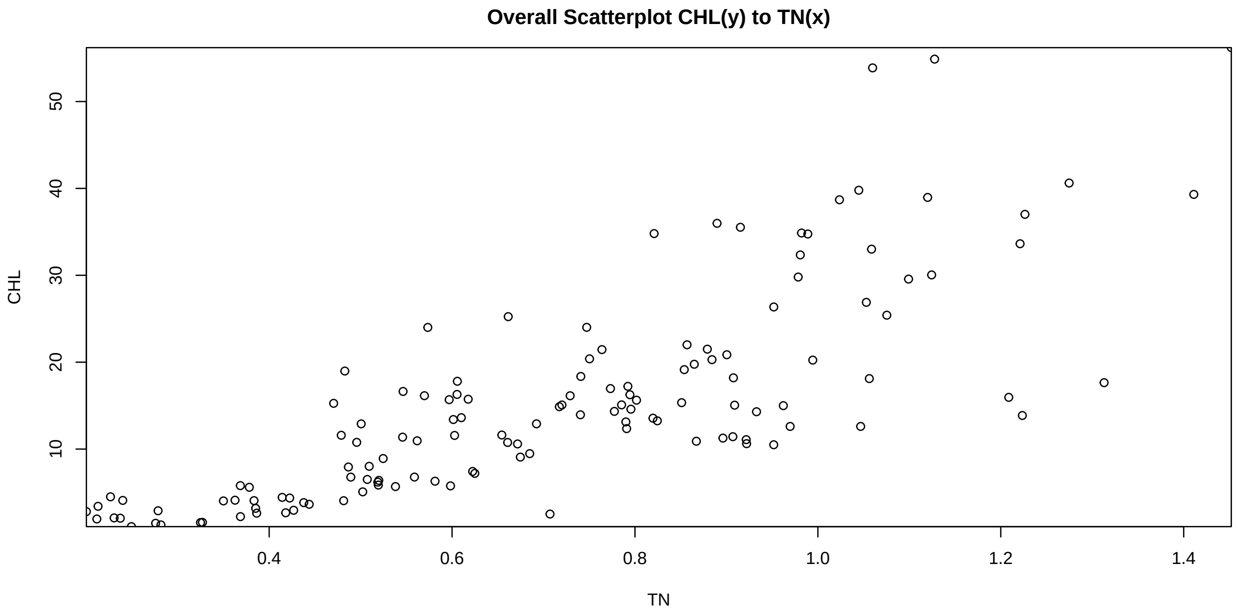
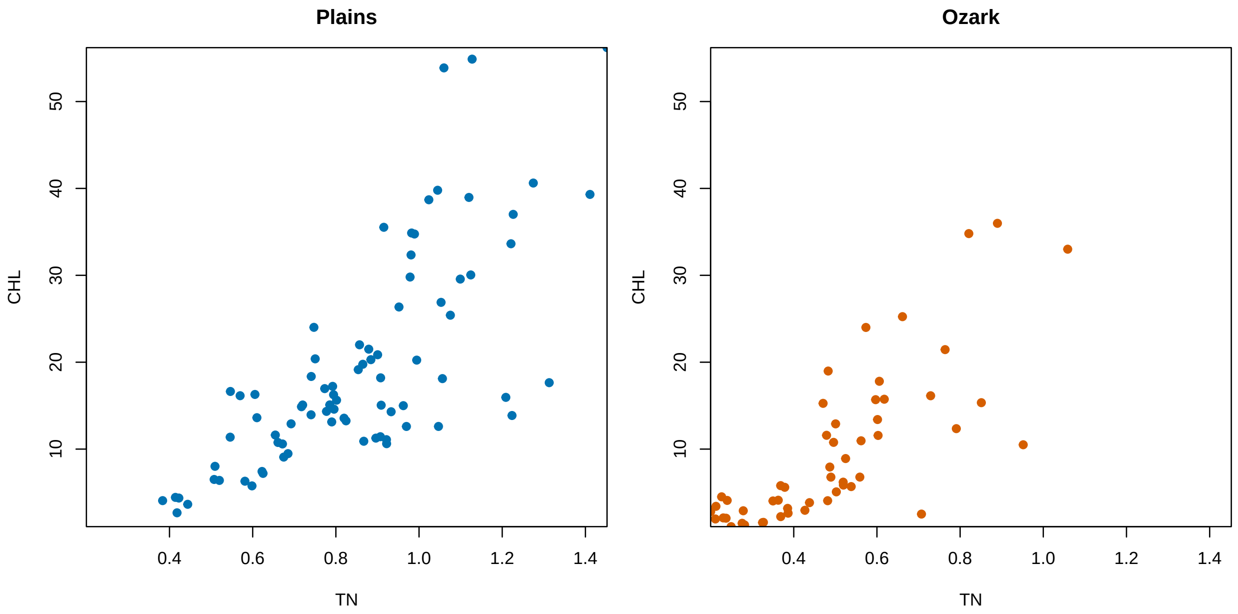
5200 Take-Home

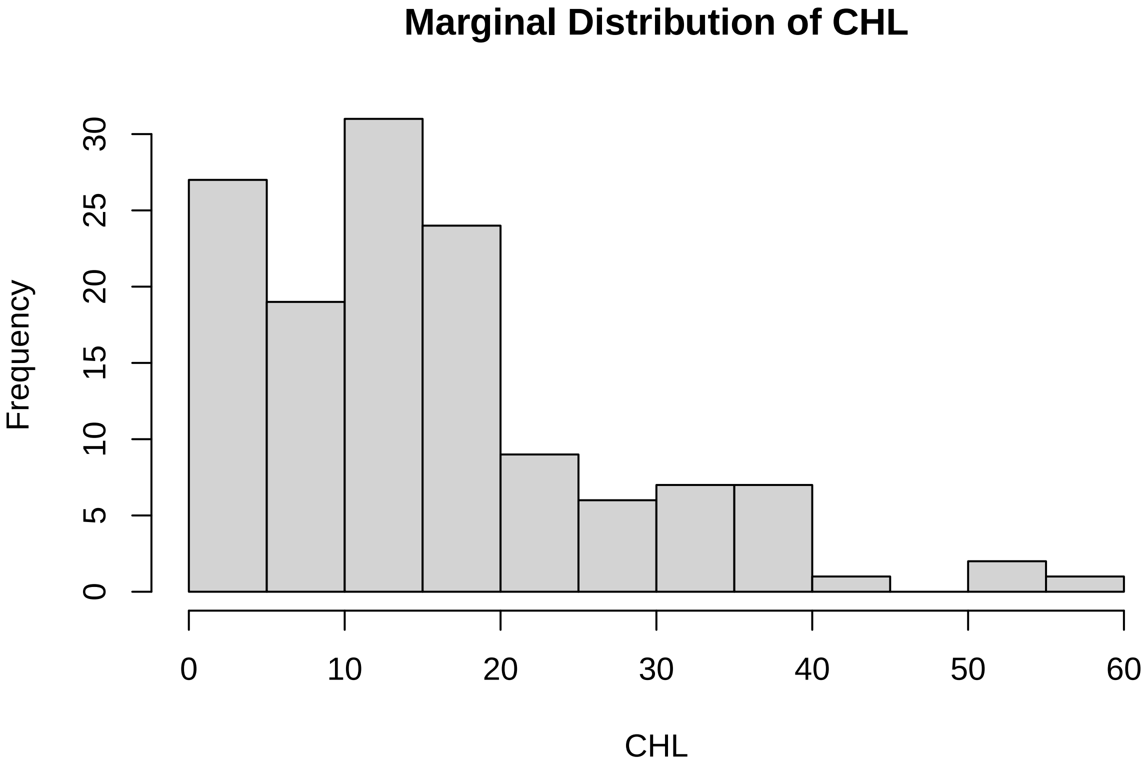
Sam Olson

Q1: CHL & TN (Plains vs. Ozarks)

Overall Distribution & Approach







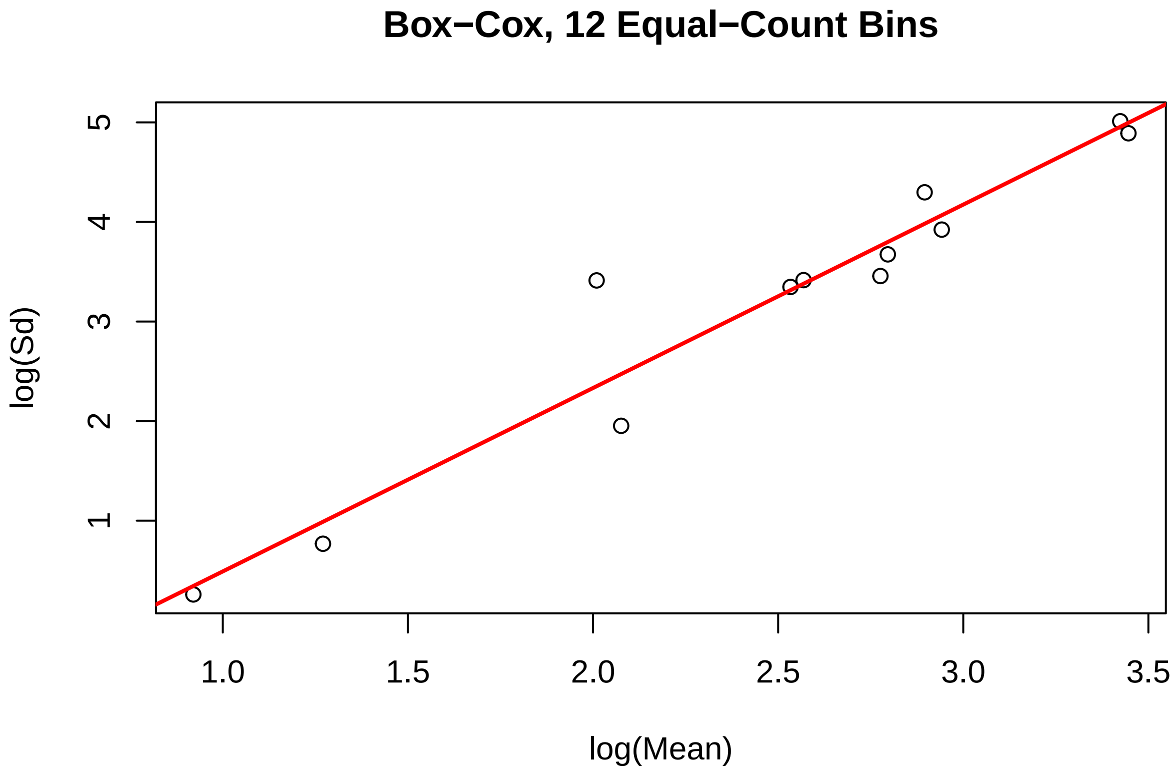
The distribution of CHL is right-skewed, which is consistent with ecological expectations for nutrient–algal processes. There is a clear positive, potentially nonlinear relationship between TN and CHL, with variance increasing at larger TN values. Scatterplots stratified by region show the same general patterns (positive trend, possible nonlinearity, increasing variance), though the ranges for the Plains and Ozarks differ while still overlapping.

Because chlorophyll–TN dynamics arise from the same underlying biological mechanism across all Missouri reservoirs, we first identify a global mean–variance model for CHL as a function of TN. This provides the shared functional form of the TN–CHL relationship that applies across regions.

After identifying this overall model, we fit the same model structure separately to the Plains and Ozarks. Differences between the region-specific fits—such as changes in intercept, slope, curvature, or dispersion—then reflect true regional differences in the strength or level of the TN–CHL relationship rather than differences introduced by choosing different model families or transformations.

This approach ensures that regional comparisons reflect actual ecological differences, avoids artifacts from inconsistent modeling, and allows direct comparison of parameters, confidence bands, and derived quantities such as $\Pr(Y > Z \mid x)$.

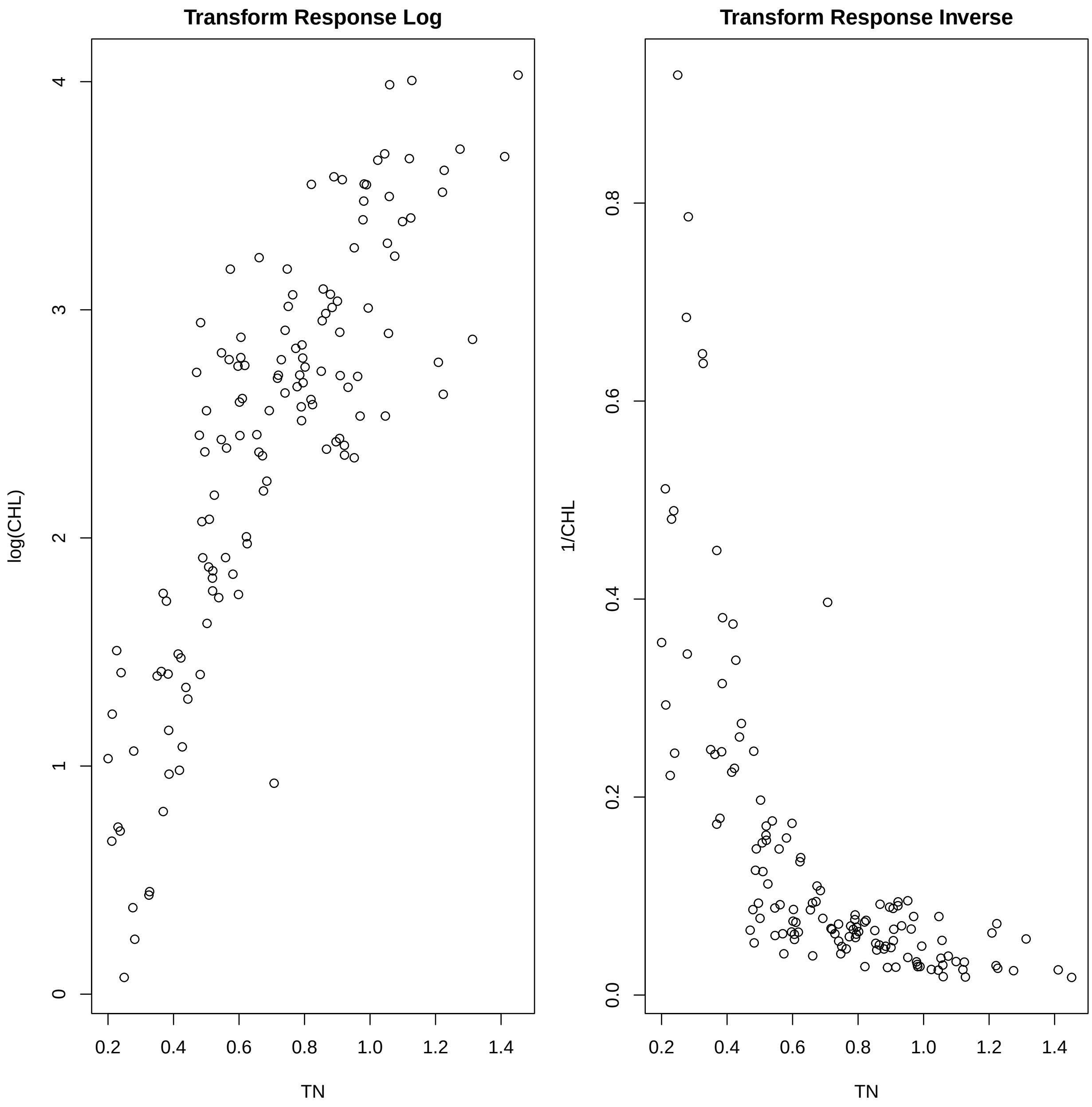
Generalized Linear Models



nbins	Equal.Count	Equal.Spaced
6	1.76	1.85
10	2.04	1.95
12	1.84	1.94
14	1.97	2.10
16	1.82	1.86
18	1.77	1.87
22	1.88	1.61

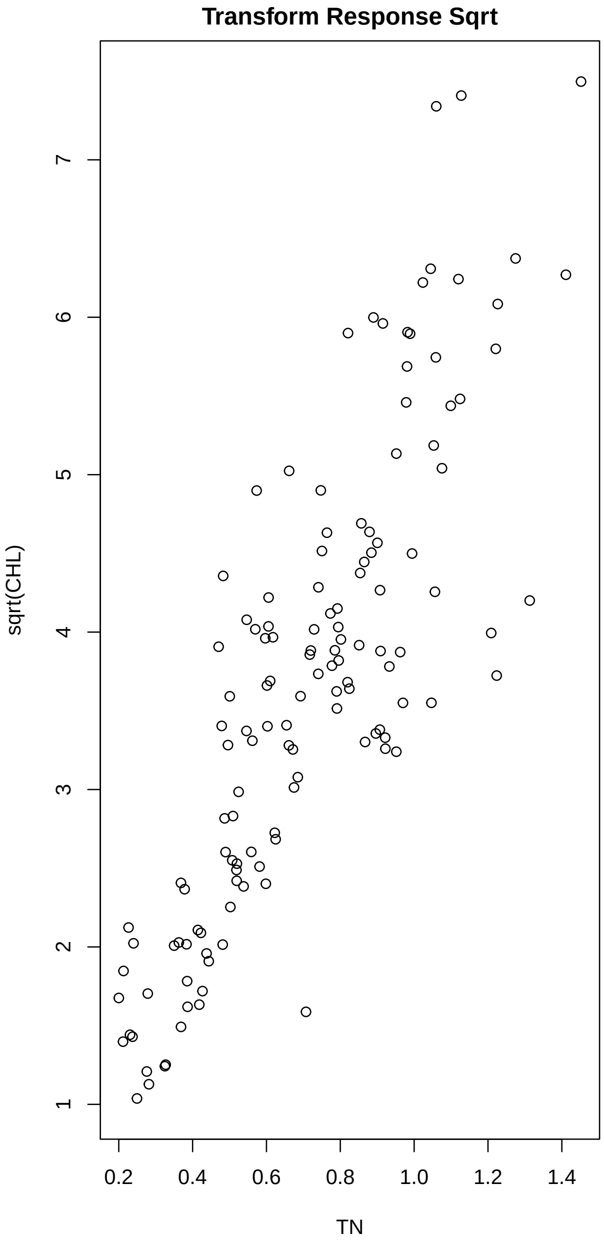
Based on the initial examination, we begin by considering suitable generalized linear models. Using the provided `boxcoxfctns` functions, I computed Box–Cox mean–variance slopes to identify an appropriate random component. The slopes were consistently between 1.6 and 2.1 across binning schemes, equally-spaced or equal-counts, indicating a variance pattern approximately proportional to μ^2 to μ^3, suggesting that either a Gamma or Inverse Gaussian random component is appropriate for the overall model.

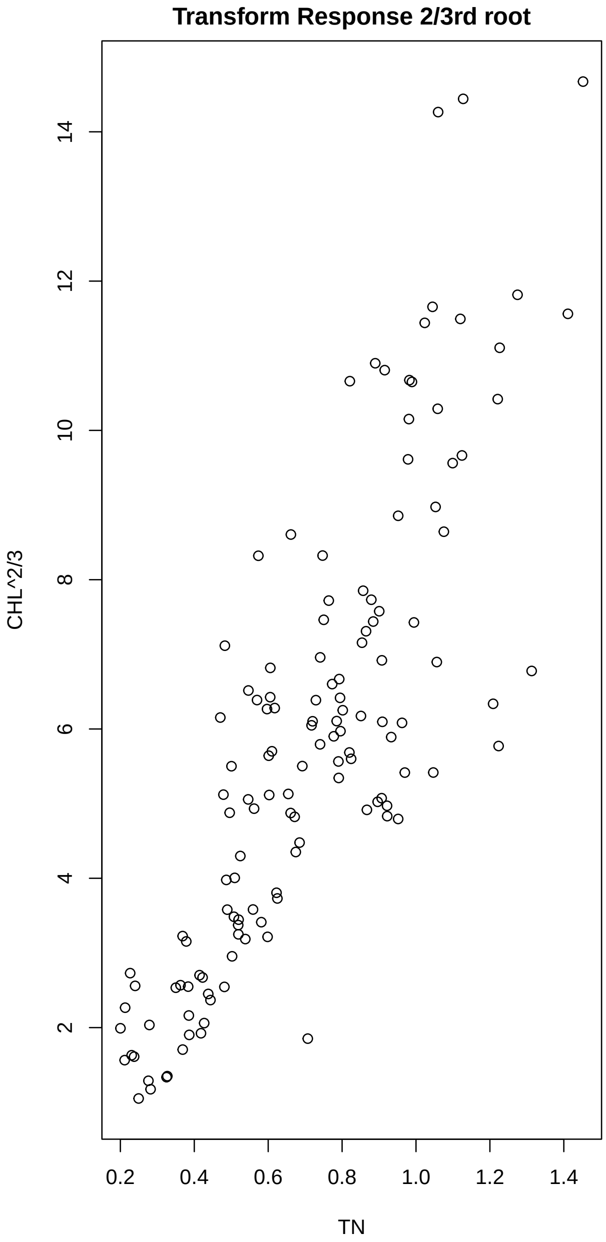
With suitable random component(s) identified, the next step is to identify a suitable systematic link function for modeling the mean relationship between TN and CHL.

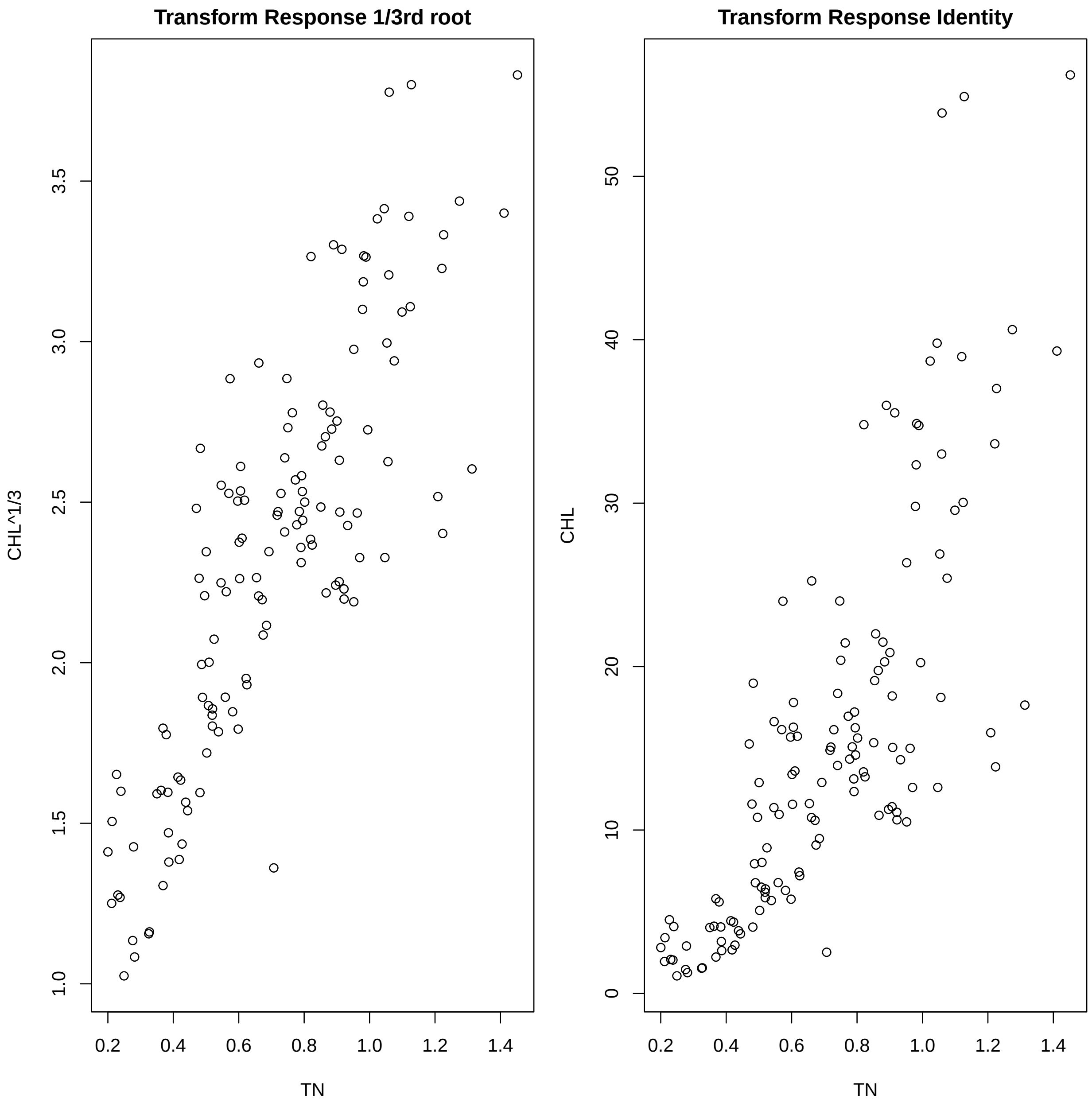


Transform Response Log

Transform Response Inverse


Transform Response Sqrt
sqrt(CHL)
1
2
3
4
5
6
7
TN
0.2
0.4
0.6
0.8
1.0
1.2
1.4

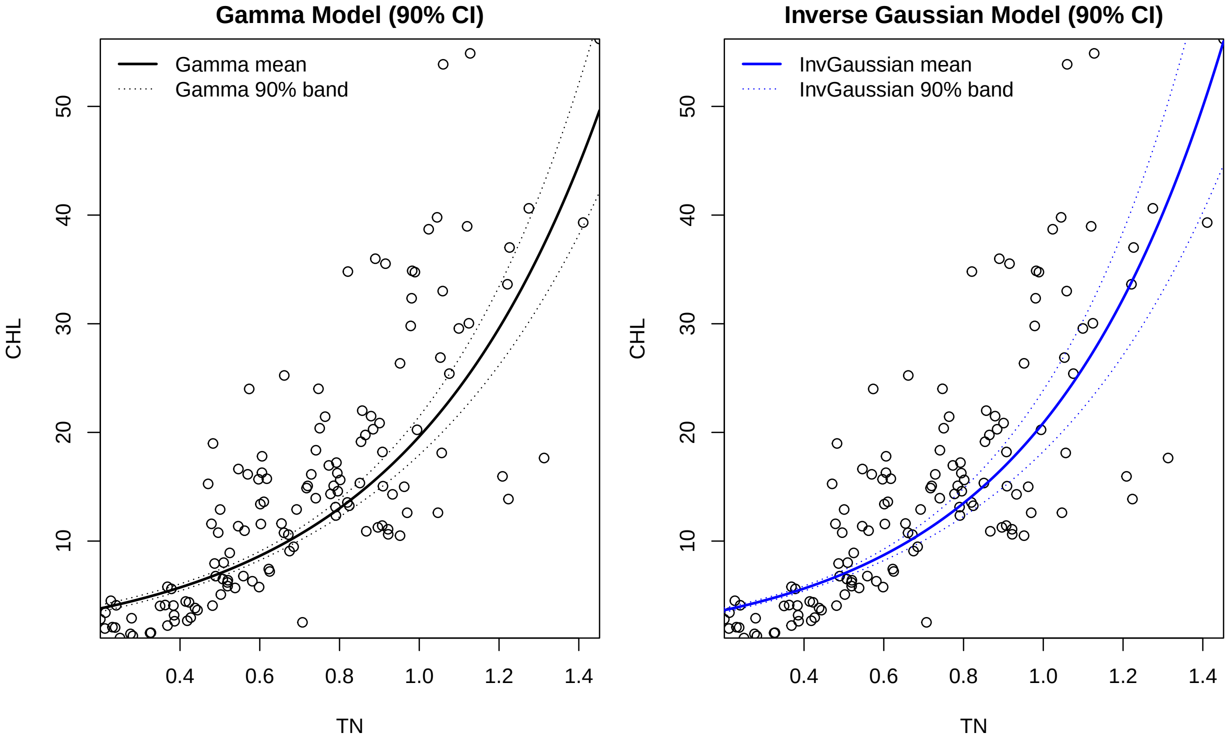

Transform Response 2/3rd root
CHL^2/3
2
4
6
8
10
12
14
TN
0.2
0.4
0.6
0.8
1.0
1.2
1.4

