Forecast probabilities - work in progress

Nicolas Blöchliger, Institute of Medical Microbiology, University of Zurich

12/05/2016

Contents

1 Goals

1. Compute the probability that a strain is pseudo-WT given an observed diameter y.
2. Compute the probability that a strain is susceptible according to official breakpoint given an observed diameter y.

2 Model

We assume:

- The distribution of the true diameter X is a mixture of three components with weights $w_i = p(C = i)$, where C encodes the component. The true diameter is 6 mm for the first component and normally distributed for the other two components:

$$p_i(x) = f_X(x|C = i) = \begin{cases} \delta_6(x) & \text{if } i = 1, \\ \phi(x; \mu_i, \sigma^2) & \text{else,} \end{cases}$$

  where

$$\delta_6(x) = \begin{cases} \infty & \text{if } x = 6 \text{ mm}, \\ 0 & \text{else.} \end{cases}$$

  Thus,

$$f_X(x) = w_1\delta_6 + \sum_{i=2}^{3} w_i\phi(x; \mu_i, \sigma^2).$$

- We observe $Y = X + E$, where E models technical error. E is zero for the first component and normally distributed and independent of X with mean $\mu_E = 0$ and constant variance σ_E^2 for the other two components. The conditional density of Y given the component C is therefore

$$f_Y(y|C = i) = \begin{cases} \delta_6(y) & \text{if } i = 1, \\ p_i * \phi(\cdot; 0, \sigma_E^2) = \phi(y; \mu_i, \sigma^2 + \sigma_E^2) & \text{else.} \end{cases}$$

  Thus,

$$f_Y(y) = w_1\delta_6 + \sum_{i=2}^{3} w_i\phi(y; \mu_i, \sigma^2 + \sigma_E^2).$$

  Note that we do not account for the fact that the observed data are rounded to integer values.

2.1 Estimation of model parameters

- We estimate w_1 as the fraction of data points in the sample that are equal to 6 mm.
- The parameters of the second and third component of Y, i.e. w_i, μ_i, and $\sigma^2 + \sigma_E^2$, are estimated by fitting a normal mixture model of two components to the data in the sample with diameters greater than 6 mm. We use the R package `mclust`.
- Estimates for the variance of the error σ_E^2 will be taken from independent work in order to obtain σ^2.

2.2 Limitations of the model

- The model does not account for the fact that $X \geq 6$ mm. As long as the means of the two components are sufficiently large (say $\mu_i - 6 \text{ mm} > 2\sigma$), this should not cause problems.
- The model does not account for the fact that $Y \leq 40$ mm. As long as the means of the two components are sufficiently small (say $40 \text{ mm} - \mu_i > 2\sigma$), this should not cause problems.
- The error is assumed to be normally distributed with constant variance. This assumption is obviously violated if X is close to 6 or 40 mm. It is also violated for antibiotics like CPD, for which diameters are distorted in order to avoid additional laborious tests.
- The distributions of the two components are assumed to be normal. This seems fine for the component corresponding to wild-type strains. However, the distribution of the component corresponding to the resistant strains with $X > 6$ mm might not be adequately modelled.

- The variances of the two components are assumed to be equal. This assumption is problematic but has the advantage of guaranteeing that there is only one decision boundary if strains are assigned to the more likely component.

3 Data

E. coli, β-lactams. *To be completed.*

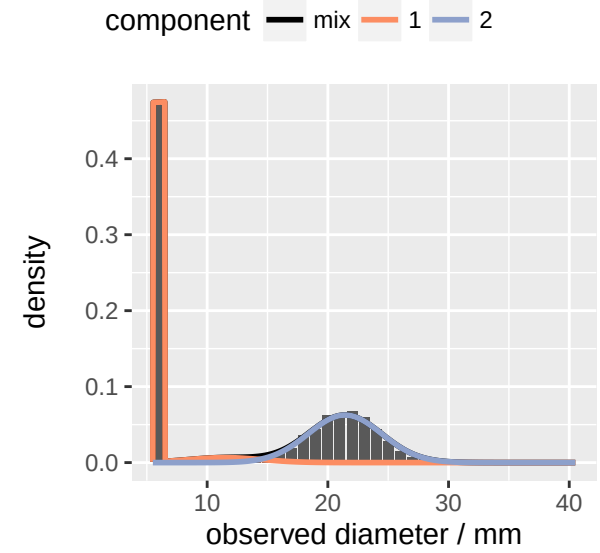
4 Results

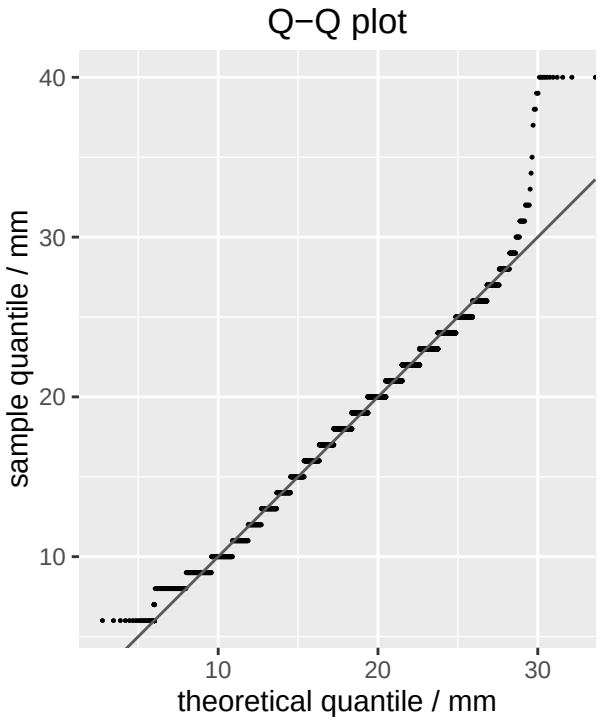
The figures in this document are organized as follows. Note that the the first and the second component are combined for visualisation.

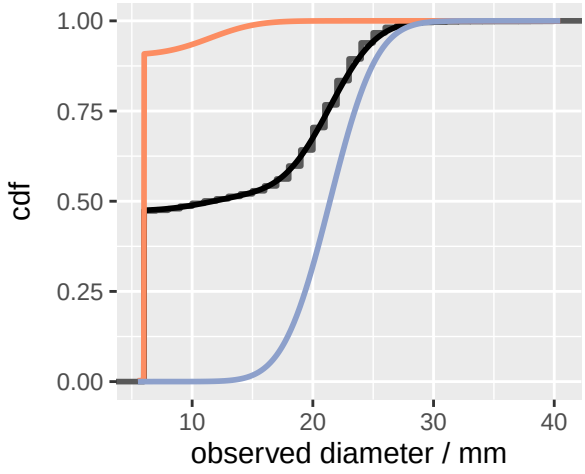
- Top-left: Histogram of sample and the estimated density of Y (black) and its components (coloured). The contribution from the first component (δ_6) is visualised as a uniform distribution with support [5.5 mm, 6.5 mm].
- Middle-left: Empirical cumulative distribution function (cdf) of Y (grey), its estimate (black) and estimated cdfs for the components of Y (coloured).
- Bottom-left: $p(C = i|Y = y)$, i.e. the probability that a data point is associated with component i given an observed diameter y. For this calculation, the first two components were grouped together.
- Top-right: Q-Q plot. If the estimated density of Y explained the data perfectly, all point would lie on the identity line (grey).
- Middle-right: Histogram of sample and the estimated density of X for various values of σ_E.
- Bottom-right: $p(X \leq t|Y = y)$, i.e. the probability that the true diameter is below a breakpoint t given an observed diameter y. For the time being, t was set such that $p(C = i|Y = t) \approx 0.5$.

