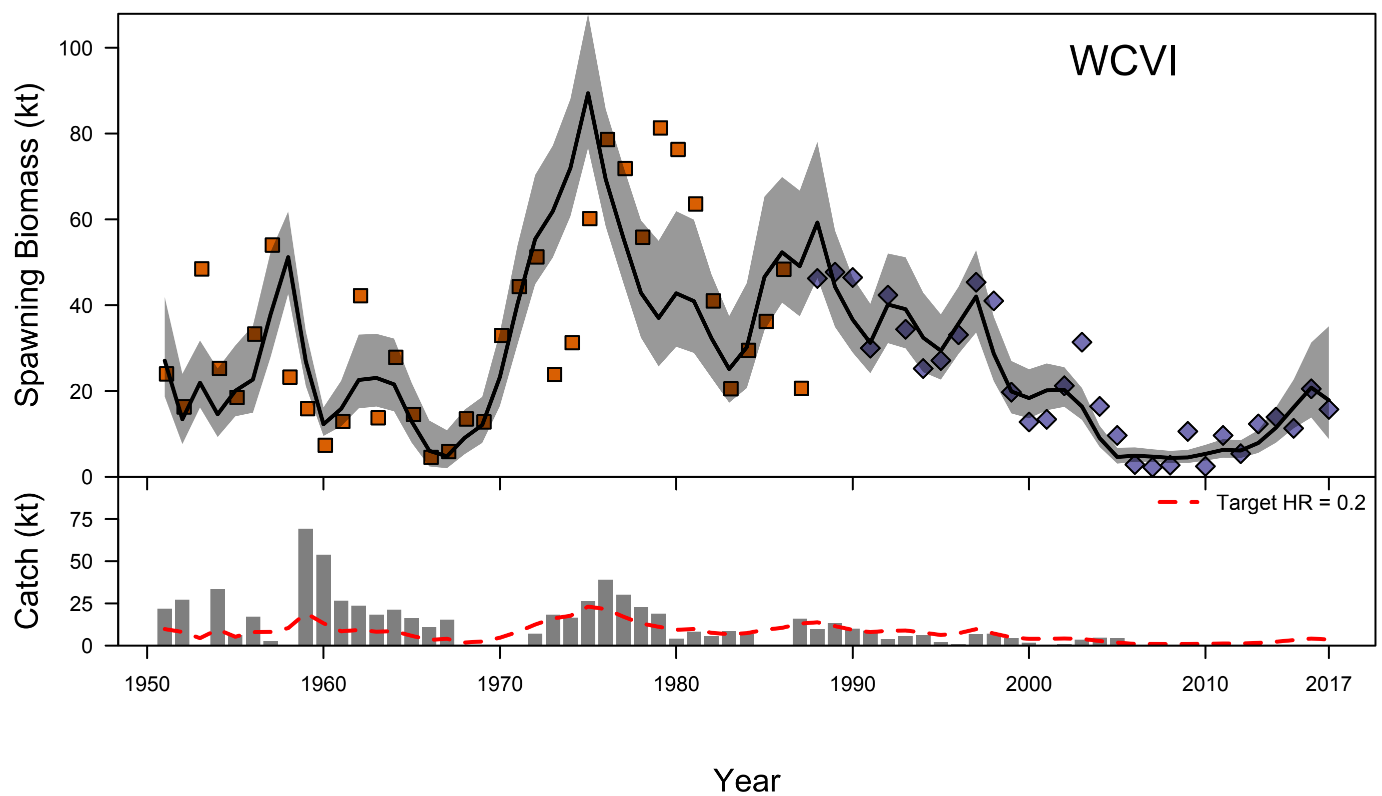


Example Decision Context: West Coast of Vancouver Island Herring



Objective:
Keep biomass above limit reference point at least 80% of the time over next 15 years

Example Decision Context: West Coast of Vancouver Island Herring



Objective:
Keep biomass above limit reference point at least 80% of the time over next 15 years

Example procedure:

- 5 year moving average of spawn index,
- 10% harvest rate
- Typical ramped control rule

Conditioning the operating model: implications for management outcomes

Operating models should be conditioned on the data

- "Operating model components… must be conditioned on the available data… so that model predictions of the data are consistent with actual data." ***(Kell et al 2006)***

- "Operating model parameters are selected (ideally by fitting or 'conditioning' the operating model(s) to data from the actual system under consideration)" ***(Punt et al 2016)***

Best practices for conditioning OM parameters on data

Fit at least one OM to the data, and choose one of the following, in order from most to least ideal *(Punt et al 2016)*

1. **Bayesian**: Produce a Bayesian posterior via your favourite MCMC method
2. **Bootstrap**: Bootstrap estimated observation and process errors, refit the OM and generate distributions of leading parameters
3. **Normal Approx**: Use covariance matrix produced by optimisation and draw from a normal approximation of the posterior

Best practices for conditioning OM parameters on data

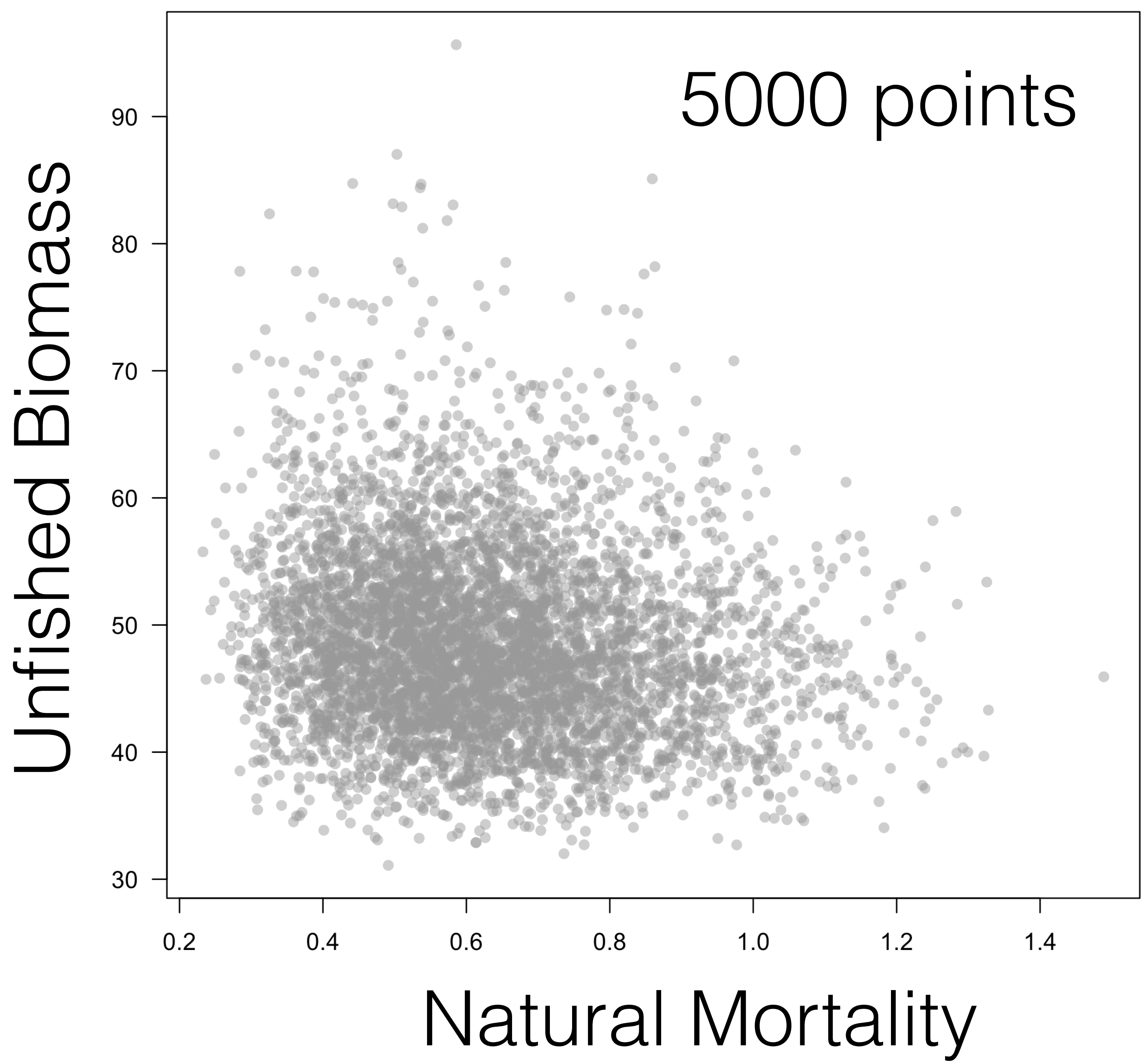
Fit at least one OM to the data, and choose one of the following, in order from most to least ideal *(Punt et al 2016)*

1. **Bayesian**: Produce a Bayesian posterior via your favourite MCMC method

2. **Bootstrap**: Bootstrap estimated observation and process errors, refit the OM and generate distributions of leading parameters

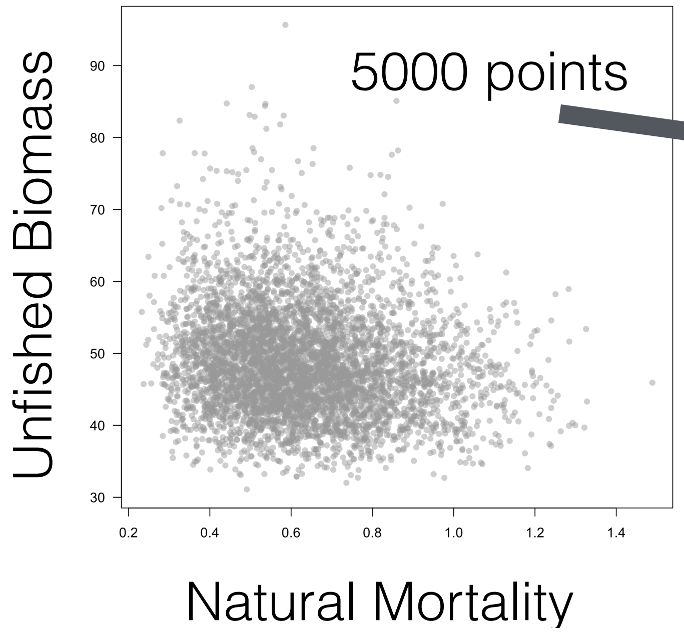
3. **Normal Approx**: Use covariance matrix produced by optimisation and draw from a normal approximation of the posterior

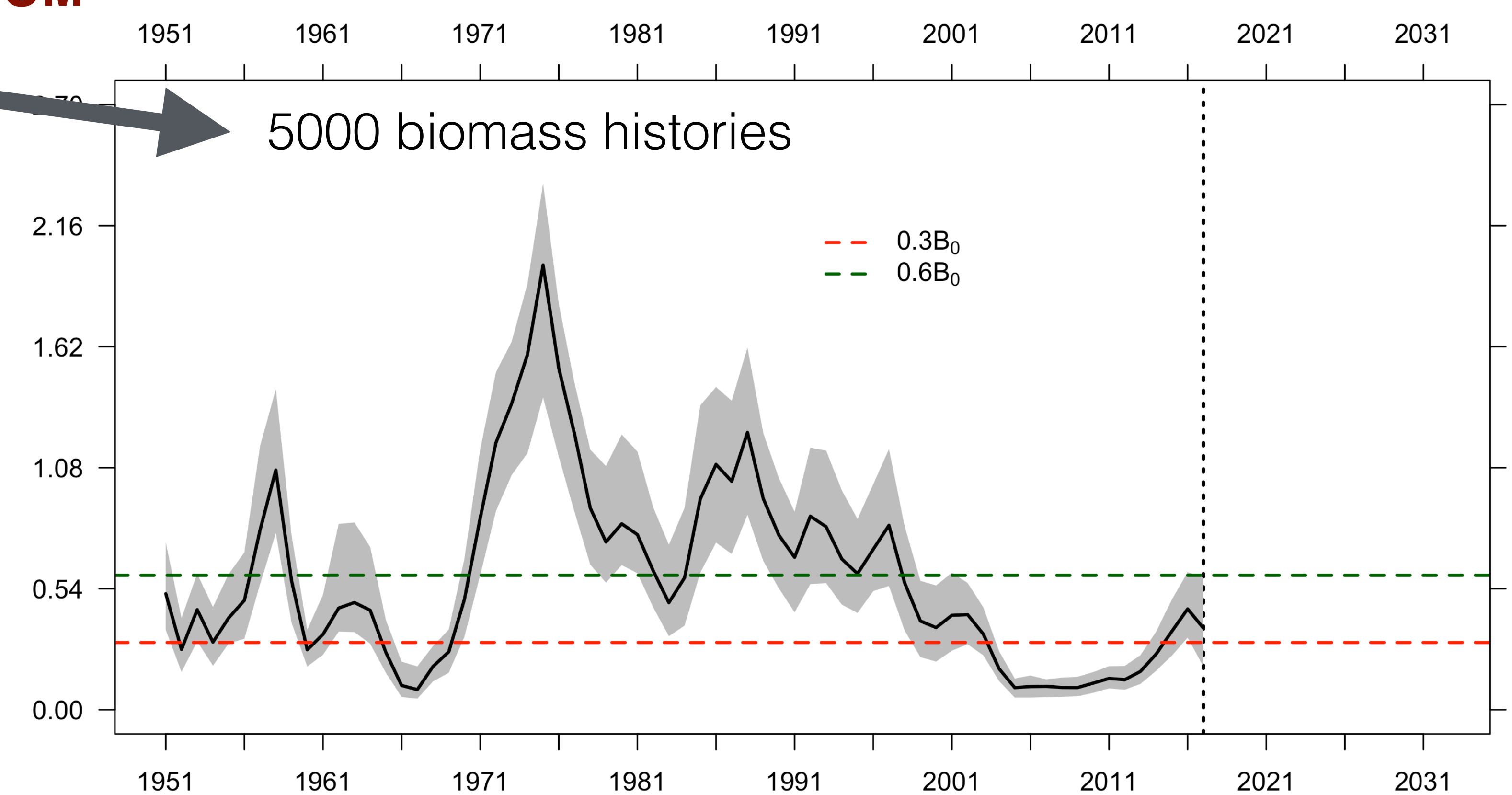
A quick description of the path from posterior to objective performance metric