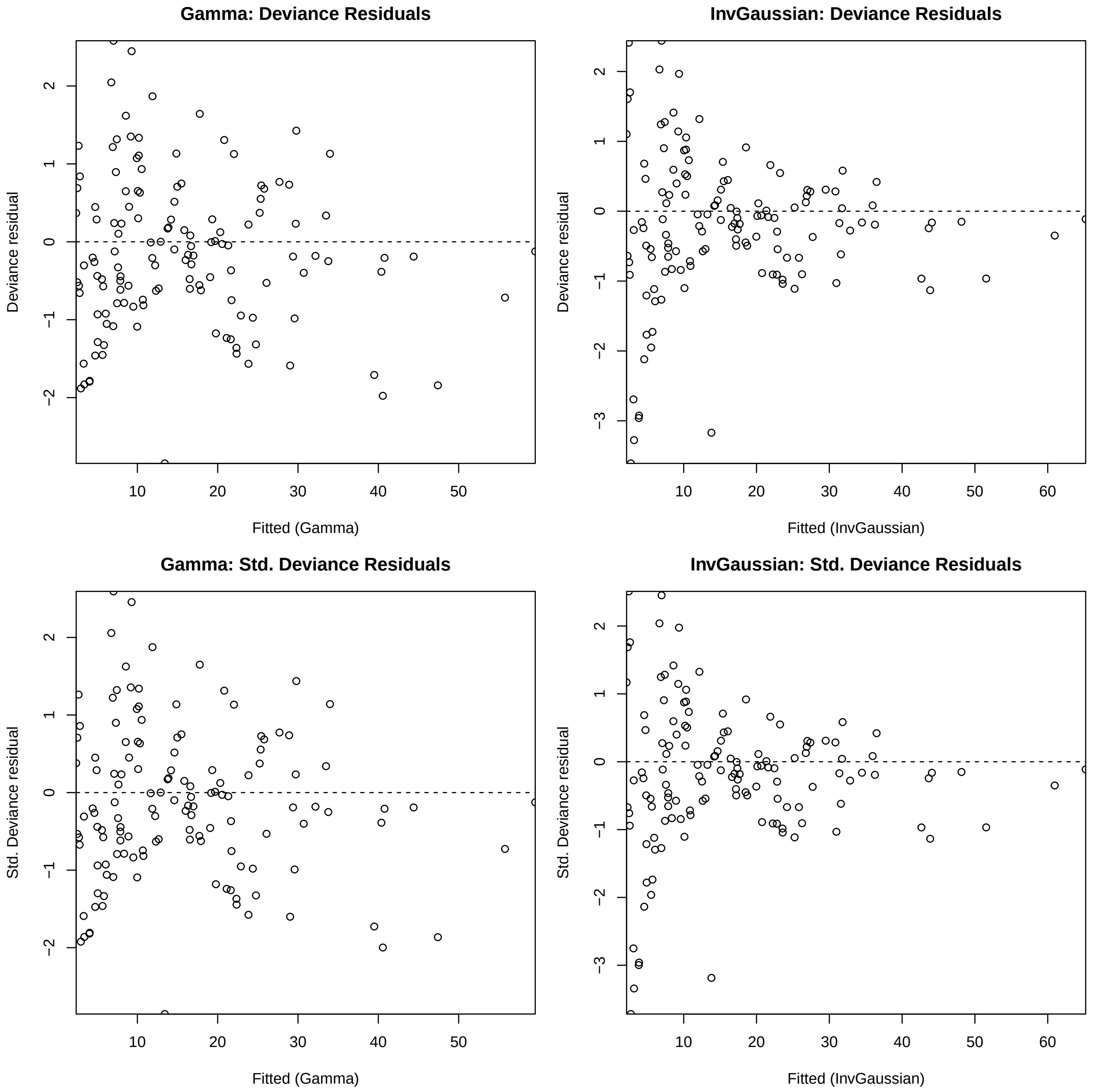


When identifying a suitable link function, the objective is to find a transformation of the mean that approximately linearizes the relationship between CHL and TN. Exploratory plots indicate that a cube-root transformation provides a reasonably linear trend, with the square-root transformation also performing adequately. Importantly, these transformations are used only to guide link selection; they do not imply transforming the CHL values themselves for the GLM.

Taken toghether with the random component selection done previously, in total: We are motivated to use Gamma and Inverse Gaussian random components ($\theta = 2, 3$), paired with power links corresponding to cube-root and square-root transformations. We then fit these models and compare.

Within a given generalized linear model family (i.e., with the random component held fixed), the scaled deviance provides a suitable likelihood-based criterion for comparing link functions. Using this measure, the cube-root power link consistently produced lower unscaled and scaled deviances than the square-root link within both the Inverse Gaussian and Gamma families. This motivates focusing on the Gamma and Inverse Gaussian models fitted with the cube-root link, and then comparing their fitted values and curve shapes to determine which provides the better overall fit. We do not compare the Gamma and Inverse Gaussian models directly using deviances, because the two random components correspond to non-nested models, making LRT-based comparisons inappropriate.



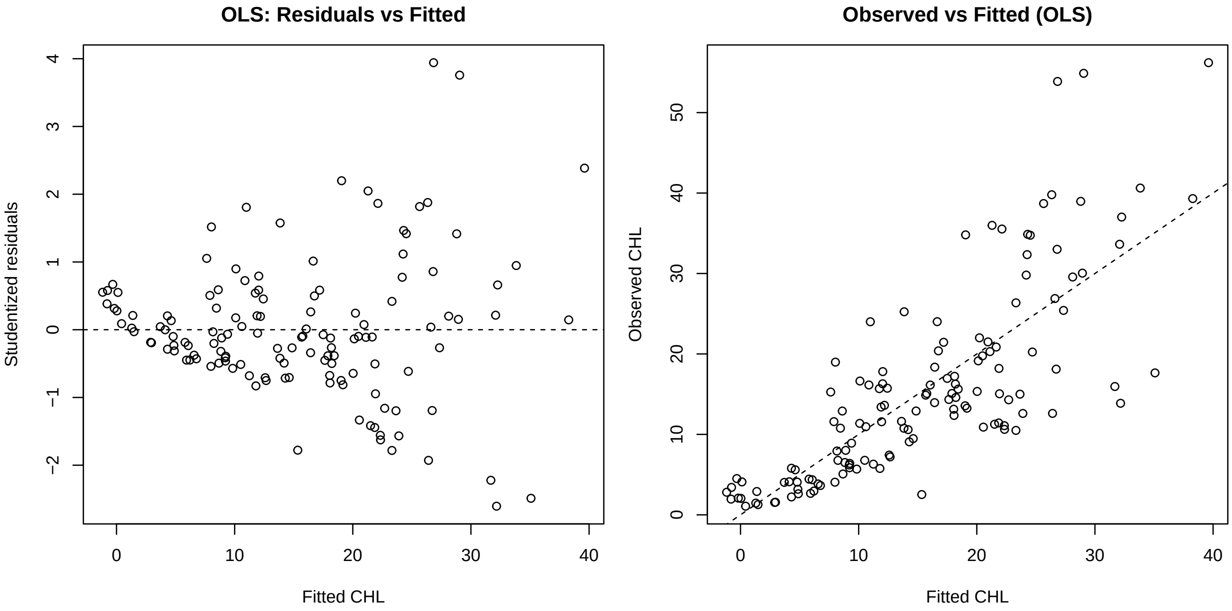


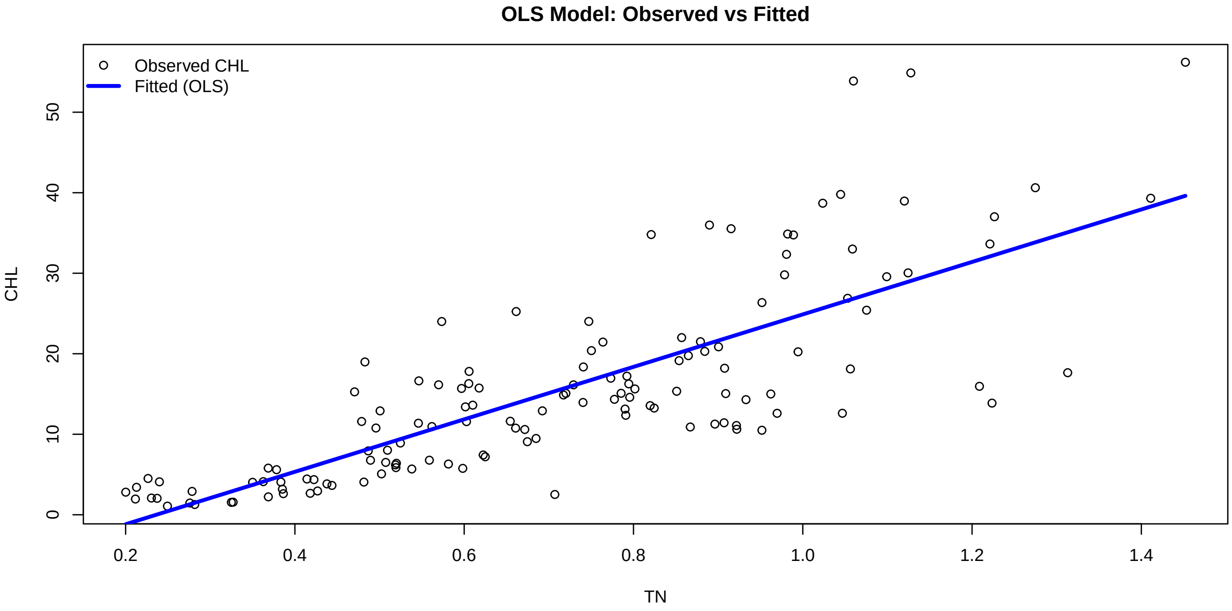
It is difficult to distinguish between the Gamma and Inverse Gaussian models using the fitted–versus–observed scatterplots alone, as both produce very similar mean curves. One might argue that the Inverse Gaussian captures the increasing variability at higher TN values (TN > 1) slightly more closely, but this visual difference is subtle. Therefore, it is more appropriate to rely on residual diagnostics to compare the models directly. Based on these diagnostics, the Gamma model ultimately provides the better overall fit.

With this preferred generalized linear model identified, we next examine a range of additive error models for completeness. By evaluating these alternatives alongside the GLM, we select a single most appropriate model to use for the subsequent region-specific comparisons.

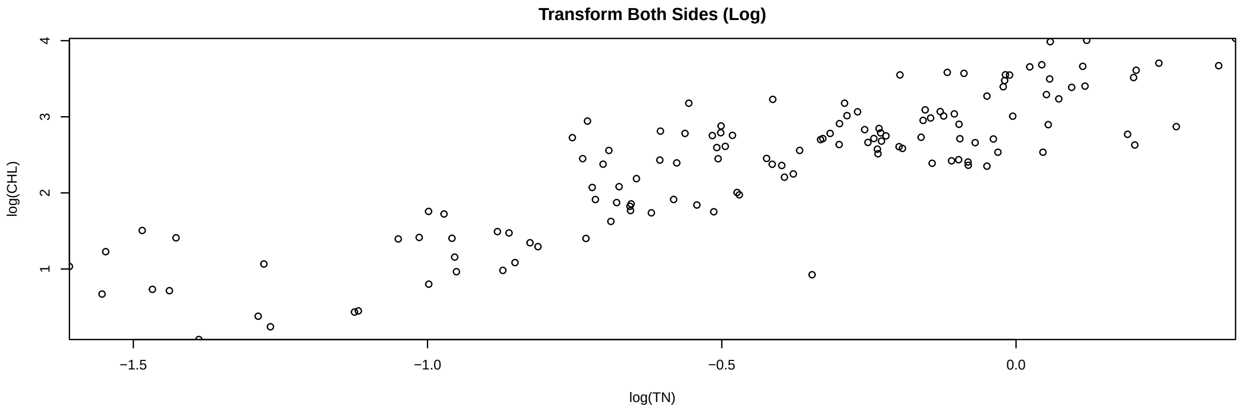
Additive Error Models

Transform Both Sides

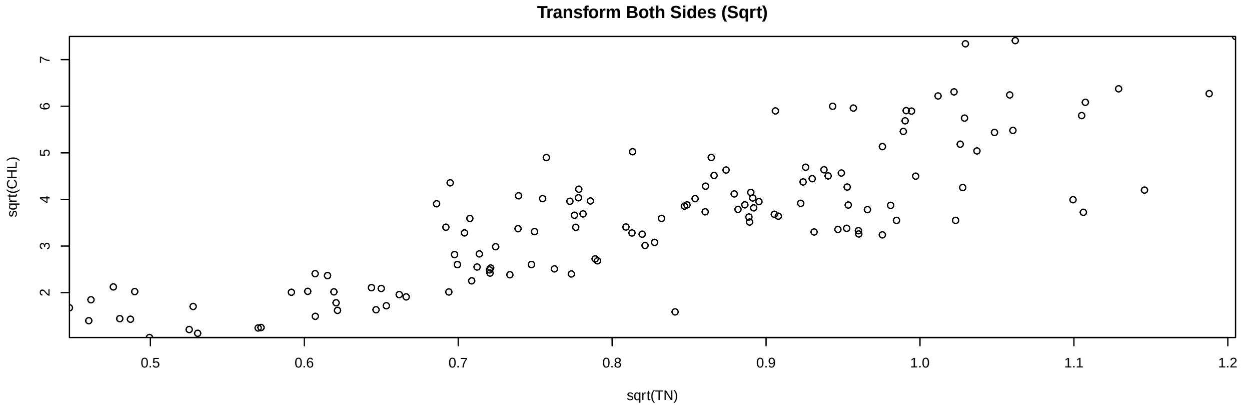




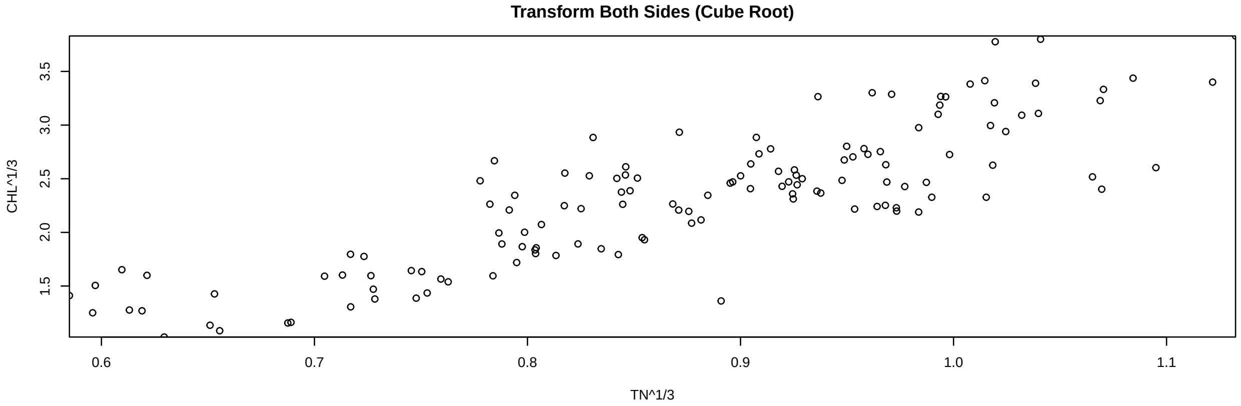
We start with a typical OLS univariate regression. While the fit seems fairly adequate (roughly linear), the studentized pearson residuals indicate possible heteroscedasticity. Because of this, we are potentially motivated to consider a transform both sides additive error model. To identify an appropriate transformation to stabilize variance, we test a number of options.



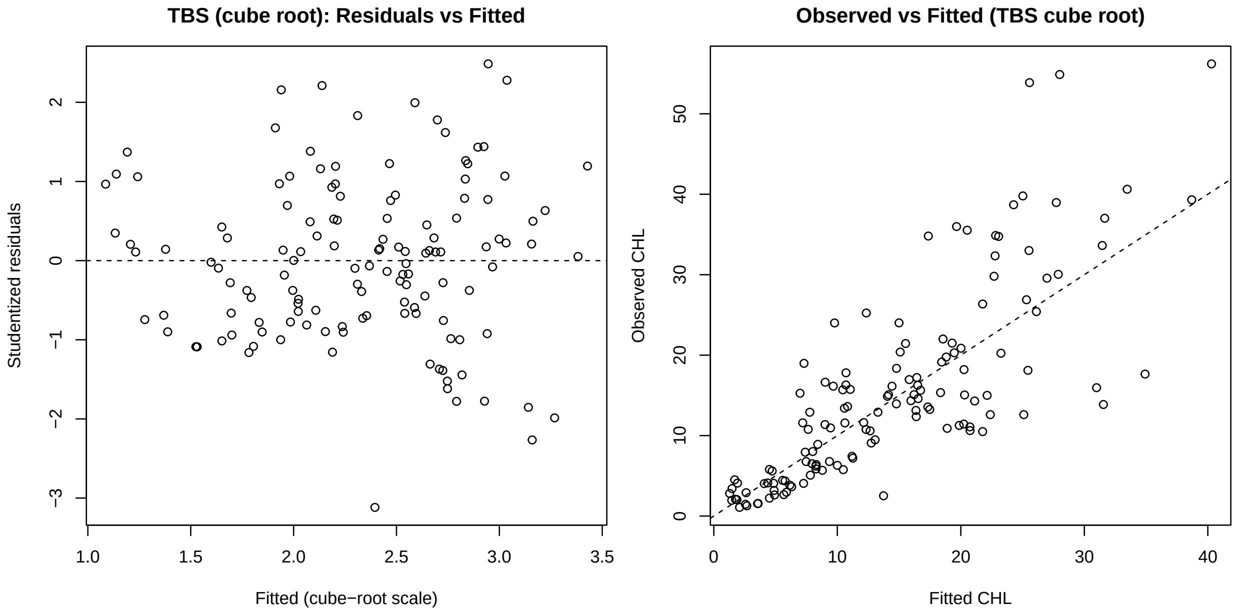
Transform Both Sides (Log)

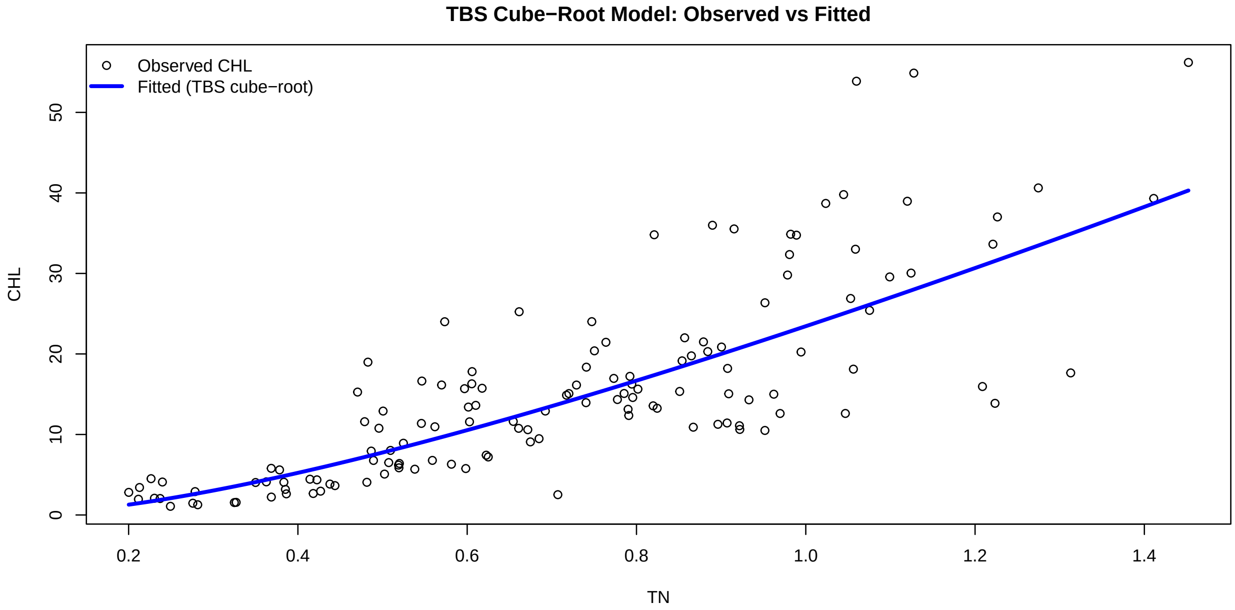


Transform Both Sides (Sqrt)



Transform Both Sides (Cube Root)



