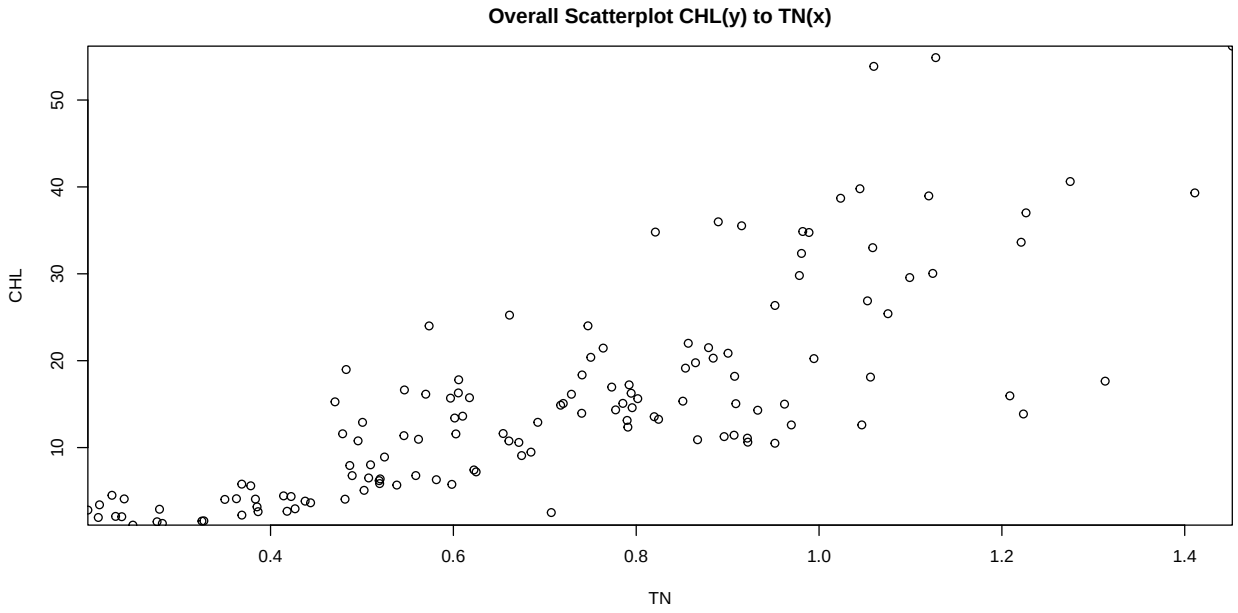
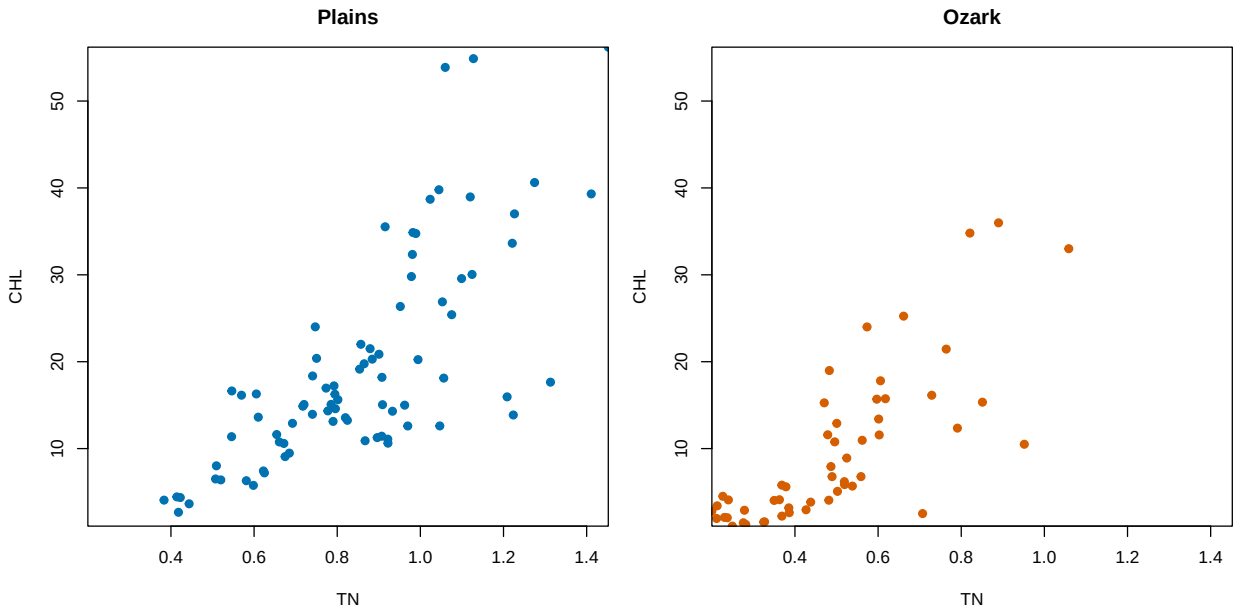
5200 Take-Home

