Replacing lognormal indices in stock assessments: a first step

Cole Monnahan

2023 September GPT

Collaborators: Megsie Siple, Pete Hulson, Paul Spencer,
Maia Kapur, Jim Thorson

Quick review of indices

- GAP/MACE give us a mean and CV, and we **assume** the distribution is lognormal
- Lognormal indices are ubiquitous
  - Globally in bespoke and generic ones like SS3, WHAM, GADGET, JABBA, MULTIFAN-CTL, CASAL
  - Likely used in all AFSC assessments (T1,3,5)
- Arise from design- or model-based (e.g., VAST) estimators
- **Both are sums of positive quantities** for smaller areas (strata, spatial cells)

$$I(c,t,l) = \sum_{x=1}^{n_x} \left(a(s,l) \times d^*(s,c,t) \right)$$

$$\hat{B}_{Tk} = \sum_i \hat{B}_{ik}$$

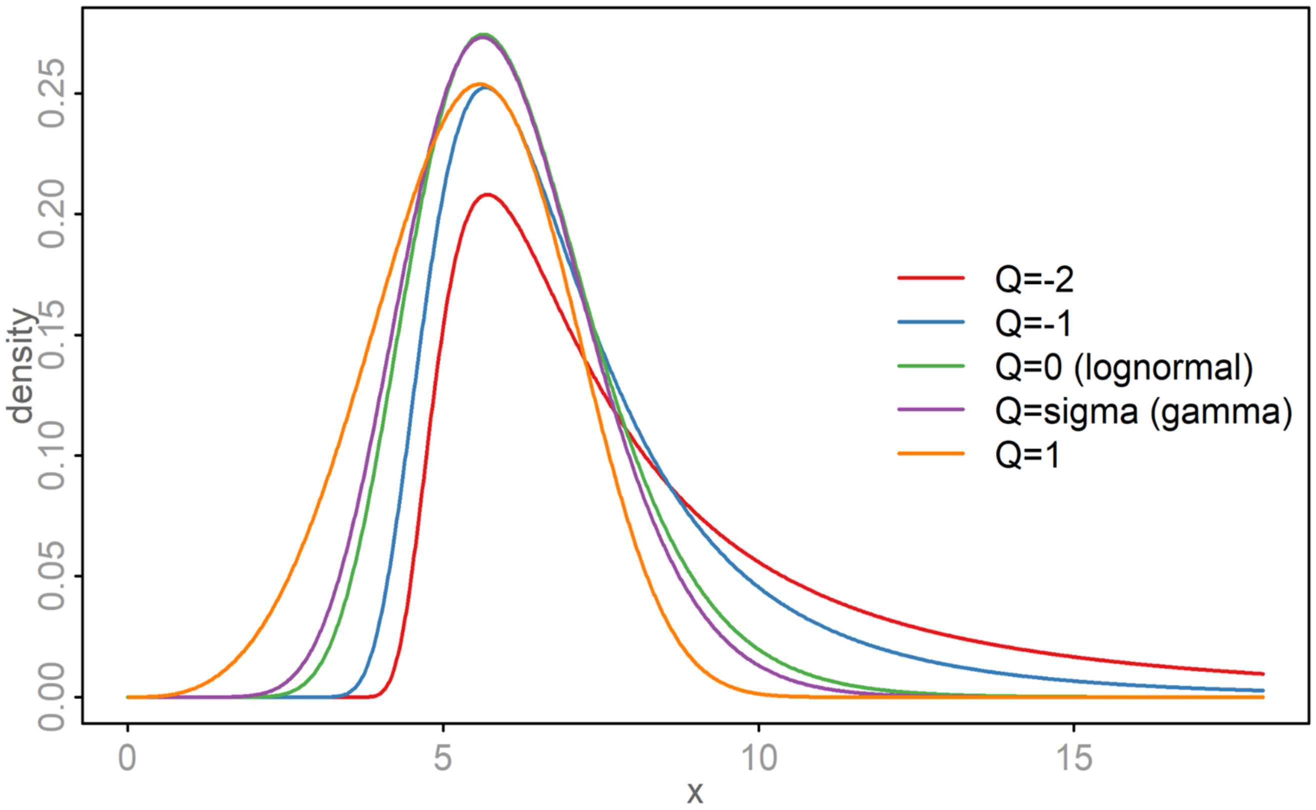
Are lognormal indices justified / does it matter?