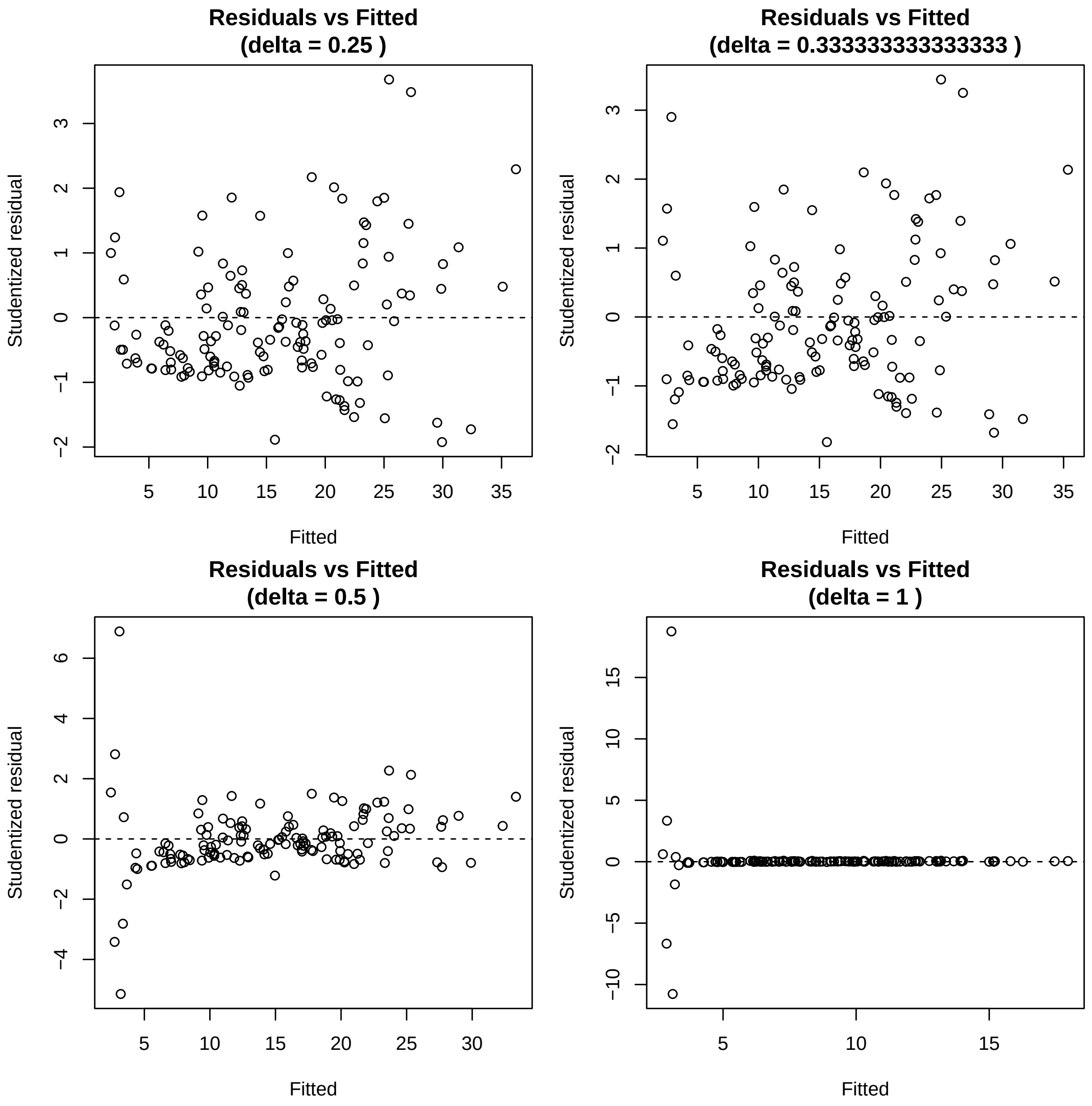


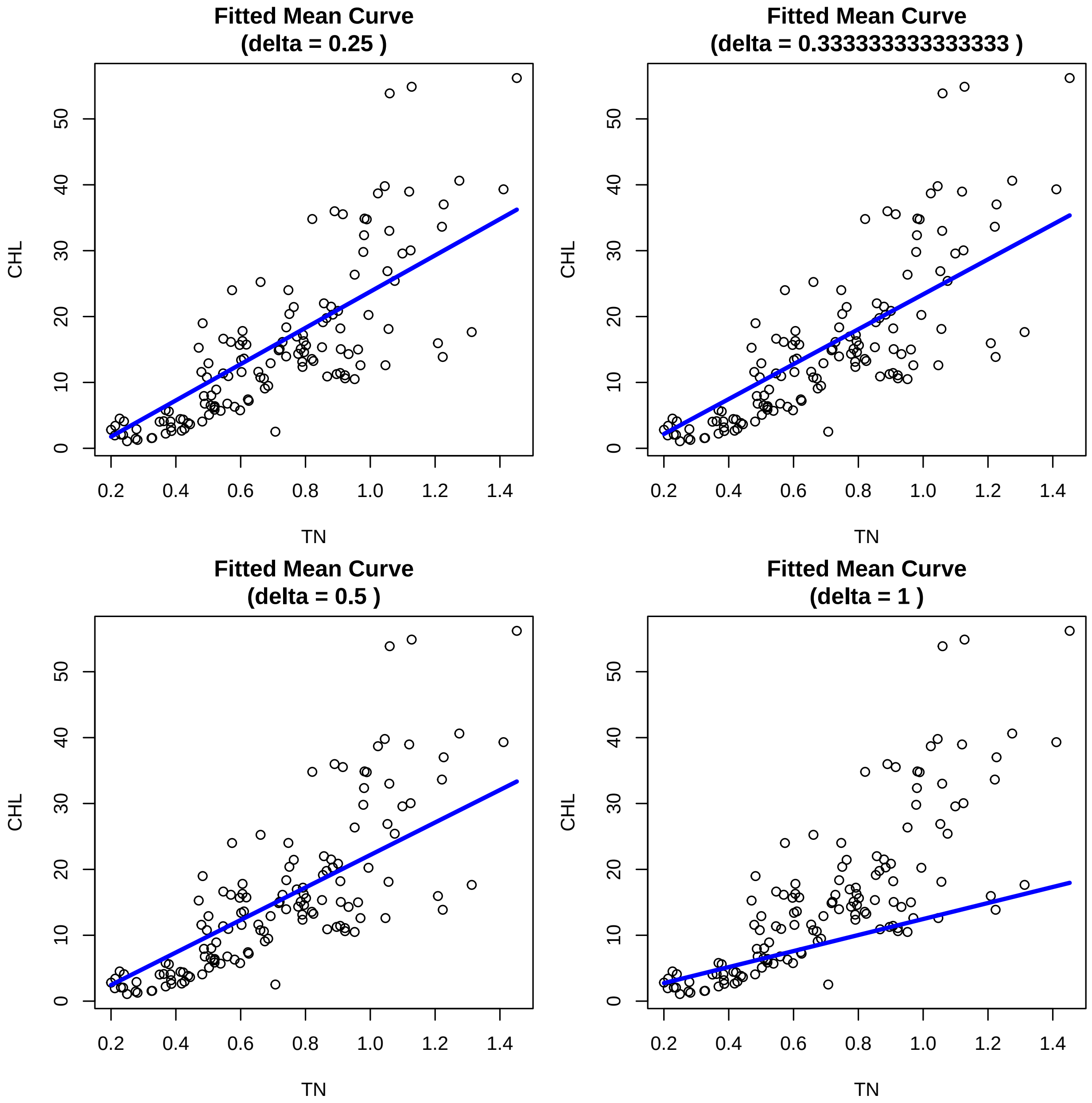
The cube-root transformation provides a reasonable variance-stabilizing effect for the transform-both-sides additive error model, consistent with our earlier findings that cube-root linearization is effective for the GLM as well. Although this approach requires back-transforming predictions to the original CHL scale, it still yields an adequate description of the TN–CHL relationship. However, only means back-transform cleanly; quantities such as variances or percentiles do not, due to issues such as Jensen's inequality, so inference for non-mean quantities is more complicated under a TBS approach.

Note: While only three transformations are given above, a number of other options were considered.

Power of the Mean

While a linear fit appears to do a half decent job at explaining the relationship between CHL and TN, we may think that the relationship is inherently non-linear. Combine this with the prior observation(s) in the GLM section, and we are motivated to believe that variance could be related to the mean, such that a power of the mean model is at least reasonable to consider in the model formulation step.





The power-of-the-mean additive error model was evaluated using $\delta = 0.25$, 1/3, 0.5, and 1. Based on the Box–Cox mean–variance slopes ($\approx$ 1.6–2.1), $\delta = 0.5$ was theoretically preferred, and indeed this value produced the most stable and uniform residuals. Smaller values of δ (0.25 and 1/3) left noticeable heteroscedasticity, while $\delta = 1$ created numerical problems due to extremely small fitted means, resulting in zero or near-zero weights.

Although the $\delta = 0.5$ fit was the best within the additive family, variance stabilization remained incomplete, and the fitted curves tended to under-represent the higher CHL values at moderate TN levels. In contrast, the Gamma GLM produced cleaner residual patterns, a coherent mean–variance structure, and avoided the numerical instability seen in the weighted additive models.

For these reasons, the additive Power of the Mean model seemed unfavorable.

Comparing Different Models

Adequate models considered were:

- Gamma GLM with cube-root power link
- Inverse Gaussian GLM with cube-root power link
- Transform-both-sides additive error model with cube-root transformation
- Power-of-the-mean additive error model with $\delta = 0.5$

Among these, I ultimately selected the Gamma GLM with the cube-root link as the working model. The power-of-the-mean additive model showed only partial variance stabilization and exhibited numerical instability from very small fitted-mean values, making it less reliable. The Inverse Gaussian GLM produced a fitted curve nearly identical to the Gamma model but had slightly less favorable residual diagnostics. The transform-both-sides cube-root model also provided a reasonable fit and captured the general trend well; however, inference under TBS requires back-transformation to the original CHL scale and becomes more complicated for quantities other than the mean (e.g., variances, quantiles, etc.), limiting its practicality for the comparisons that follow.

In contrast, the Gamma GLM with cube-root link aligns closely with the Box–Cox variance diagnostics, yields well-behaved residuals, and provides a coherent and interpretable mean–variance structure without the back-transformation complications of TBS. It therefore offers the most appropriate balance of fit quality, interpretability, and inferential coherence. I carry this model forward for the region-specific comparisons between the Plains and Ozarks.

Extending to regions

Comparing Two Regions

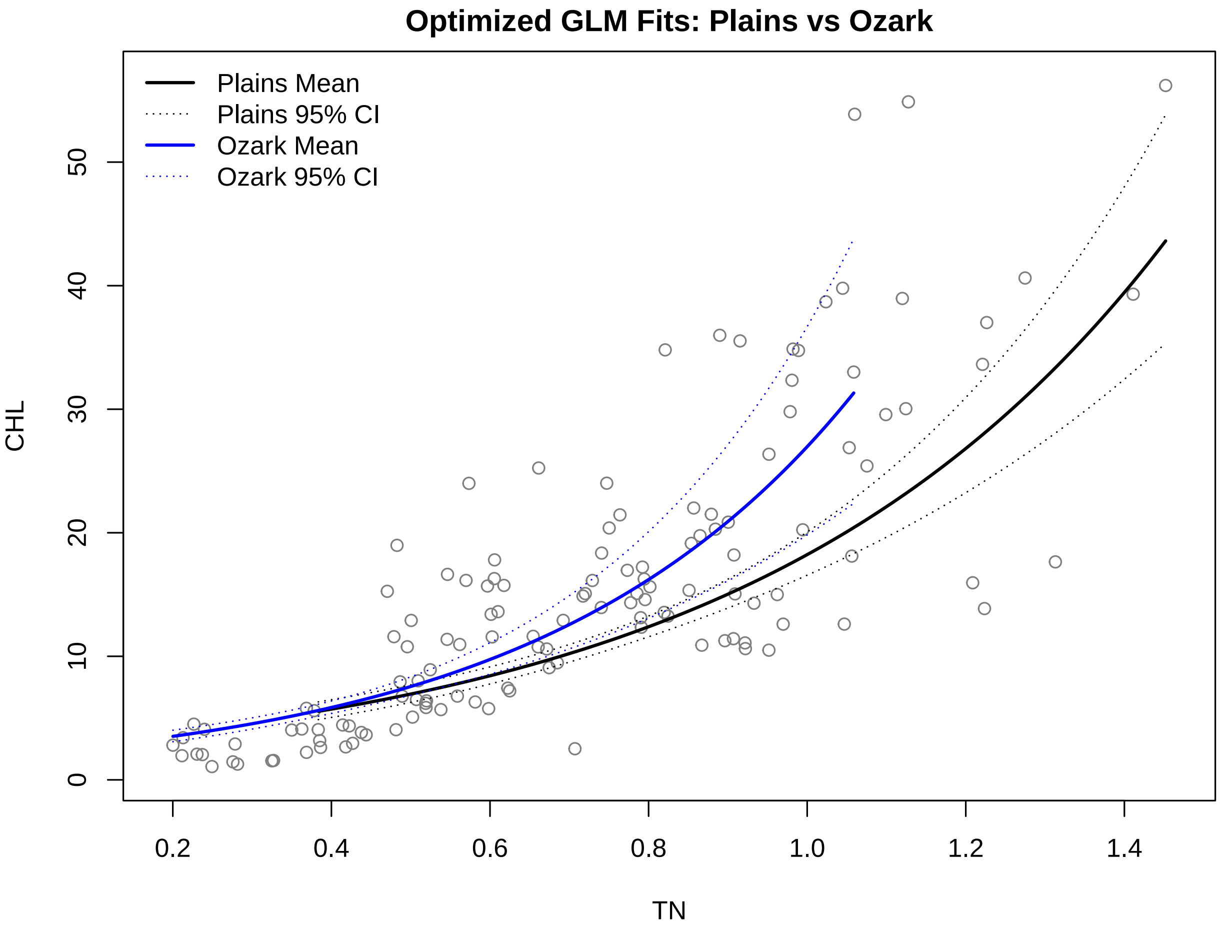
Now, using the preferred model as justified above (the generalized linear model), we'll individually calculate and plot the respective fit and confidence intervals for the two regions.

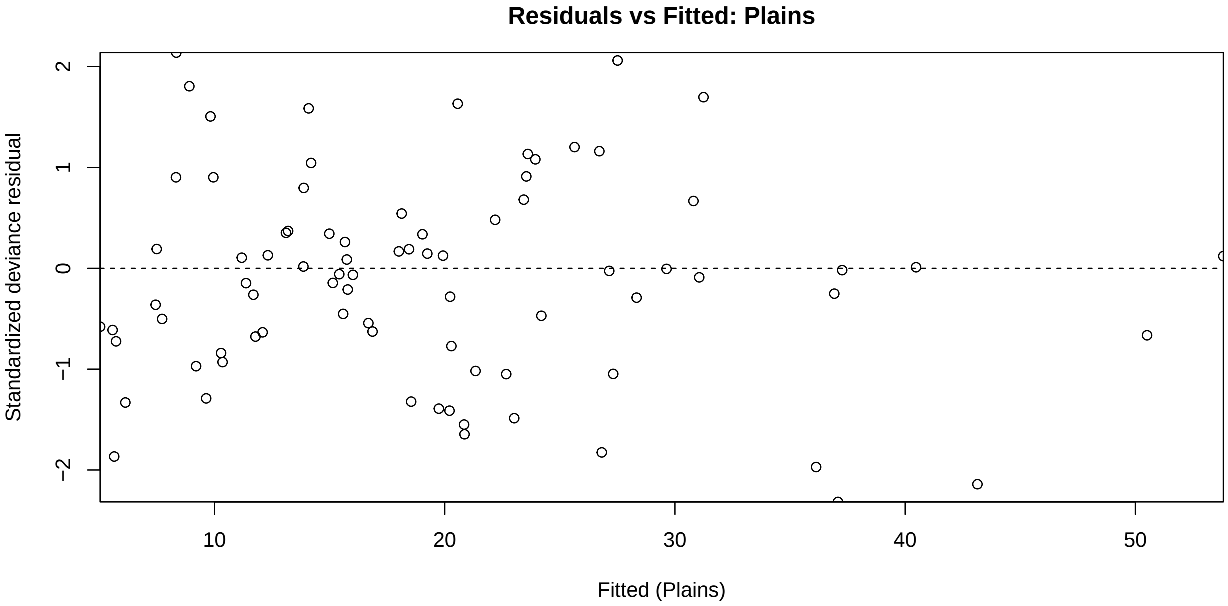
Table 2: Region-specific optimized GLM coefficients with 95% Wald CIs

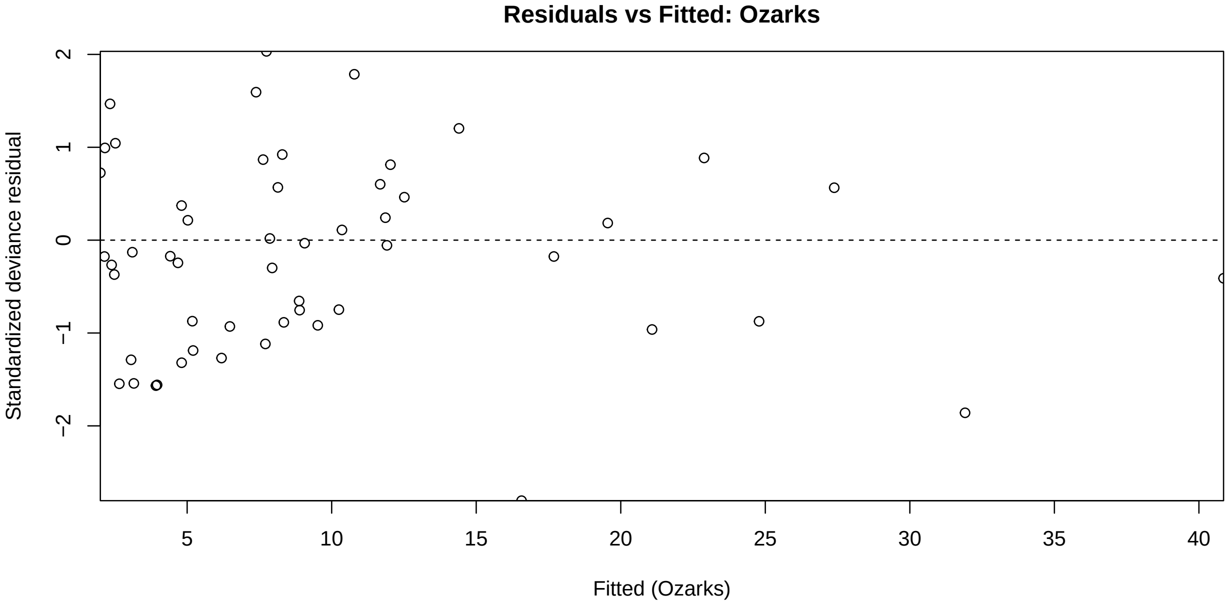
Region	Term	Estimate	SE	LCL	UCL
Plains	Intercept	0.9731	0.1180	0.7417	1.2044
Plains	TN	1.9301	0.1479	1.6402	2.2199
Ozark	Intercept	0.7500	0.1107	0.5331	0.9670
Ozark	TN	2.5447	0.2525	2.0499	3.0395

Table 3: Predicted CHL for TN = 0.4, 0.6, 0.8, 1.0 with 95% CIs by region

Region	TN	Fit	LCL	UCL
Plains	0.4	5.7264	5.0535	6.4889
Plains	0.6	8.4242	7.7594	9.1459
Plains	0.8	12.3928	11.5794	13.2633
Plains	1.0	18.2311	16.5652	20.0644
Ozark	0.4	5.8584	5.3762	6.3838
Ozark	0.6	9.7456	8.5571	11.0991
Ozark	0.8	16.2119	13.0813	20.0916
Ozark	1.0	26.9687	19.8236	36.6892


Optimized GLM Fits: Plains vs Ozark
Plains Mean
Plains 95% CI
Ozark Mean
Ozark 95% CI
CHL
0
10
20
30
40
50
TN
0.2
0.4
0.6
0.8
1.0
1.2
1.4





Region	Scaled_Deviance
Plains	79.59165
Ozarks	54.28234

Using the selected Gamma GLM with cube-root link, I fit separate models for the Plains and Ozarks. Figure X displays the fitted mean curves with pointwise 95% confidence bands, and Tables 2–3 summarize the region-specific coefficients and predicted CHL values at TN = 0.4, 0.6, 0.8, and 1.0.

At lower TN values (approximately 0.4–0.6), the fitted curves and their 95% bands overlap substantially, and the predicted means are very similar across regions. However, as TN increases toward the upper end of the Ozarks range (TN ≈ 0.8–1.0), the Ozarks curve becomes noticeably steeper and the bands begin to separate. At TN = 1.0, the pointwise 95% intervals only barely overlap, suggesting higher CHL in the Ozarks at that TN value, but with modest statistical separation given the limited covariate overlap and sample size.

The coefficient estimates in Table 2 show a larger slope for the Ozarks and a slightly lower intercept relative to the Plains. The corresponding 95% Wald intervals overlap for both parameters, so parameter-wise evidence for differences is suggestive rather than conclusive. Nonetheless, the joint pattern—steeper Ozarks slope, divergence of fitted curves at moderate TN, and consistently higher Ozarks predictions at TN values near 1.0—indicates a somewhat stronger TN–CHL response in the Ozarks.

Basic model assessment supports use of the Gamma GLM for both regions. Residual-versus-fitted plots show no strong systematic patterns for either region, and the standardized deviance residuals remain centered around zero with reasonably constant spread. The scaled deviances (Plains: 79.6; Ozarks: 54.3) are similar in magnitude relative to sample size, indicating that the Gamma mean–variance structure is adequate for each regional model. These diagnostics suggest that the same model form is appropriate in both groups, which is a prerequisite for comparing the fitted regression functions.

Because a full-versus-reduced likelihood ratio test is reserved for Part III of the exam and is not required here, conclusions in Part I rely on comparisons of fitted curves, confidence bands, and parameter intervals rather than a formal LRT. Based on these comparisons, the two regions appear to share the same general functional relationship between TN and CHL, but with evidence of a stronger response in the Ozarks at higher TN levels. This conclusion is tempered by the limited TN range observed for the Ozarks and the substantial band overlap at lower TN values, but overall the fitted curves and coefficient patterns suggest meaningful regional differences in the magnitude—though not the form—of the TN–CHL relationship.

Probability Assessments

For a fixed TN value x_i, let Y_i denote the Plains CHL response and Z_i the Ozarks CHL response. From the region–specific Gamma GLMs with cube–root link, the conditional means are

$$\mu_P(x_i) = E(Y_i \mid x_i), \qquad \mu_O(x_i) = E(Z_i \mid x_i),$$

and the Gamma variance functions are:

$$\mathrm{Var}(Y_i \mid x_i) = \phi_P \, \mu_P(x_i)^2, \qquad \mathrm{Var}(Z_i \mid x_i) = \phi_O \, \mu_O(x_i)^2,$$

where ϕ_P and ϕ_O are the dispersion parameters for the Plains and Ozarks models.

The quantity of interest is

$$\Pr(Y_i > Z_i \mid x_i)$$

Using a plug–in Normal approximation to the Gamma distribution, we write

$$Y_i \mid x_i \approx N\big(\mu_P(x_i), \phi_P \mu_P(x_i)^2\big), \qquad Z_i \mid x_i \approx N\big(\mu_O(x_i), \phi_O \mu_O(x_i)^2\big),$$

and assume Y_i and Z_i are conditionally independent given x_i.

Define the difference:

$$D_i = Y_i - Z_i$$

Then

$$E(D_i \mid x_i) = \mu_P(x_i) - \mu_O(x_i)$$

And, noting conditional independence assumption:

$$\mathrm{Var}(D_i \mid x_i) = \mathrm{Var}(Y_i \mid x_i) + \mathrm{Var}(Z_i \mid x_i) = \phi_P \mu_P(x_i)^2 + \phi_O \mu_O(x_i)^2$$

So:

$$D_i \mid x_i \approx N\Big(\mu_P(x_i) - \mu_O(x_i), \phi_P \mu_P(x_i)^2 + \phi_O \mu_O(x_i)^2\Big)$$

Therefore,

$$\Pr(Y_i > Z_i \mid x_i) = \Pr(D_i > 0 \mid x_i) \approx \Phi\left(\frac{\mu_P(x_i) - \mu_O(x_i)}{\sqrt{\phi_P \mu_P(x_i)^2 + \phi_O \mu_O(x_i)^2}}\right)$$

where Φ is the standard Normal CDF.

Under the cube–root power link used in `basic.glm` (with `pwr = 1/3`), the linear predictor is

$$\eta(x) = \mu(x)^{1/3}, \qquad \eta(x) = x^\top \hat{\beta},$$

so the fitted mean at covariate value x is

$$\hat{\mu}(x) = \left(x^\top \hat{\beta}\right)^{1/\frac{1}{3}} = \left(x^\top \hat{\beta}\right)^3$$

For each region and each TN value x_i,

$$\hat{\mu}_P(x_i) = \left((1, x_i)^\top \hat{\beta}_P\right)^3, \qquad \hat{\mu}_O(x_i) = \left((1, x_i)^\top \hat{\beta}_O\right)^3,$$

with corresponding dispersion estimates $\hat{\phi}_P$ and $\hat{\phi}_O$.