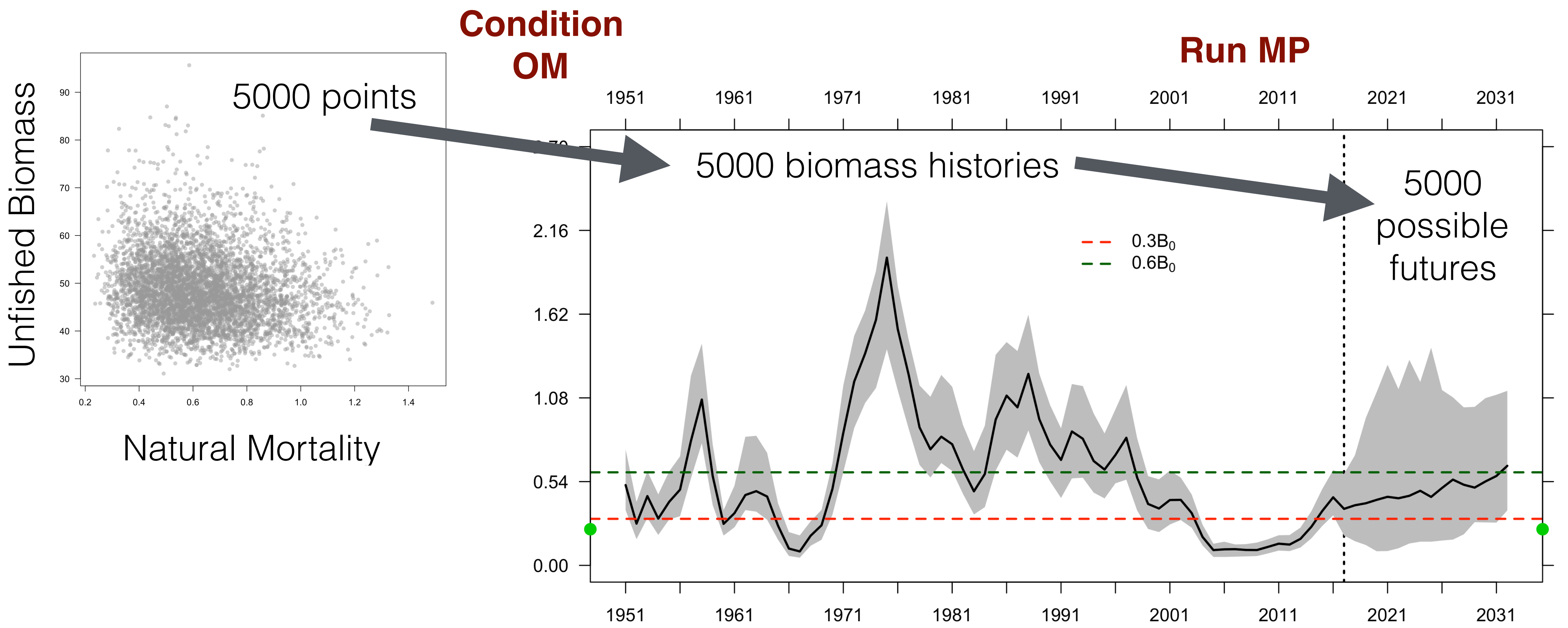
A quick description of the path from posterior to objective performance metric

Condition OM

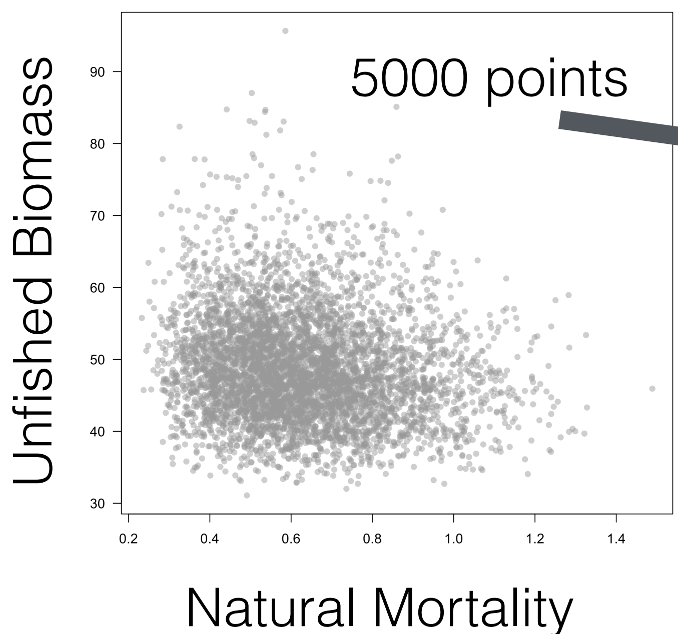




A quick description of the path from posterior to objective performance metric

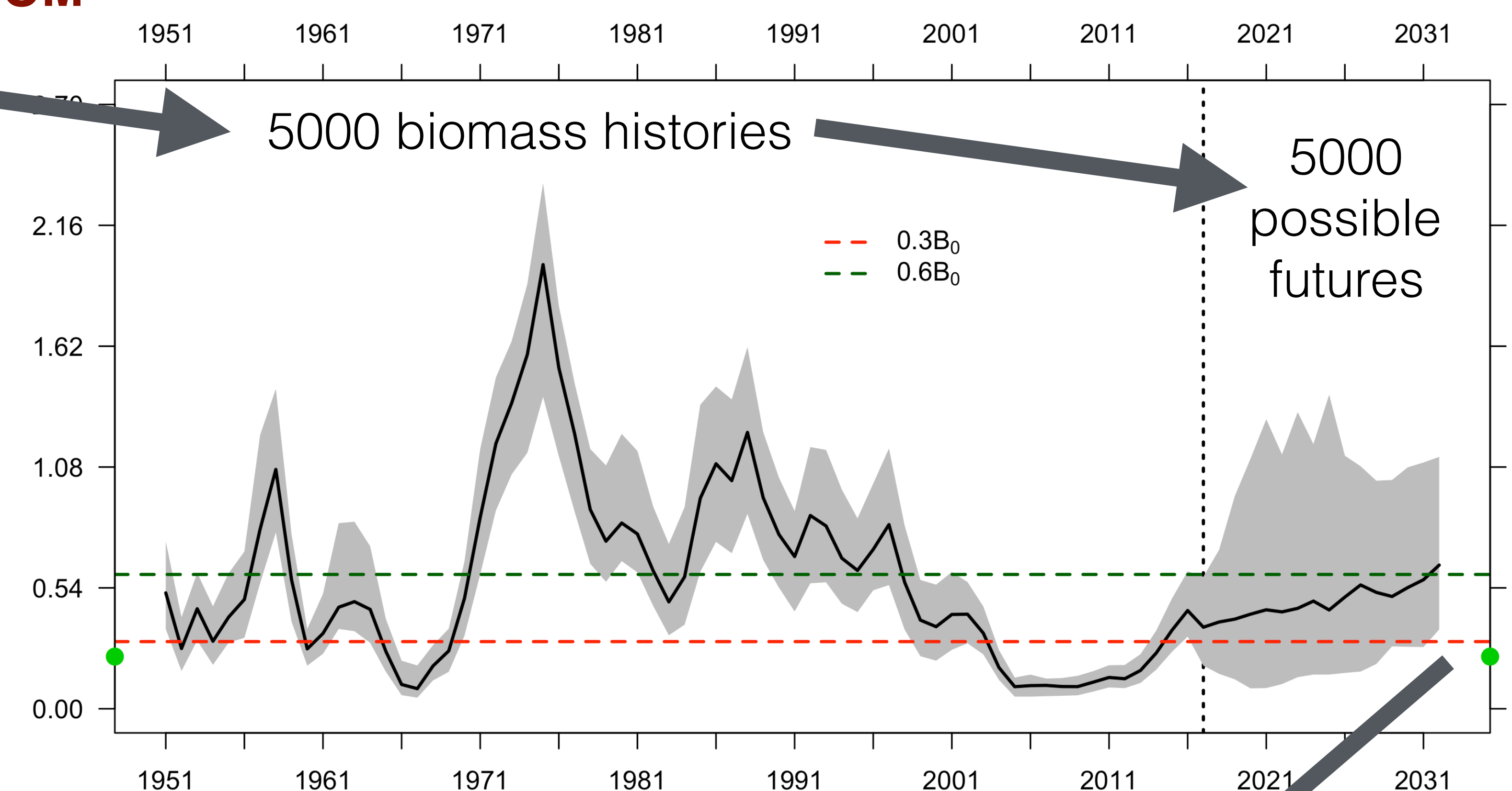


A quick description of the path from posterior to objective performance metric



Condition OM

Run MP

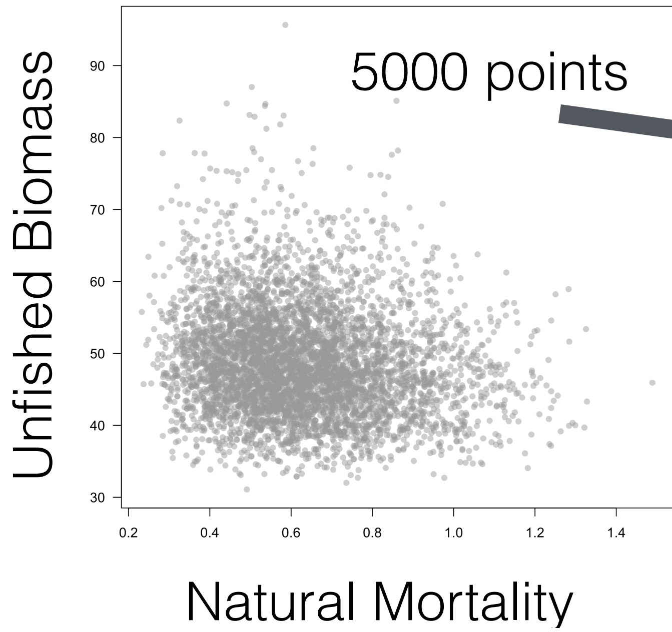


Objective:
Keep biomass above limit reference point at least 80% of the time over next 15 years

Calculate metric

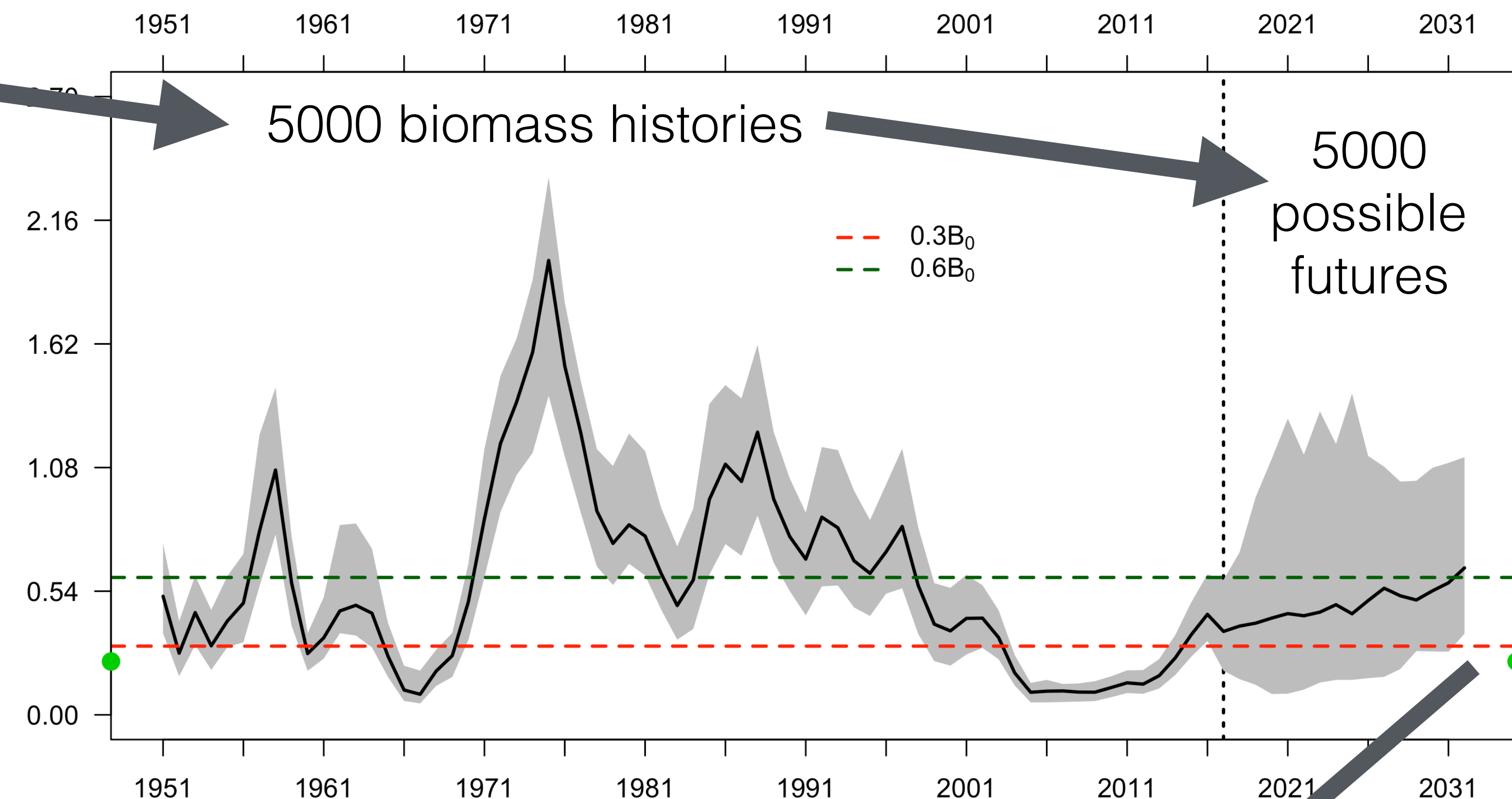
$$P(B_t > B_{lim}) = 81\%$$

A quick description of the path from posterior to objective performance metric



Condition OM

Run MP



Calculate metric

$$P(B_t > B_{lim}) = 81\%$$

Objective:
Keep biomass above limit reference point at least 80% of the time over next 15 years