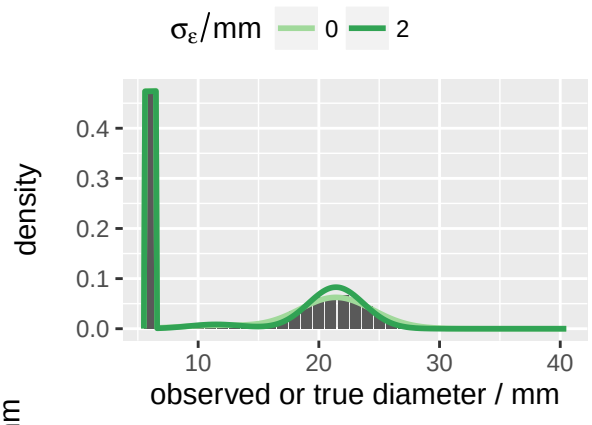
4.1 AM10

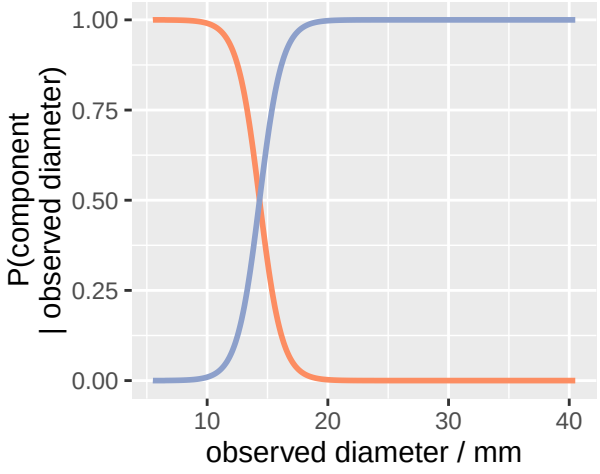
$\mu_{2,3} = 11.5$ mm, 21.4 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 3$ mm. $w = 0.47, 0.09, 0.91$.

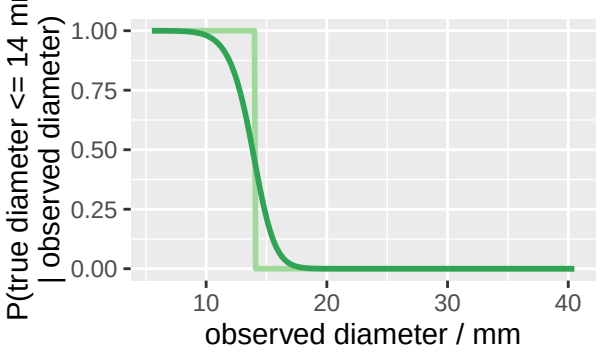






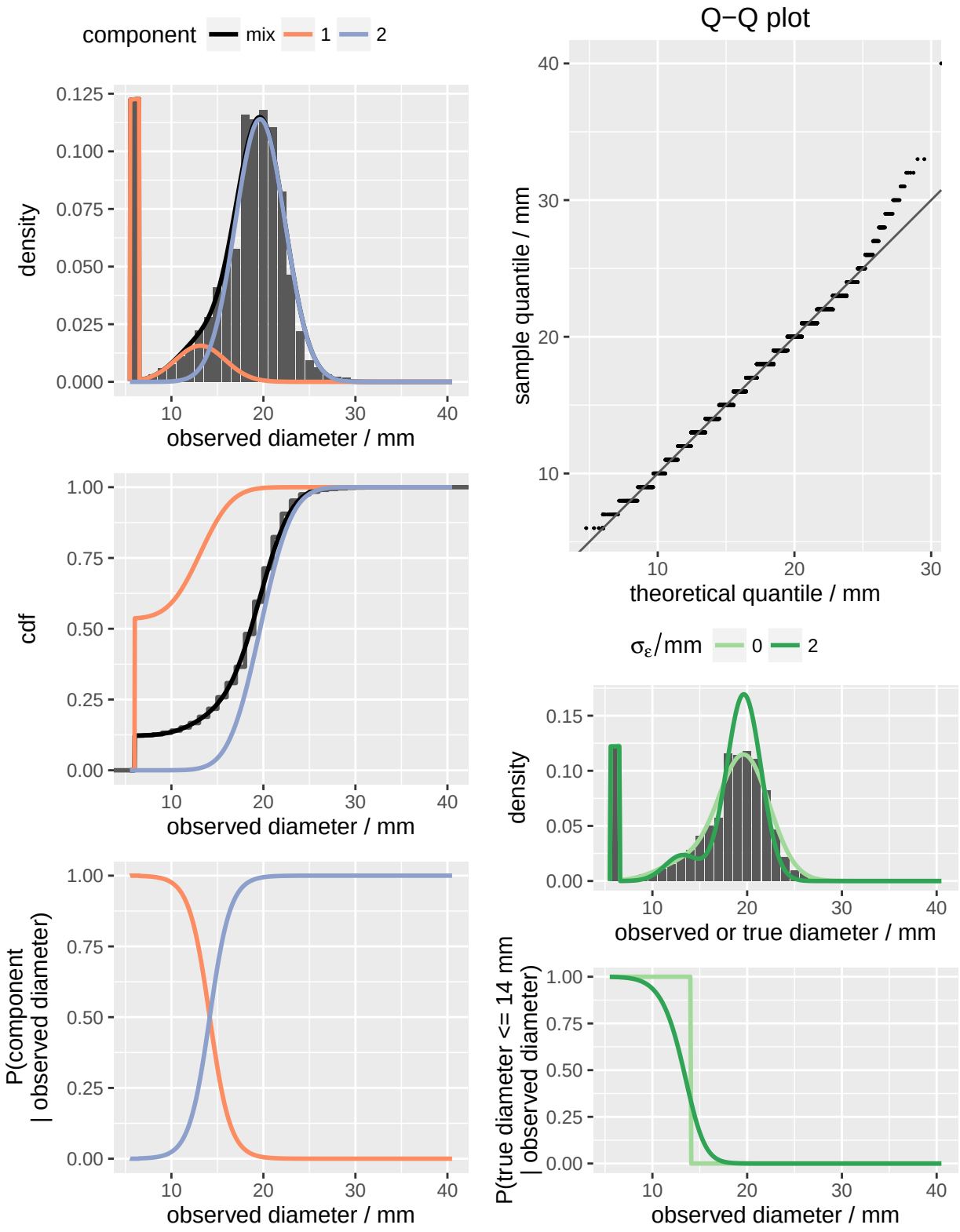






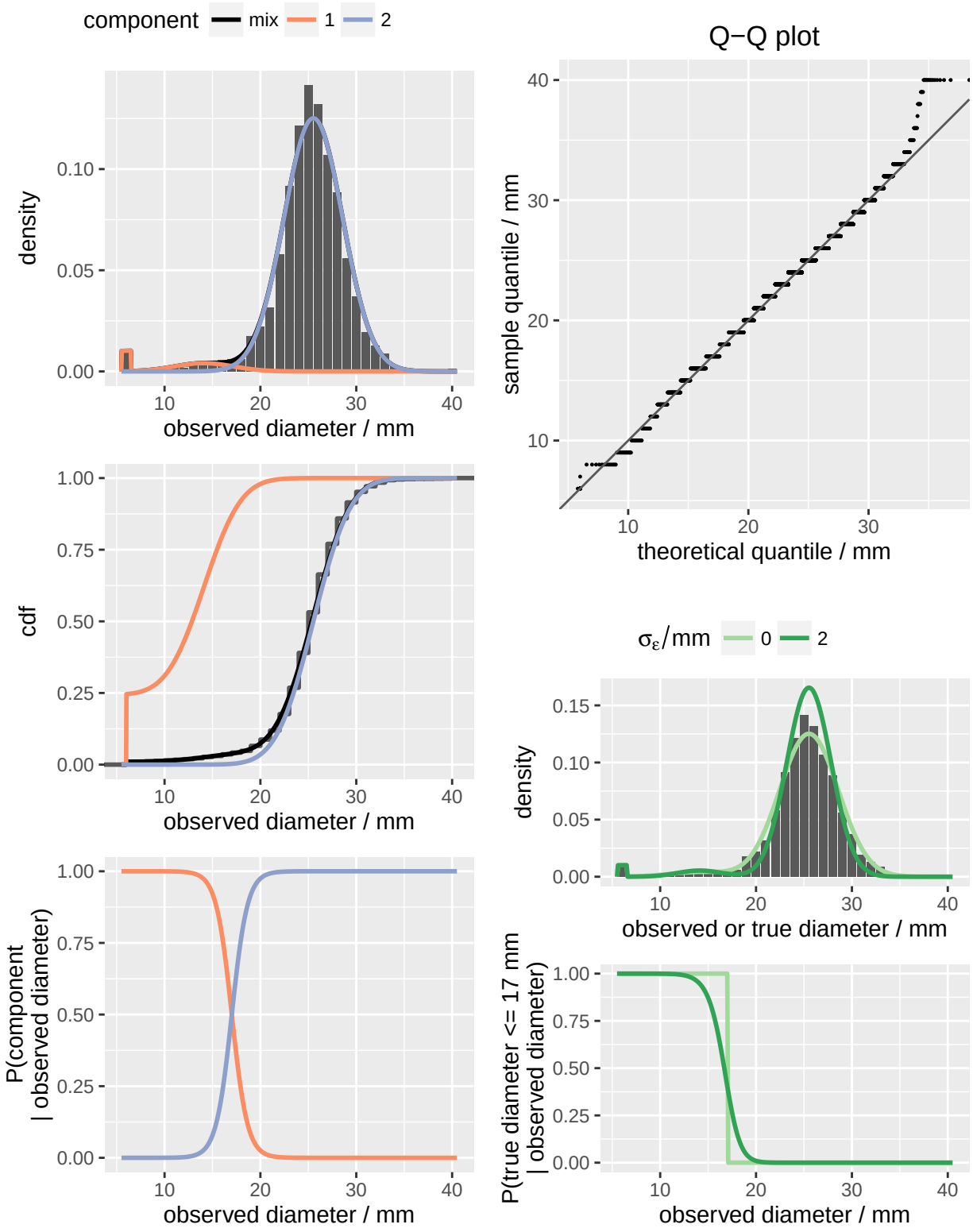
4.2 KF

$\mu_{2,3} = 13.2$ mm, 19.6 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 2.7$ mm. $w = 0.12, 0.12, 0.88$.