The plug–in estimator of the desired probability is then:

$$\widehat{\Pr}(Y_i > Z_i \mid x_i) = \Phi\left(\frac{\hat{\mu}_P(x_i) - \hat{\mu}_O(x_i)}{\sqrt{\hat{\phi}_P \hat{\mu}_P(x_i)^2 + \hat{\phi}_O \hat{\mu}_O(x_i)^2}}\right)$$

Implementing this:

```r
# Models fitted already
# mod_gamma_13_plains
# mod_gamma_13_ozark

# Helper: get fitted mean mu(x) from a basic.glm
predict_mu_power <- function(mod, x, pwr = 1/3) {
  # regression coefficients (column of estb)
  beta_hat <- mod$estb[, 1]
  # design matrix: intercept + TN
  X <- cbind(1, x)
  eta <- as.vector(X %*% beta_hat)
  # because eta = mu^pwr
  mu <- eta^(1 / pwr)
  mu
}
```

```r
# Compute plug-in Pr(Y > Z | x) via Normal approximation
prob_YgtZ_normal <- function(x, mod_P, mod_O, pwr = 1/3) {
  # Plains mean and dispersion
  mu_P <- predict_mu_power(mod_P, x, pwr = pwr)
  phi_P <- mod_P$ests$phi
  # Ozarks mean and dispersion
  mu_O <- predict_mu_power(mod_O, x, pwr = pwr)
  phi_O <- mod_O$ests$phi
  # Mean and variance of the difference D = Y - Z
  mean_D <- mu_P - mu_O
  var_D <- phi_P * mu_P^2 + phi_O * mu_O^2
  sd_D <- sqrt(var_D)
  # Probability that D > 0
  pnorm(mean_D / sd_D)
}

# Point estimate at x = 0.70
prob_0.70 <- prob_YgtZ_normal(0.70,
                              mod_gamma_13_plains,
                              mod_gamma_13_ozark)
cat("Estimated probability is:", prob_0.70, "\n")
```

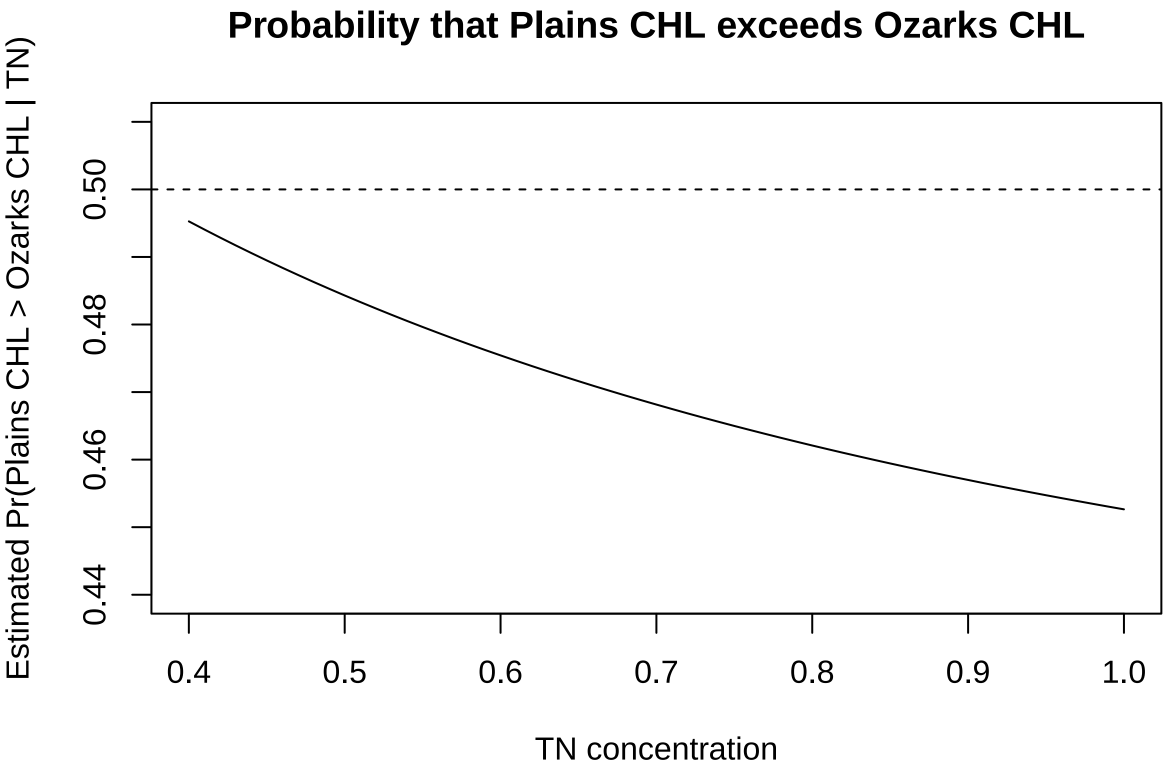
```
## Estimated probability is: 0.4681497
```

Our estimate is then: $\Pr(Y_i > Z_i \mid x_i = 0.70) = 0.46815$.

Now, the sequence of values $x_i \in \{0.4, 0.41, 0.42, \cdots, 1.0\}$ is given by:

```r
# Grid x_i in {0.40, 0.41, ..., 1.00}
x_grid    <- seq(0.40, 1.00, by = 0.01)
prob_grid <- sapply(x_grid, prob_YgtZ_normal,
                    mod_P = mod_gamma_13_plains,
                    mod_O = mod_gamma_13_ozark)

# Plot
# TN on x-axis
# estimated probabilities on y-axis
plot(x_grid, prob_grid, type = "l",
     ylim = c(0.44, 0.51),
     xlab = "TN concentration",
     ylab = "Estimated Pr(Plains CHL > Ozarks CHL | TN)",
     main = "Probability that Plains CHL exceeds Ozarks CHL")
# reference line at 0.5
abline(h = 0.5, lty = 2)
```



So, as TN concentration increases, the probability that Plains CHL exceeds Ozarks CHL decreases (from roughly 0.5 to 0.45).

Relation Between CHL and TN within Regions

Across both regions, chlorophyll exhibits the same basic pattern with respect to total nitrogen: CHL increases as TN increases, with variance rising alongside the mean. The Gamma GLM with a cube-root link provides an adequate and coherent description of this mean–variance relationship in both regions, with well-behaved residuals and no indication that different model forms are required. Thus, the two regions appear to follow the same underlying functional relationship between TN and CHL.

When fitted separately, however, the region-specific models show meaningful differences in magnitude of the TN response. At lower TN values (approximately 0.4–0.6), the fitted curves and their 95% confidence bands overlap substantially, and the predicted means are nearly identical across regions. As TN increases toward the upper end of the Ozarks data range (0.8–1.0), the Ozarks curve becomes noticeably steeper. Predicted CHL diverges upward in the Ozarks, and the confidence bands begin to separate, with only slight overlap at TN = 1.0.

Coefficient estimates reinforce this pattern: both regions have positive slopes, but the Ozarks slope is larger, indicating a stronger increase in CHL per unit increase in TN. Although the 95% Wald intervals overlap and do not independently confirm a difference, the joint behavior of the fitted curves, slope estimates, and widening separation at higher TN values together suggest that the Ozarks exhibit a stronger TN–CHL response.

Finally, the probability assessment provides an interpretable summary: for moderate TN values (0.4–0.7), the Plains and Ozarks have nearly equal probability of producing higher CHL. But as TN approaches 1.0, the probability that Plains CHL exceeds Ozarks CHL drops below 0.5, consistent with a steeper Ozarks response.

In summary:

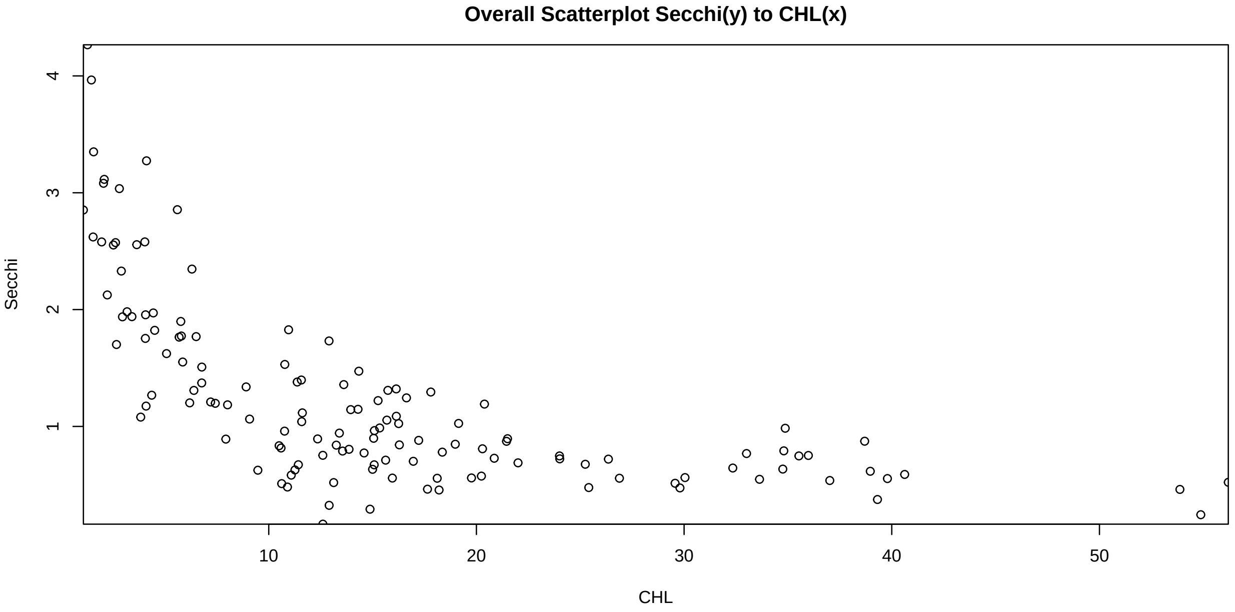
- The form of the TN–CHL relationship is the same in both regions.
- The magnitude of the response differs: CHL rises more sharply with TN in the Ozarks, particularly at higher TN values.
- The evidence is suggestive but not definitive due to overlapping bands at low TN and limited high-TN data in the Ozarks.

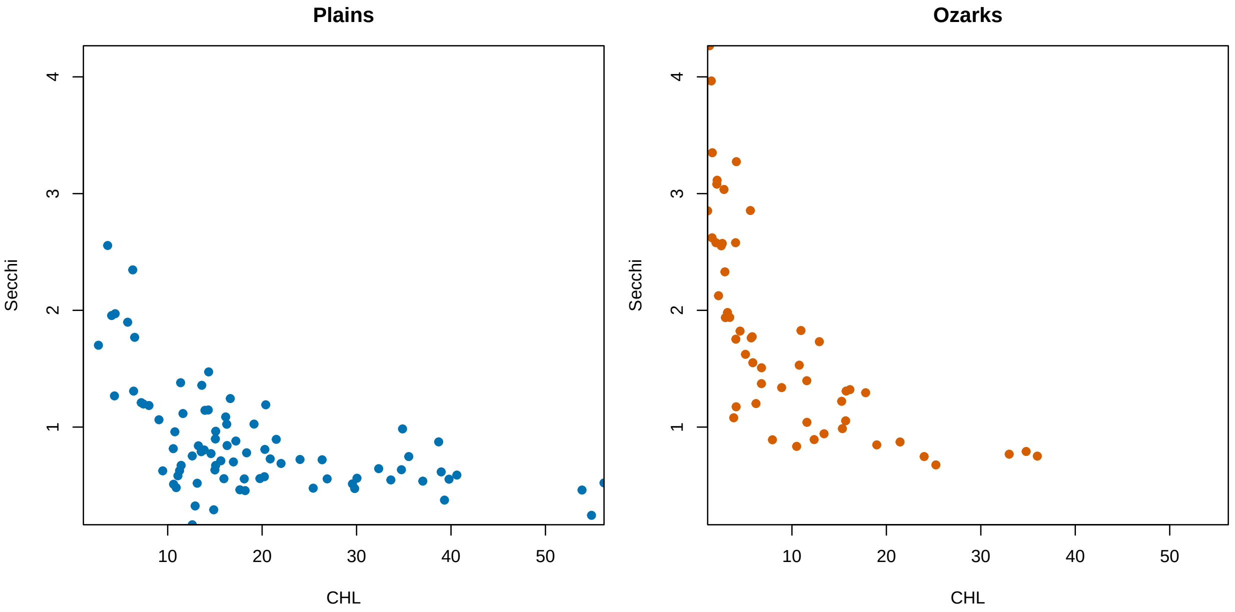
Overall, the two regions share a common biological relationship between TN and CHL, but the Ozarks appear to show a somewhat stronger response to increasing TN, especially towards the upper end of the observed TN range.

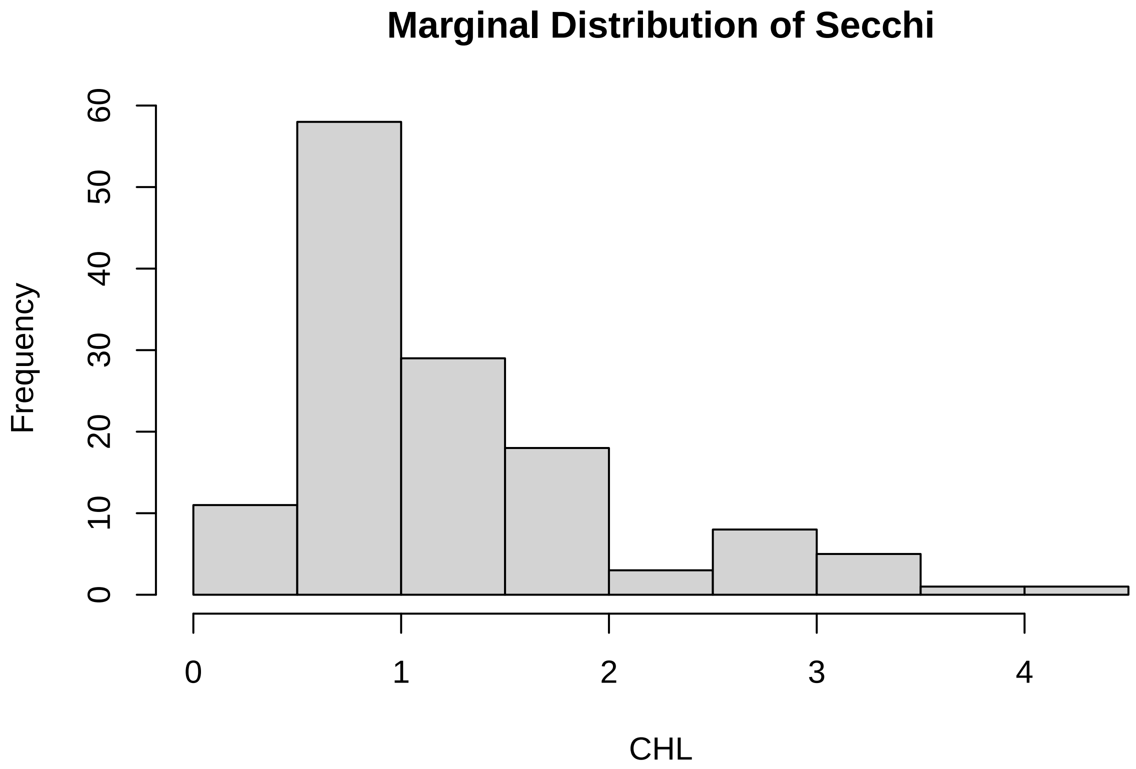
Q2: SECCHI & CHL (Plains vs. Ozarks)

Overall Distribution & Approach

We'll again start as in Part I, looking at the scatterplot of Secchi and CHL, in addition to the marginal distribution of Secchi.







Henceforth, I'll remove the references to "As in Part I", for the sake of brevity. That being said:

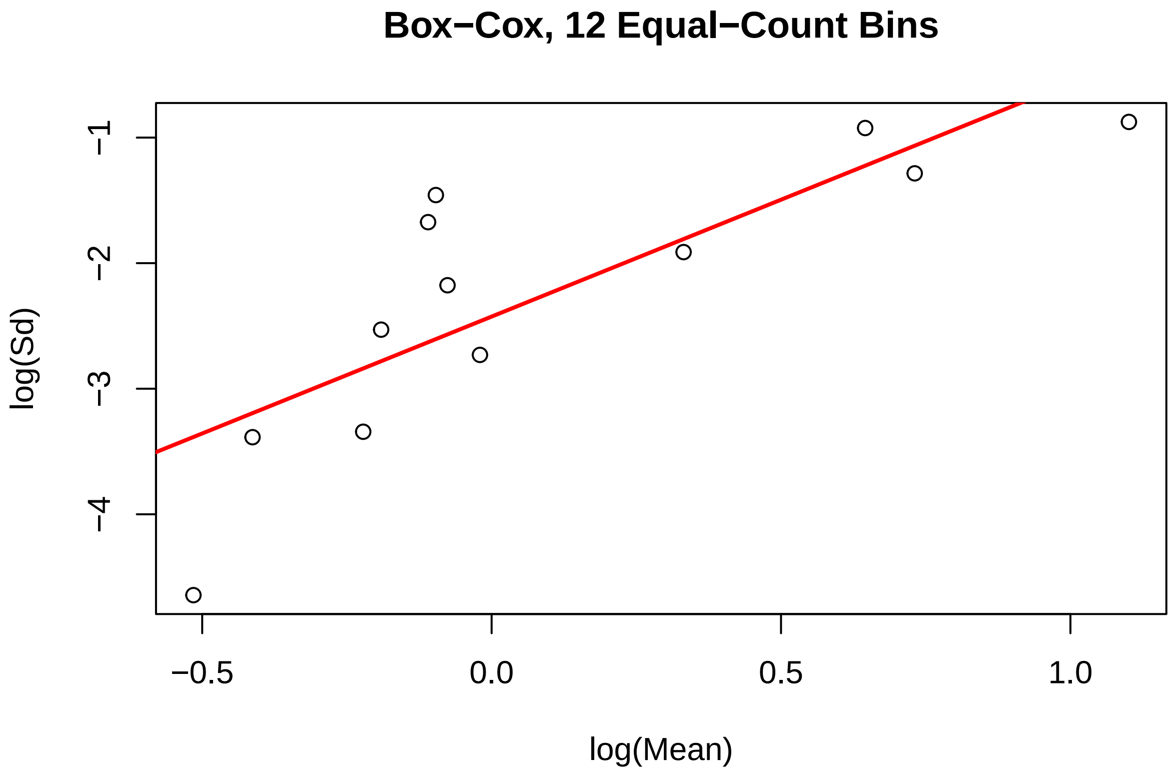
The distribution of Secchi Depth is right-skewed, consistent with ecological theory and historical empirical results. There is a clear negative, potentially nonlinear relationship between CHL and Secchi, with possible decreasing variance for larger CHL values. Scatterplots stratified by region show the same general patterns (negative trend, possible nonlinearity, decreasing variance), though the ranges for the Plains and Ozarks differ in distribution while still overlapping in terms of CHL range.

After identifying an appropriate overall model, we fit the same model structure separately to the Plains and Ozarks. Differences between the region-specific fits—such as changes in intercept, slope, curvature, or dispersion—then reflect true regional differences in the strength or level of the CHL–Secchi relationship rather than differences introduced by choosing different model families or transformations.

This approach ensures that regional comparisons reflect actual ecological differences, avoids artifacts from inconsistent modeling, and allows direct comparison of parameters, confidence bands, and other quantities of interest to assess the possibility of regional differences.

Generalized Linear Model

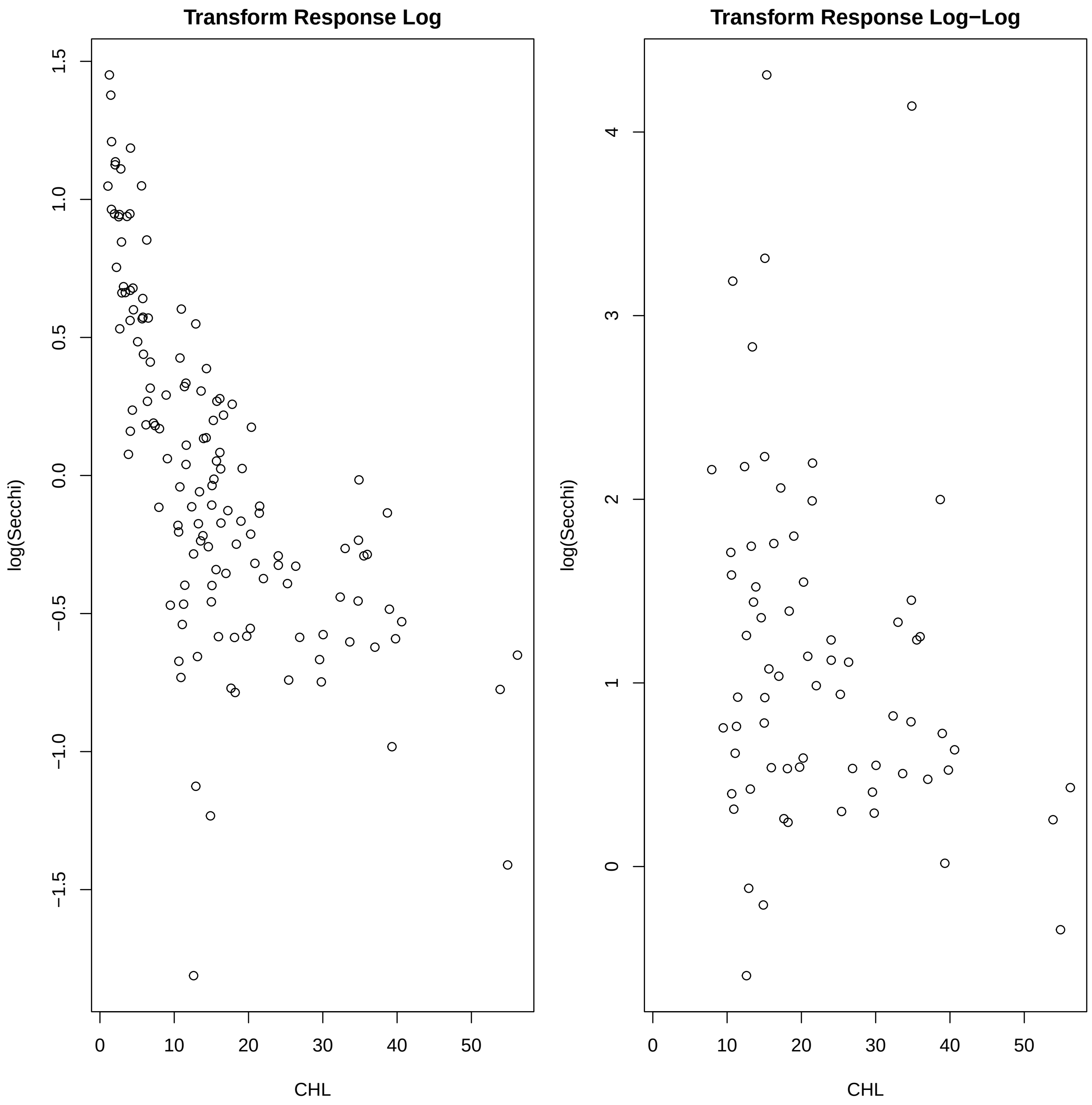
We'll start by assessing potential generalized linear models that adequately describe the relationship between Secchi and CHL.