PASSED

MSE models must run fast enough to provide timely advice

- MSE work is by its nature iterative, and within a single cycle models are potentially fit thousands of times
- Analysts learn as each iteration brings new information, leading to changes (tweaks) in OMs and MPs
- 5000 replicates takes time!
  - For these data based MPs, 6 hours
  - For the model based MP tested in the last MSE cycle, 16 hours for 100 reps **(=> 800 hours for 5000!)**

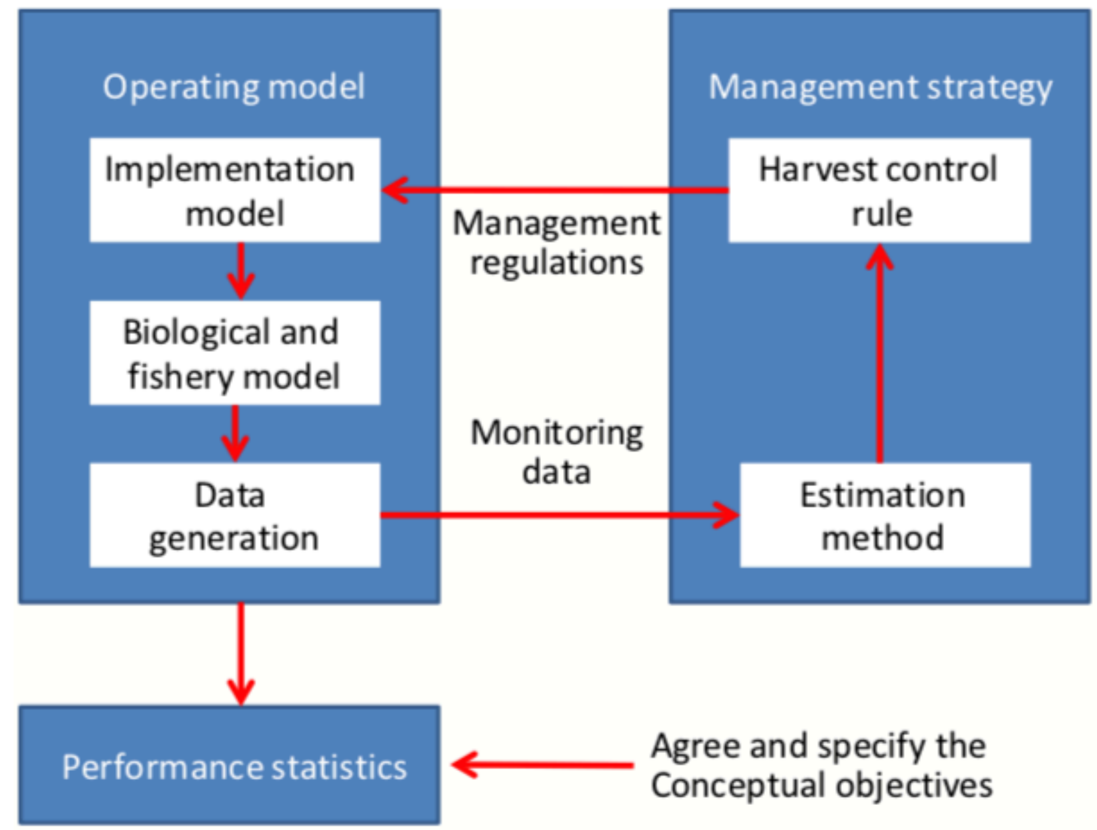
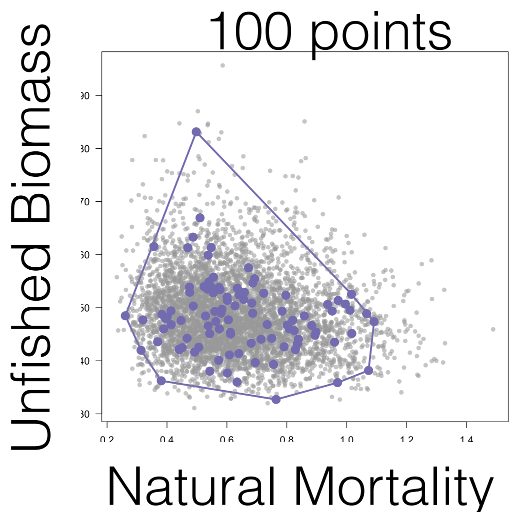
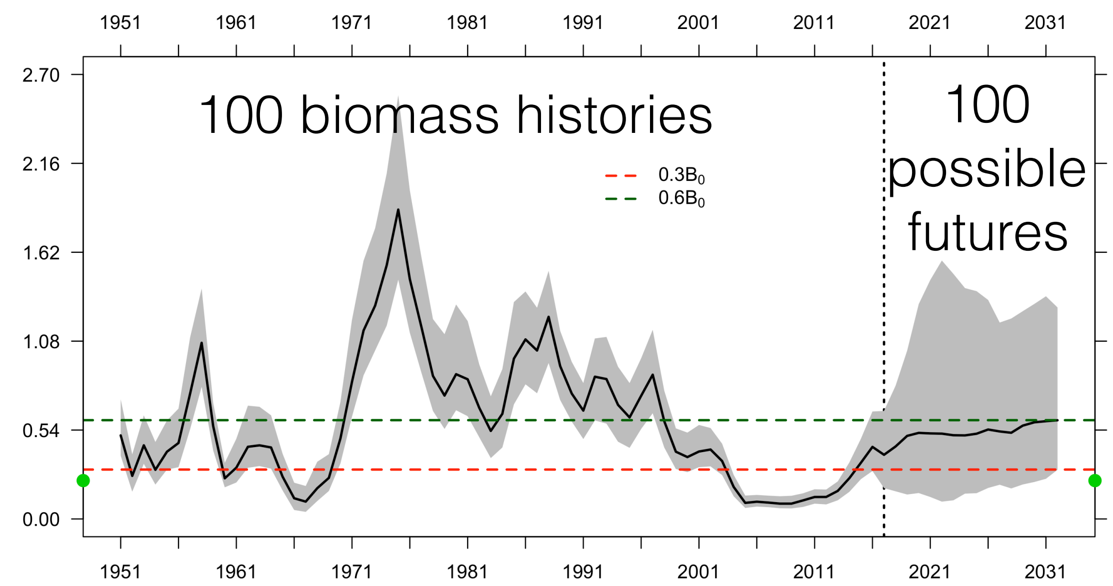


Figure 1 Conceptual overview of the management strategy evaluation modelling process.

Punt et al 2016

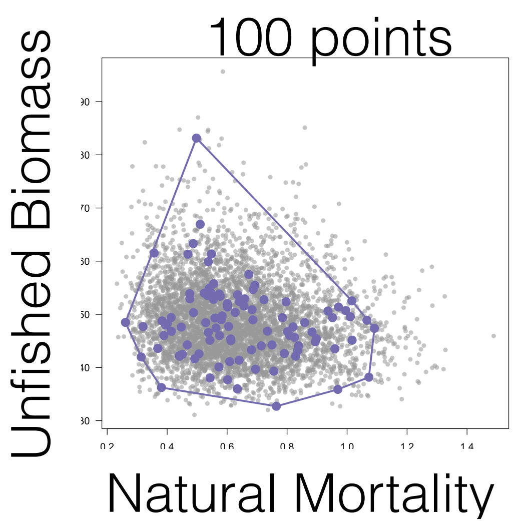
Sampling from the posterior is used to reduce the number of replicates required

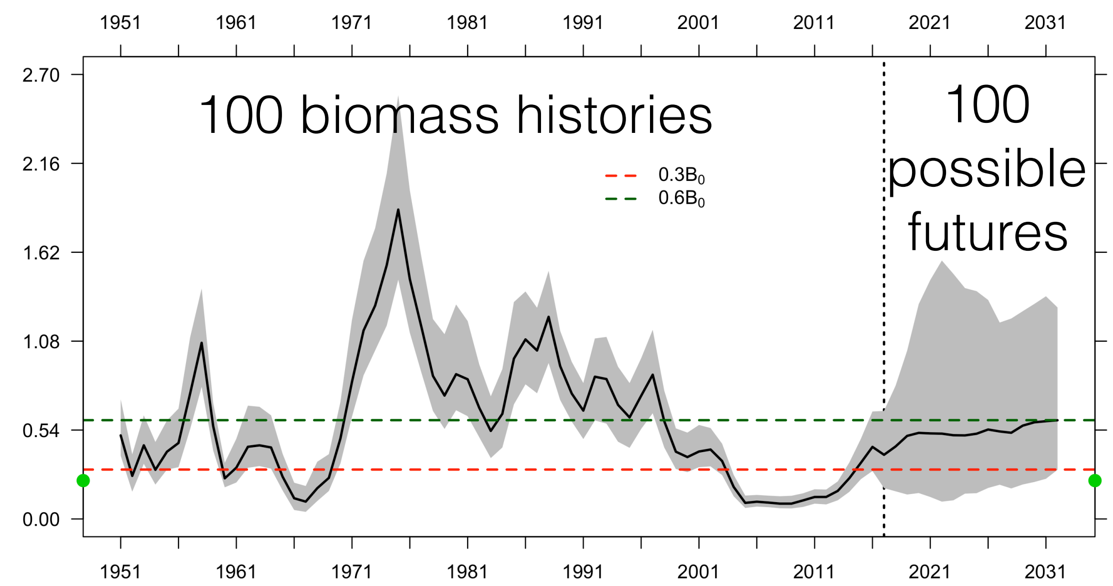




$$P(B_t > B_{lim}) = 81.5\%$$

Sampling from the posterior is used to reduce the number of replicates required

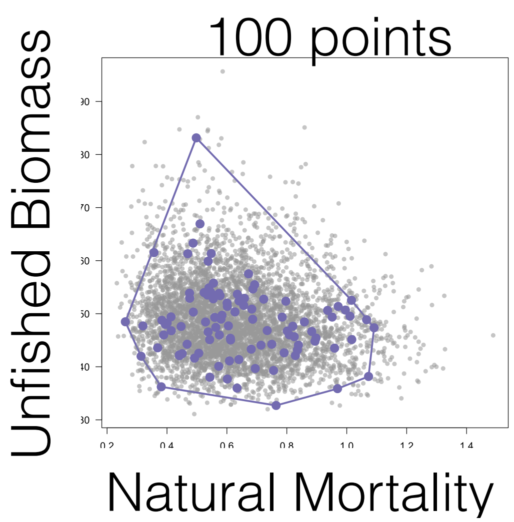


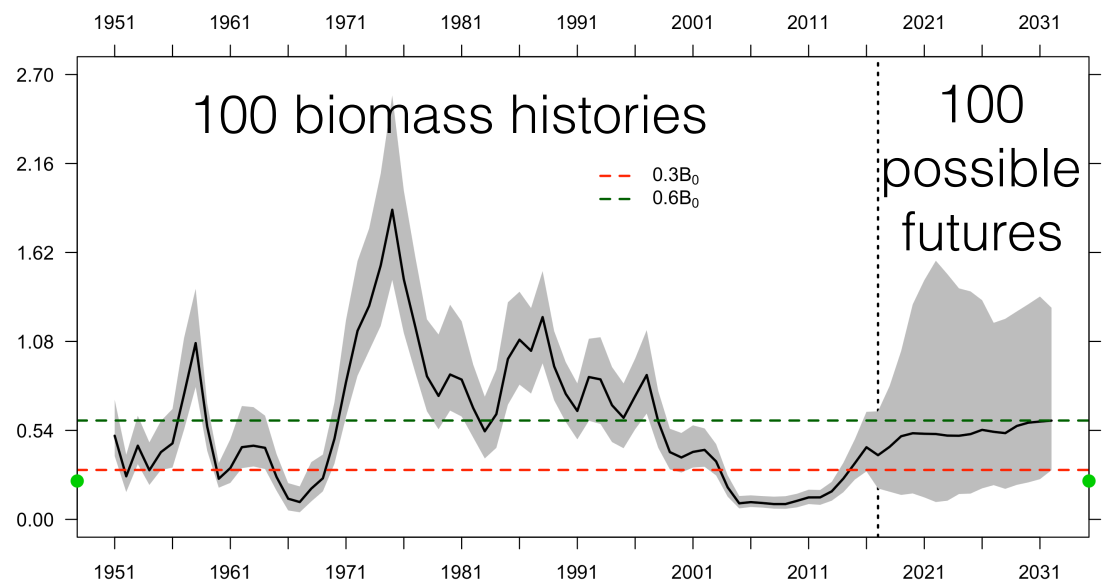


$P(B_t > B_{lim}) = 81.5\%$

PASSED

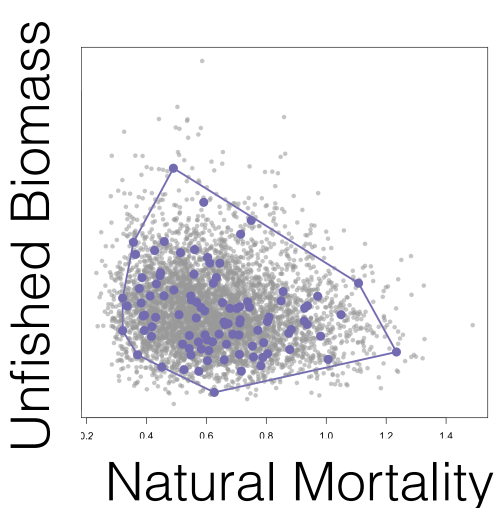
But performance metrics are sensitive to the random seed used to take samples