- No, lognormal indices have no statistical justification
- Indices are most important data (get the trend right; Francis 2011)
- Our goals are:
  a. To understand how common non-lognormality is at AFSC
  b. Find a more appropriate/flexible distribution
  c. Test implications of replacing the lognormal in assessments
- Enter the generalized gamma distribution (GGD)

Beaulieu (2012); Beaulieu et al. (1995); Dufrense (2004); Romeo et al. (2003), etc.

Basics of the generalized gamma

- Flexible distribution characterized by 3 parameters
  - Mean, variance, skewness (Q)
- Special cases
  - Q=0 lognormal
  - Q=sigma gamma
  - Also Weibull, exponential, half-normal



$$g \sim Gamma(Q^{-2}, 1)$$
$$w = log(Q^{2} g)/Q$$
$$x = exp(\mu + \sigma w)$$
$$f(x|\mu, \sigma, Q) = |Q| \, (Q^{-2})^{Q^{-2}} / (\sigma \, x \, \Gamma(Q^{-2})) \, exp(Q^{-2}(Qw - exp(Qw)))$$

Sums of lognormals are not lognormal

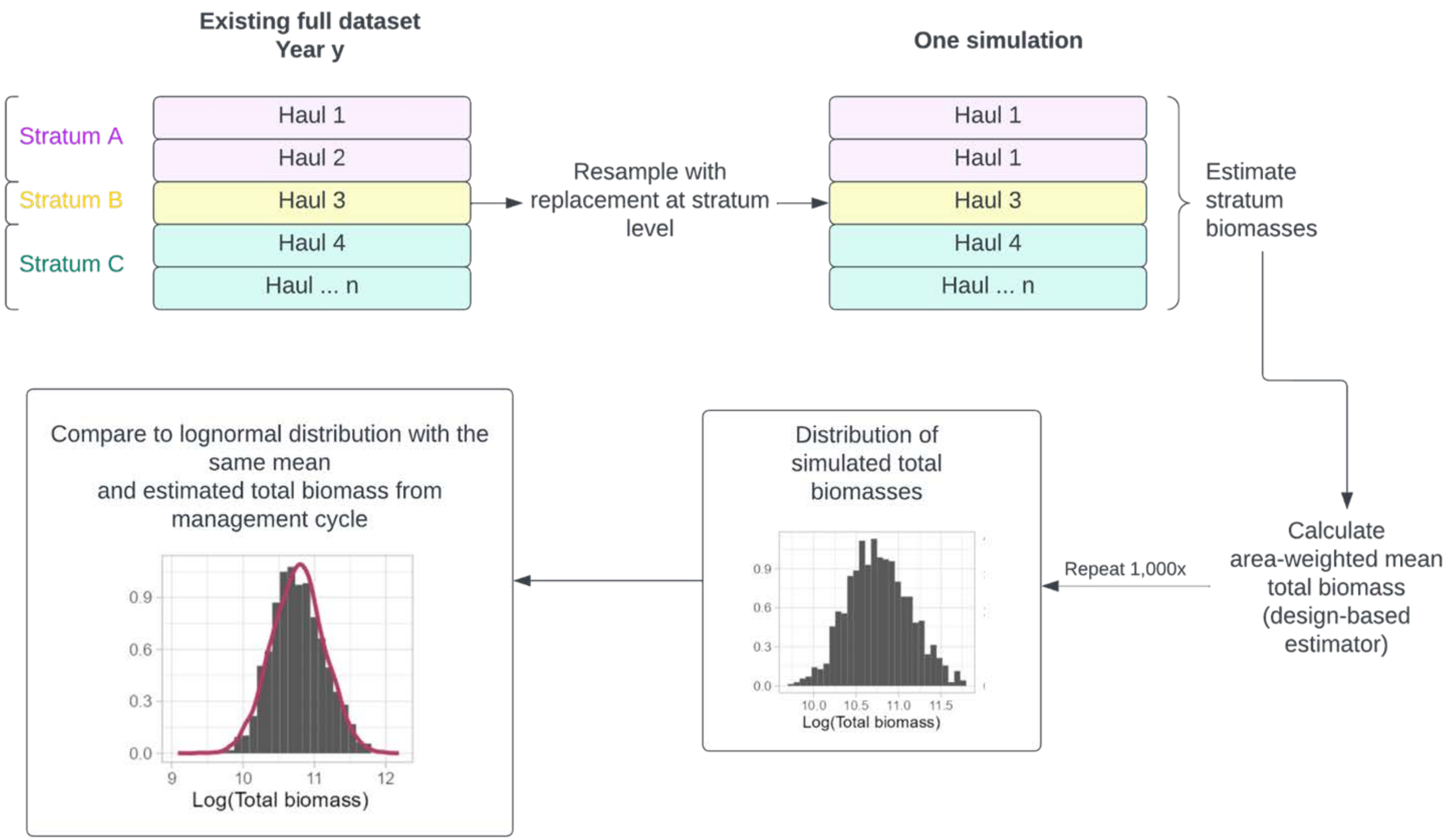
Experiment:

- Simulate draws from 4 normal rvs, X1,...,Xn then let Y=log(e^X1+..+e^Xn)
- Is Y~normal?

- No, it is not, except in very narrow circumstances
- A more flexible generalized gamma distribution (GGD) fits better than a lognormal
- Y is not GGD, but it can better approximate it
- Stats literature very clear **that sums of lognormals are not lognormal** and in fact have no known analytical form

Monnahan et al. in prep

Design-based bootstrapping procedure



Proposed procedure:

1. Simulate samples from the distribution of biomass
2. Fit the GGD to those samples to get an estimate for the mean, SD, and Q
3. Read those into stock assessment and use GGD pdf in place of lognormal

This is done for each year separately

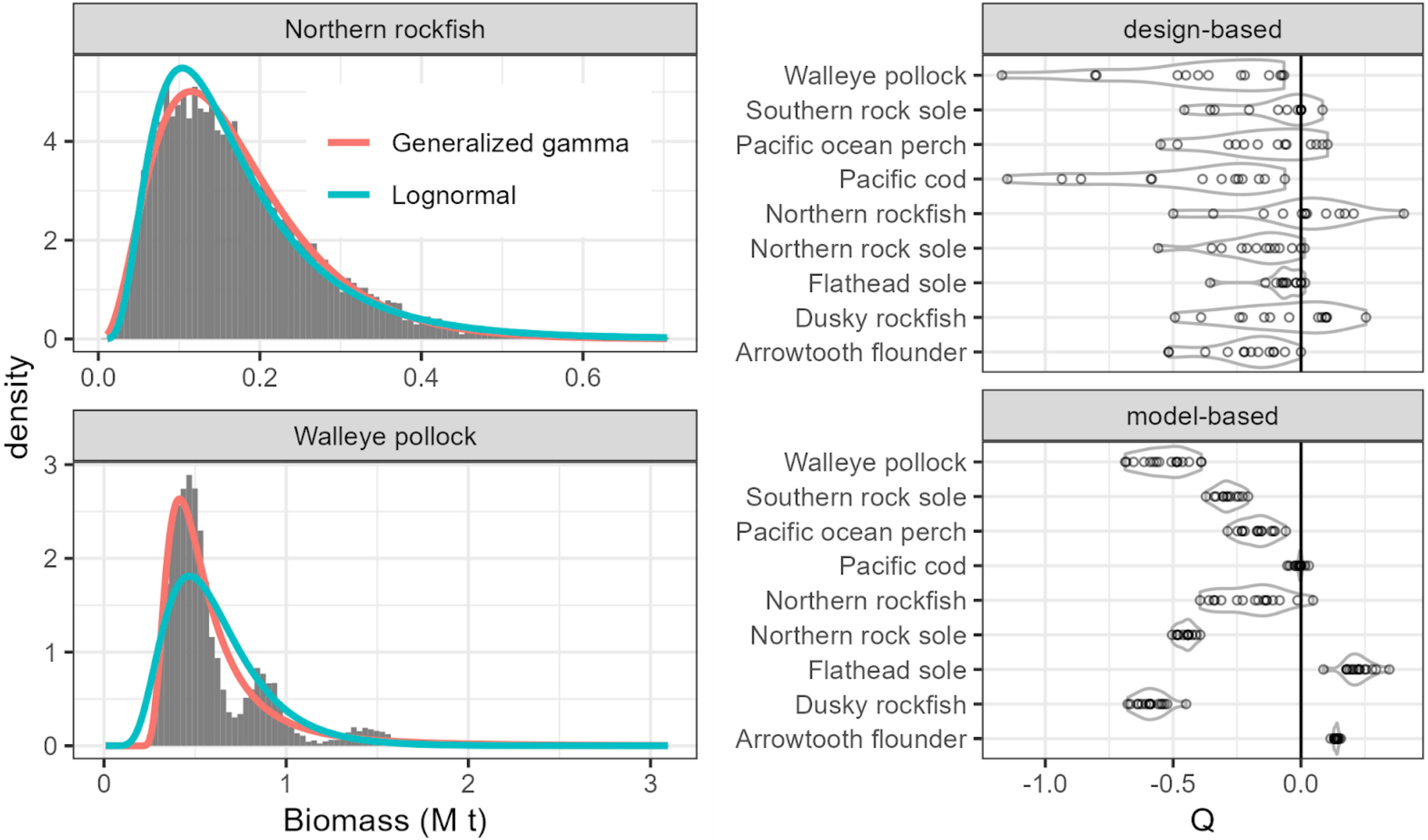
Model-based "bootstrapping" procedure

- VAST estimates biomass B, then uses the delta method to get σ=SE(log(B)), and we **then assume N(log(B),σ)**

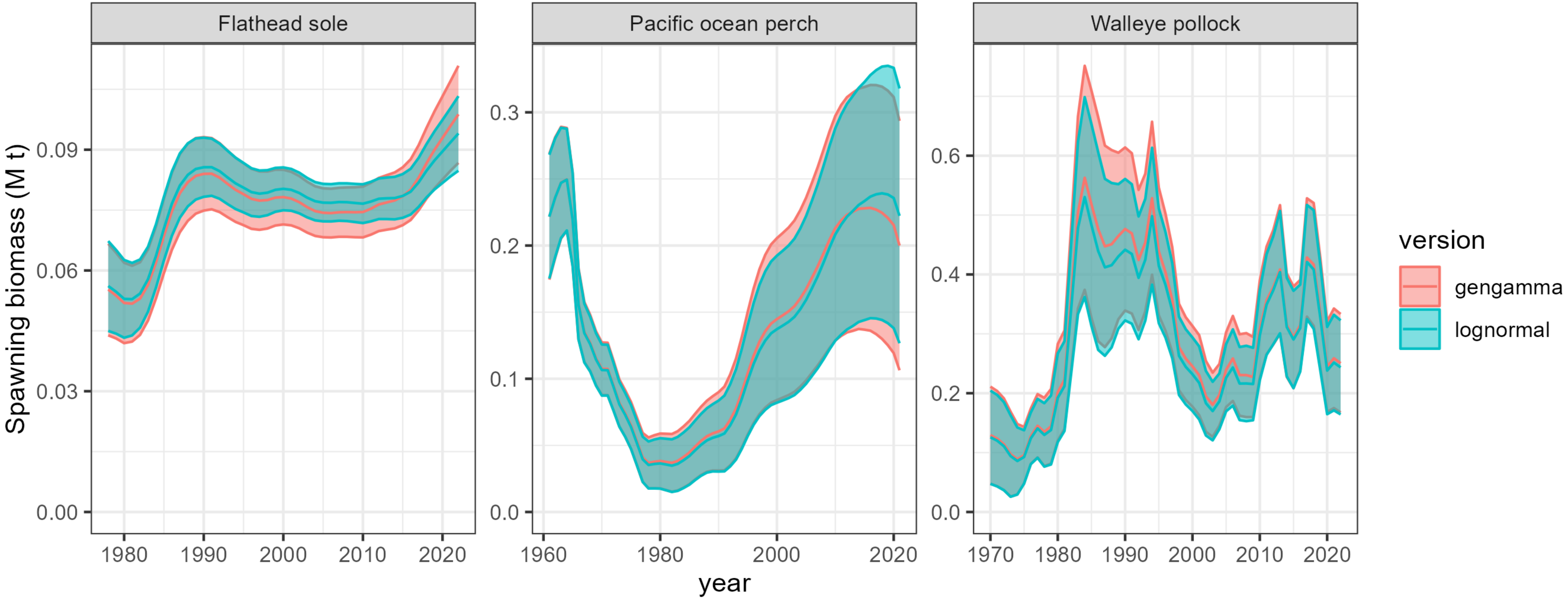
But what is the real distribution of B? Three options to quantify it:

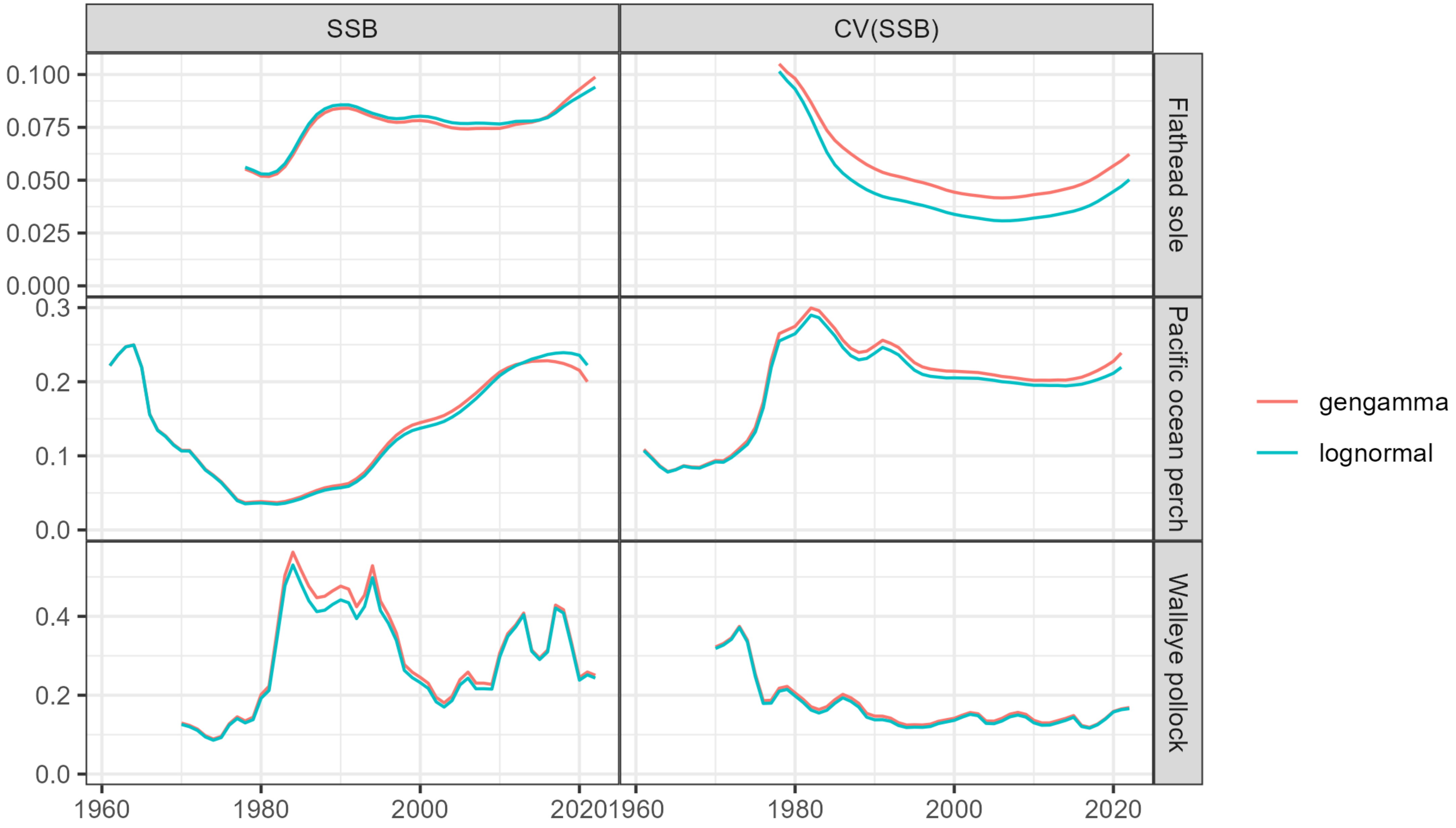
- Data bootstrapping – too slow!
- Posterior sampling w/ MCMC – too slow!
- Model-based bootstrapping-ish – promising!
  - Assume MLE of all parameters is MVN and resample
  - Uses "joint precision matrix" of fixed and random effects
  - Is this valid to do?