Sam Olson

Q1: CHL & TN (Plains vs. Ozarks)

Overall Distribution & Approach





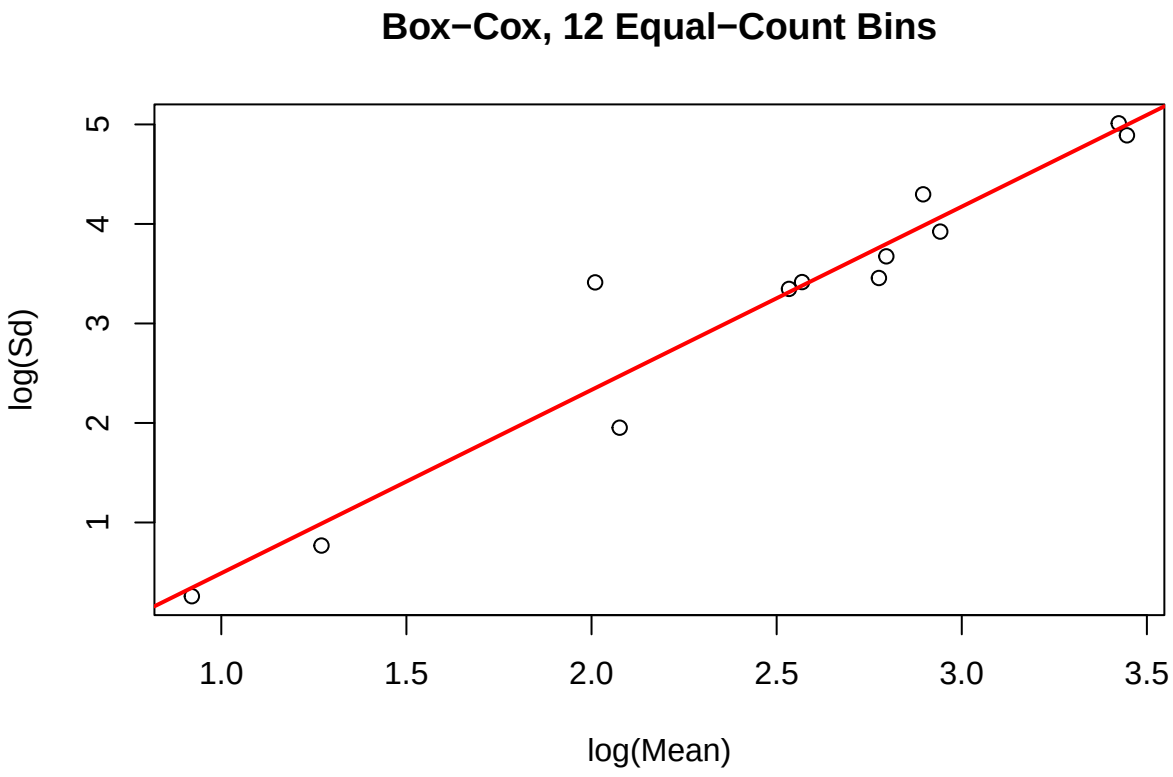
CHL is right-skewed and increases with TN, with variance also rising at higher TN values. The regional scatterplots (Plains vs. Ozarks) show the same qualitative pattern but suggest possible differences in the strength of the TN–CHL relationship.

Let Y_i denote CHL for lake i ($i = 1, \ldots, 134$), with corresponding TN value x_i. Each lake appears once, so the observations may be treated as independent. Specifying a conditional model for $Y_i \mid x_i$ yields a joint distribution for the full collection Y_i.

Assuming a common underlying biological mechanism across Missouri reservoirs, we first fit an overall TN–CHL model to determine an appropriate mean–variance structure and link function. We then apply this same model structure separately to the Plains and Ozarks so that any differences in fitted curves or parameters reflect genuine regional differences rather than artifacts of using different model families.

This framework supports meaningful comparison of regression functions, confidence bands, and derived quantities such as $\Pr(Y > Z \mid x)$.