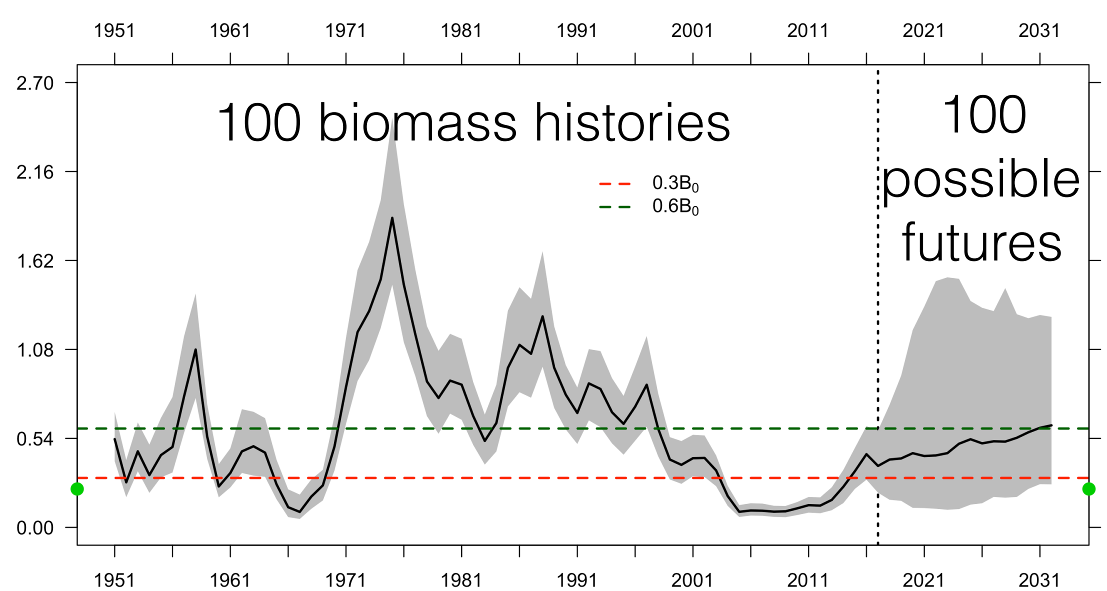




PASSED

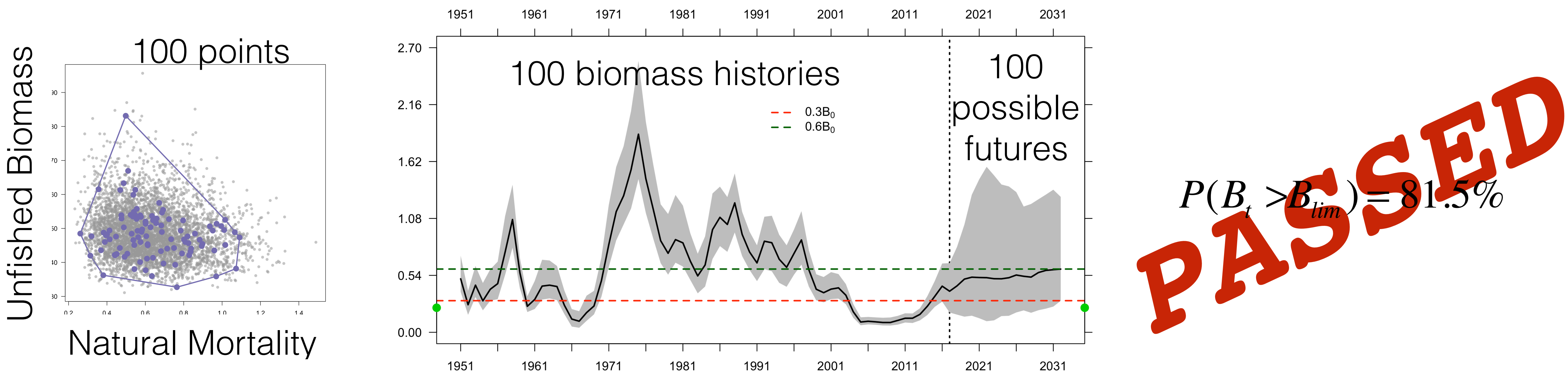
$P(B_t > B_{lim}) = 81.5\%$

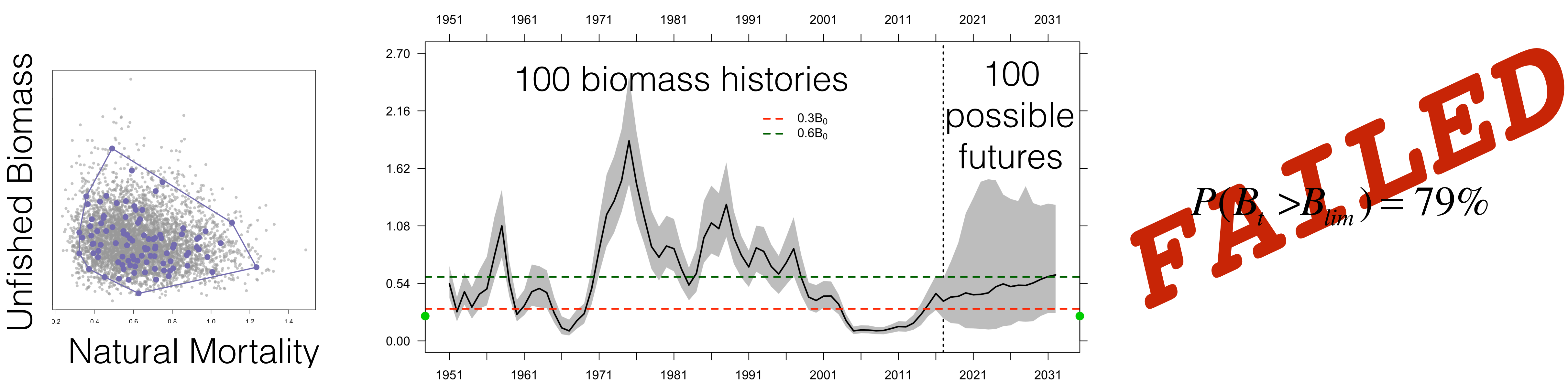




$P(B_t > B_{lim}) = 79\%$

Sampling from the posterior is used to reduce the number of replicates required





Conditioning may affect the final choice of MP

To summarise the previous slide, MP objective performance metrics are *random variables*

They are drawn from a distribution that is conditioned on

1. The data, via the posterior,

2. The method used to sample posterior parameter distributions, and

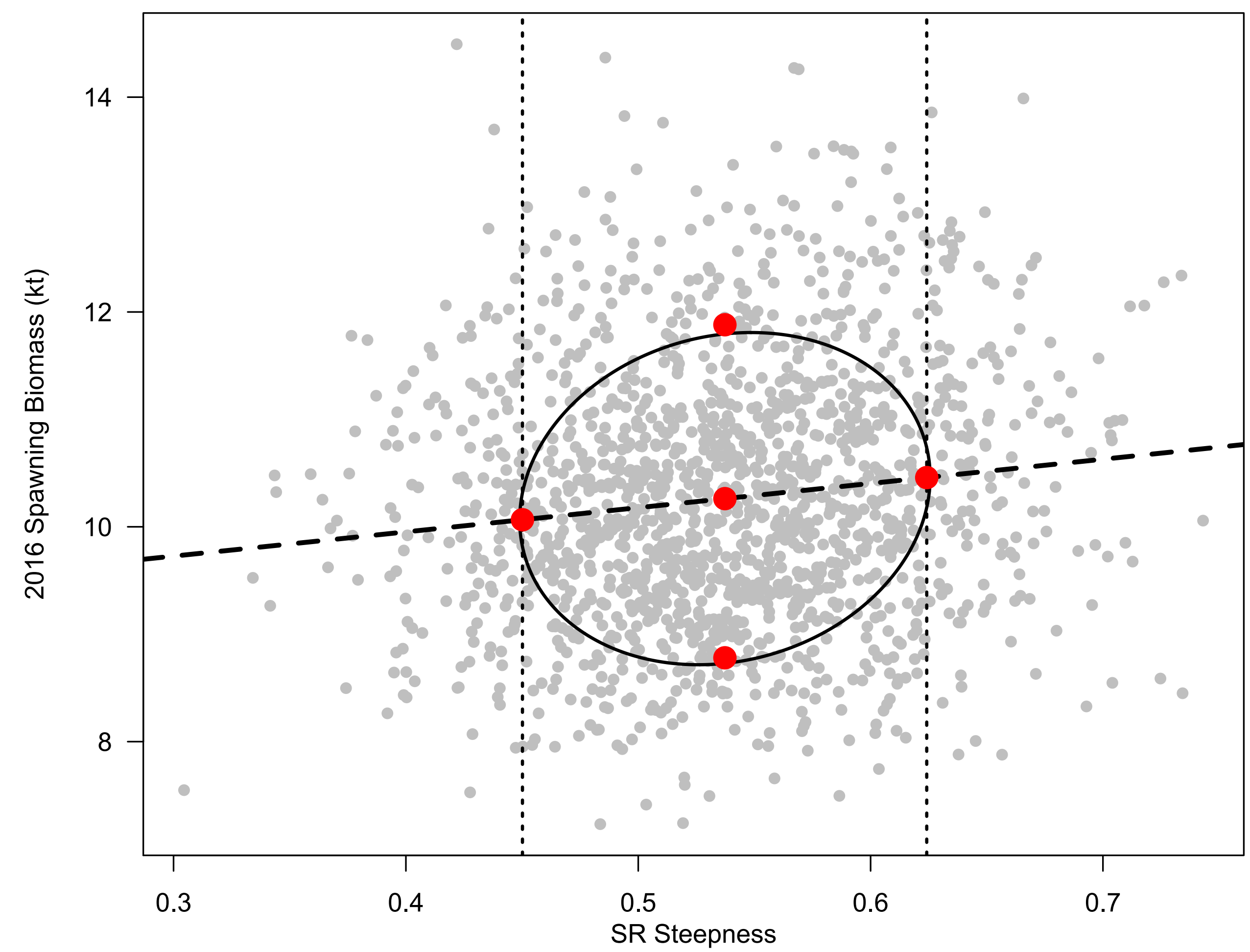
3. Process and observation errors in the projections

Differences in performance metrics caused by these three effects could expand or shrink the pool of acceptable MPs