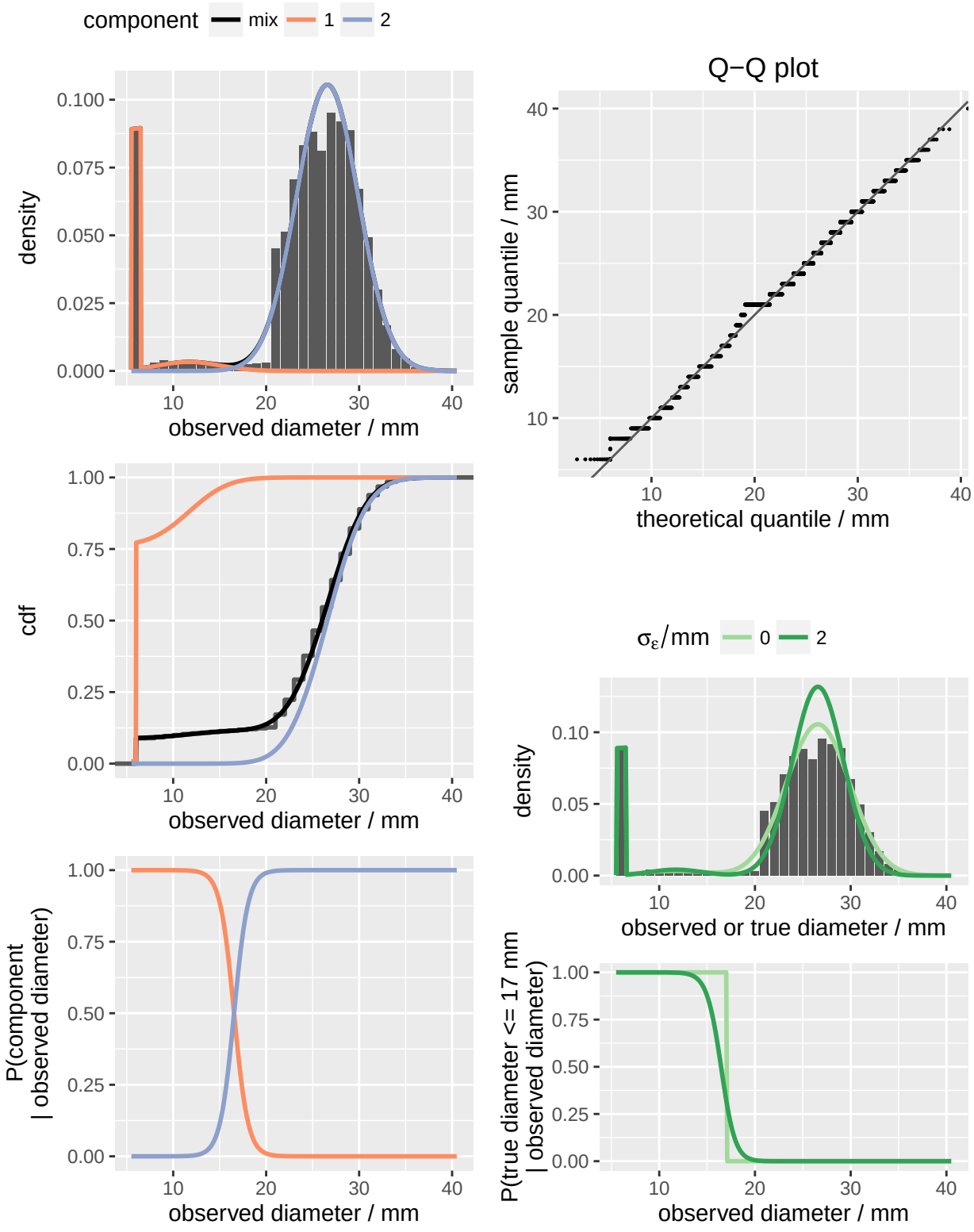
4.3 FOX

$\mu_{2,3} = 14.1$ mm, 25.5 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 3.1$ mm. $w = 0.01, 0.03, 0.97$.



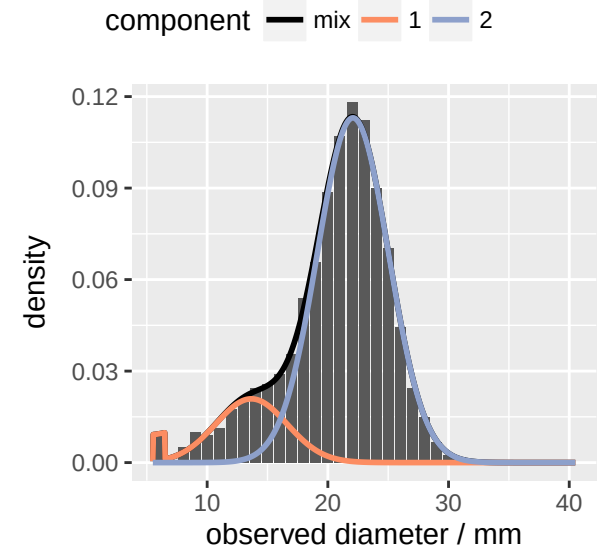
4.4 CPD

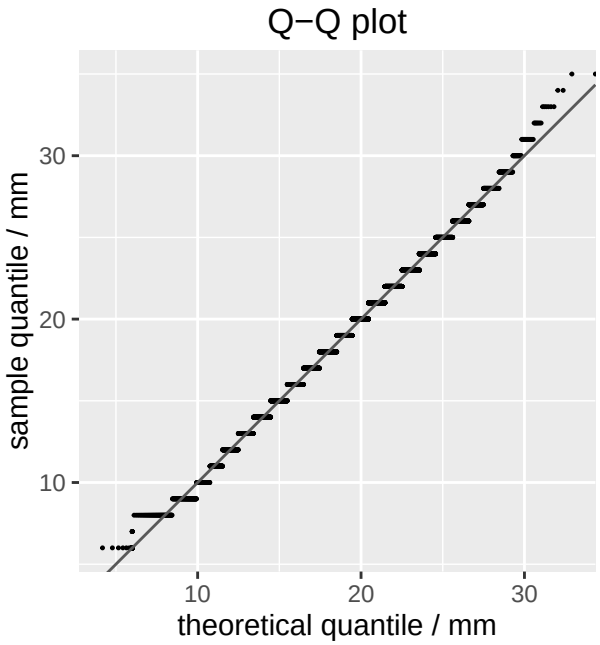
$\mu_{2,3} = 11.7$ mm, 26.6 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 3.3$ mm. $w = 0.09, 0.03, 0.97$.