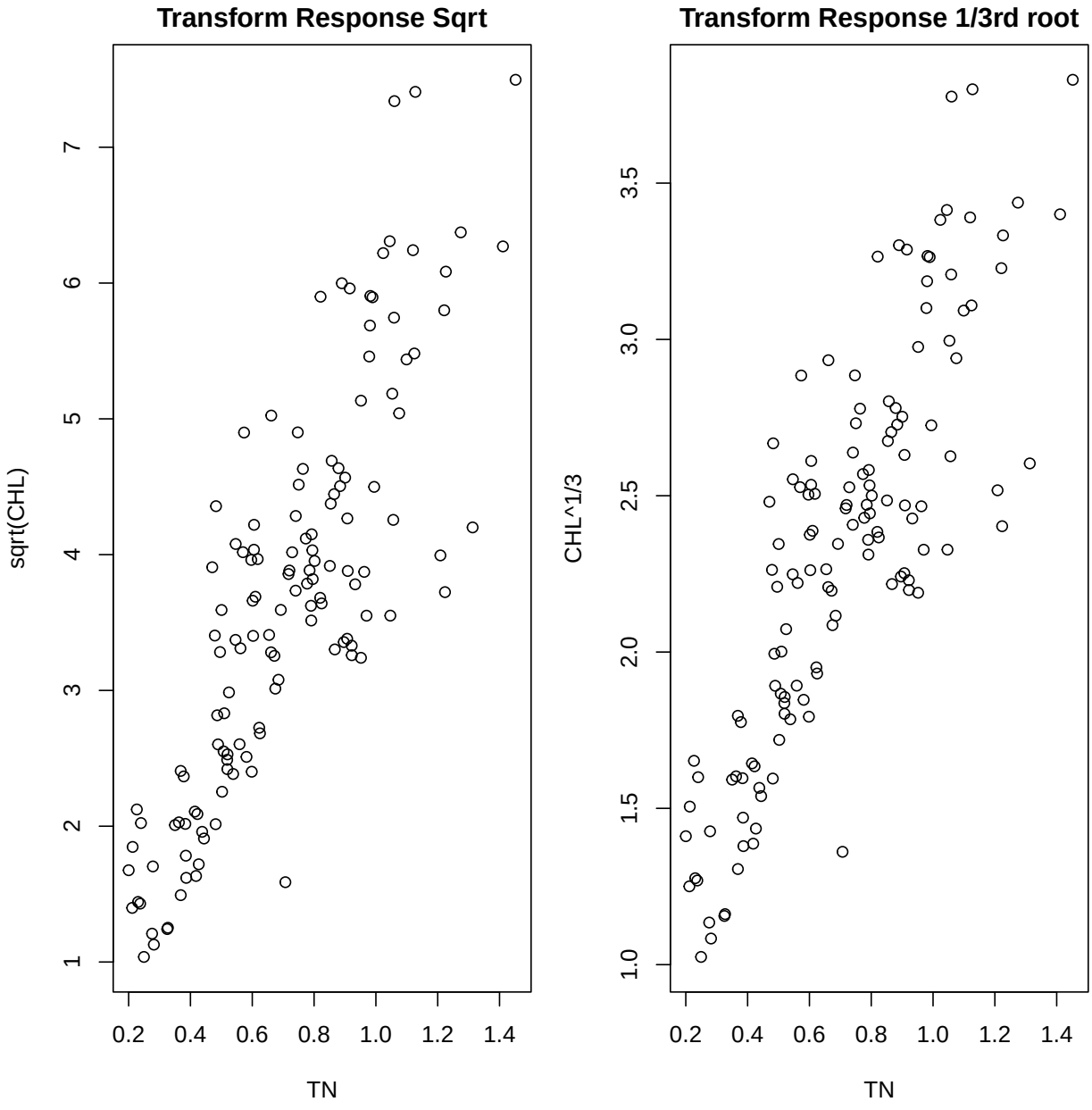
Generalized Linear Models



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

To select an appropriate GLM family, I used the `boxcoxfctns` routines to estimate Box–Cox mean–variance slopes. Across all binning schemes, the slopes lay between 1.6 and 2.1, indicating that $\mathrm{Var}(Y \mid x)$ grows roughly like μ^2–μ^3. This variance pattern is consistent with Gamma or Inverse Gaussian random components, so these families are the natural GLM candidates.