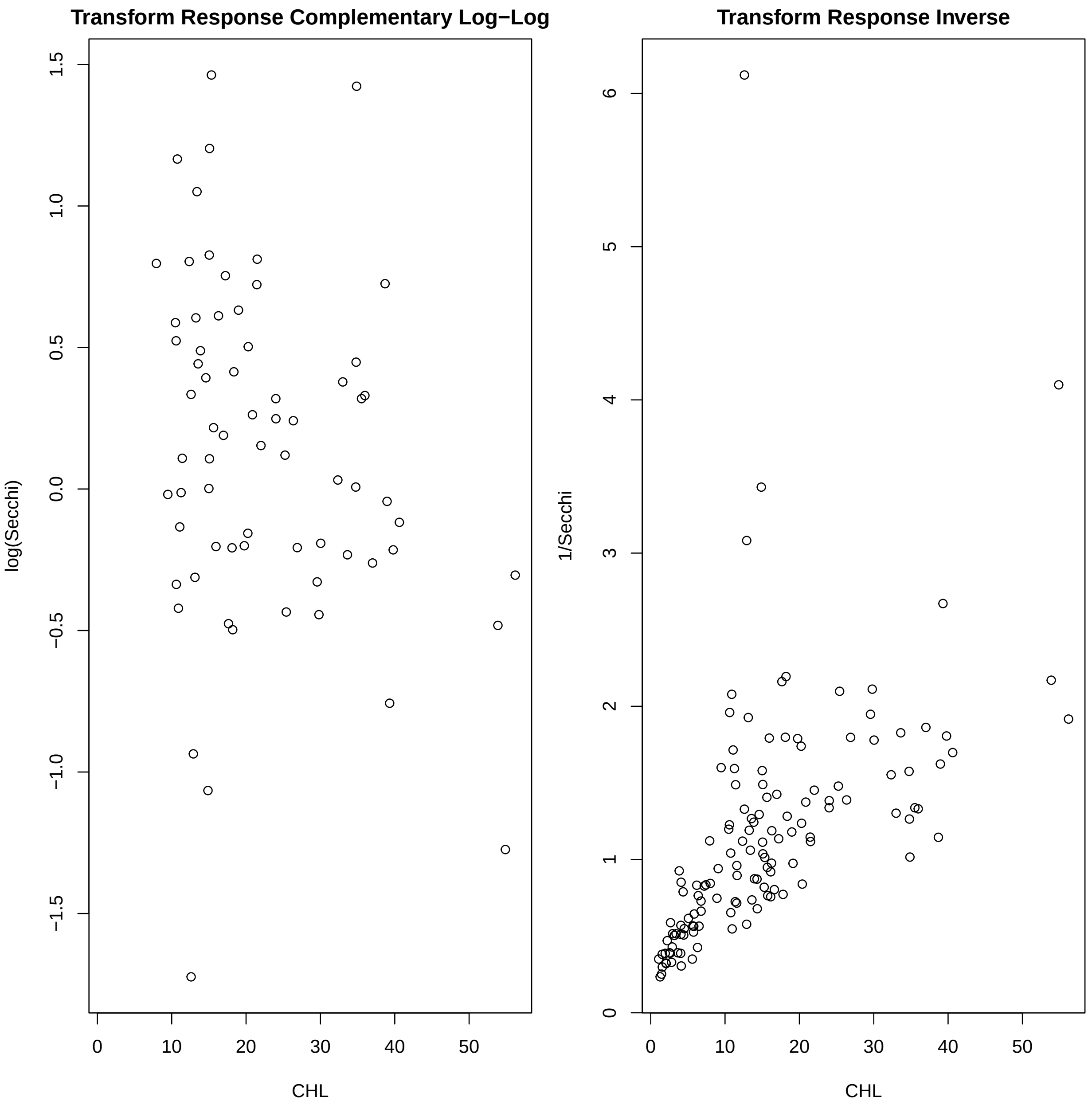
nbins	Equal.Count	Equal.Spaced
6	2.13	2.13
10	1.86	2.71
12	1.86	2.12
14	1.69	2.29
16	1.75	2.69
18	1.73	2.37
22	1.77	2.42

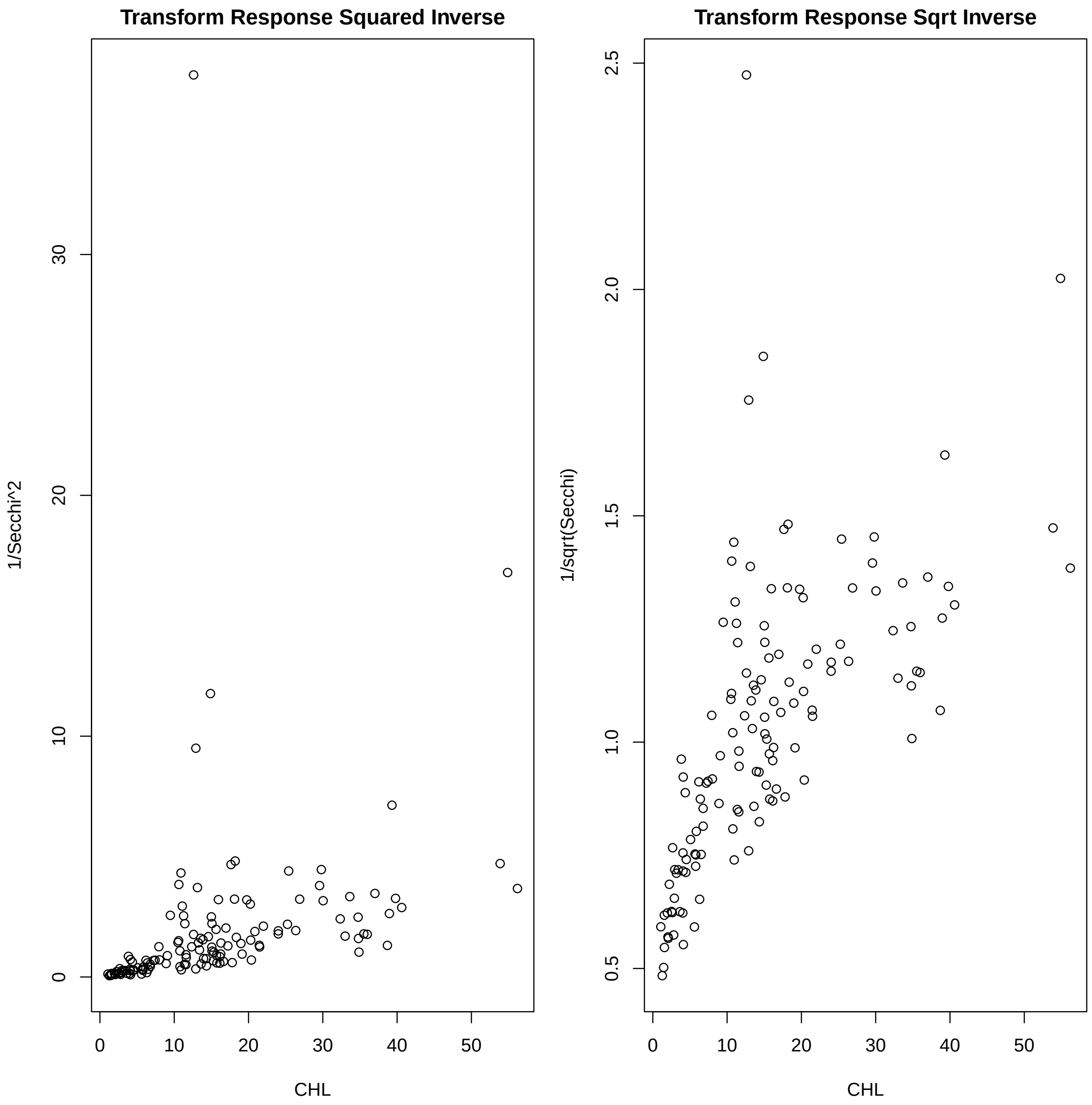
Based on the initial examination, we begin by considering suitable generalized linear models. Using the provided `boxcoxfctns` functions, I computed Box–Cox mean–variance slopes to identify an appropriate random component. The slopes were consistently between 1.6 and 2.1 across binning schemes, equally-spaced or equal-counts, indicating a variance pattern approximately proportional to μ^2 to μ^3, suggesting that either a Gamma or Inverse Gaussian random component is appropriate for the overall model.

With suitable random component(s) identified, the next step is to identify a suitable systematic link function for modeling the mean relationship between CHL and Secchi.



Transform Response Log
log(Secchi)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
10
20
30
40
50
CHL
Transform Response Log−Log
log(Secchi)
4
3
2
1
0
0
10
20
30
40
50
CHL


Transform Response Complementary Log−Log
log(Secchi)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
10
20
30
40
50
CHL
Transform Response Inverse
1/Secchi
6
5
4
3
2
1
0
0
10
20
30
40
50
CHL



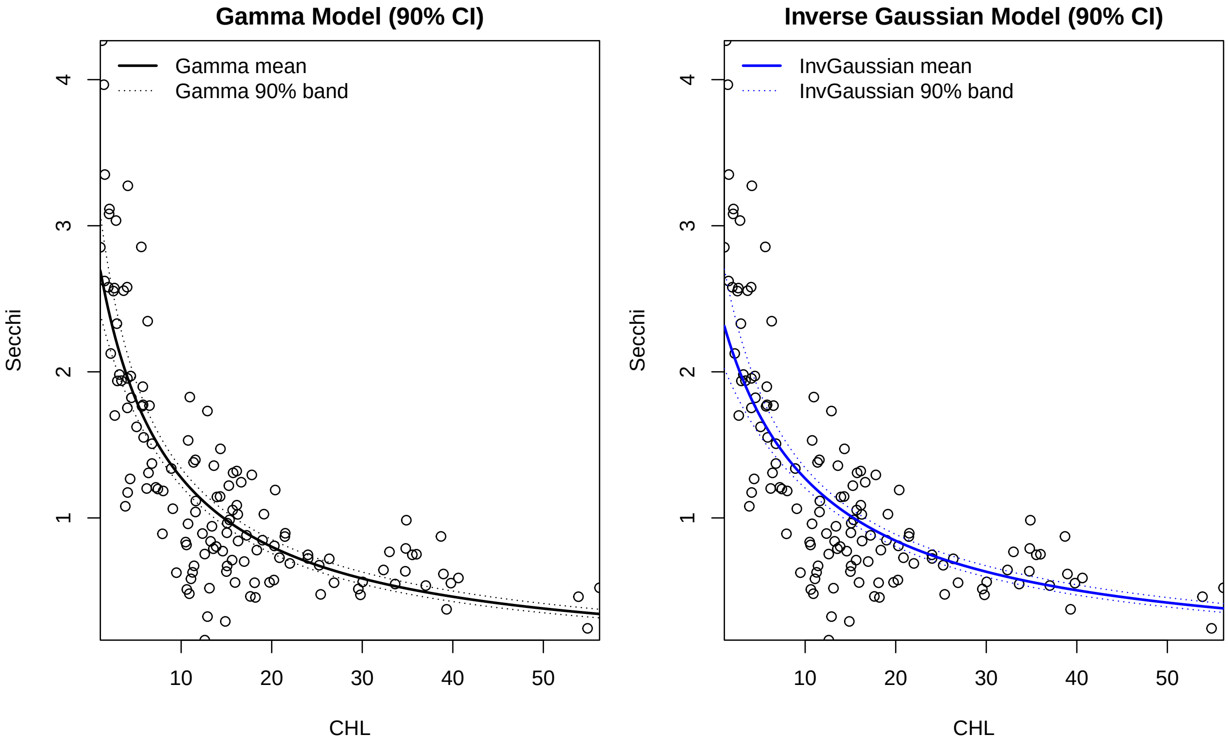
When identifying a suitable link function, the objective is to find a transformation of the mean that approximately linearizes the relationship between Secchi and CHL. More importantly however, we need a link function that is monotonically decreasing, since such an interpretation aligns both with theory and the expected relationship shown in the data.

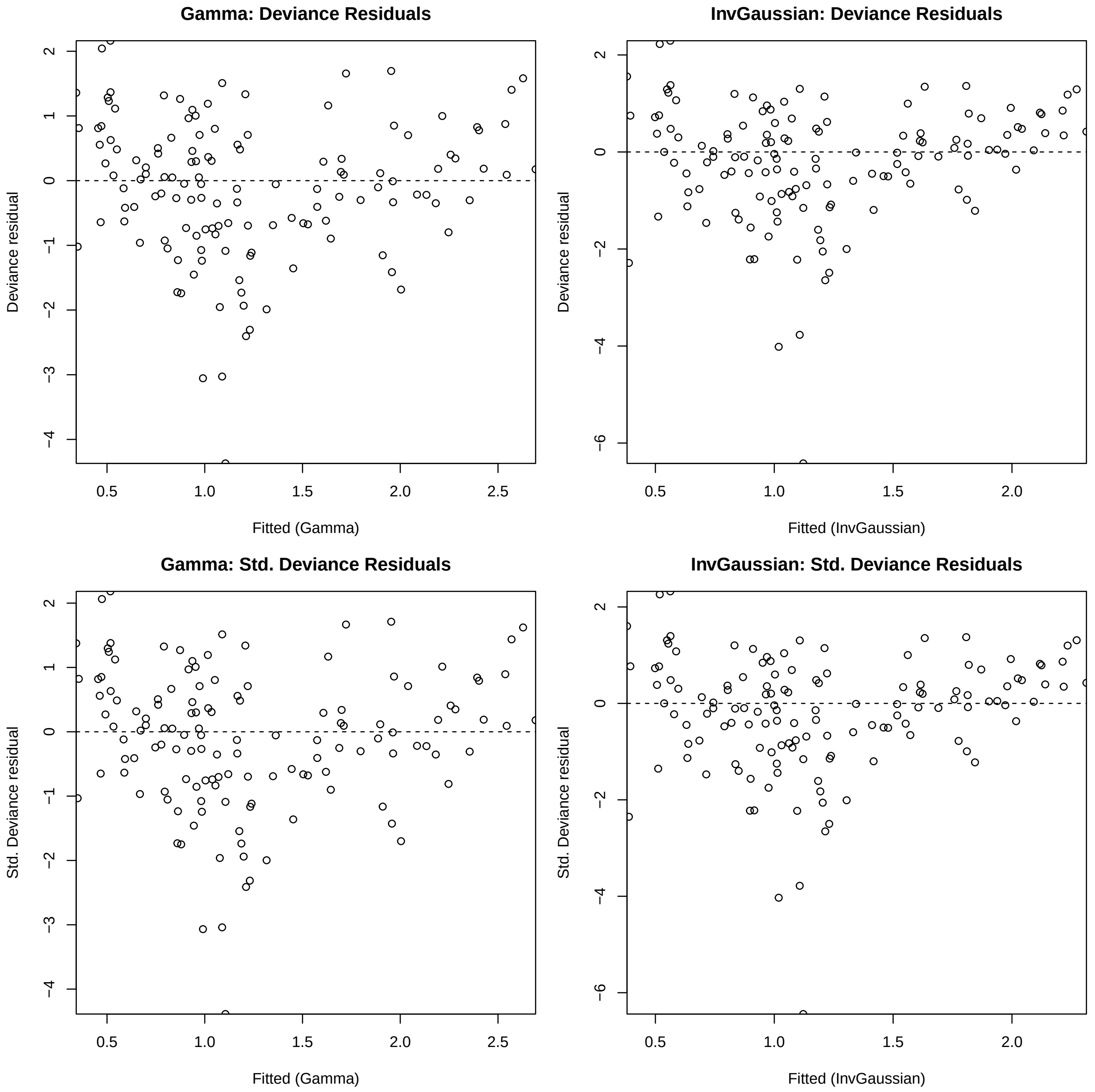
The exploratory plots indicate that an inverse transformation provides a reasonably linear trend, with the square-inverse transformation also performing adequately. Importantly, these transformations are used only to guide link selection; they do not imply transforming the Secchi values themselves for the GLM.

Taken together with the random component selection done previously, in total: We are motivated to use Gamma and Inverse Gaussian random components ($\theta = 2, 3$), paired with inverse and square inverse links. We then fit these models and compare.

We attempted to fit an inverse Gaussian GLM with the canonical inverse-squared link. Using reasonable starting values from a corresponding `glm()` fit and enforcing positive initial linear predictors, the `basic.glm` routine still failed to converge due to numerical instabilities (iterations producing invalid η-values). We therefore selected a Gamma and Inverse Gaussiam GLMs with inverse link, which converged and provided an adequate fit by residual diagnostics.

We then look at the fitted curve on the scatterplot in addition to the deviance residuals to compare the two different models.



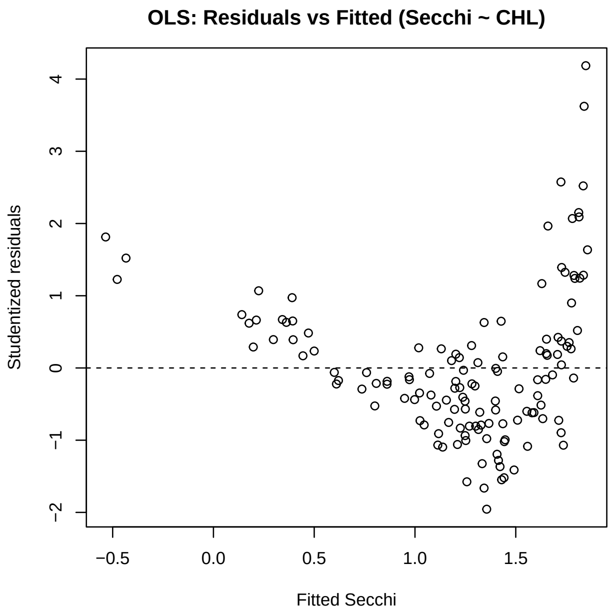


It is difficult to distinguish between the Gamma and Inverse Gaussian models using the fitted–versus–observed scatterplots alone, as both produce very similar mean curves. One might argue that the Inverse Gaussian captures the increasing variability at lower CHL values (CHL < 10) slightly more closely, though at the expense of having a worse fit for these values (the mean curve misses nearly all observations in this range. Combine this with the residual plots, which indicate the Inverse Gaussian deviance residuals exhibit possibly greater heteroscedasticity in addition to lacking symmetry, and the choice of a Gamma generalized linear model with an inverse link seems most appropriate.

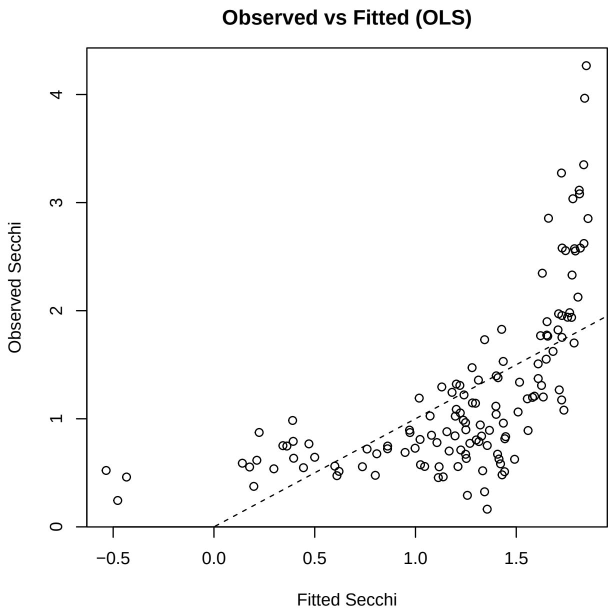
With this preferred generalized linear model identified, we next examine a range of additive error models for completeness. By evaluating these alternatives alongside the GLM, we select a single most appropriate model to use for the subsequent region-specific comparisons.

Additive Error Models

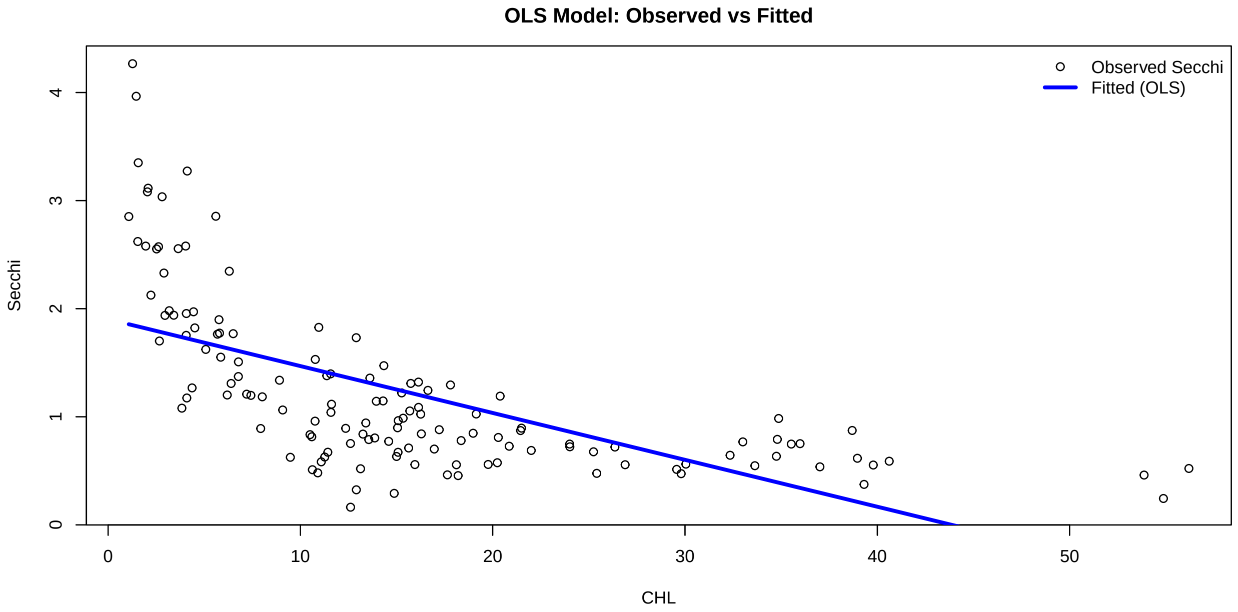
Transform Both Sides



OLS: Residuals vs Fitted (Secchi ~ CHL)

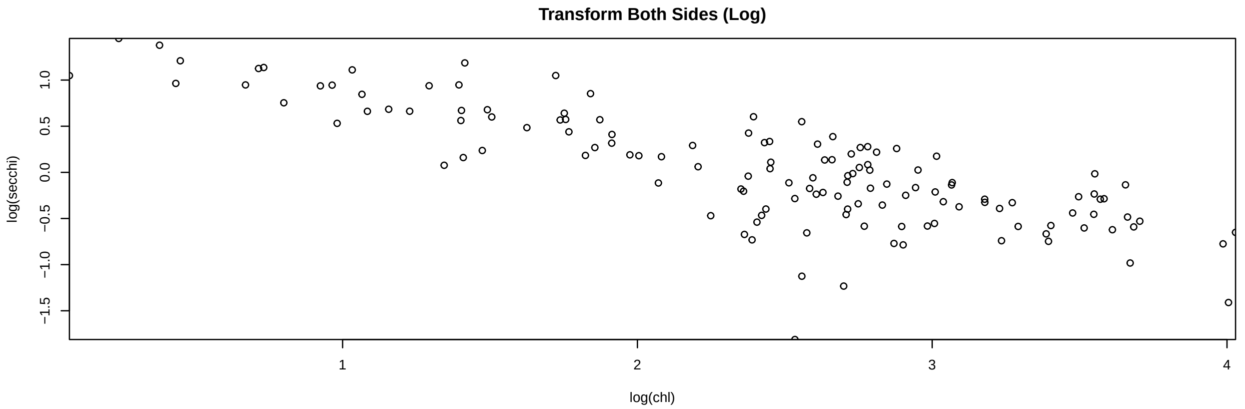


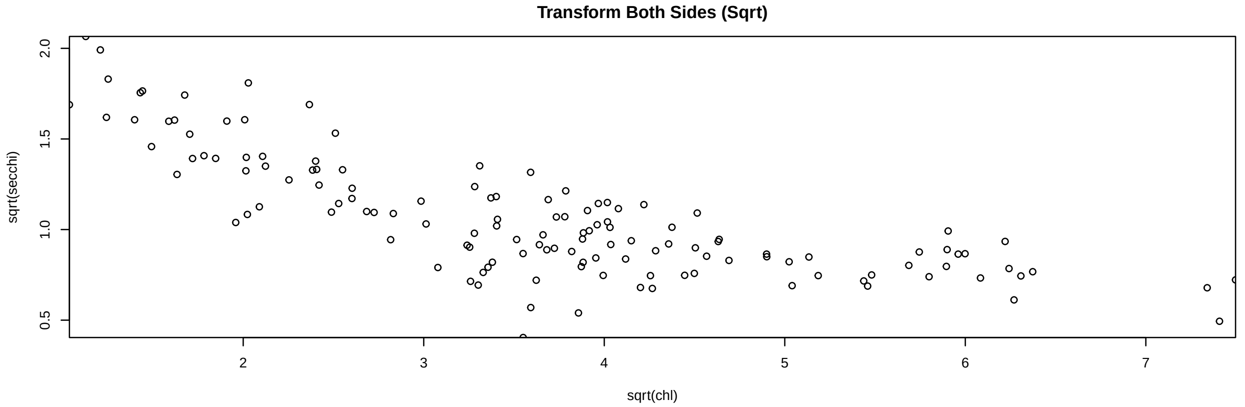
Observed vs Fitted (OLS)

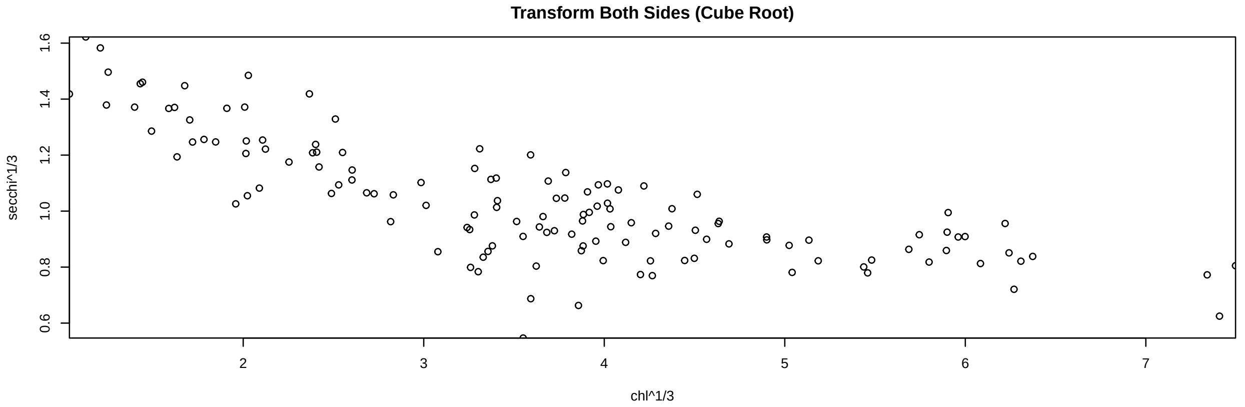


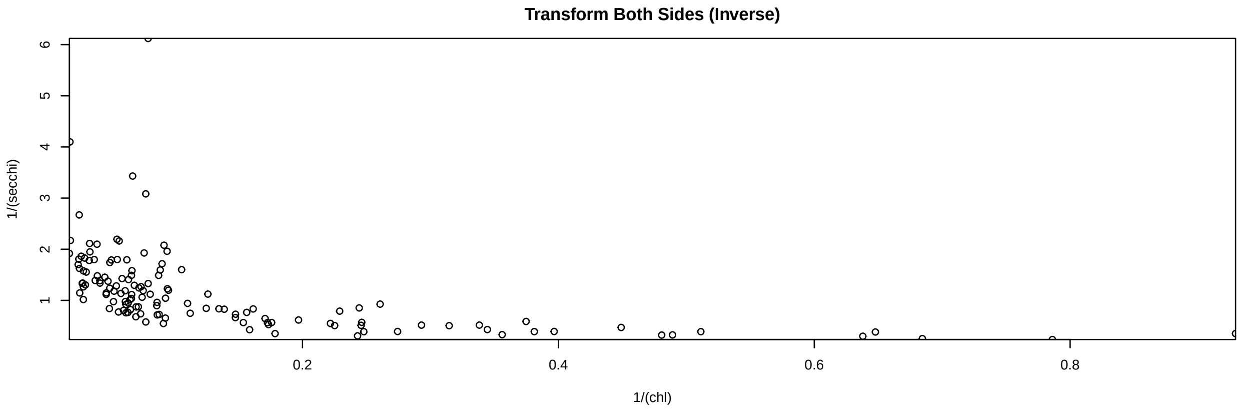
OLS Model: Observed vs Fitted

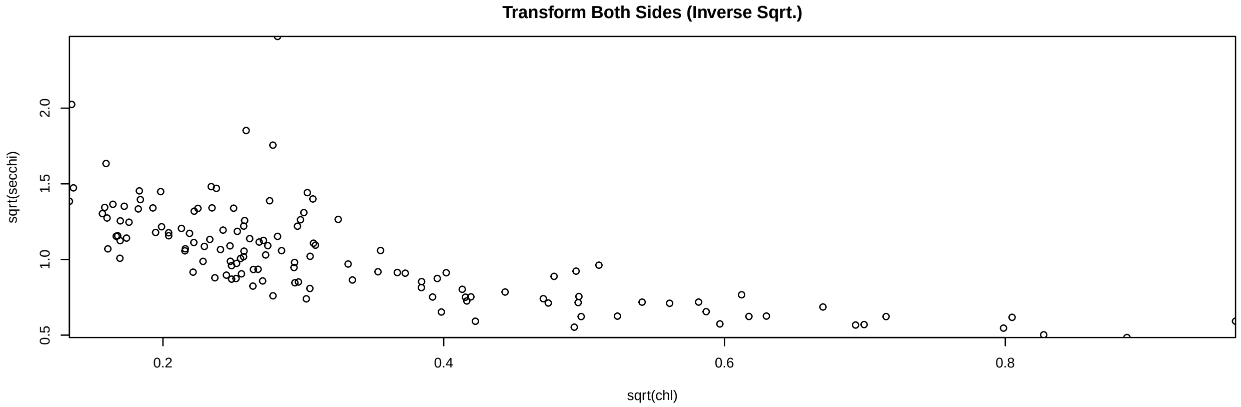
Though we eventually want to adopt a more fitting curve, a key consideration of the above linear fit is the clear heteroscedasticity in the model. This motivates a potential transformat both sides model.