Our previous attemps to improve sampling

Sablefish: weighted sampling of the joint marginal of two posterior dimensions

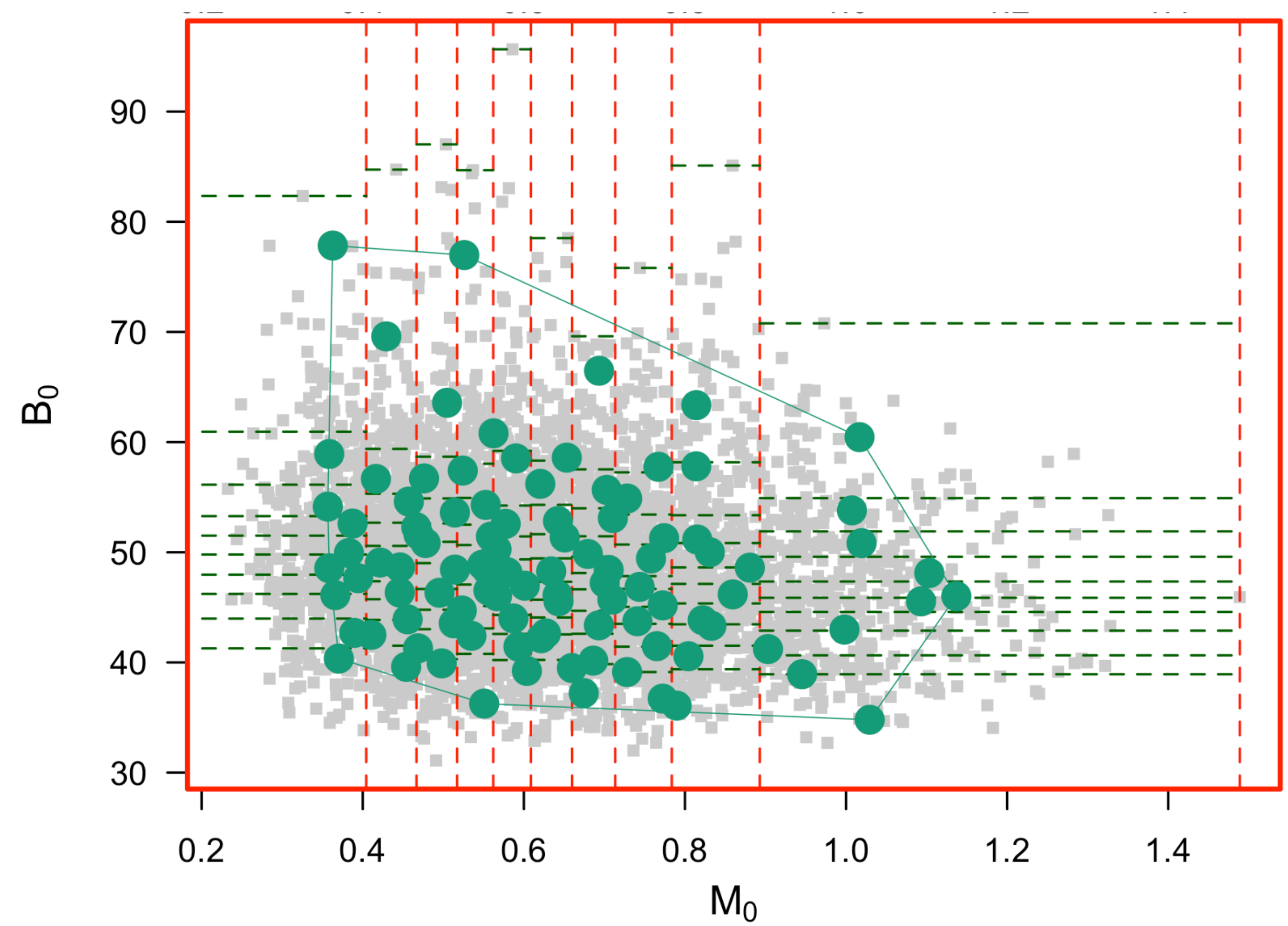
- We sampled 5 points from the Bayesian posterior to create 5 OMs
    - posterior mean **(ref OM)**
    - 10th and 90th precentiles of the marginal B_{2017} and steepness distributions **(optimistic and pessimistic OMs)**
- Ran 100 replicates at each point
- Sampled each OM weighted by relative posterior density at corresponding points
- We weren't sure if this made a large difference or not - *which is why we started this work*



BC Sablefish, 2016
Cox, Holt and Johnson, CSAS In Press

WCVI Herring: stratified sampling of joint marginal posterior into conditional centiles.

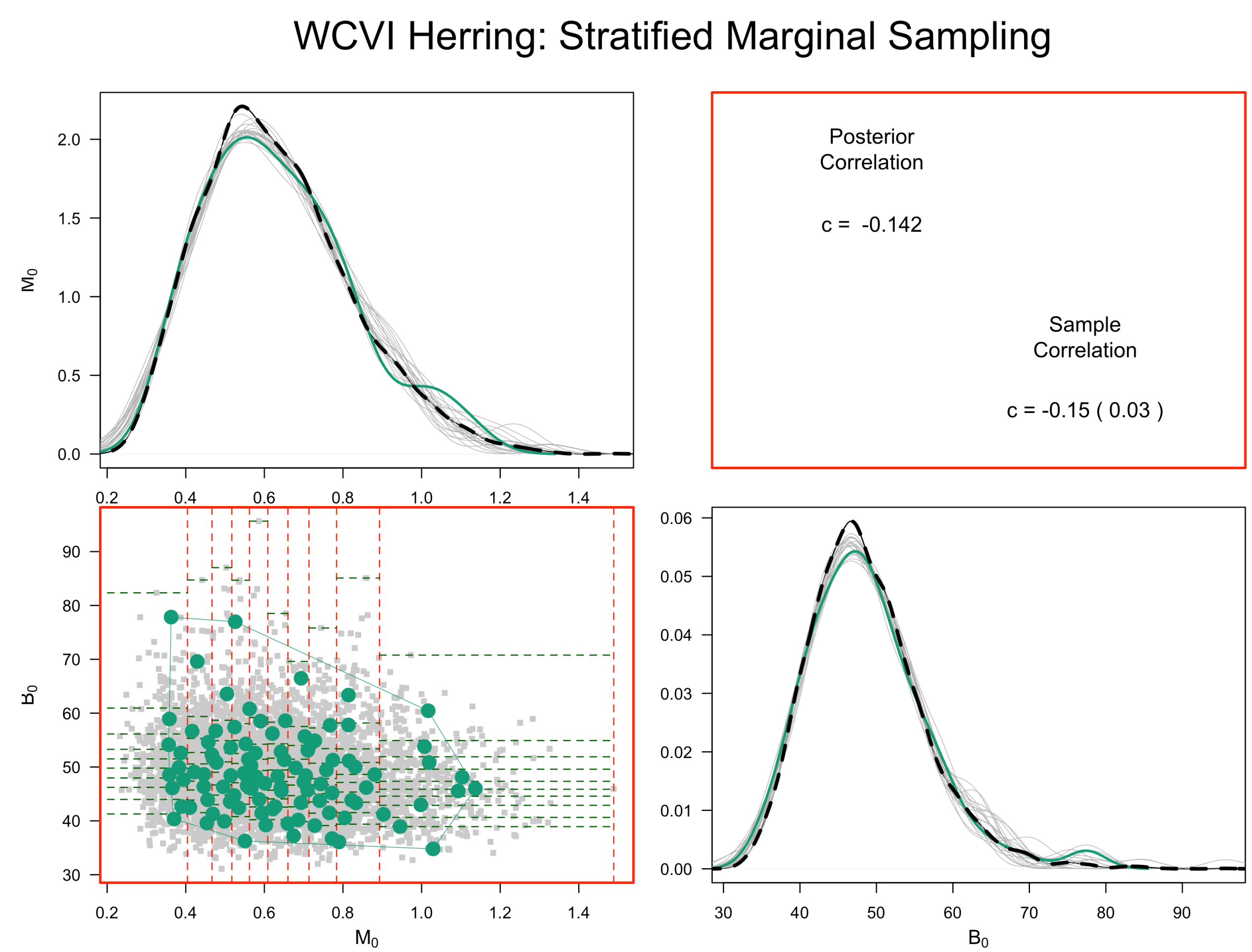
- Noticed sensitivity to seed values when conditioning the operating model
- Designed a conditional stratified sampling design which broke joint marginal posterior of 2 dimensions into centiles
- Stratified joint marginal of M and B0
- Randomly sampled 1 point within each centile



WCVI Herring, 2018
Cox, Johnson, Cleary, and Benson, CSAS In Press

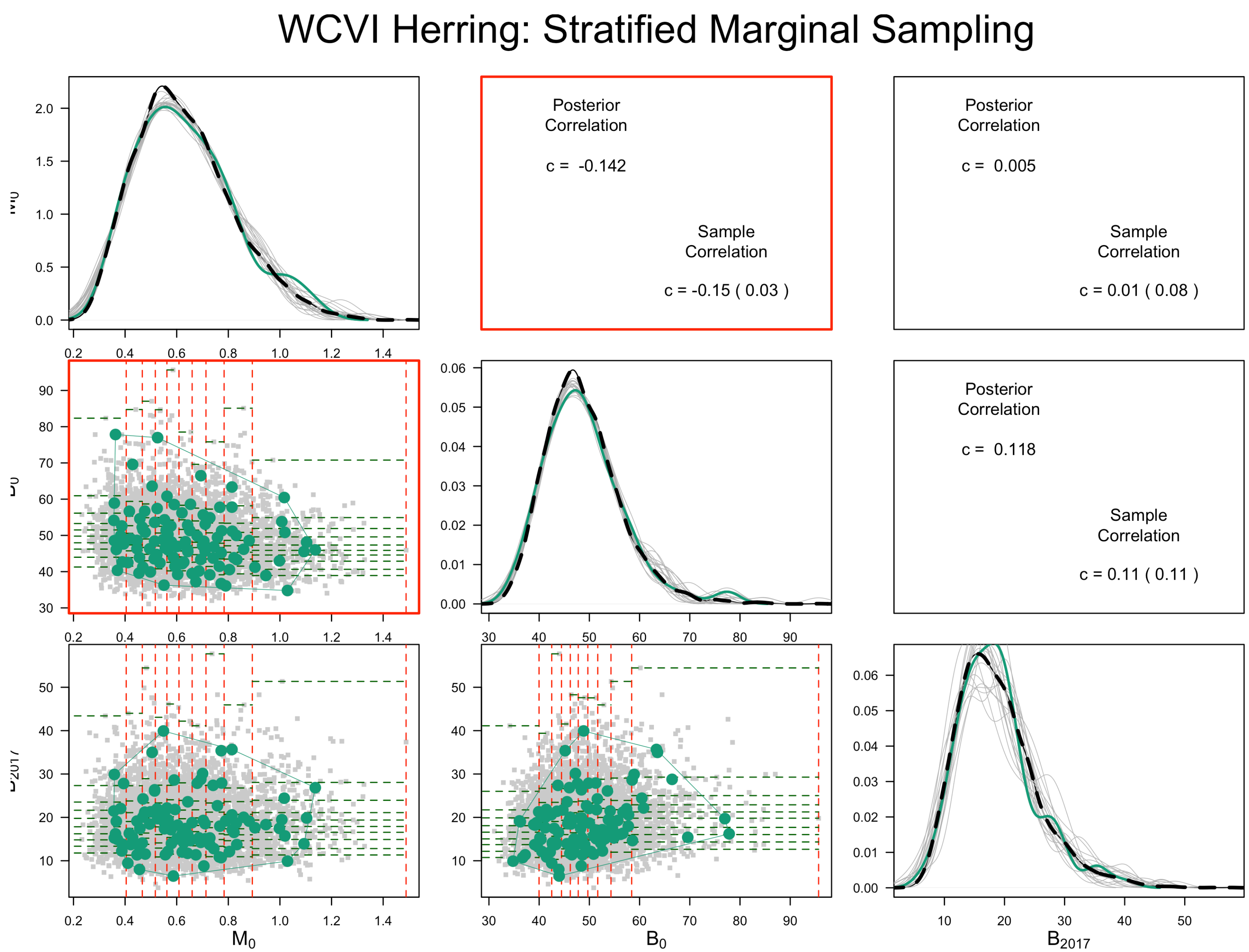
Stratification of joint marginal posterior into conditional centiles.

- Noticed sensitivity to seed values when conditioning the operating model
- Designed a conditional stratified sampling design which broke joint marginal posterior of 2 dimensions into centiles
- Stratified joint marginal of M and B0
- Randomly sampled 1 point within each centile



WCVI Herring, 2018
Cox, Johnson, Cleary, and Benson, CSAS In Press

Stratification of joint marginal posterior into conditional centiles.



WCVI Herring, 2018
Cox, Johnson, Cleary, and Benson, CSAS In Press

Testing different sampling designs

Best practices for conditioning OM parameters on data

Fit at least one OM to the data, and choose one of the following, in order from most to least ideal *(Punt et al 2016)*

1. **Bayesian**: Produce a Bayesian posterior via your favourite MCMC method

2. **Bootstrap**: Bootstrap estimated observation and process errors, refit the OM and generate distributions of leading parameters

3. **Normal Approx**: Use covariance matrix produced by optimisation and draw from a normal approximation of the posterior

Conditioning may affect the final choice of MP

- To summarise the previous slide, MP objective performance metrics are *random variables*

- They are drawn from a distribution that is conditioned on

    1. The data, via the posterior,

    2. **The method used to sample posterior parameter distributions,** and

    3. Process and observation errors in the projections

- Differences in performance metrics caused by these three effects could expand or shrink the pool of acceptable MPs

Research questions: can we define some best practices for taking samples to condition the OM?

1. What sampling methods are best suited to reducing the sensitivity of the objective performance metrics to sample size and random seed?

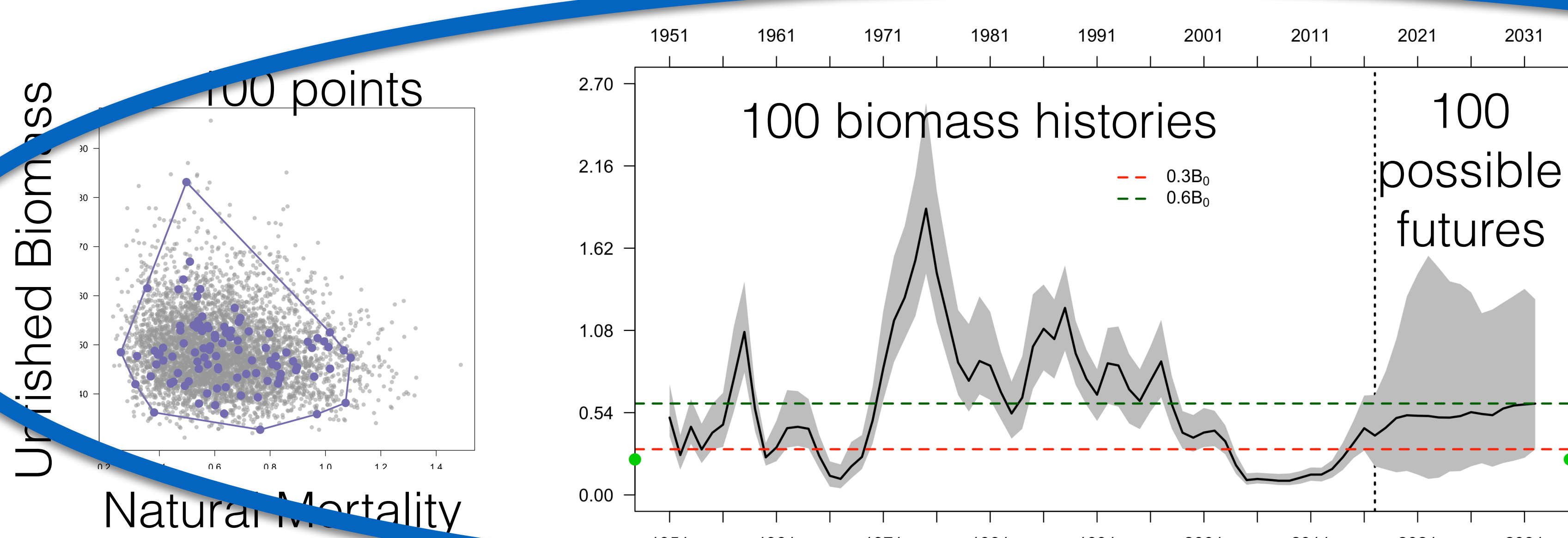
2. What is the minimum number of samples required to

   A. fix the ranking of MPs for each sampling method?

   B. reduce variance of metrics within some tolerance?

3. What qualities can be used to identify a good sample before significant time is spent in simulations, i.e. filter samples so that variance of metrics is within some tolerance?