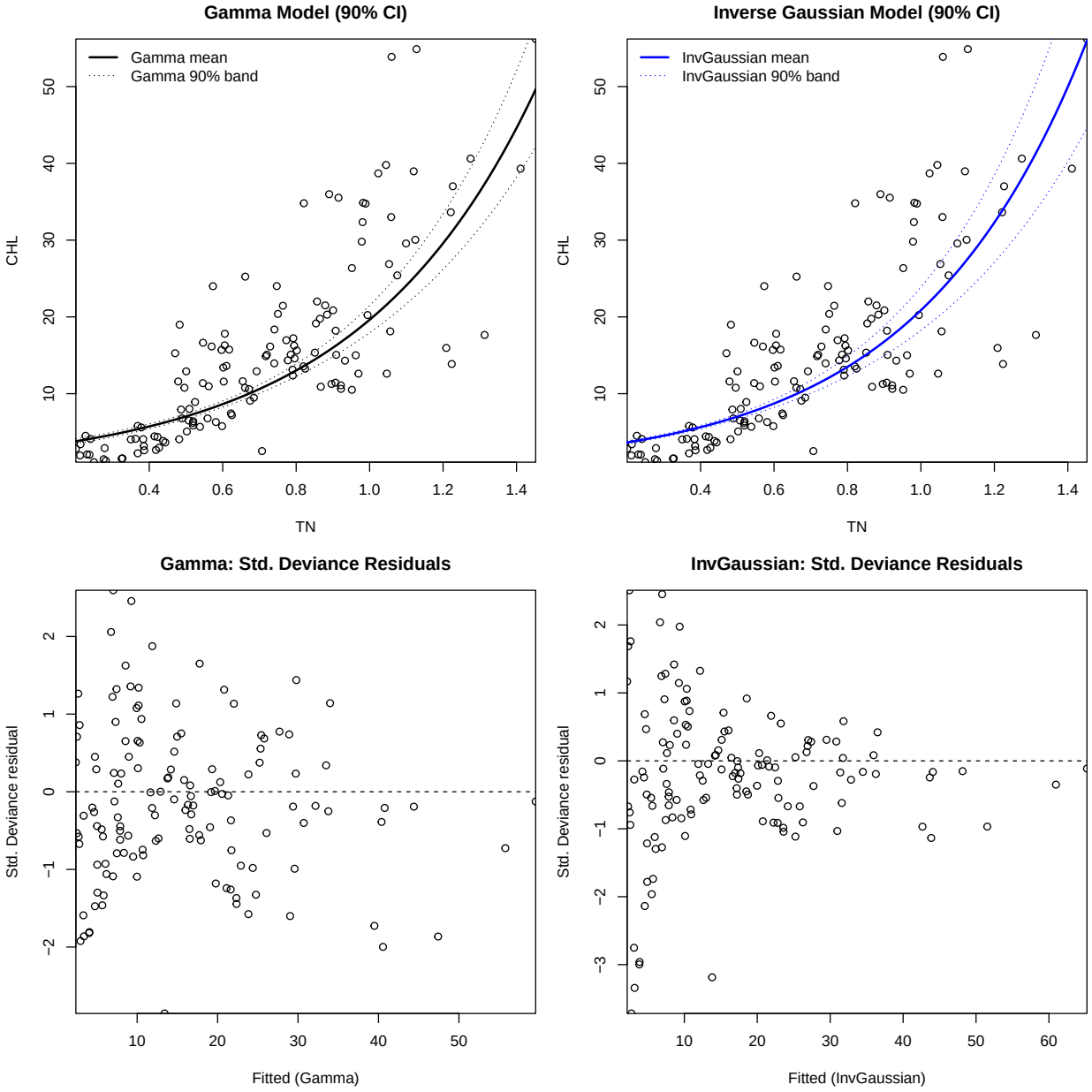
With the random component identified, the next step is to determine a suitable link function for modeling the TN–CHL mean relationship.



To select a link function, I examined transformations of CHL that approximately linearize its relationship with TN. Exploratory plots showed that the cube-root transformation yielded the most linear trend, with the square-root transformation performing reasonably well. These transformations were used only to guide link choice—not to transform the response in the GLM itself.

Combined with the earlier variance diagnostics, this motivated fitting Gamma and Inverse Gaussian GLMs paired with power links corresponding to the cube-root and square-root transformations.

Within each GLM family, the scaled deviance provides an appropriate criterion for comparing link functions. The cube-root power link consistently produced lower deviances than the square-root link in both the Gamma and Inverse Gaussian families, so attention is restricted to the cube-root versions of these two models. The Gamma and IG models themselves are not compared by deviance, as they are non-nested.



The Gamma and Inverse Gaussian models yield nearly identical fitted curves, so scatterplots alone cannot distinguish them. Residual diagnostics provide clearer evidence: the Gamma model exhibits more stable variance and better-behaved deviance residuals, whereas the IG model shows mild heteroscedasticity. Thus, the Gamma GLM is selected as the preferred model.

With the GLM choice established, I briefly examine additive error models for completeness before selecting a final model for the region-specific comparisons.

Other Models Considered – Additive Error Models

Transform Both Sides

A simple OLS fit shows a roughly linear TN–CHL trend but clear heteroscedasticity in the studentized residuals, motivating consideration of transform-both-sides (TBS) additive error models. Several variance-stabilizing transformations were examined; the cube-root transformation performed best and produced a reasonable linearization of the relationship.

However, TBS models require back-transformation, and only the mean back-transforms cleanly—variance and quantile inference become distorted. Combined with the remaining heteroscedasticity in the residuals, the TBS approach is not competitive with the GLM.

Power of the Mean

Given the apparent nonlinearity and mean–variance relationship in the data, I also considered power-of-the-mean (POM) additive error models. Several values of the variance-stabilizing power δ were tested, including models that allowed the mean curve itself to be nonlinear. Although $\delta = 0.5$ performed best—consistent with the Box–Cox slopes—its residuals still showed heteroscedasticity, and small fitted means made the weighting scheme unstable for larger δ.

Overall, even the best POM fits failed to achieve satisfactory variance stabilization and tended to misrepresent higher CHL values. In contrast, the Gamma GLM provided cleaner residuals and a coherent mean–variance structure, so the POM family was not retained as a candidate model.