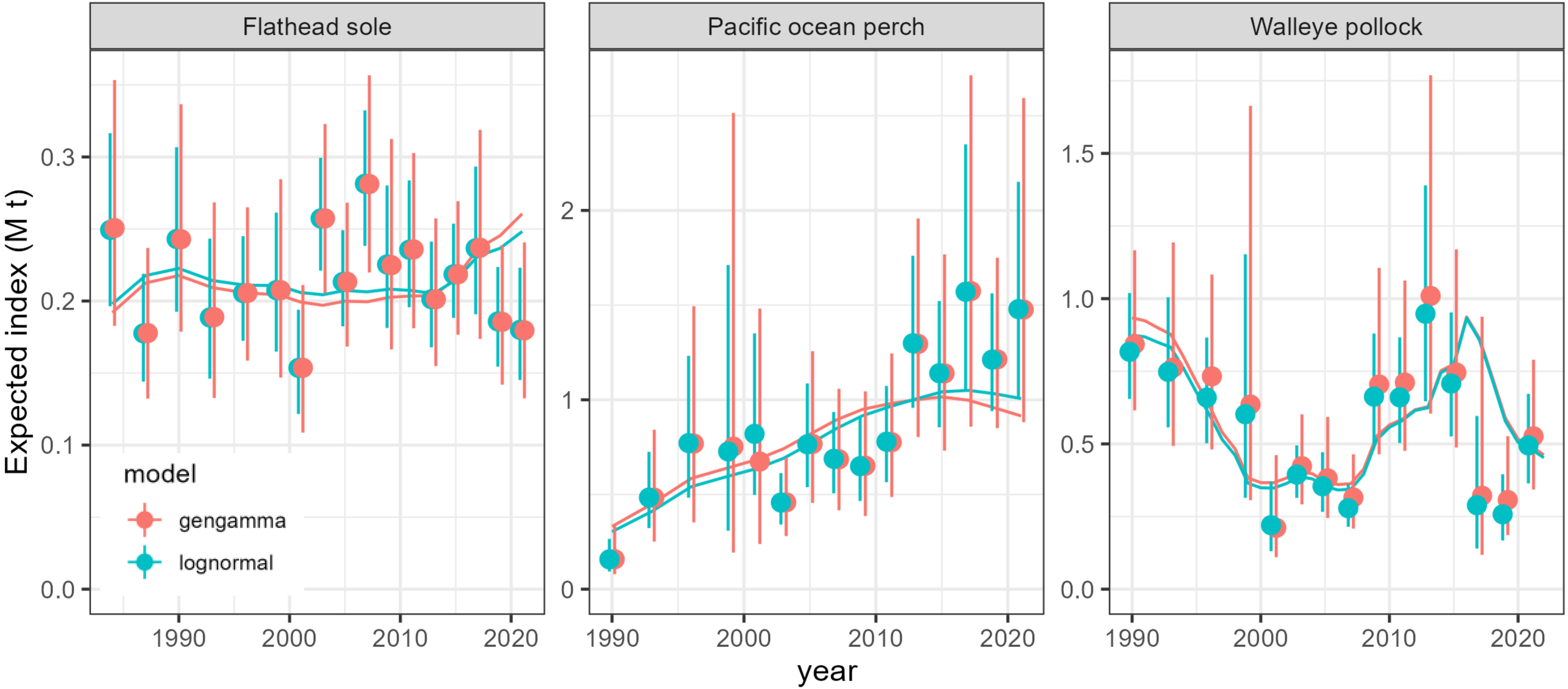
Results of lognormality



Impact on stock assessments




SSB
CV(SSB)
0.100
0.075
0.050
0.025
0.000
0.3
0.2
0.1
0.0
0.4
0.2
0.0
1960
1980
2000
2020
Flathead sole
Pacific ocean perch
Walleye pollock
gengamma
lognormal


Flathead sole
Pacific ocean perch
Walleye pollock
Expected index (M t)
year
model
gengamma
lognormal
0.0
0.1
0.2
0.3
0
1
2
0.5
1.0
1.5
1990
2000
2010
2020

Implications and recommendations for model-based estimators

- adding 1 extra parameter (vs. lognormal, gamma, etc.) gains dexterity of fit
- Not too computationally burdensome
- GAP/MACE would have to operationalize and provide GGD estimates to SSMA/MESA
- Need to refine bootstrapping procedure
- What about multimodal bootstrap distributions?