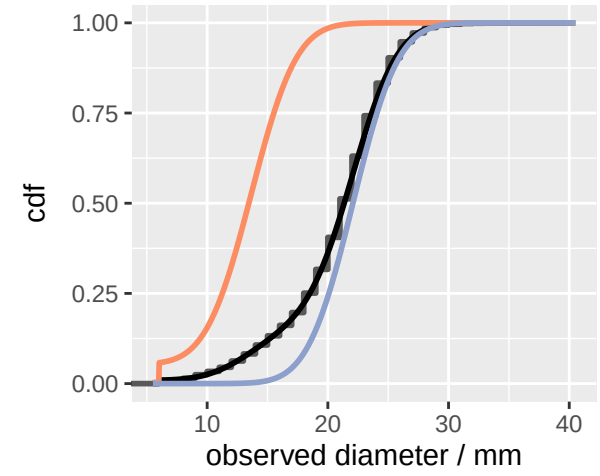


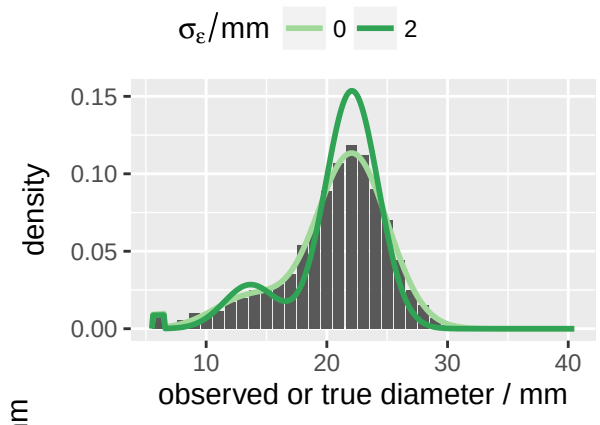
4.5 AMC

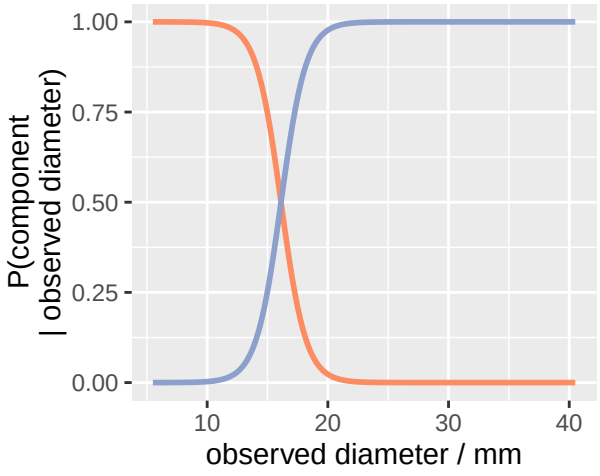
$\mu_{2,3} = 13.7$ mm, 22.1 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 3$ mm. $w = 0.01, 0.16, 0.84$.