Comparing Different Models

The two GLM candidates were the Gamma and Inverse Gaussian models with cube-root links. Both produced nearly identical mean curves, but the Gamma model showed slightly better residual behavior. The additive models (TBS and POM) were not retained: TBS requires back-transformation and still showed heteroscedasticity, while POM models exhibited numerical instability and incomplete variance stabilization.

Overall, the Gamma GLM with a cube-root link best matched the Box–Cox variance diagnostics, produced well-behaved residuals, and offered the most coherent and interpretable mean–variance structure. I therefore use this model for the region-specific comparisons.

Extending to regions

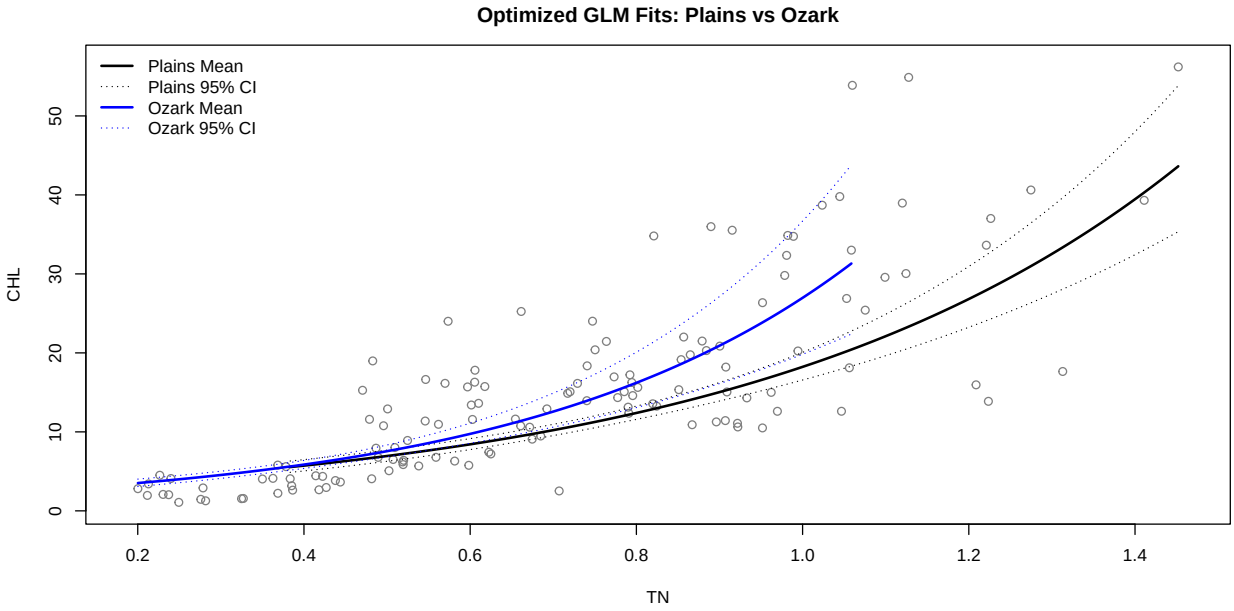
Comparing Two Regions

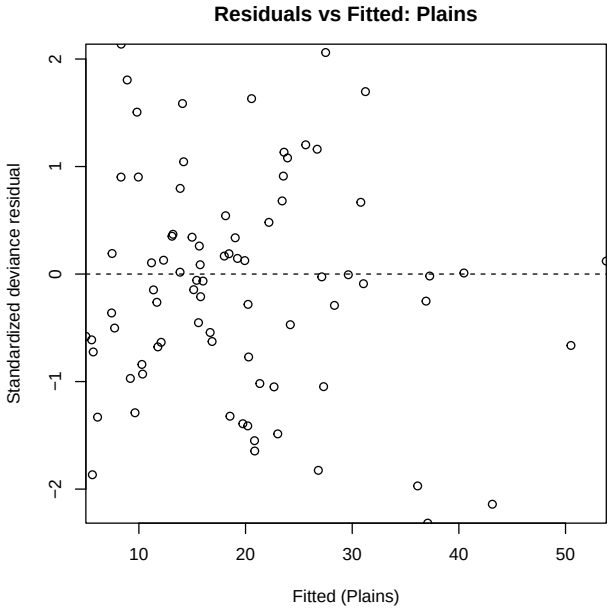
Table 2: Region-specific optimized GLM coefficients with 95% Wald CIs