We're fisheries scientists. We do simulation experiments.



$$P(B_t > B_{lim}) = 81\%$$

x100

Varying

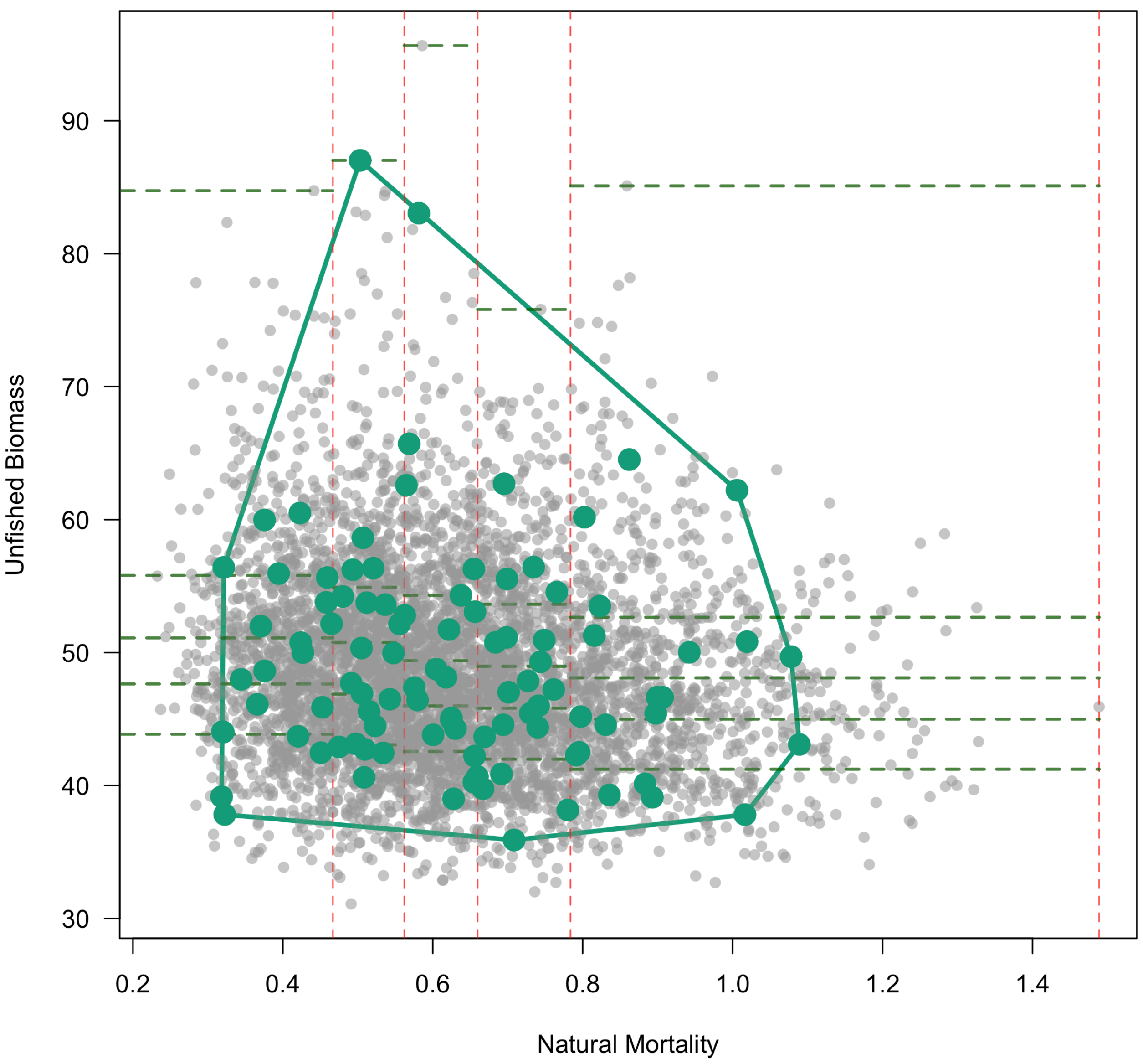
- Sample Size (25, 50, 100, 200, 500, 1000)
- Sampling design - simple random sampling plus two stratified methods

Other sampling methods: joint marginal stratification - jmarg

(McKay, Beckman, and Conover 2000)

This is the same method as used in the previous Herring MSE, except:

- Now splitting into 25 equal density regions - 5x5
- Depending on sample size, may take multiple draws from each cell
- Shown: 100 samples



Other sampling methods: latin hypercube samping (space filling design) - LHS

(McKay, Beckman, and Conover 2000)

Latin hypercube sampling exploits the properties of a latin square - think Sudoku

Every entry appears in each row and column exactly once.

Sampling from cells marked with the same entry will spread sampling evenly across the dimensions

1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1

Other sampling methods: latin hypercube samping (space filling design) - LHS

(McKay, Beckman, and Conover 2000)

Latin hypercube sampling exploits the properties of a latin square - think Sudoku

Every entry appears in each row and column exactly once.

Sampling from cells marked with the same entry will spread sampling evenly across the dimensions

1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1

Other sampling methods: latin hypercube samping (space filling design) - LHS

(McKay, Beckman, and Conover 2000)

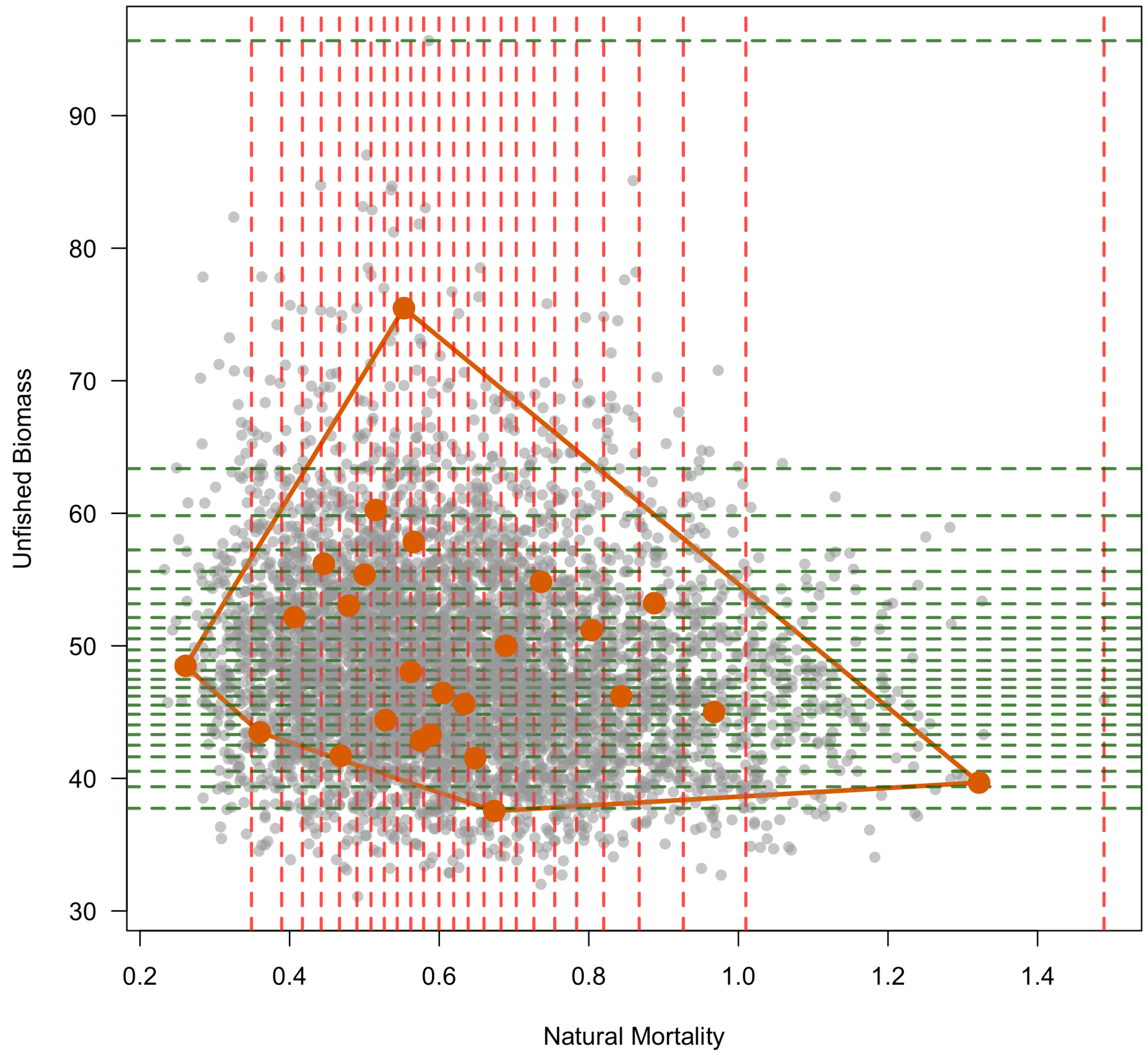
For a posterior distribution,

1. stratify each margin into strata of equal density

2. Label resulting hypercubic design with the latin property

3. Sample within each cell

We need to approximate because we have a discrete posterior ***(`clhs` package, Roudier 2011)***



(Roudier 2011)

Other sampling methods: latin hypercube samping (space filling design) - LHS

(McKay, Beckman, and Conover 2000)

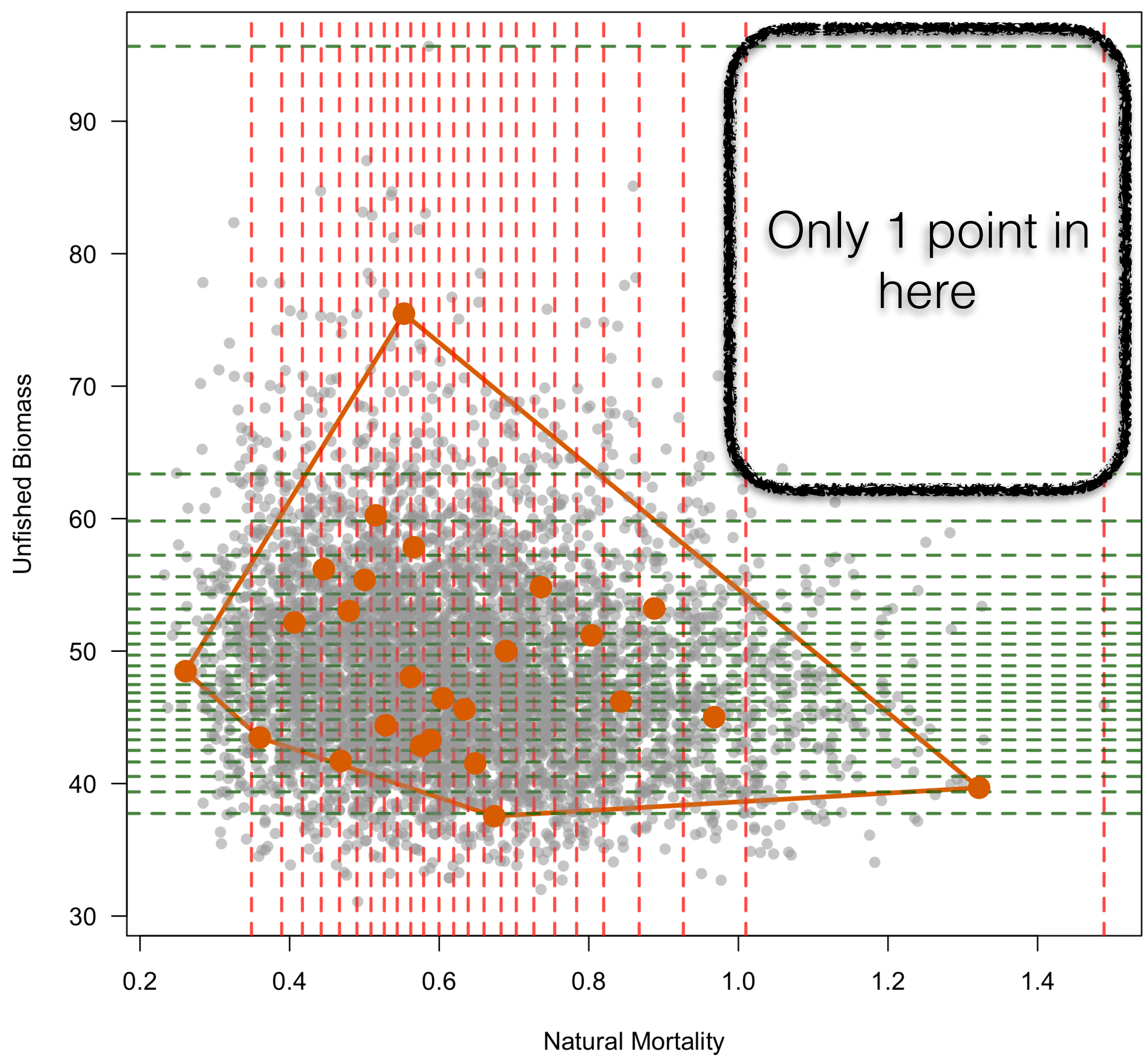
For a posterior distribution,

1. stratify each margin into strata of equal density

2. Label resulting hypercubic design with the latin property

3. Sample within each cell

We need to approximate because we have a discrete posterior ***(`clhs` package, Roudier 2011)***