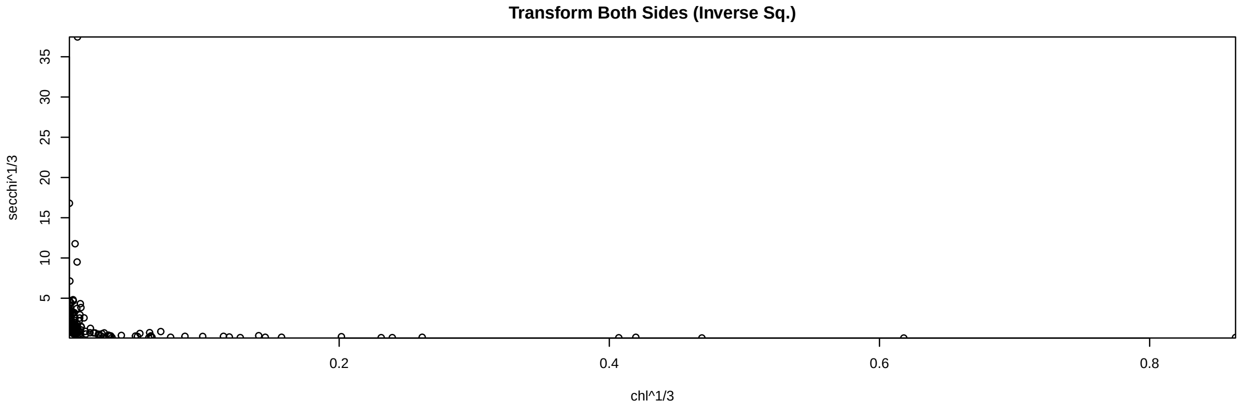

Transform Both Sides (Log)
log(secchi)
1.0
0.5
0.0
−0.5
−1.0
−1.5
1
2
3
4
log(chl)

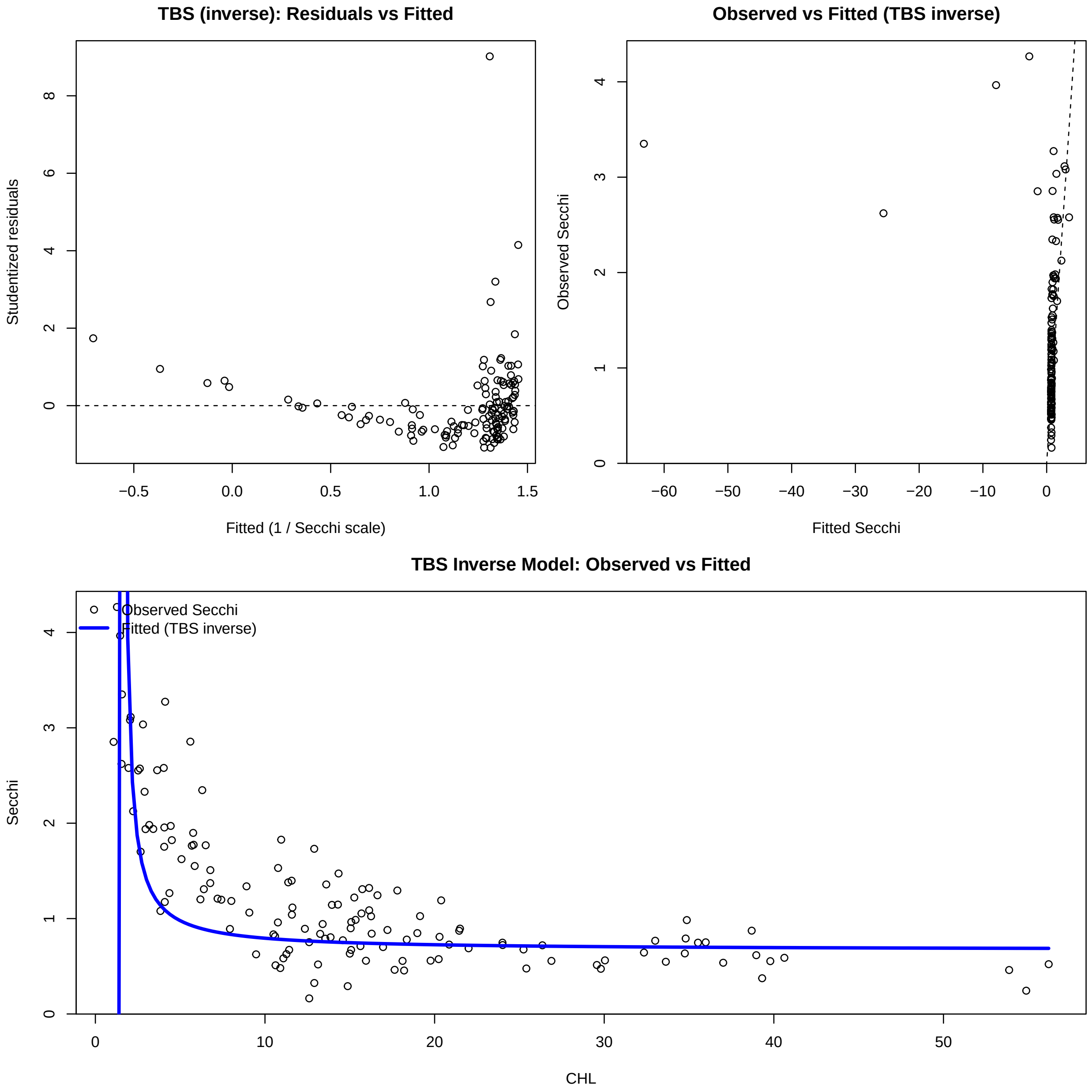

Transform Both Sides (Sqrt)
sqrt(secchi)
2.0
1.5
1.0
0.5
2
3
4
5
6
7
sqrt(chl)

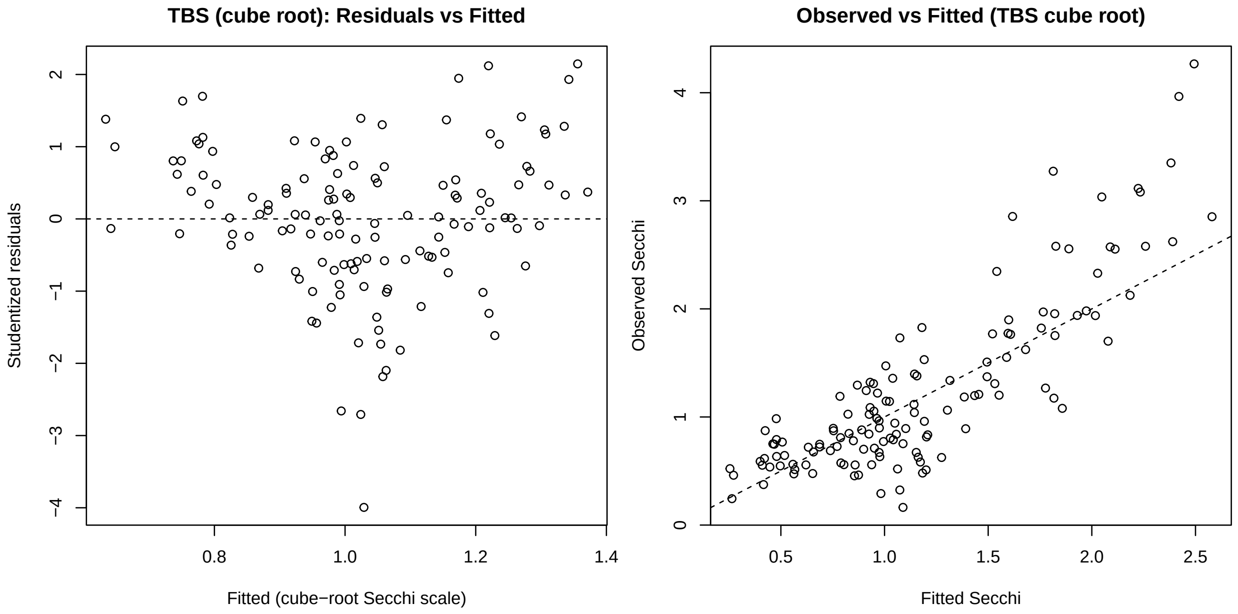

Transform Both Sides (Cube Root)
secchi^1/3
1.6
1.4
1.2
1.0
0.8
0.6
2
3
4
5
6
7
chl^1/3

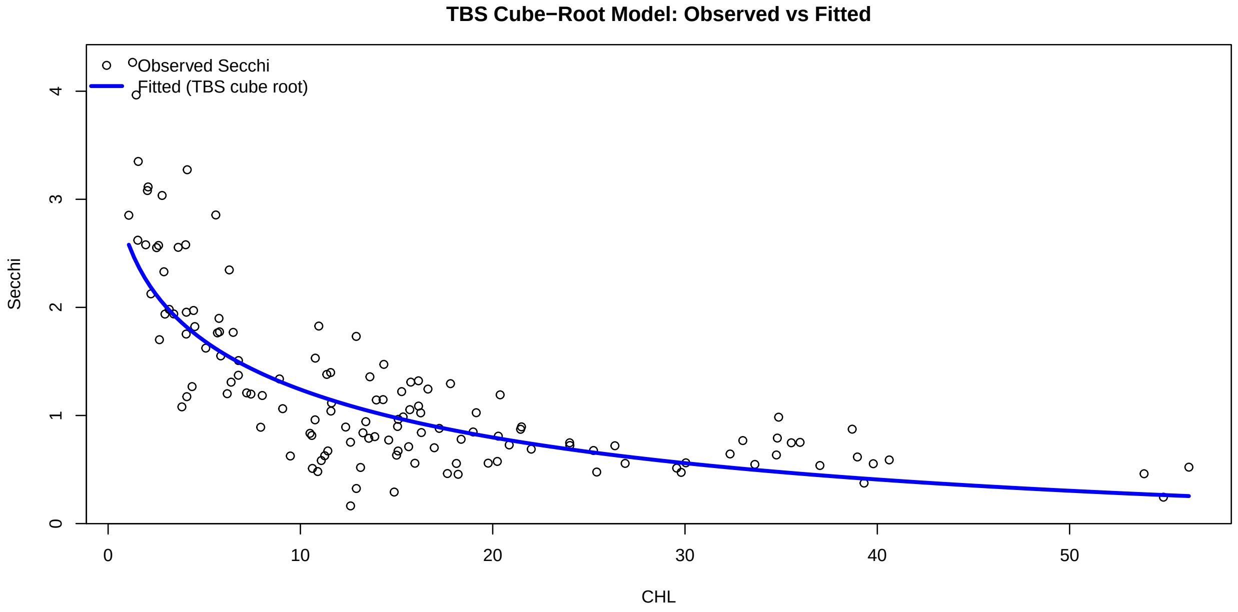

Transform Both Sides (Inverse)
1/(secchi)
1
2
3
4
5
6
1/(chl)
0.2
0.4
0.6
0.8


Transform Both Sides (Inverse Sqrt.)
sqrt(secchi)
0.5
1.0
1.5
2.0
sqrt(chl)
0.2
0.4
0.6
0.8


Transform Both Sides (Inverse Sq.)
secchi^1/3
5
10
15
20
25
30
35
chl^1/3
0.2
0.4
0.6
0.8


TBS (inverse): Residuals vs Fitted
Studentized residuals
0
2
4
6
8
−0.5
0.0
0.5
1.0
1.5
Fitted (1 / Secchi scale)
Observed vs Fitted (TBS inverse)
Observed Secchi
0
1
2
3
4
−60
−50
−40
−30
−20
−10
0
Fitted Secchi
TBS Inverse Model: Observed vs Fitted
Observed Secchi
Fitted (TBS inverse)
Secchi
0
1
2
3
4
0
10
20
30
40
50
CHL





We considered several possible variance–stabilizing transformations for a transform-both-sides (TBS) additive error model. Among these, the inverse and cube-root transformations appeared most promising based on their scatterplots.

The fitted TBS models and their diagnostic plots (shown above) make clear, however, that a TBS approach is not adequate for this application. Even with the best-performing transformation, the fitted curve fails to capture the observed curvature in the Secchi–CHL relationship, and the residual plots exhibit pronounced patterns and heteroscedasticity.

That said, the cube-root transformation performs noticeably better than the inverse transformation: it avoids the explosive behavior near low Secchi values and provides somewhat improved variance stabilization. Nonetheless, appreciable funneling remains in the residuals, particularly at the lower end of CHL, and the overall fit is still inferior to the GLM approaches. Consequently, the TBS additive error model is not competitive as an overall model for Secchi depth.

Power of the Mean

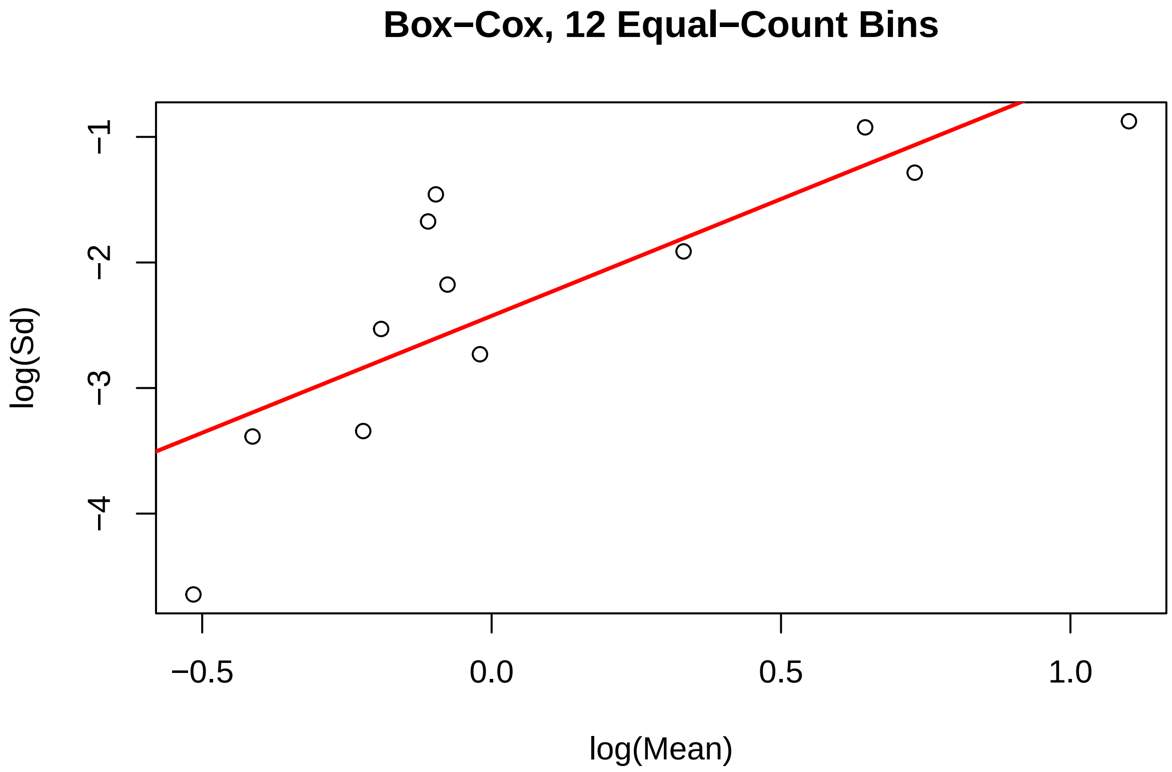
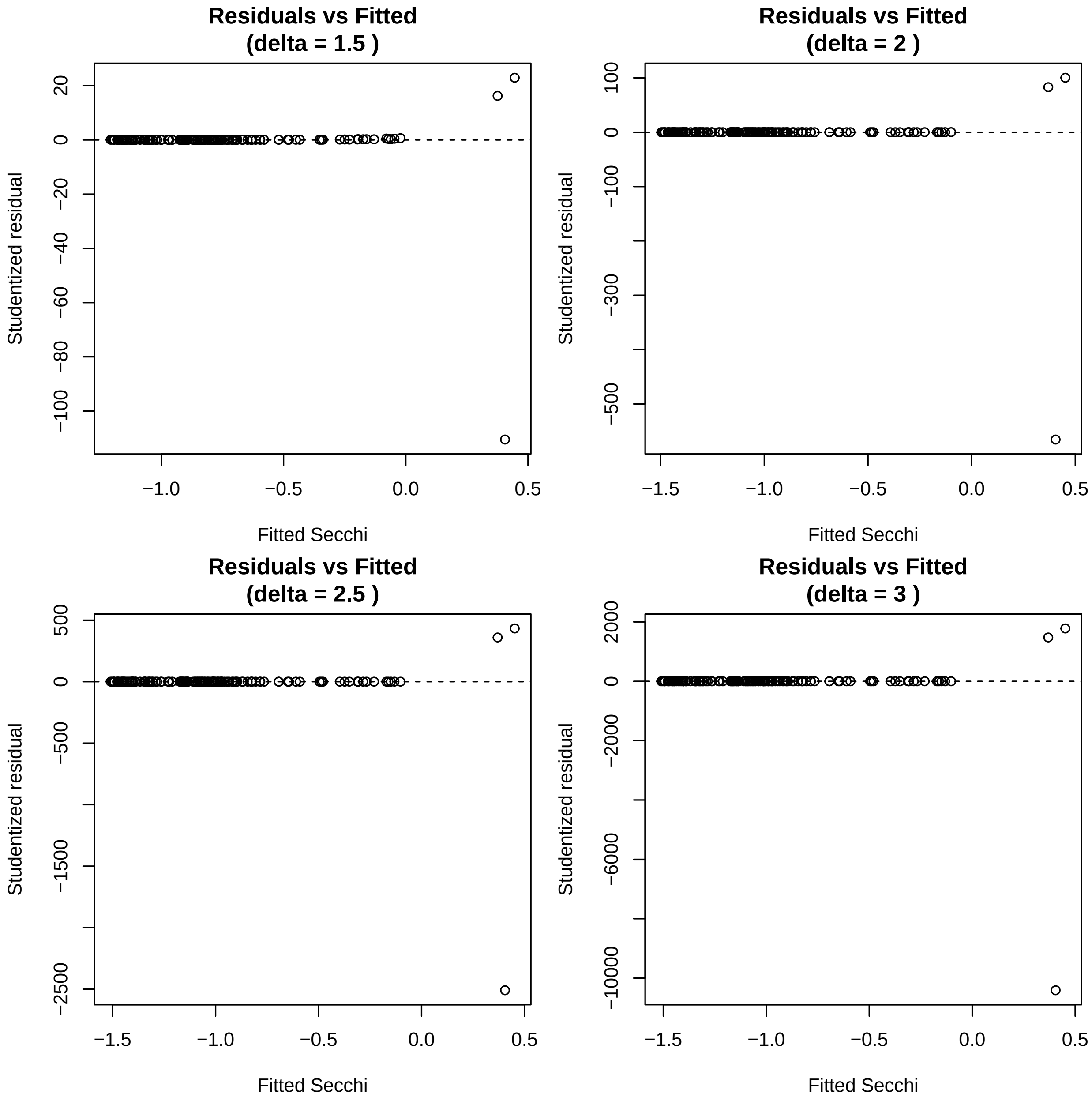
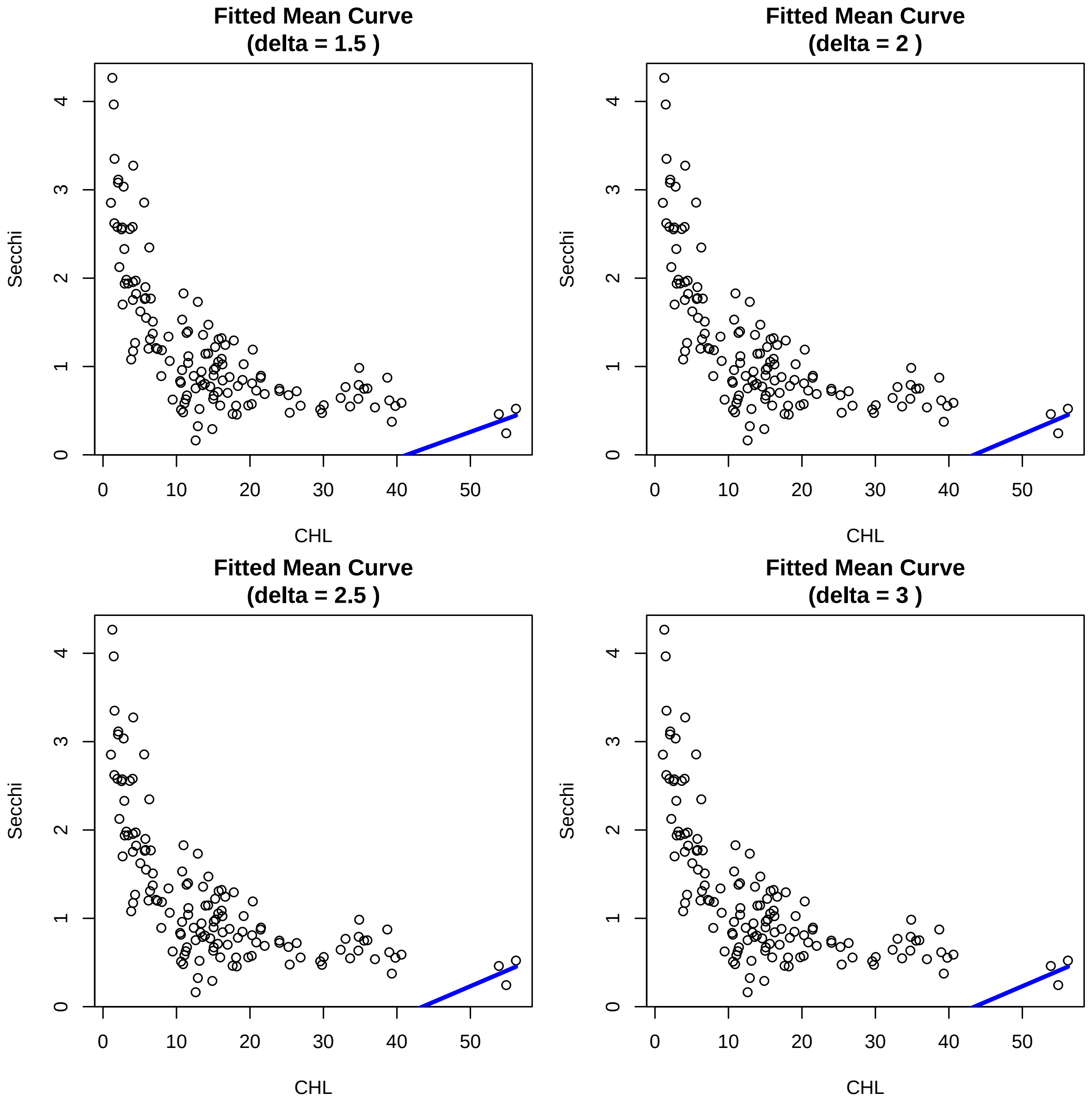


Table 6: Estimated delta (power of the mean) from Box-Cox SD–Mean plots

nbins	Equal.Count.delta	Equal.Spaced.delta
6	2.129	2.126
10	1.863	2.709
12	1.862	2.118
14	1.693	2.294
16	1.748	2.690
18	1.734	2.372
22	1.773	2.422


Residuals vs Fitted
(delta = 1.5)
Studentized residual
20
0
−20
−40
−60
−80
−100
−1.0
−0.5
0.0
0.5
Fitted Secchi
Residuals vs Fitted
(delta = 2)
Studentized residual
100
0
−100
−300
−500
−1.5
−1.0
−0.5
0.0
0.5
Fitted Secchi
Residuals vs Fitted
(delta = 2.5)
Studentized residual
500
0
−500
−1500
−2500
−1.5
−1.0
−0.5
0.0
0.5
Fitted Secchi
Residuals vs Fitted
(delta = 3)
Studentized residual
2000
0
−2000
−6000
−10000
−1.5
−1.0
−0.5
0.0
0.5
Fitted Secchi



Fitted Mean Curve (delta = 1.5)

Fitted Mean Curve (delta = 2)

Fitted Mean Curve (delta = 2.5)

Fitted Mean Curve (delta = 3)

We next consider power-of-the-mean (PoM) additive error models as an alternative to the GLM framework. Recall that the Box–Cox mean–standard-deviation plots for Secchi versus CHL suggested a variance relationship of approximately

$$\mathrm{Var}(Y \mid x) \propto \mu(x)^{\theta}$$

with θ in the range 1.5–3. This motivates exploring PoM models with variance weights of the form

$$w_i \propto \mu(x_i)^{-2\delta},$$

where the power δ is chosen to counteract the apparent growth in variance with the mean. Based on the Box–Cox summaries, we examine PoM models for

$$\delta = 1.5, 2, 2.5, \text{and } 3,$$

corresponding to increasingly aggressive variance stabilization.