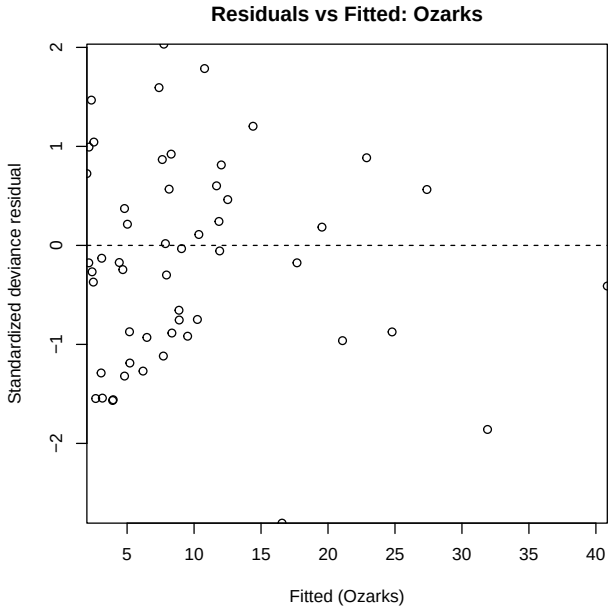
Region	Term	Estimate	SE	LCL	UCL
Plains	Intercept	0.9731	0.1180	0.7417	1.2044
Plains	TN	1.9301	0.1479	1.6402	2.2199
Ozark	Intercept	0.7500	0.1107	0.5331	0.9670
Ozark	TN	2.5447	0.2525	2.0499	3.0395

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







Region	Scaled_Deviance
Plains	79.59165
Ozarks	54.28234

Using the selected Gamma GLM with a cube-root link, I fit the model separately to the Plains and Ozarks. The fitted curves and pointwise 95% confidence bands (Figure X) show substantial overlap at lower TN values ($\approx$ 0.4–0.6), with very similar predicted means. At higher TN ($\approx$ 0.8–1.0), the Ozarks curve becomes noticeably steeper, and the bands begin to separate, indicating higher CHL in the Ozarks at the upper end of the TN range.

The coefficient estimates reflect this pattern: the Ozarks model has a larger slope and a slightly smaller intercept, though the 95% Wald intervals overlap for both parameters. Taken together—steeper slope, upward divergence of fitted curves, and higher predicted CHL near TN = 1.0—the results suggest a stronger TN–CHL response in the Ozarks, even if parameter-wise evidence is not conclusive.

Model diagnostics support using the same GLM form in both regions. Residual-versus-fitted plots show no systematic patterns, standardized deviance residuals are well-behaved, and the scaled deviances are similar relative to sample size. Thus, regional differences can be interpreted as biological rather than artifacts of model fit.

Overall, both regions share the same functional TN–CHL form, but the Ozarks appear to exhibit a stronger response at higher TN levels.

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,

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

with variance functions

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

where ϕ_P and ϕ_O are the dispersion parameters for the Plains and Ozarks models. The target quantity is

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

Using a plug-in Normal approximation to the Gamma distribution,

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

and assuming conditional independence of Y_i and Z_i given x_i, define

$$D_i = Y_i - Z_i.$$

Then

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

and

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

Hence

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

so that

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

where Φ is the standard Normal CDF.

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

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

so the fitted mean at covariate value x is

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

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

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

```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
}

# 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_O.70 <- prob_YgtZ_normal(0.70,
                              mod_gamma_13_plains,
                              mod_gamma_13_ozark)
cat("Estimated probability is:", prob_O.70, "\n")
```

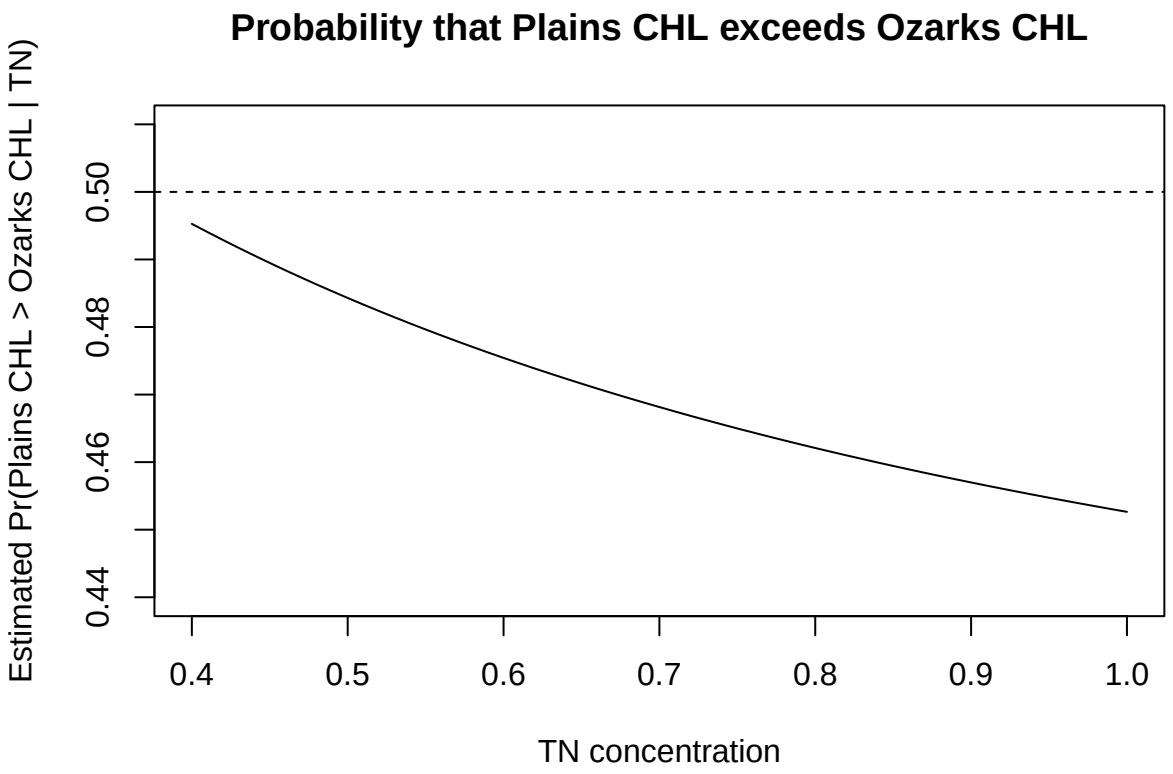
```
## Estimated probability is: 0.4681497
```