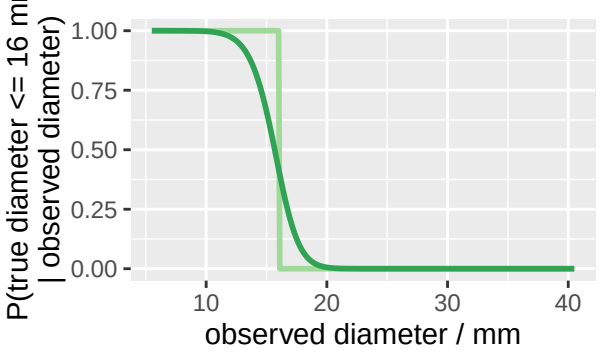






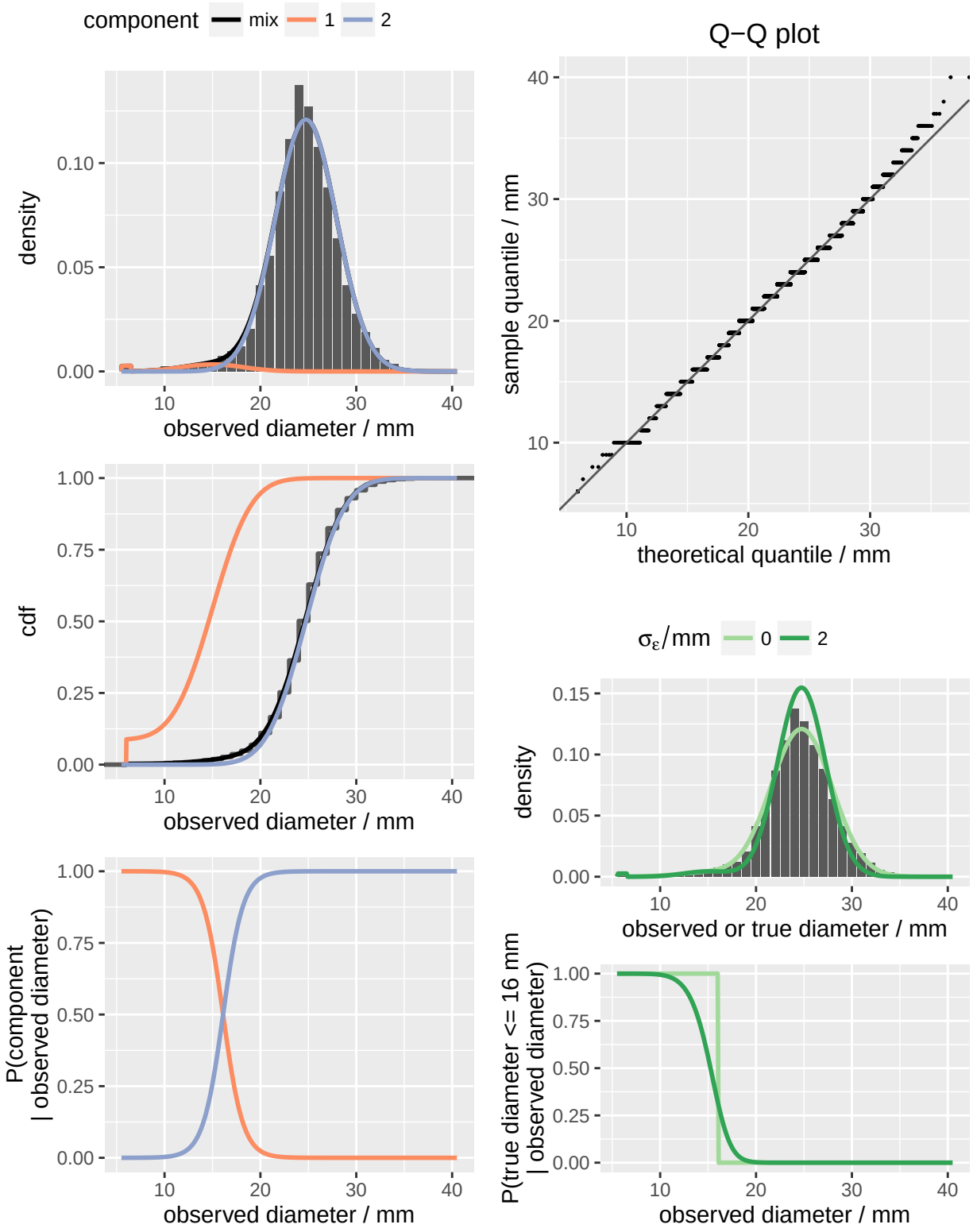






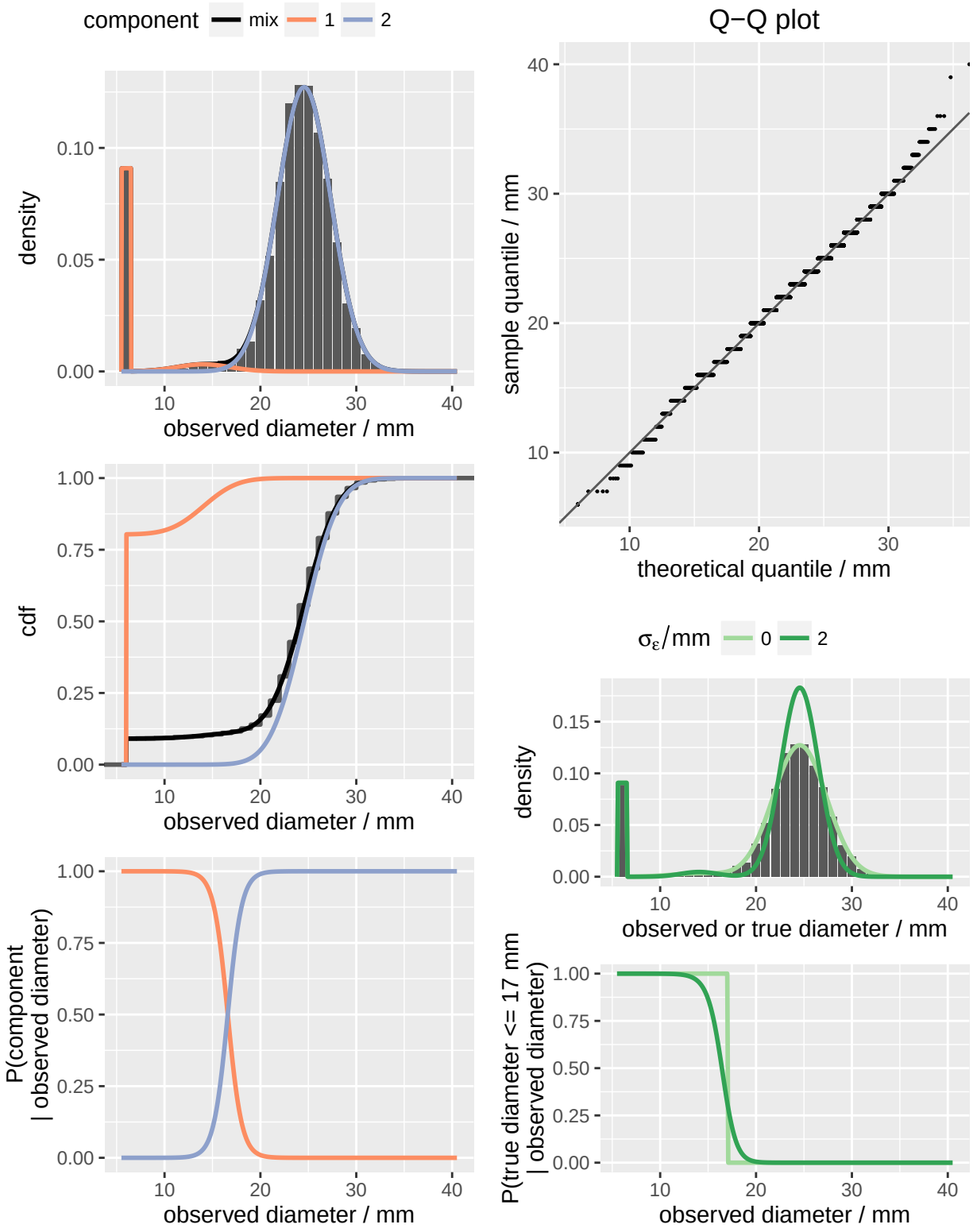
4.6 TPZ

$\mu_{2,3} = 15$ mm, 24.7 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 3.2$ mm. $w = 0$, 0.03, 0.97.



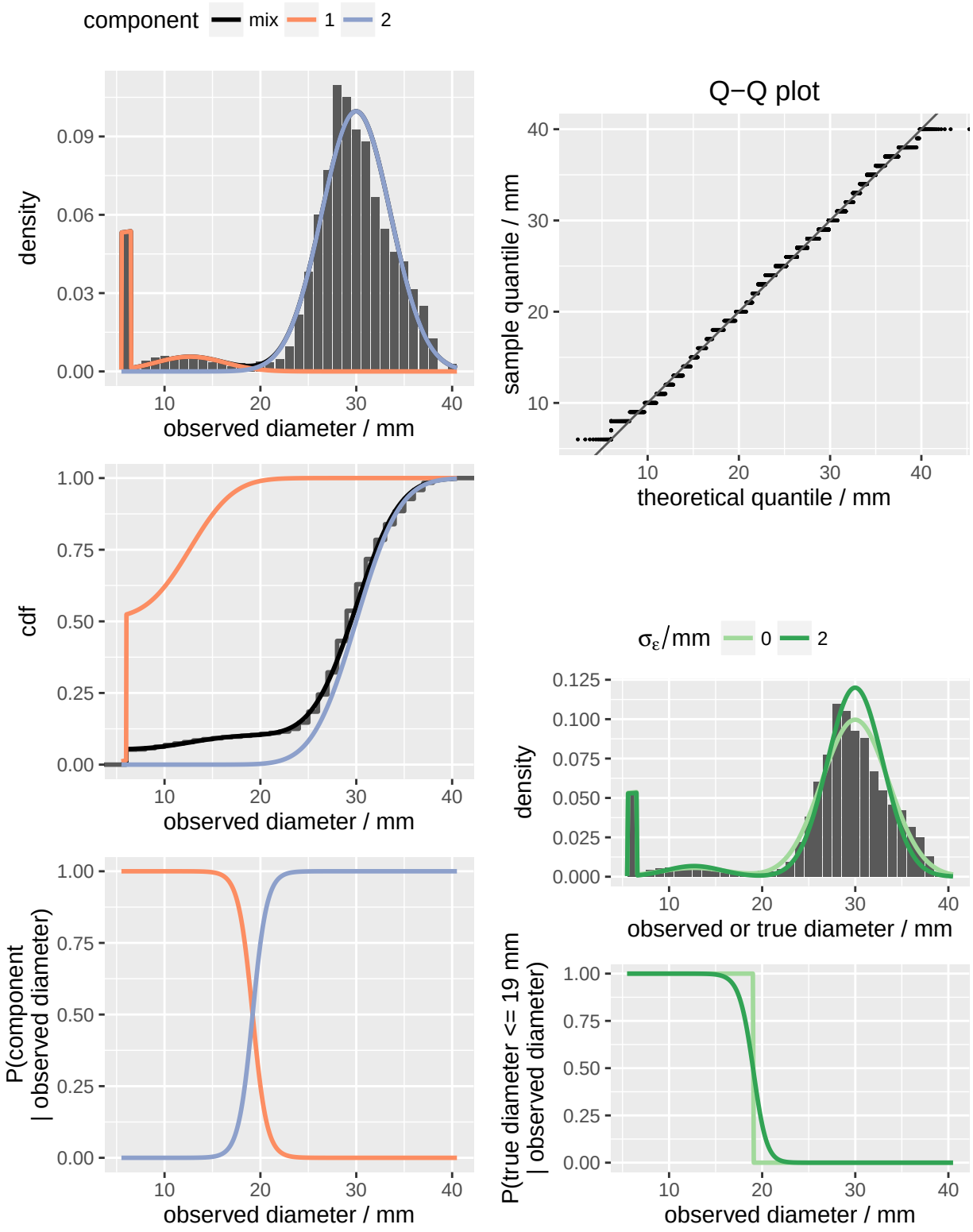
4.7 CXM

$\mu_{2,3} = 14.1$ mm, 24.6 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 2.8$ mm. $w = 0.09, 0.02, 0.98$.



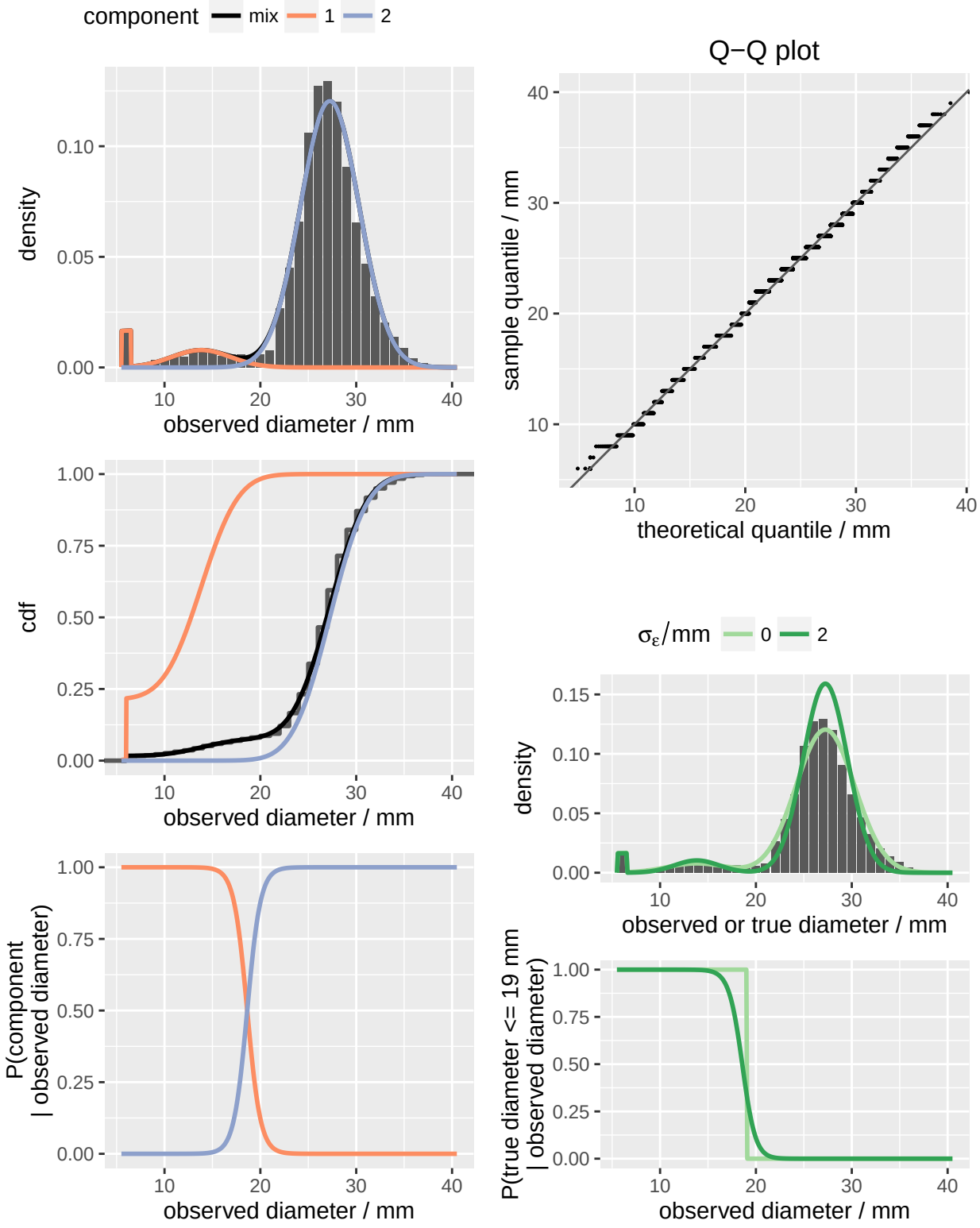
4.8 CTX

$\mu_{2,3} = 12.7$ mm, 30 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 3.6$ mm. $w = 0.05, 0.05, 0.95$.