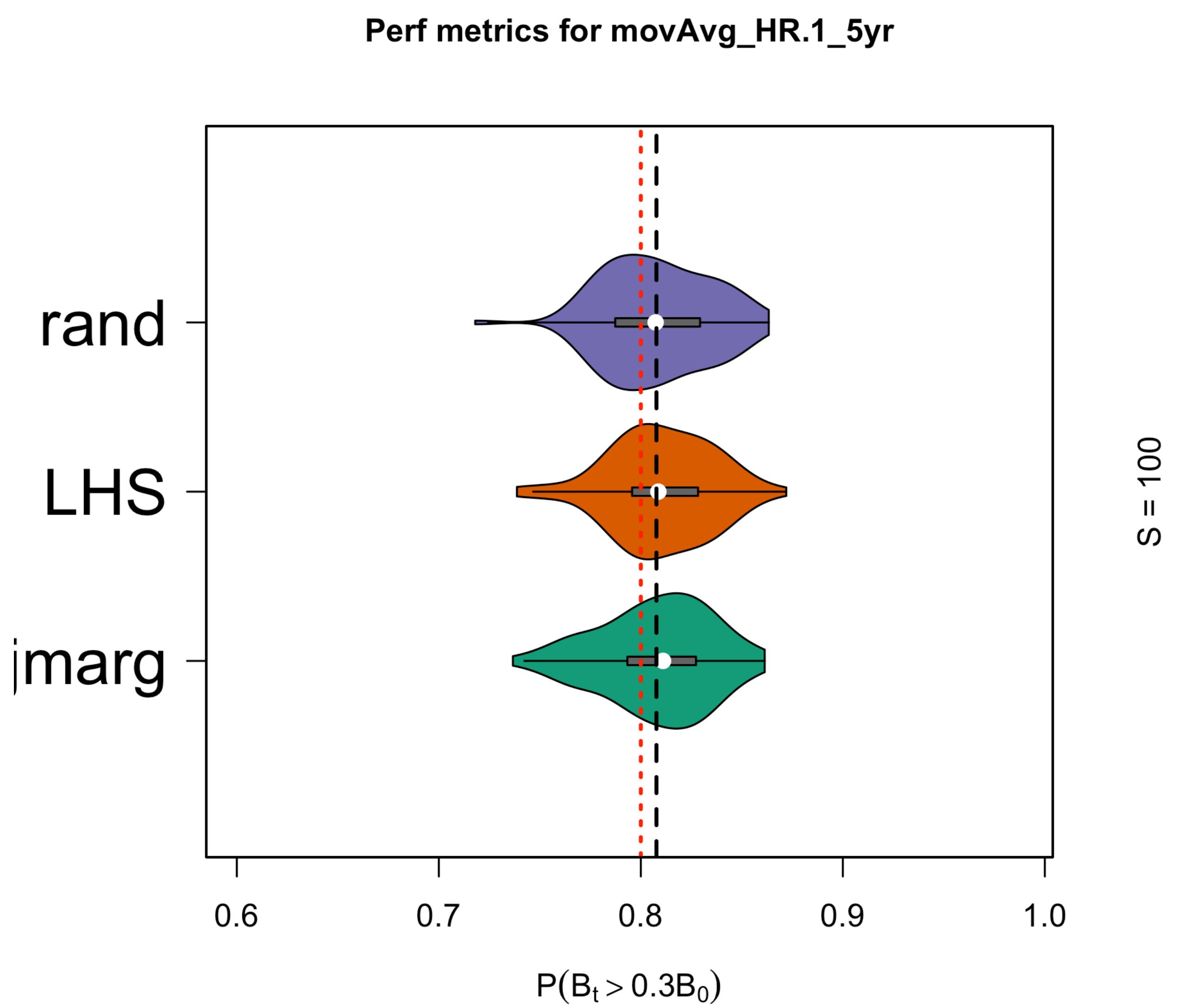


(Roudier 2011)

Results

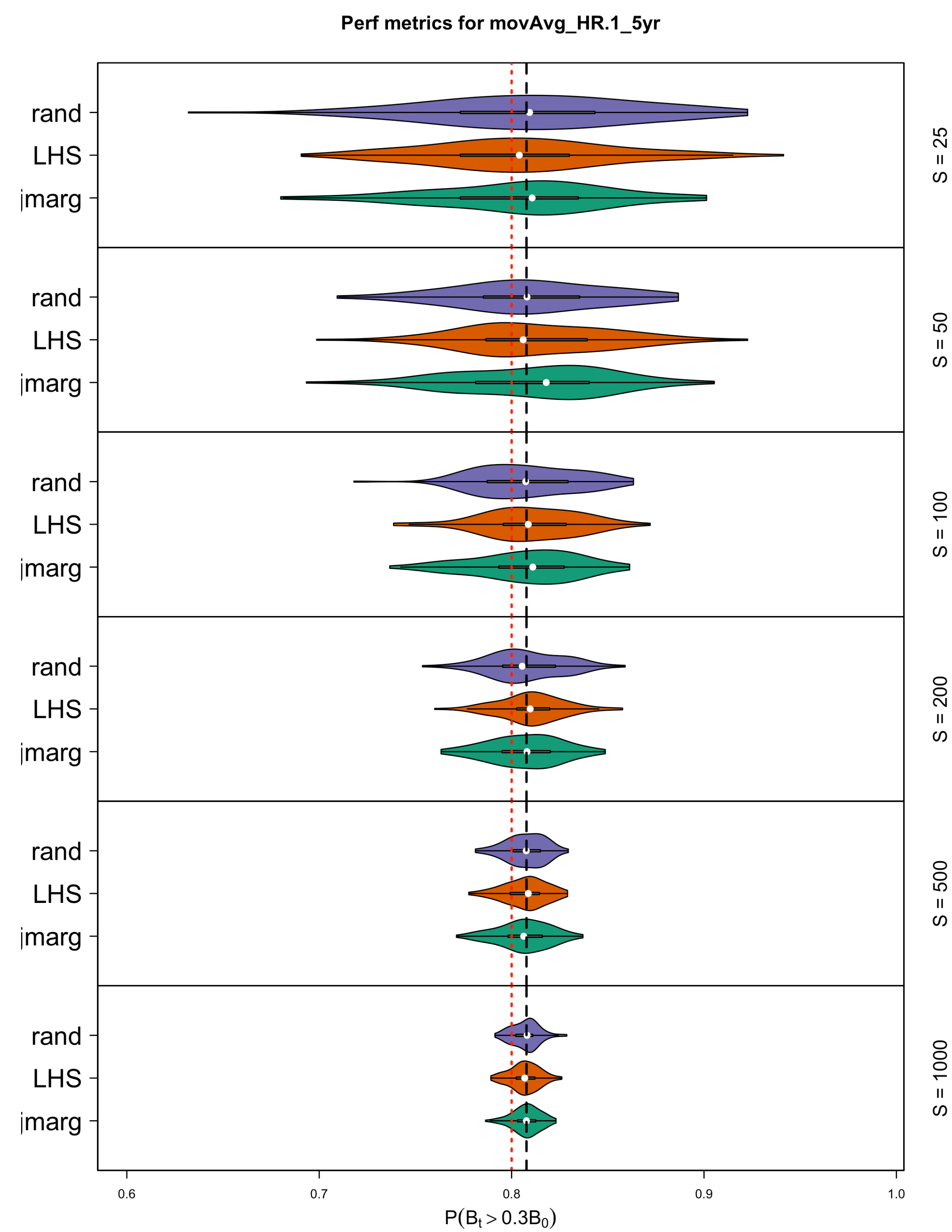
Random sampling holds its own for the standard sample size

- Ranking is stable
- At 100 samples the three methods don't appear to be differentiated for our example MP and objective
- All appear to be asymptotically unbiased (phew)
- Some skew apparent in the joint marginal and simple random sampling designs, less in LHS



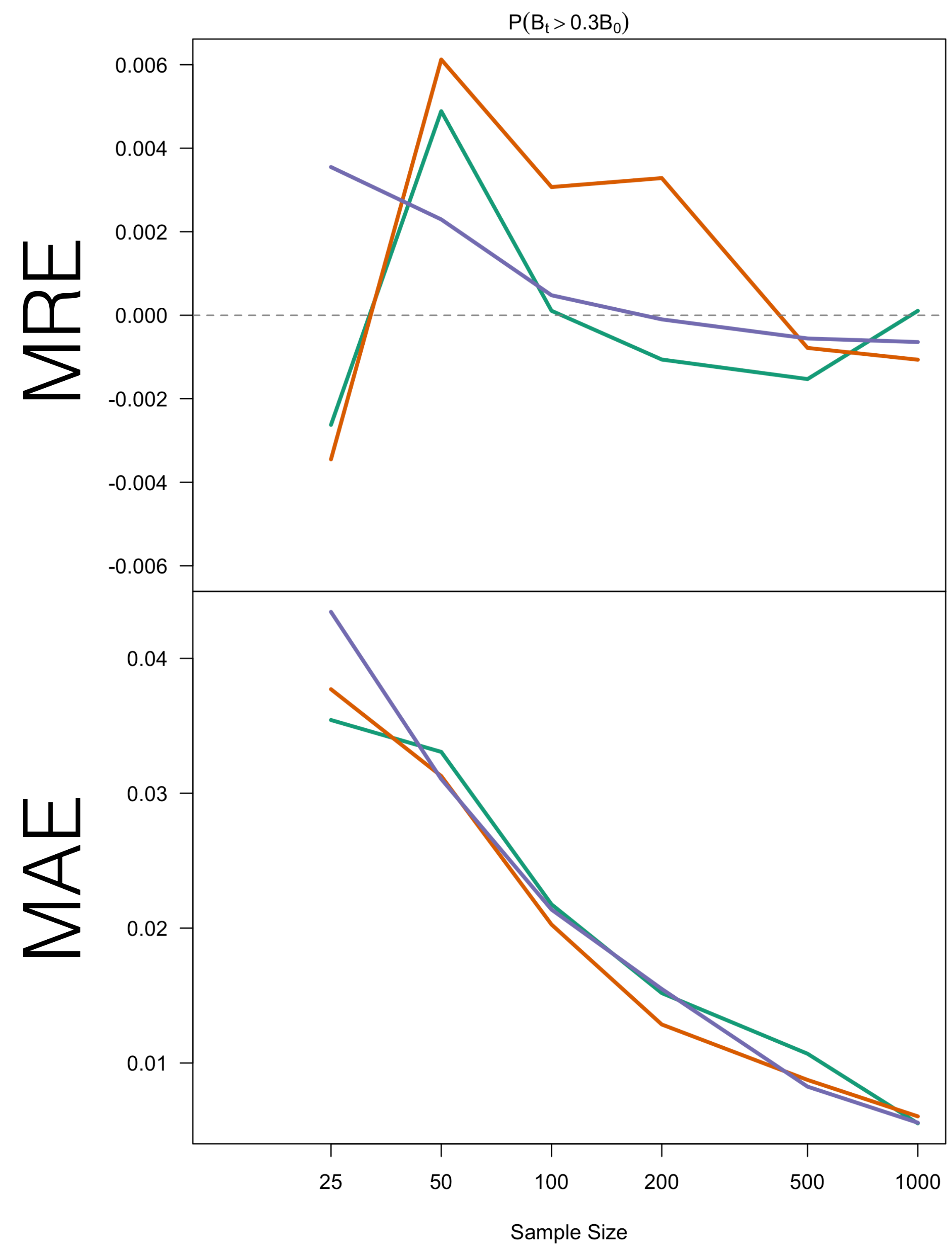
Random sampling holds its own across all sample sizes

- Ranking was stable above 50 samples
- Random sampling is actually less variable at S = 50
- LHS, which I expected would do the best, doesn't appear to outperform in any significant way until S = 200 - this may be too large for most applications, especially model based MPs
- Some skew apparent in distributions, possible causes
    - approximate LHS
    - choice of margins for jmarg