However, the fitted curves and residual diagnostics for these PoM models reveal substantial problems. For several choices of δ, the fitted Secchi–CHL relationship becomes biologically and statistically implausible: the fitted curve is essentially flat or even slightly increasing over much of the CHL range, with a pronounced and unrealistic "upturn" only appearing for CHL values above about 40. This behavior contradicts the clearly decreasing pattern seen in the raw data and in the GLM fits.

The residual plots tell a similar story. Even after weighting by $\mu(x)^{-2\delta}$, the residuals exhibit strong heteroskedasticity and systematic patterns across CHL, indicating that the assumed PoM variance structure does not adequately capture the true mean–variance relationship. In particular, large residuals persist at both low and moderate CHL values, and the spread of residuals does not appear stable across the range of fitted means.

Taken together, these issues suggest that, although the Box–Cox diagnostics motivated exploring PoM models, the resulting fits are not adequate descriptions of the Secchi–CHL relationship. The power-of-the-mean additive error models considered here fail to reconcile the observed curvature and variance pattern, and are therefore not competitive with the Gamma GLM with inverse link as candidates for the overall model.

These PoM models, however, were all specified directly in terms of CHL as the predictor. Given that Secchi depth is itself a measure of light penetration through the water column, it may be more natural to re-express the model in terms of a volumetric or "optical" argument that directly links depth to the amount of material (CHL) per unit volume. This motivates a further exploration based on a simple disk–volume construction.

A Possibly Better Approach? Disk & Volume

The initial motivation for a different additive formulation comes from a simple geometric argument relating Secchi depth to the volume of water above the disk. Suppose the visible water column above the Secchi disk can be idealized as a cone with radius r and height h, where h is the Secchi depth.

The volume of this cone is

$$V = \pi r^2 \frac{h}{3}.$$

Solving for depth gives

$$h = \frac{3V}{\pi r^2}.$$

Chlorophyll concentration, CHL, is measured as mass per unit volume (e.g., $\mu g/L$). If we think of V as the volume of water above the disk, then the total amount of chlorophyll in that column is proportional to CHL $\times V$. Under a highly simplified view where "visibility" is determined by a roughly fixed amount of material above the disk (i.e., a fixed effective mass of chlorophyll obscuring the disk), we would have

$$\text{CHL} \times V \approx \text{constant},$$

so that $V \propto 1/\text{CHL}$.

Substituting back into the expression for h yields

$$h \propto \frac{V}{r^2} \propto \frac{1}{\text{CHL}},$$

suggesting that Secchi depth should be approximately inversely related to CHL.

This heuristic argument supports modeling Secchi depth as a decreasing function of 1/CHL rather than CHL itself.

In the additive error framework, this leads us to consider models of the form

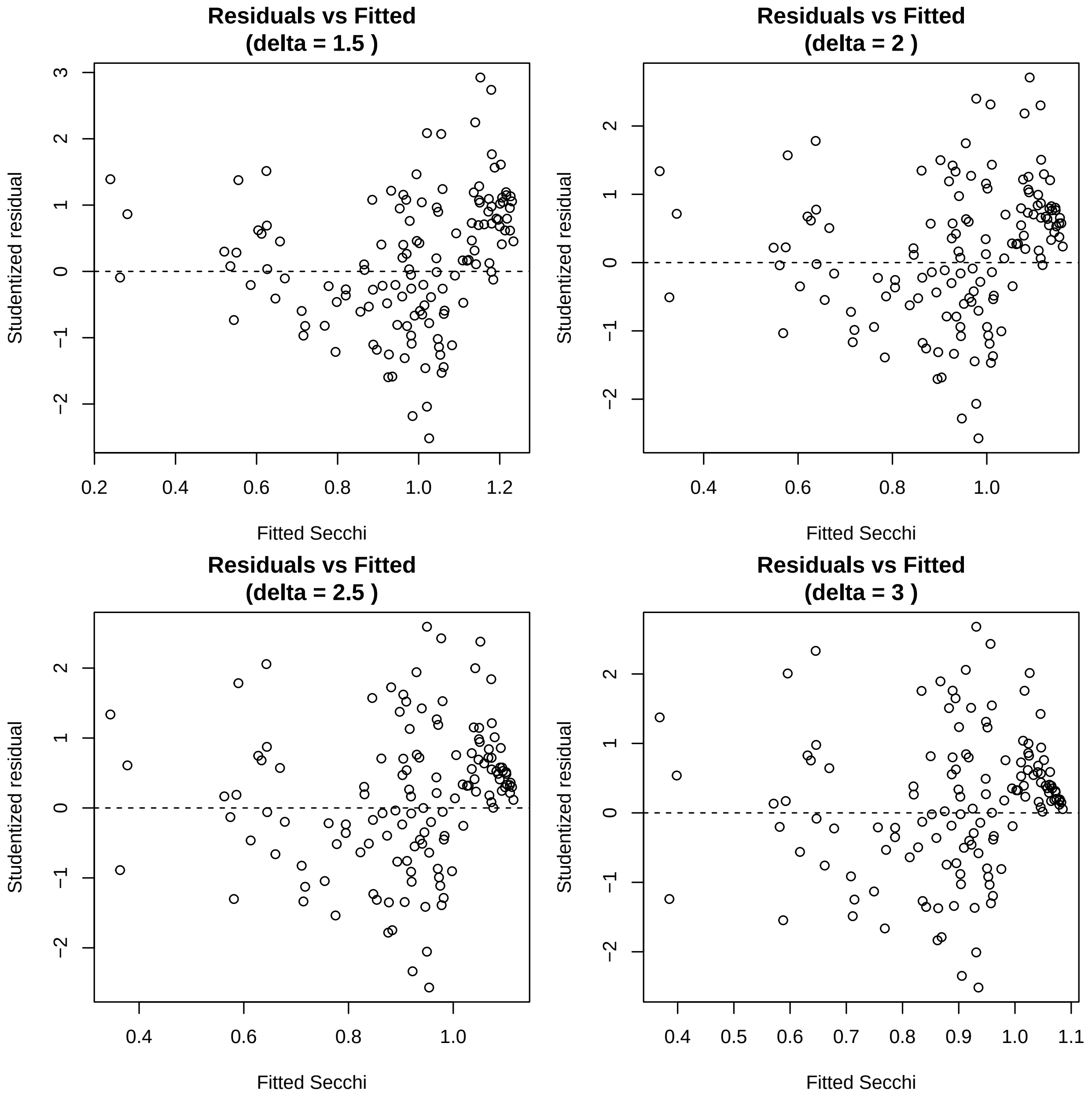
$$\text{Secchi}_i = \beta_0 + \beta_1 \frac{1}{\text{CHL}_i} + \varepsilon_i,$$

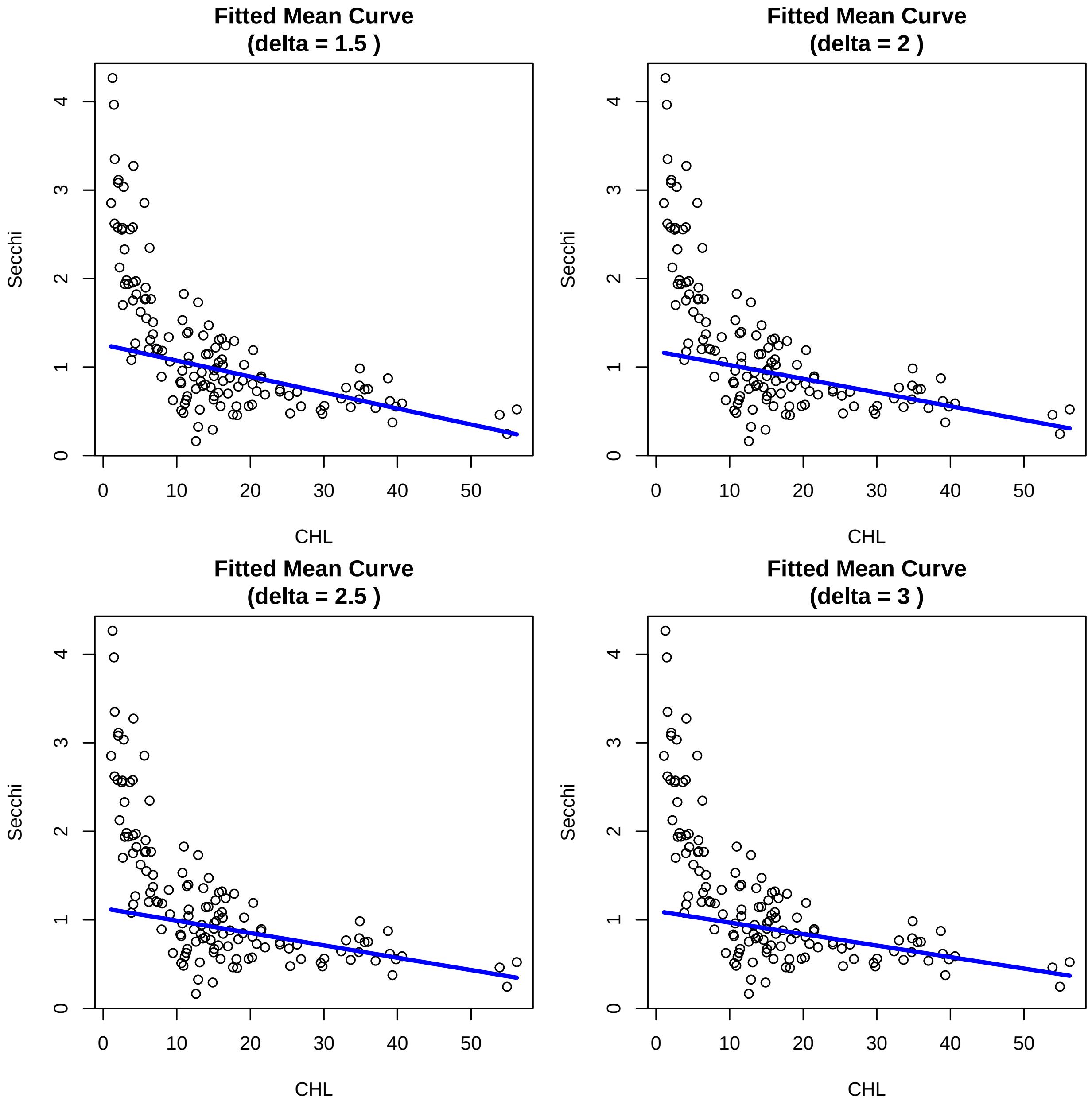
and, in the transform-both-sides setting,

$$\text{Secchi}_i^{1/3} = \gamma_0 + \gamma_1 \left(\frac{1}{\text{CHL}_i} \right)^{1/3} + e_i,$$

which retain the decreasing relationship, avoid explosive behavior at low Secchi depths, and align with the reciprocal structure suggested by the disk–volume mechanism.

In practice, these reciprocal and transform-both-sides models produce smoother fitted curves than the earlier PoM fits and offer some improvement in variance stabilization on the transformed scale. However, their residual diagnostics are still not as clean as those of the Gamma GLM with inverse link, so we treat them as useful supporting models rather than as the primary overall model for Secchi depth.


Residuals vs Fitted
(delta = 1.5)
Studentized residual
Fitted Secchi
3
2
1
0
−1
−2
0.2
0.4
0.6
0.8
1.0
1.2
Residuals vs Fitted
(delta = 2)
Studentized residual
Fitted Secchi
2
1
0
−1
−2
0.4
0.6
0.8
1.0
Residuals vs Fitted
(delta = 2.5)
Studentized residual
Fitted Secchi
2
1
0
−1
−2
0.4
0.6
0.8
1.0
Residuals vs Fitted
(delta = 3)
Studentized residual
Fitted Secchi
2
1
0
−1
−2
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1



The development of this model parallels the geometric reasoning used in earlier examples, such as:

- the tree-trunk cylinder model, where diameter and height jointly determine volume; and
- the walleye cuboid example, where body depth and length jointly determine weight.

In both cases, the response variable is interpreted as a geometric function of the underlying physical structure, and a seemingly nonlinear biological curve becomes linear after expressing the response against an appropriate geometric predictor.

Here, we apply the same logic to Secchi depth. If the visible water column above the disk is idealized as a cone (as argued in the previous section), then Secchi depth is inversely related to the amount of material (i.e., chlorophyll) in the water column. This motivates an additive model built on 1/CHL rather than CHL itself, just as the tree-diameter and fish-body-depth examples motivated transforming their predictors before fitting.

We have:

$$x = \text{CHL}, \qquad \phi = \frac{1}{x}, \qquad Y = \text{Secchi depth}.$$

We fit an additive linear model in the transformed predictor

$$Y = \beta_0 + \beta_1, \phi + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid \phi] = 0.$$

so that

$$\mathbb{E}[Y \mid \phi] = \beta_0 + \beta_1, \phi.$$

This is directly analogous to the straight-line model in the tree-diameter or walleye-body-depth examples, except the biologically meaningful predictor is now 1/CHL.

Now we express the mean as a function of CHL

$$m(x) = \mathbb{E}[Y \mid x] = \beta_0 + \frac{\beta_1}{x},$$

a reciprocal curve that is steep when CHL is low and gradually flattens, just like the concave-down fish-weight and tree-volume curves in the earlier examples.

To stabilize variance, we also apply the same cube-root transform-both-sides logic used in the walleye example, where weight was linearized after transforming both sides.

Let

$$Z = Y^{1/3}, \qquad \psi = \phi^{1/3} = x^{-1/3}.$$

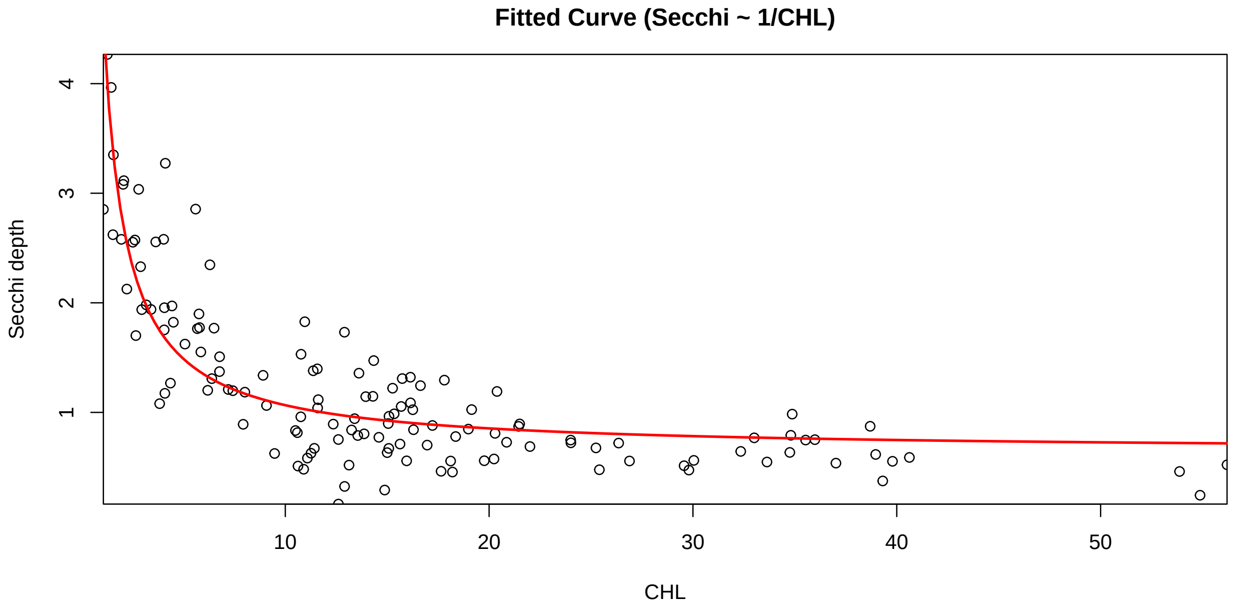
Fit

$$Z = \alpha_0 + \alpha_1, \psi + \eta, \qquad \mathbb{E}[\eta \mid \psi] = 0.$$

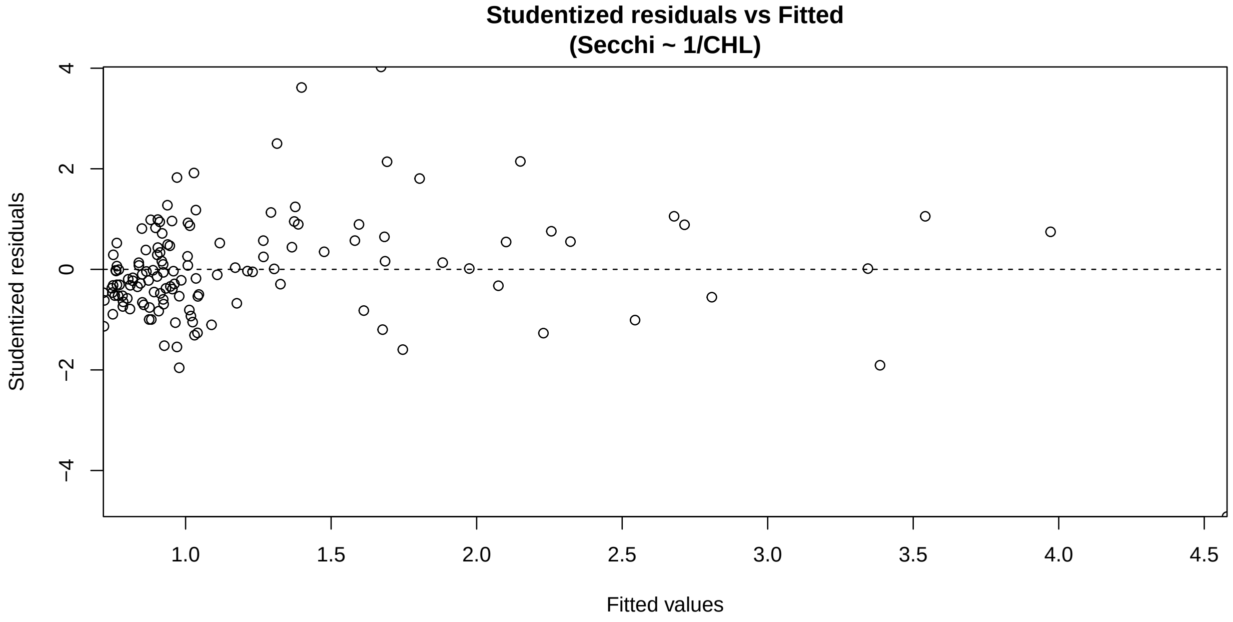
Then

$$\mathbb{E}[Z \mid x] = \alpha_0 + \alpha_1 x^{-1/3} \quad \Longrightarrow \quad \mathbb{E}[Y \mid x] \approx \left(\alpha_0 + \alpha_1 x^{-1/3}\right)^3.$$

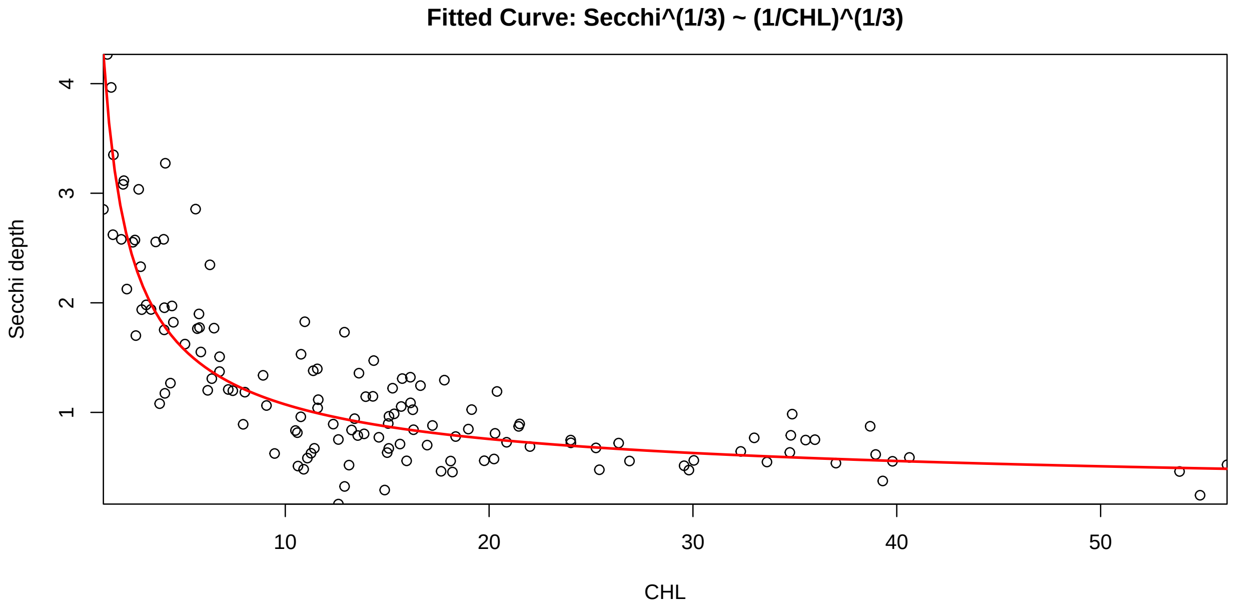
This reproduces the same modeling pattern as the cube-root fish-weight example: a straight line on the transformed scale becomes a smooth, strictly decreasing nonlinear curve on the original scale.