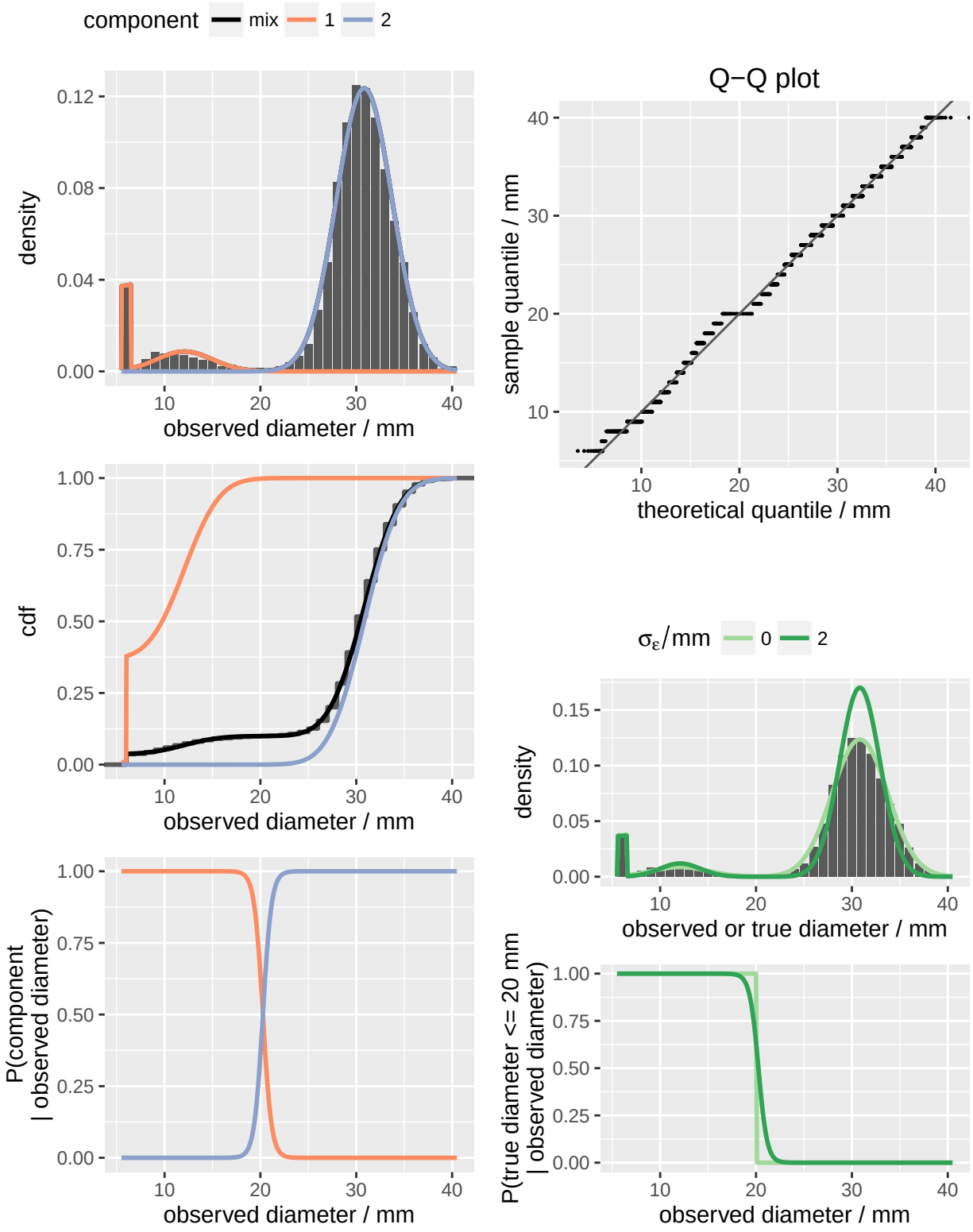
4.9 CAZ

$\mu_{2,3} = 13.8$ mm, 27.2 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 3.1$ mm. $w = 0.02, 0.06, 0.94$.



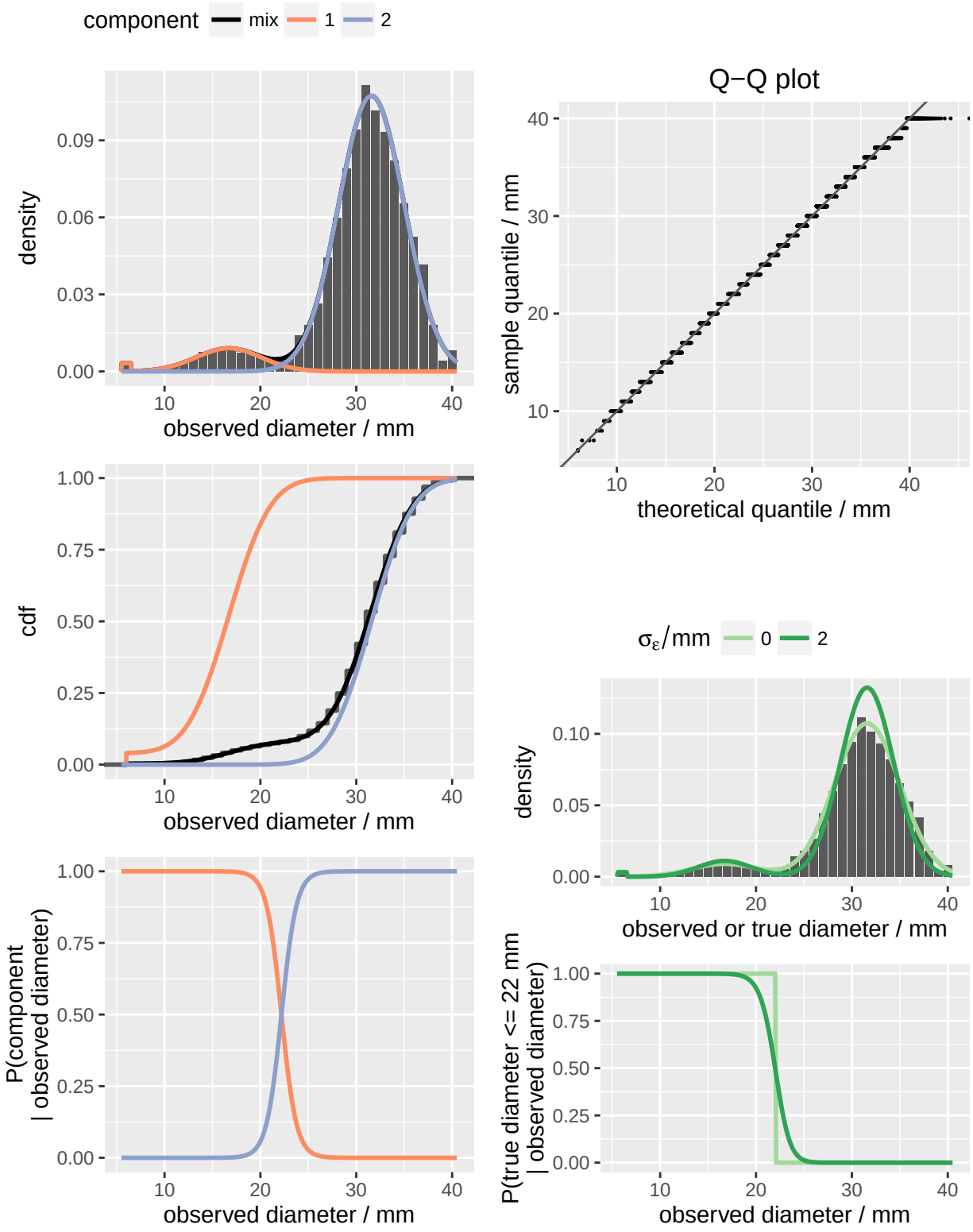
4.10 CRO

$\mu_{2,3} = 12.1$ mm, 30.8 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 2.9$ mm. $w = 0.04, 0.07, 0.93$.