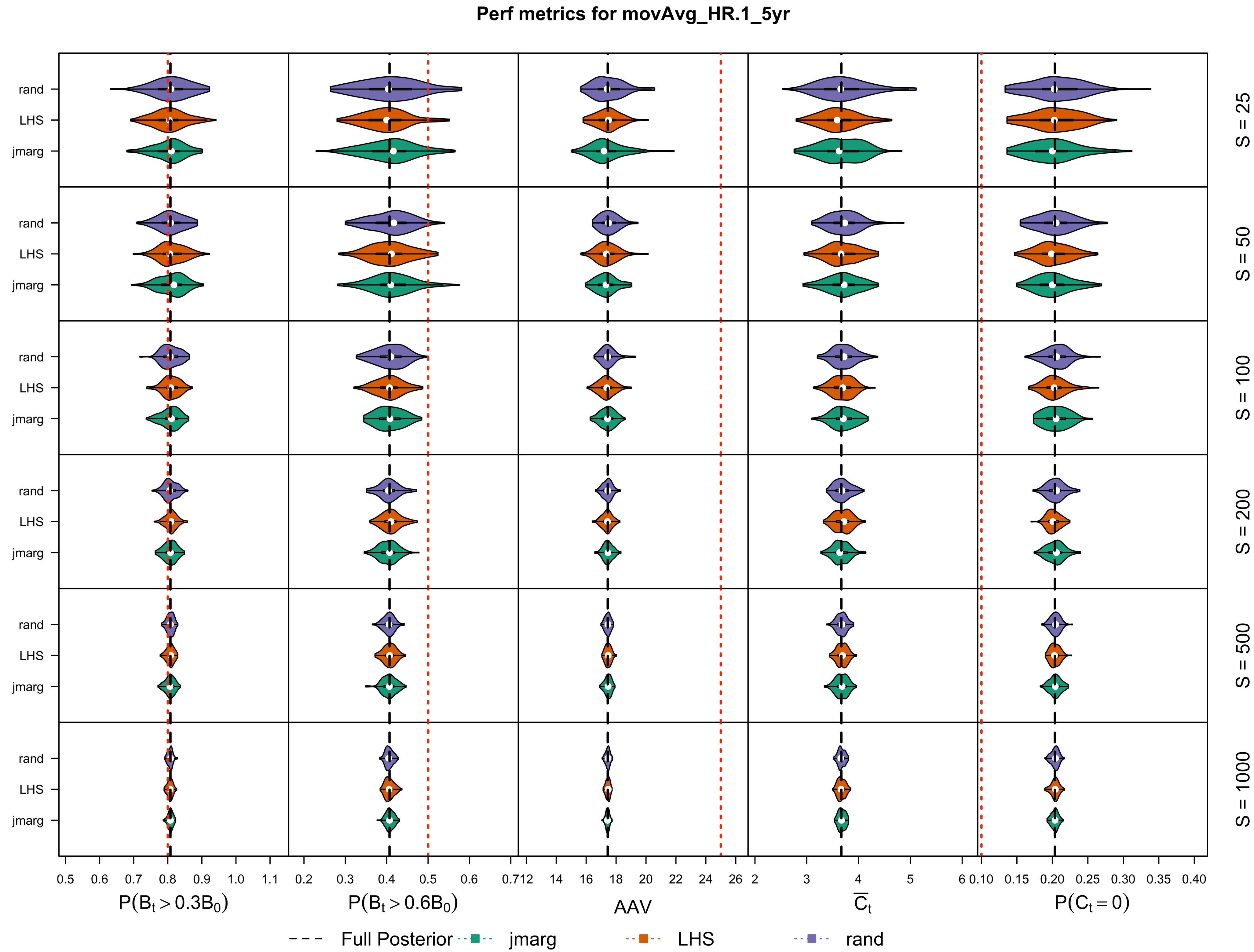


Loss functions make it easier to see the trend in performance

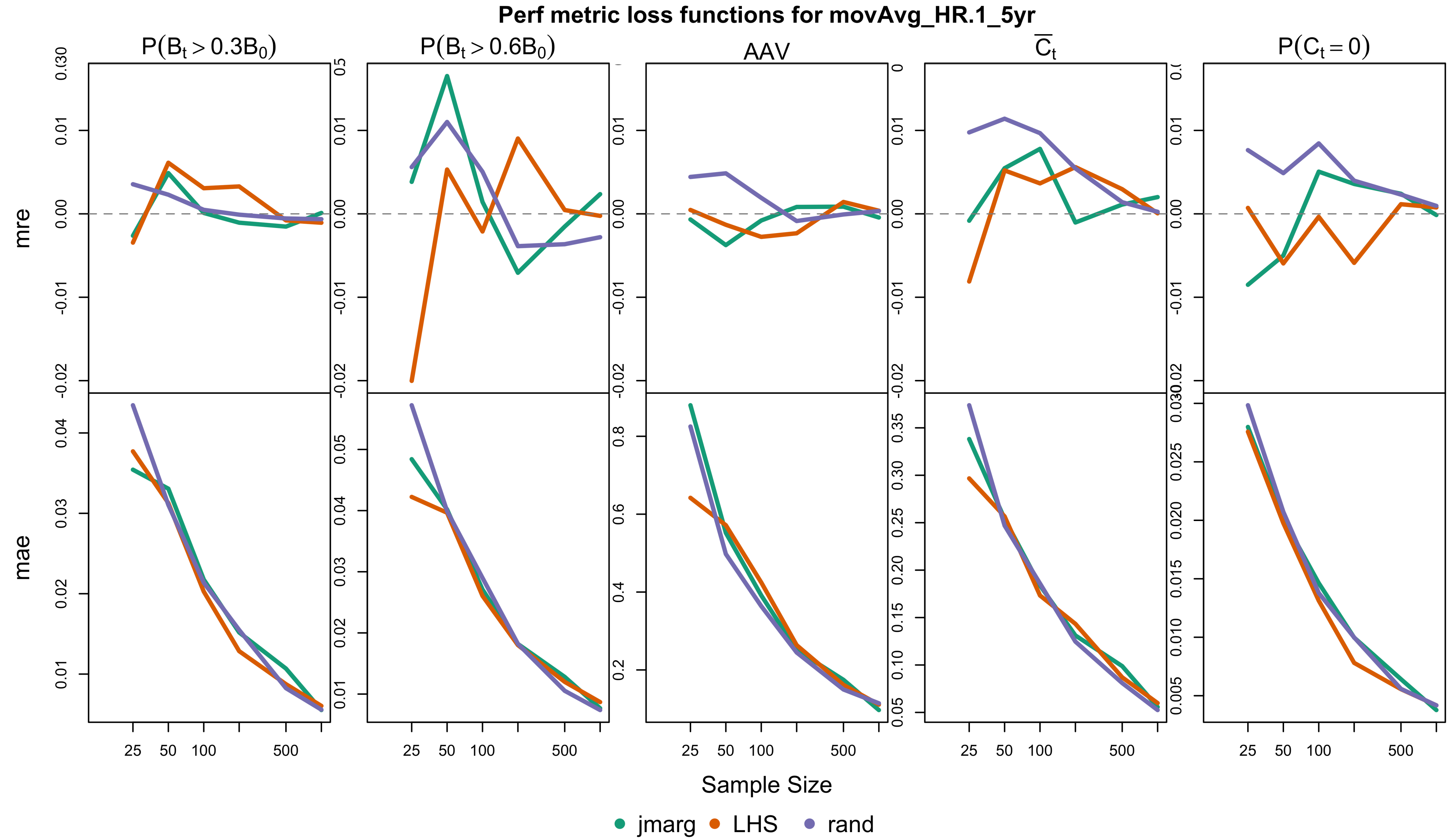
- Loss curves show the mean relative error and mean absolute error of the objective metrics as a function of sample size
- Fluctuations in mean relative error for each method (~ 0.5% of the true value)
- As expected, mean absolute error declines with sample size
- Shows real benefit of stratified sampling is at lower sample sizes, or above 100 samples, with marginal gains otherwise



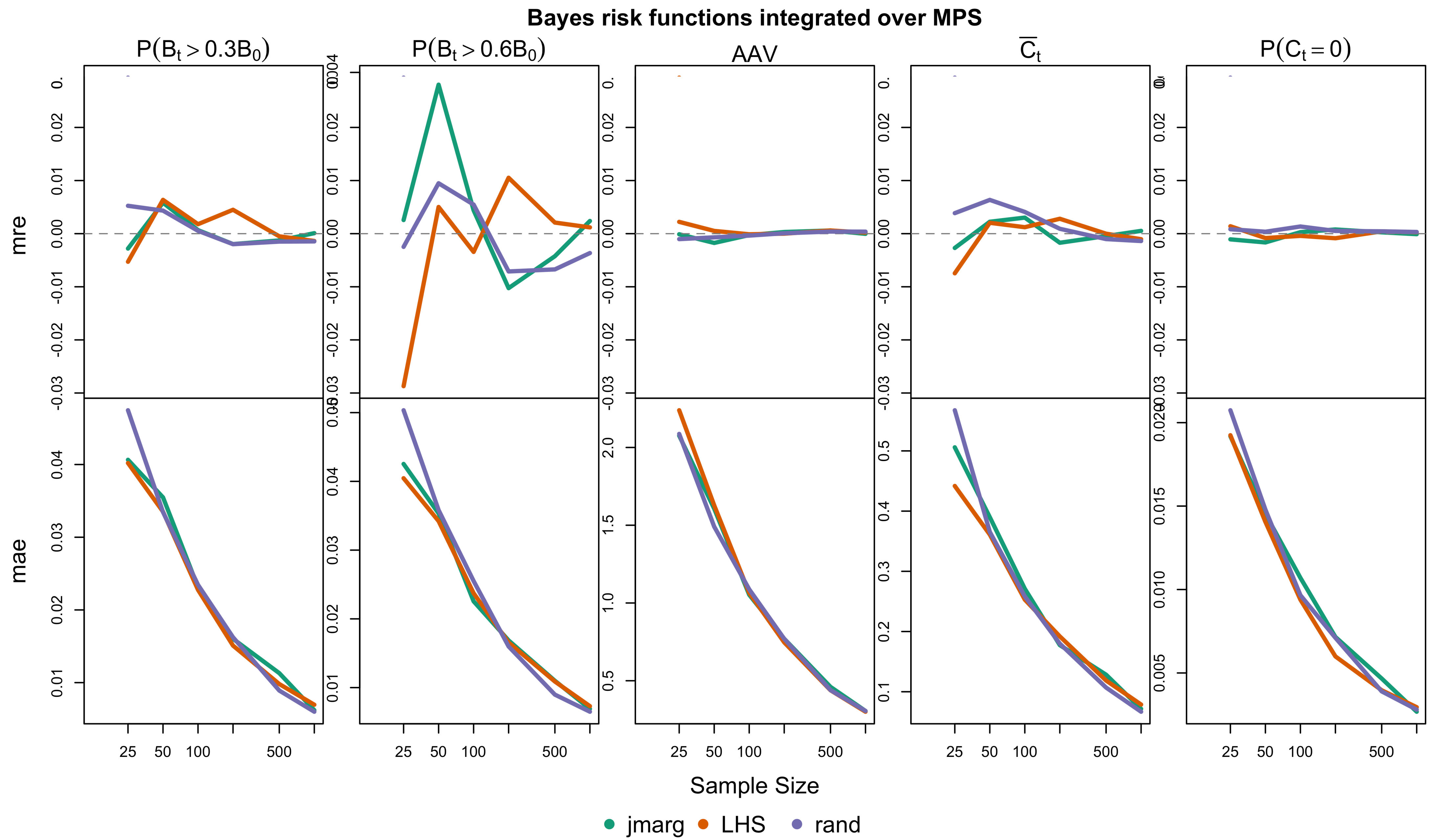
Some sensitivity to the objectives as well



Loss pattern remains similar over objectives



Averaging loss over multiple MPs shows a similar picture



Discussion

Research questions: can we define some best practices for taking samples to condition the OM?

1. What sampling methods are best suited to reducing the sensitivity of the objective performance metrics to sample size and random seed?

2. What is the minimum number of samples required to

   A. fix the ranking of MPs for each sampling method?

   B. reduce variance of metrics within some tolerance?

3. What qualities can be used to identify a good sample before significant time is spent in simulations, i.e. filter samples so that variance of metrics is within some tolerance?

Partway to discovering the effect of sampling design on MP performance metrics (Q1)

1. Differences between sampling methods is largest at small sample sizes, marginal at 100 points, and then better again at larger samples

2. Marginal improvements may not seem like much, but 1 or 2 percentage points could be the difference between passing or failing MSC certification (is P(Bt > Blim) >= .95?), or passing or failing the MSE process

3. Number of samples where difference is made may be too low for differentiating ranking, or too high for practical applications

Partway to discovering the effect of sampling design on MP rankings and variance (Q2)

1. understanding the effect of sample size and method on obj perf metric sensitivity can help make MSE more efficient - if ranking is all we care about, then smaller sample sizes may be adequate

2. Next step: understand what it is about the samples that lie in the middle of each distribution - is there some quality we can detect and control for?

Future work

Still need to discover qualities of a “good” sample from the Bayesian posterior (Q3)

Hypothesis:

Low KL divergence will likely be correlated with low bias of obj performance metrics

Normal approximations are next, with some guiding questions below.

Normal approximation of the posterior

- Test sampling designs under approximation
- Does the number of samples required to adequately approximate the posterior increase? (symm. KL divergence)
- If so, does increase in compute time outweigh the benefit of approximating the posterior (assuming MCMC can actually be run)?

Thanks for listening!

Questions/Comments?

1. What sampling methods are best suited to reducing the sensitivity of the objective performance metrics to sample size and random seed?

2. What qualities can be used to a priori identify a good sample, i.e. bias within some tolerance?

3. What is the minimum number of samples required to

   A. fix the ranking of MPs for each sampling method?

   B. reduce variance of metrics within some tolerance?