Implications and recommendations for stock assessment

- Q!=0 changes index leverage & thus its statistical weight
- **We should get the likelihood right**
- The burden is on justifying a lognormal index
- Both bootstrap approaches are straightforward & easy-ish
- GGD seems to work in assessments (bespoke+SS3)
- When will this impacts assessments? Hard to know but if:
  - Q<0 and there are large residuals
  - Conflicting indices, interacts w/ data weighting

Discussion points and future work

- Evidence that Q<0 often in haul-level data and indices
- What causes Q<0?
  - A fundamental property of fish biology? Population characteristics?
  - Survey design/execution? Traits of haul-level data?
  - Something inherent in the design- and model-based estimators?
- Do the bootstrap procedures accurately estimate distribution?
- Does the precautionary principle come into play?
- We need to apply this to more GOA assessments

Request for Plan Team feedback

Should we

- Expand to AI and BS BT surveys?
- Expand to MACE surveys?
  (Plays particularly well with Sam Urmy's work)
- Abandon?

Thanks for listening!

Thanks to collaborators

References

- N. C. Beaulieu, "An Extended Limit Theorem for Correlated Lognormal Sums," in IEEE Transactions on Communications, vol. 60, no. 1, pp. 23-26, January 2012, doi: 10.1109/TCOMM.2011.091911.110054.
- N. C. Beaulieu, A. A. Abu-Dayya and P. J. McLane, "Estimating the distribution of a sum of independent lognormal random variables," in IEEE Transactions on Communications, vol. 43, no. 12, pp. 2869-, Dec. 1995, doi: 10.1109/26.477480.
- Dufresne, Daniel. "The log-normal approximation in financial and other computations." Advances in applied probability 36.3 (2004): 747-773.
- Francis R.I.C. Chris. 2011. Data weighting in statistical fisheries stock assessment models. Canadian Journal of Fisheries and Aquatic Sciences. 68(6): 1124-1138. https://doi.org/10.1139/f2011-025
- Romeo, M., Da Costa, V. & Bardou, F. Broad distribution effects in sums of lognormal random variables. Eur. Phys. J. B 32, 513–525 (2003). https://doi.org/10.1140/epjb/e2003-00131-6
- Lawless, J.F., 1980. Inference in the Generalized Gamma and Log Gamma Distributions. Technometrics 22, 409–419. https://doi.org/10.1080/00401706.1980.10486173
- Prentice, R.L., 1974. A Log Gamma Model and Its Maximum Likelihood Estimation. Biometrika 61, 539–544. https://doi.org/10.2307/2334737
- Stacy, E.W., 1962. A Generalization of the Gamma Distribution. The Annals of Mathematical Statistics 33, 1187–1192.
- Cox, C., Chu, H., Schneider, M.F., Muñoz, A., 2007. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Statistics in Medicine 26, 4352–4374. https://doi.org/10.1002/sim.2836
- Crooks, G.E., 2019. Field guide to continuous probability distributions: v.1.0.0. Berkeley Institute for Theoretical Sciences (BITS), Berkeley.

TMB code

template<class Type>

```
Type dgengamma( Type x, Type mean, Type sigma, Type Q, int give_log=0){
  Type k = pow(Q, -2);
  Type Beta = pow(sigma, -1)*Q;
  Type log_theta = log(mean) - lgamma((k*Beta+1)/Beta)+lgamma(k);
  Type mu = log_theta + log(k)/Beta;
  Type w = (log(x) - mu) / sigma;
  Type abs_q = sqrt(Q*Q);  // = abs(Q); not differentiable at 0!!
  Type qi = 1/square(Q);
  Type qw = Q*w;
  Type logres = -log(sigma*x) + log(abs_q)*(1-2*qi) + qi*(qw-exp(qw)) - lgamma(qi);
  if(give_log) return logres; else return exp(logres);
}
```

ADMB code

```
FUNCTION dvariable dgengamma(const double& x, dvariable mean, const double& sigma, const
double& Q)
  RETURN_ARRAYS_INCREMENT();
  double k = pow( Q, -2 );
  double Beta = pow( sigma, -1 ) * Q;
  dvariable log_theta = log(mean) - lgamma( (k*Beta+1)/Beta ) + lgamma( k );
  dvariable mu = log_theta + log(k) / Beta;
  dvariable w = (log(x) - mu) / sigma;
  double abs_q = sqrt(Q*Q);  // = abs(Q); not differentiable!
  double qi = 1/square(Q);
  dvariable qw = Q*w;
  dvariable logres = -log(sigma*x)+log(abs_q)*(1-2*qi) + qi * (qw-exp(qw))-lgamma(qi);
  RETURN_ARRAYS_DECREMENT();
  return(logres);
```