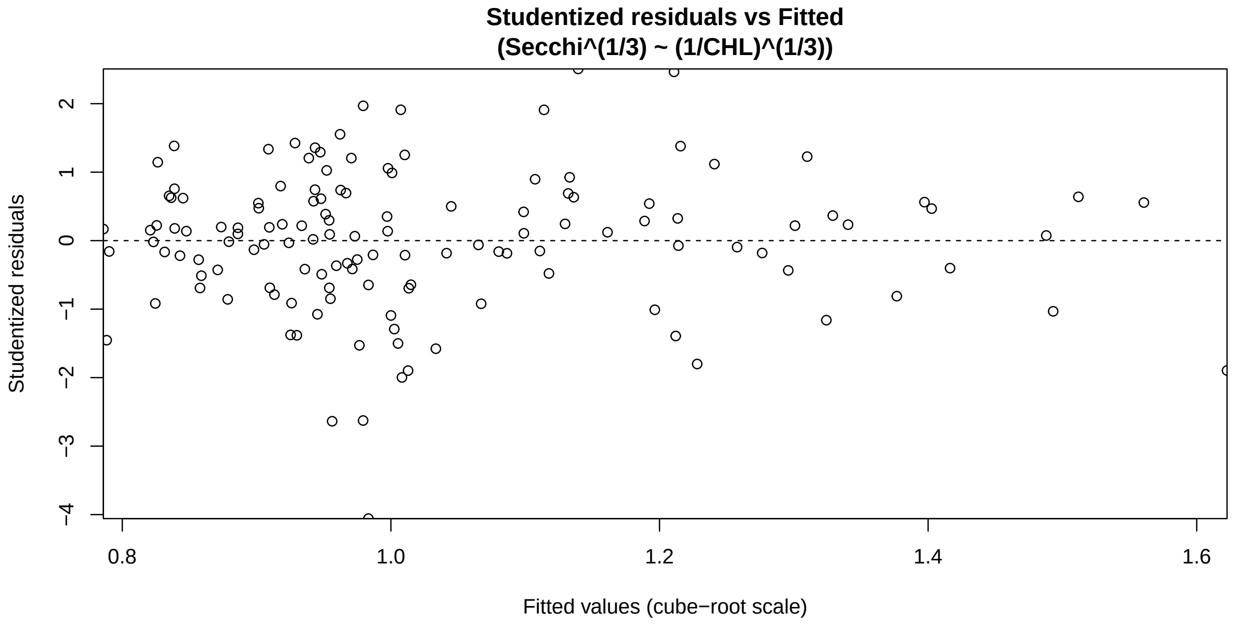


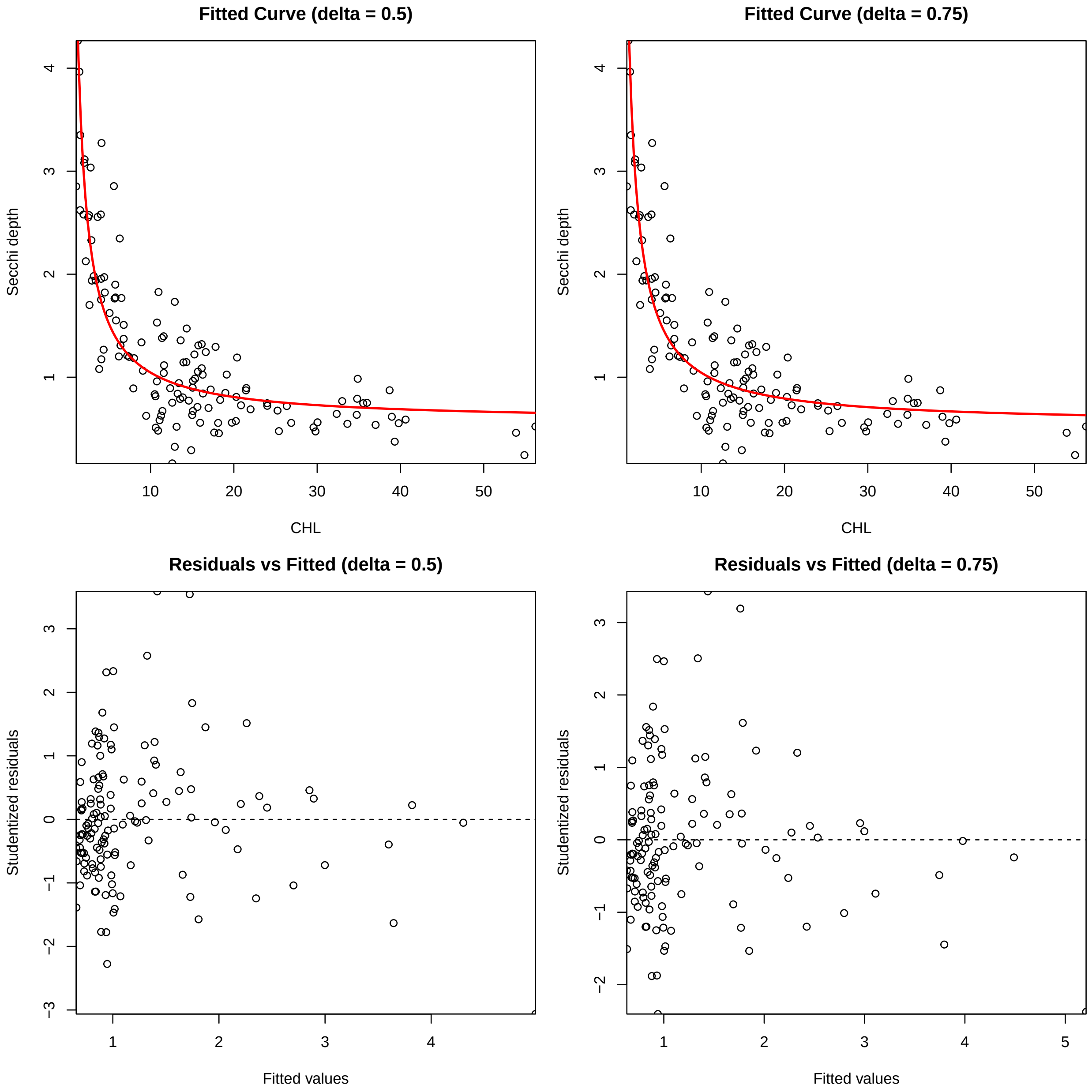
Fitted Curve (Secchi ~ 1/CHL)



Studentized residuals vs Fitted
(Secchi ~ 1/CHL)


Fitted Curve: Secchi^(1/3) ~ (1/CHL)^(1/3)
Secchi depth
1
2
3
4
10
20
30
40
50
CHL


Studentized residuals vs Fitted
(Secchi^(1/3) ~ (1/CHL)^(1/3))
Studentized residuals
−4
−3
−2
−1
0
1
2
0.8
1.0
1.2
1.4
1.6
Fitted values (cube−root scale)



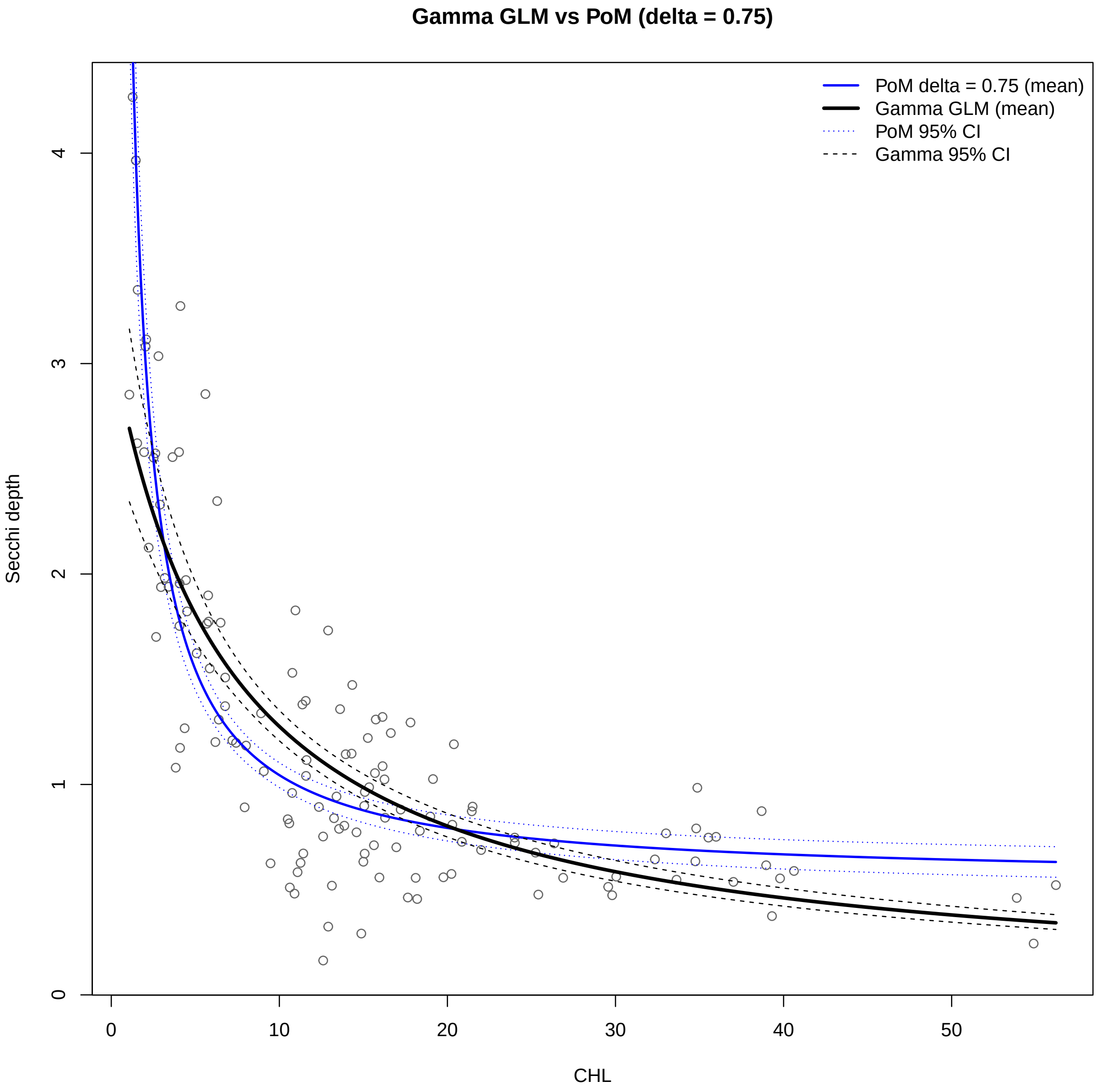
Although the simple reciprocal additive model and its cube-root transform-both-sides version both capture the basic decreasing shape, neither achieves satisfactory variance stabilization on its own. The Power-of-the-Mean extensions with $\delta = 0.5$ and $\delta = 0.75$ provide the best performance among the additive models we explored, offering partial variance stabilization and smooth reciprocal curves. However, even these PoM-weighted fits retain visible structure in the residuals and do not match the quality of the Gamma GLM with inverse link. As a result, the only additive models that might reasonably be considered "adequate" are the reciprocal TBS model and the reciprocal PoM models with $\delta = 0.5$ or $\delta = 0.75$, but they remain secondary to—rather than competitors with—the Gamma GLM selected as the overall model.

Between the two power-of-mean specifications considered, $\delta = 0.75$ produced a marginally better high-end (large CHL) fit: the tail of the fitted curve adhered more closely to the empirical Secchi pattern than the $\delta = 0.5$ model.

Comparing Models

Adequate models considered were:

- Gamma GLM with inverse link
- Power-of-the-mean additive error model with $\delta = 0.75$ with cube root variance stabilizing transformation (TBS)



Overall, the Gamma GLM provides the best combination of fit quality, stability, and interpretability, and is therefore the preferred model. The rationale is as follows: Although both the Gamma GLM (inverse link) and the PoM–TBS model with $\delta = 0.75$ produce broadly similar fitted curves, several features favor the Gamma GLM:

1. **Correct mean–variance structure.** The Gamma model naturally encodes $\mathrm{Var}(Y \mid x) = \phi, \mu(x)^2$,

which matches the heteroscedasticity in Secchi depth. The PoM–TBS model imposes a variance structure through estimated weights and is more sensitive to instability in $\hat{\mu}$.

2. **More stable fitted curve and bands.** The Gamma GLM yields a smooth monotone decreasing curve with tight, regular confidence intervals. The PoM curve flattens at high CHL and its CI flares at both ends because of back-transformation and weight sensitivity.

3. **Better residual behavior.** Earlier diagnostics showed the Gamma GLM's deviance residuals were well-behaved, while the PoM model retained curvature and uneven spread.

4. **Simpler interpretation.** The GLM mean $\mu(x) = 1/(\beta_0 + \beta_1 x)$ is biologically interpretable and does not require back-transformation. The PoM–TBS model requires transforming both sides and back-transforming the mean $(\alpha_0 + \alpha_1 x^{-1/3})^3$, making inference more cumbersome.

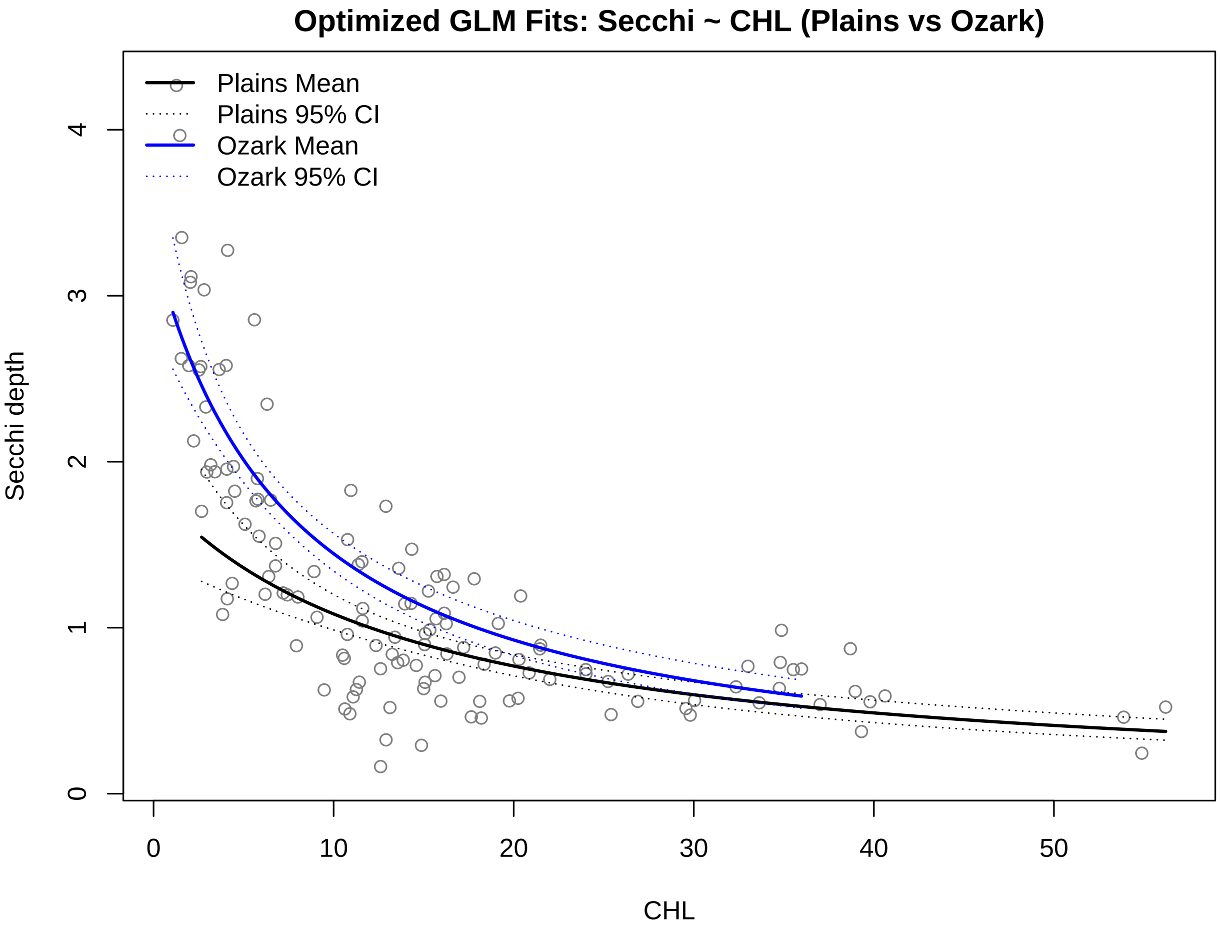
Applying to Regions

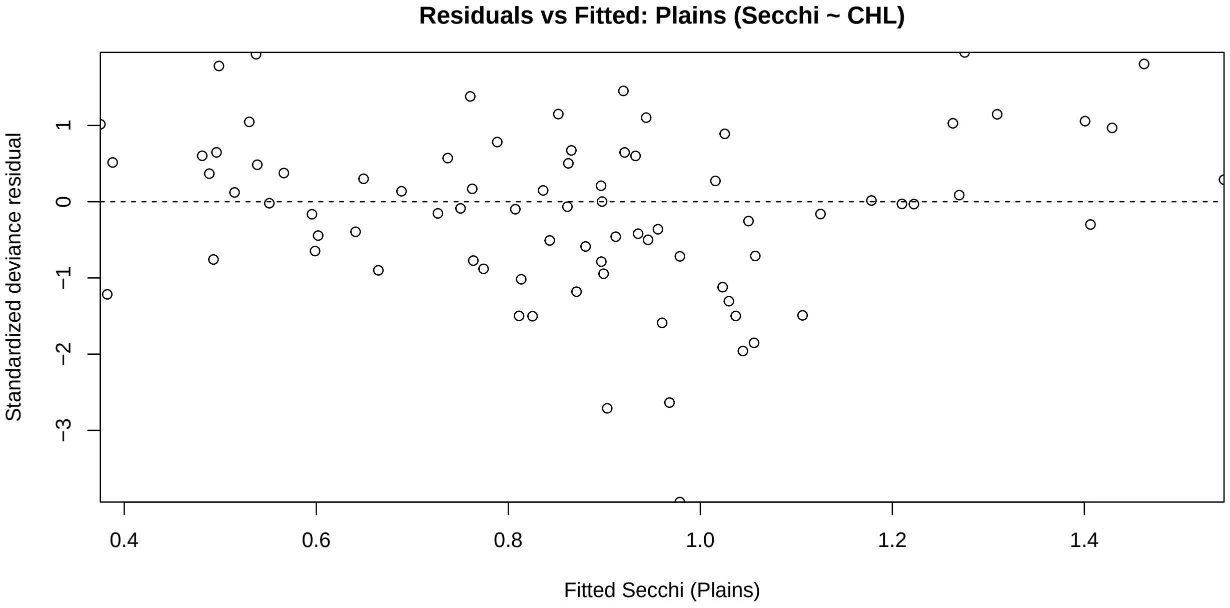
Table 7: Region-specific optimized GLM coefficients for Secchi CHL with 95% Wald CIs

Region	Term	Estimate	SE	LCL	UCL
Plains	Intercept	0.5464	0.0799	0.3898	0.7030
Plains	CHL	0.0377	0.0051	0.0277	0.0477
Ozark	Intercept	0.3030	0.0263	0.2514	0.3547
Ozark	CHL	0.0389	0.0039	0.0312	0.0465

Table 8: Predicted Secchi for CHL = 1, 5, 10, 20 with 95% CIs by region

Region	CHL	Fit	LCL	UCL
Plains	1	1.7119	1.3653	2.2944
Plains	5	1.3606	1.1728	1.6200
Plains	10	1.0829	0.9865	1.2001
Plains	20	0.7689	0.7101	0.8384
Ozark	1	2.9250	2.5747	3.3856
Ozark	5	2.0108	1.8740	2.1692
Ozark	10	1.4460	1.3418	1.5677
Ozark	20	0.9258	0.8322	1.0431


Optimized GLM Fits: Secchi ~ CHL (Plains vs Ozark)
Plains Mean
Plains 95% CI
Ozark Mean
Ozark 95% CI
Secchi depth
0
1
2
3
4
CHL
0
10
20
30
40
50



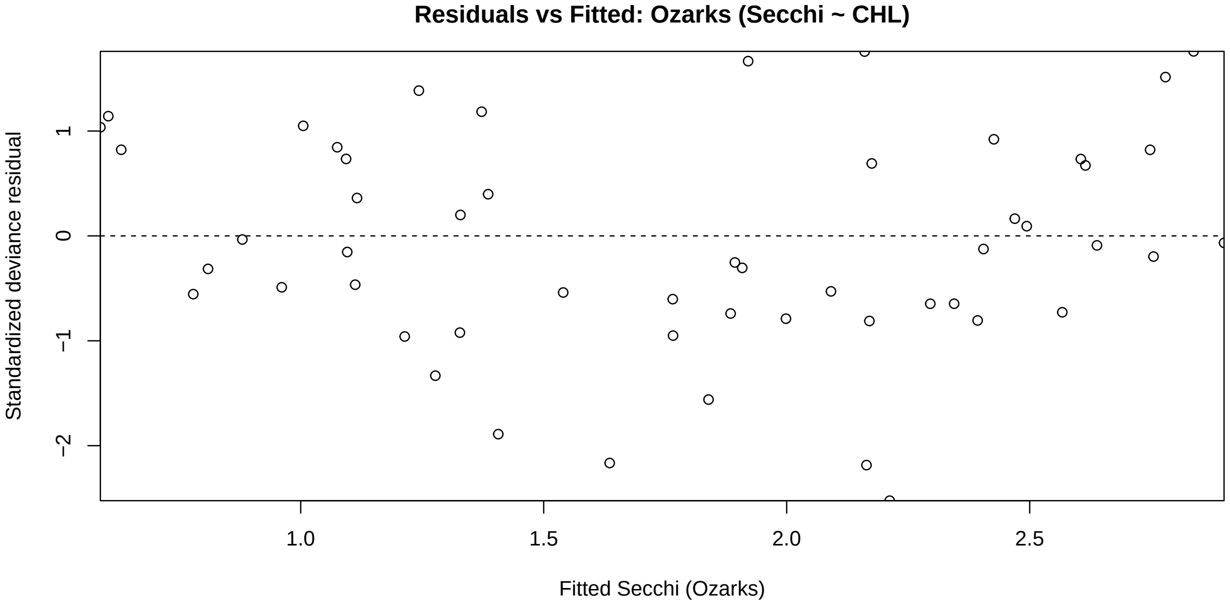


Table 9: Scaled deviance by region for Secchi ~ CHL Gamma GLMs

Region	Scaled_Deviance
Plains	93.27254
Ozarks	53.66591

Interpreting the Model & Results

Across both regions, Secchi depth shows the same basic negative and nonlinear relationship with chlorophyll. Secchi decreases as CHL increases, and the Gamma GLM with an inverse link provides a clear and biologically reasonable description of this pattern. When the model is fit separately to the Plains and the Ozarks using the same structure, the shape of the Secchi to CHL curve is the same in both regions. This indicates that the underlying process linking light attenuation to chlorophyll concentration is shared across Missouri lakes.

However, the strength of the relationship differs in a meaningful way. The fitted curves show a clear vertical separation between the two regions. For any fixed CHL value, the Plains consistently have higher Secchi depth than the Ozarks. This difference appears even at low CHL values, where Secchi in the Plains is roughly 1.7 to 2.5 meters compared to about 1.3 to 2.0 meters in the Ozarks. The separation grows as CHL increases. By CHL values around 20 to 30, the regional curves are clearly distinct, and their 95 percent confidence bands show little overlap.

The coefficient estimates reinforce this pattern. The Plains and Ozarks have almost identical slopes on the inverse link scale, which means the rate at which Secchi decreases with CHL is similar. But the Ozarks have a smaller fitted intercept, which corresponds to lower overall Secchi depth across the entire CHL range. This suggests that Ozark lakes experience stronger light attenuation for the same amount of chlorophyll.

Residual diagnostics show that the model fits both regions well and does not point to different functional forms or variance patterns. This means the observed separation reflects true regional differences rather than modeling issues.

In summary, the two regions share the same basic biological relationship between Secchi depth and chlorophyll, but differ in overall clarity. Ozark lakes have consistently lower Secchi depth than Plains lakes for the same CHL values.

Q3: LRT SECCHI & CHL (Plains vs. Ozarks)

Let Y_{gi} denote Secchi depth for observation i in region $g \in P, O$ (Plains, Ozarks), with covariate $x_{gi} = \text{CHL}$. Assume a Gamma model with common shape $\alpha > 0$ and mean $\mu_{gi} > 0$:

$$Y_{gi} \mid x_{gi} \sim \text{Gamma}(\alpha, \mu_{gi}).$$

The density (shape/mean form) is

$$f(y_{gi} \mid \mu_{gi}, \alpha) = \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu_{gi}}\right)^{\alpha} y_{gi}^{\alpha-1} \exp\left(-\frac{\alpha y_{gi}}{\mu_{gi}}\right), \qquad y_{gi} > 0.$$

The GLM uses the inverse link

$$g(\mu) = \frac{1}{\mu}, \qquad \eta_{gi} = \frac{1}{\mu_{gi}}.$$

Reduced model (common regression)

$$\eta_{gi} = \beta_0 + \beta_1 x_{gi}, \qquad \mu_{gi} = \frac{1}{\beta_0 + \beta_1 x_{gi}}.$$

The reduced log–likelihood is

$$\ell_0(\beta_0, \beta_1, \alpha) = \sum_{g,i} \Big[(\alpha - 1)\log y_{gi} - \alpha y_{gi}(\beta_0 + \beta_1 x_{gi}) + \alpha \log(\beta_0 + \beta_1 x_{gi}) + \alpha \log \alpha - \log \Gamma(\alpha)\Big].$$

Full model (region-specific regressions)

$$\eta_{Pi} = \beta_{P0} + \beta_{P1} x_{Pi}, \qquad \eta_{Oi} = \beta_{O0} + \beta_{O1} x_{Oi},$$

$$\mu_{Pi} = \frac{1}{\beta_{P0} + \beta_{P1} x_{Pi}}, \qquad \mu_{Oi} = \frac{1}{\beta_{O0} + \beta_{O1} x_{Oi}}.$$

Parameters:

$$\theta_1 = (\beta_{P0}, \beta_{P1}, \beta_{O0}, \beta_{O1}, \alpha).$$

The full log–likelihood is

$$\begin{aligned}\ell_1 = &\sum_{i=1}^{n_P} \Big[(\alpha - 1)\log y_{Pi} - \alpha y_{Pi}(\beta_{P0} + \beta_{P1} x_{Pi}) + \alpha \log(\beta_{P0} + \beta_{P1} x_{Pi}) + \alpha \log \alpha - \log \Gamma(\alpha)\Big] \\ &+ \sum_{i=1}^{n_O} \Big[(\alpha - 1)\log y_{Oi} - \alpha y_{Oi}(\beta_{O0} + \beta_{O1} x_{Oi}) + \alpha \log(\beta_{O0} + \beta_{O1} x_{Oi}) + \alpha \log \alpha - \log \Gamma(\alpha)\Big]\end{aligned}$$

Likelihood ratio test then is of the form:

$$-2\{\ell_1(\hat{\theta}_0) - \ell_0(\hat{\theta}_1)\} \leq \chi^2_{2,1-\alpha}$$

since the full model has 5 parameters and the reduced model has 3.

For:

Table 10: Model Fit, Log-Likelihoods, and LRT Results

Quantity	Value
Convergence (Reduced)	0
Convergence (Full)	0
LogLik (Reduced MLE)	-58.30
LogLik (Full MLE)	-58.30
Lambda (raw)	-1e-05
LRT p-value	1.000

- H_0: reduced model is true (no region difference)
- H_1: full model is true (regions differ)

In short,

- Definitive Question: Do the regions differ in the relation between Chlorophyll and Secchi depth?
- Definitive Answer: No!