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:

```
# 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

In both regions, CHL increases with TN and the variance rises with the mean. The Gamma GLM with a cube-root link adequately captures this mean–variance relationship, with well-behaved residuals and no indication that different model forms are needed. Thus, the TN–CHL relationship has the same functional form in the Plains and Ozarks.

Fitting the model separately by region shows differences in magnitude. For TN $\approx$ 0.4–0.6, fitted curves and 95% bands nearly coincide, and predicted CHL is similar across regions. As TN increases toward 0.8–1.0, the Ozarks curve steepens, predicted CHL in the Ozarks diverges upward, and the bands separate, indicating a stronger TN response in the Ozarks at higher TN.

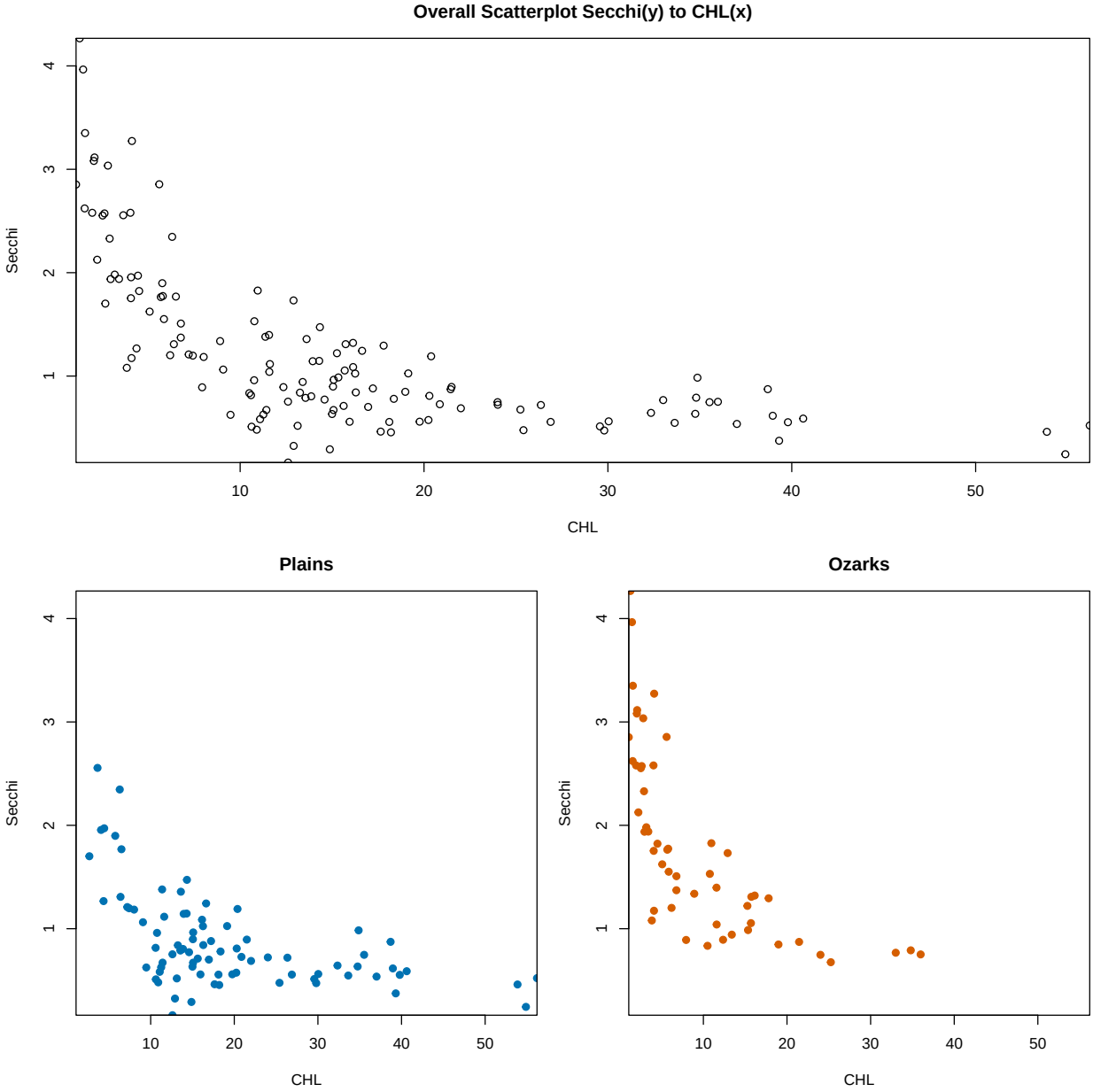
The coefficient estimates are consistent with this picture: both slopes are positive, but the Ozarks slope is larger. Although the 95% Wald intervals overlap and do not provide definitive parameter-wise separation, the combined evidence from curves, bands, and slopes suggests a stronger TN–CHL response in the Ozarks.