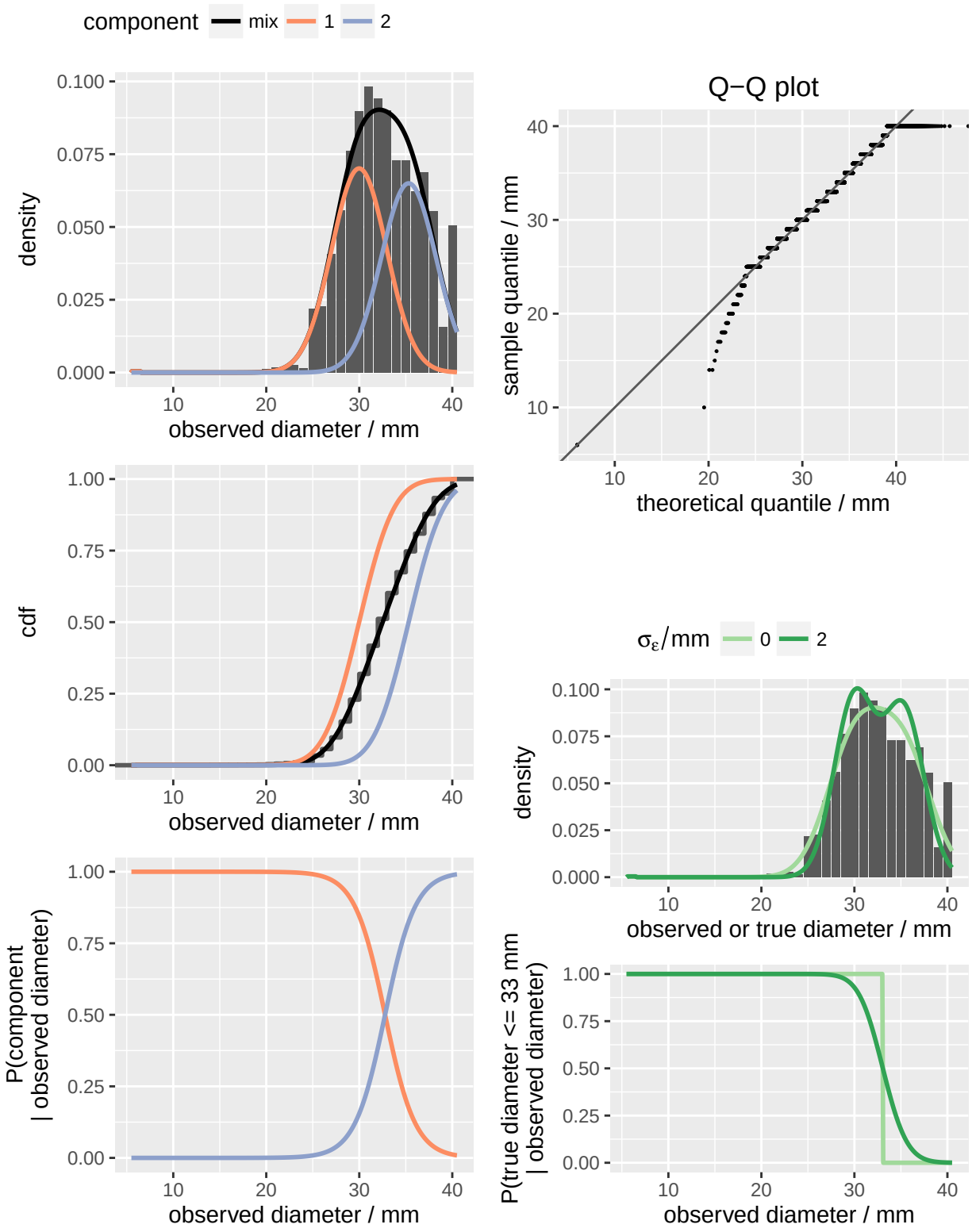
4.11 FEP

$\mu_{2,3} = 16.7$ mm, 31.6 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 3.4$ mm. $w = 0, 0.08, 0.92$.



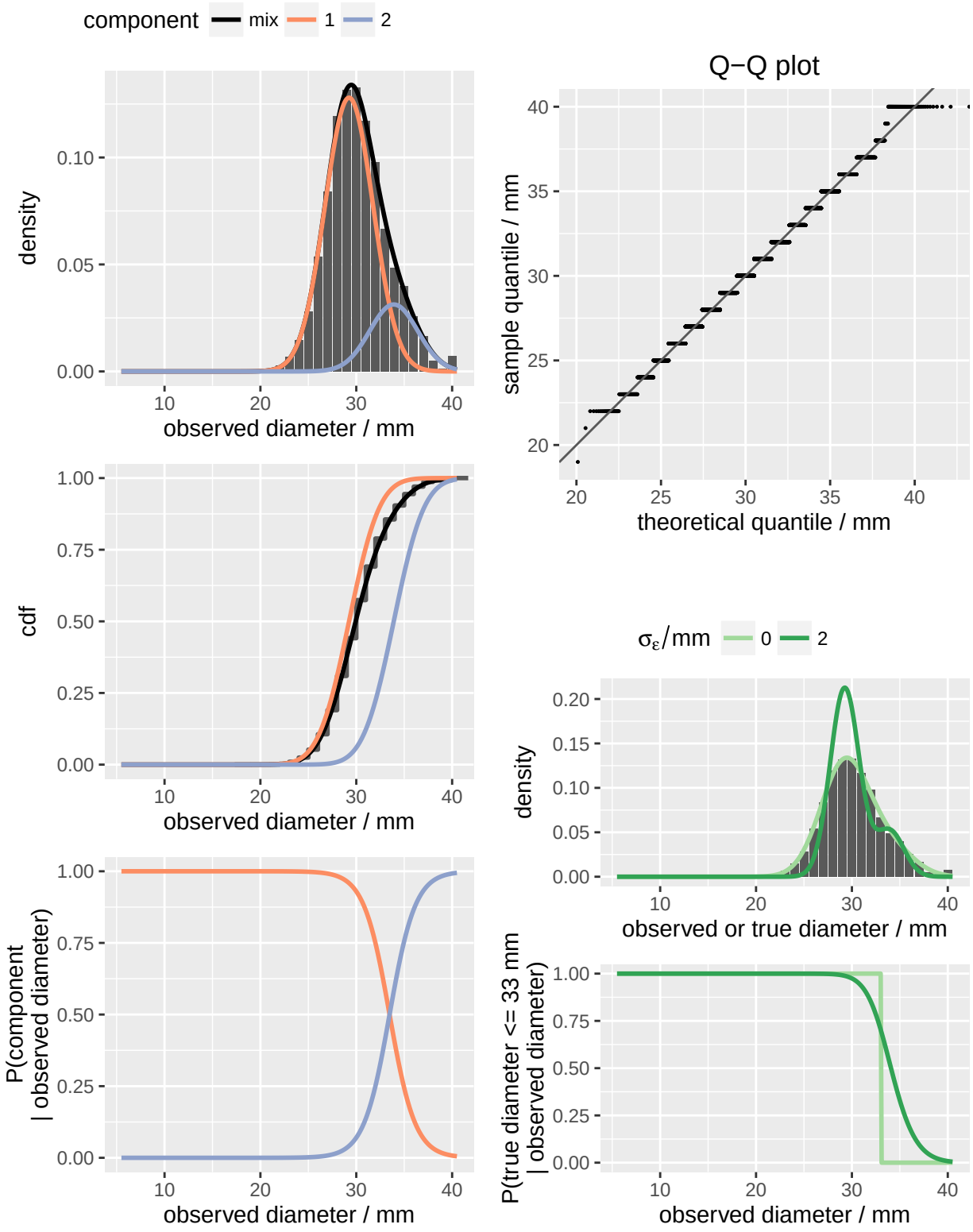
4.12 ETP

$\mu_{2,3} = 30$ mm, 35.3 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 3$ mm. $w = 0, 0.52, 0.48$.



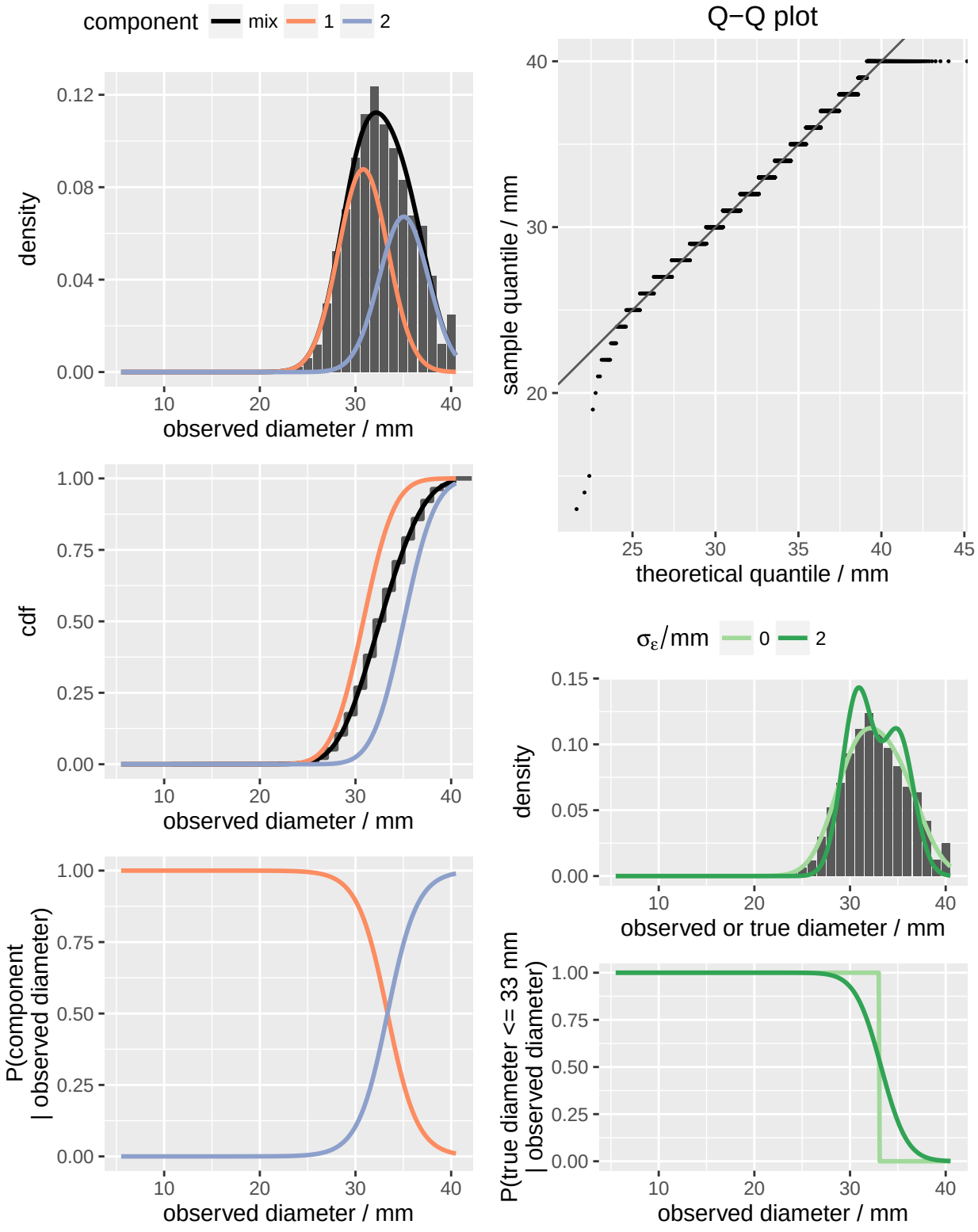
4.13 IPM

$\mu_{2,3} = 29.2$ mm, 33.9 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 2.5$ mm. $w = 0.8, 0.2$.



4.14 MEM

$\mu_{2,3} = 30.8$ mm, 35.1 mm. $\sqrt{\sigma^2 + \sigma_E^2} = 2.6$ mm. $w = 0.57, 0.43$.



5 Conclusion

In discussions with Peter Keller and Michael Hombach on May, 11th and 12th, 2016, we decided to split this project according to the goals stated in Sec. 1. Roadmap:

- Compare p(pseudo-WT|observed diameter) with ground truth.
- Compute the probability of very major errors for CBP and CBP + 2 mm.
- Investigate robustness of model p(S|observed diameter).
- Extend analysis to quinolones, tetracyclines, aminoglycosides, cholistin, etc. using data cleaned by Giorgia Valsesia.
- Plan manuscript.
- Meet with Marc Schmid regarding implementation.

6 Appendix

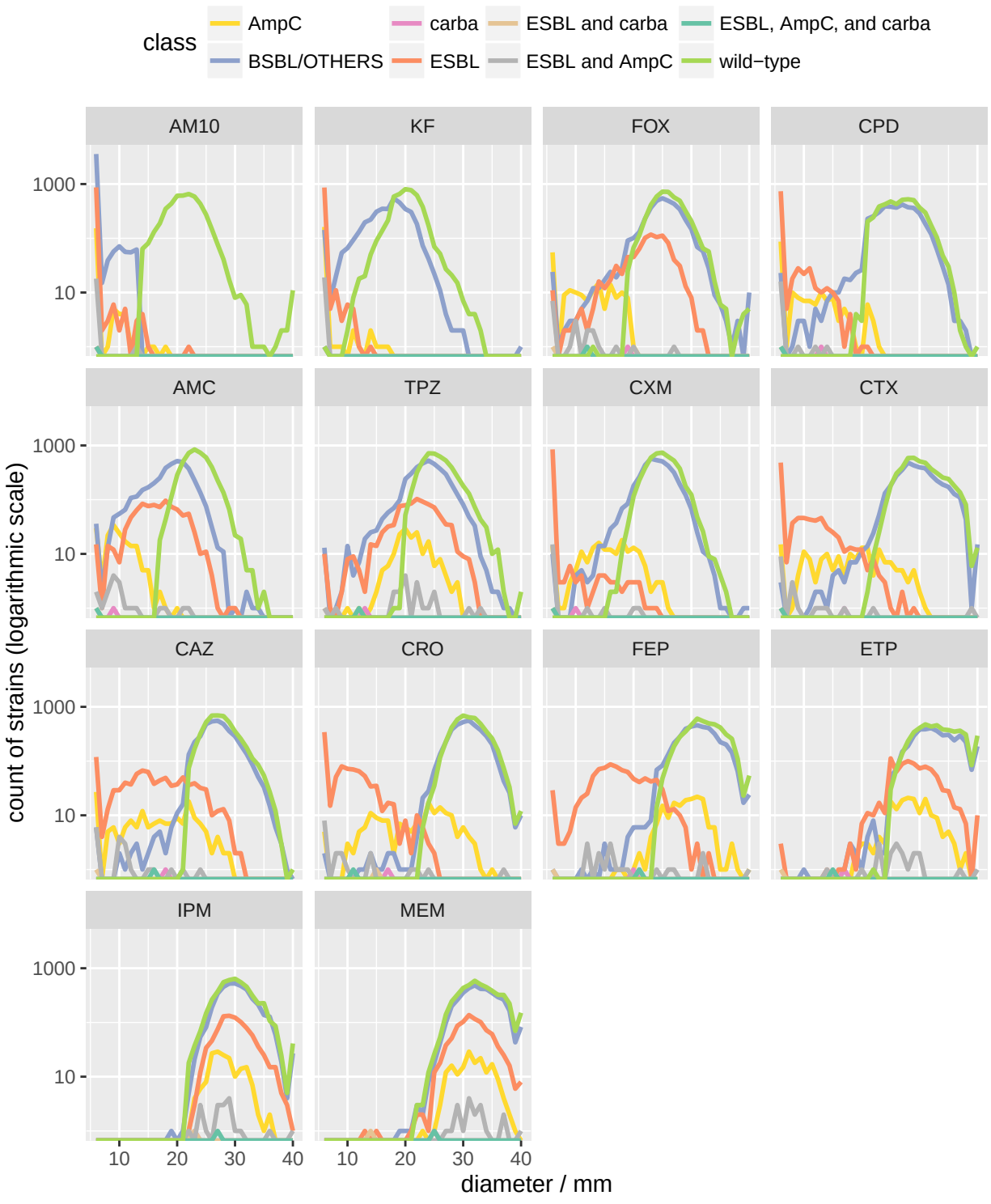


Figure 1: Distributions of diameters.