The probability assessment summarizes this difference: for moderate TN (0.4–0.7), Pr(Plains CHL $>$ Ozarks CHL | TN) $\approx$ 0.5, but as TN approaches 1.0 this probability falls below 0.5, indicating that Ozarks CHL is more likely to exceed Plains CHL at higher TN.

Overall, the regions share the same basic TN–CHL form, but the Ozarks show a stronger response to increasing TN, especially near the upper end of the observed TN range.

Q2: SECCHI & CHL (Plains vs. Ozarks)

Overall Distribution & Approach



We begin by examining the scatterplot of Secchi depth versus CHL and the marginal distribution of Secchi. Secchi depth is right-skewed, consistent with ecological expectations, and shows a clear negative, potentially nonlinear relationship with CHL, with variance decreasing at larger CHL values. Stratifying by region (Plains vs. Ozarks) reveals the same basic pattern—negative trend, possible curvature, and decreasing variance—while still leaving open the possibility that the CHL–Secchi relationship differs in magnitude between regions.

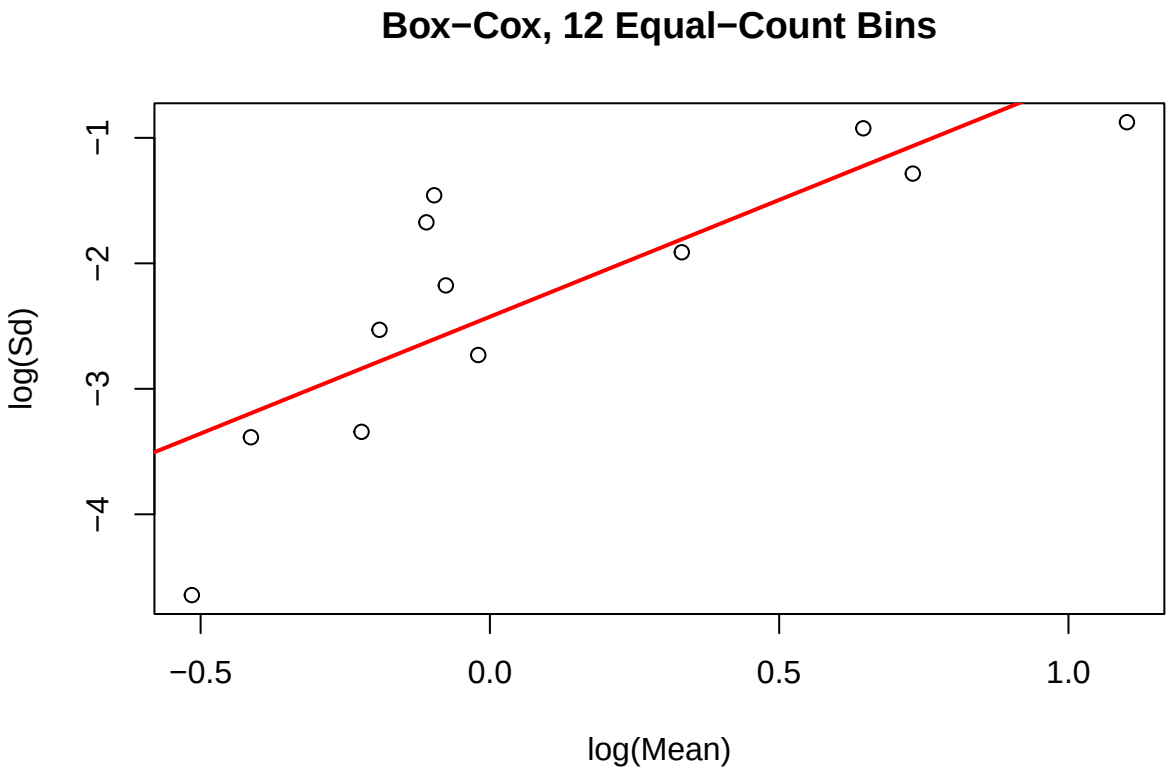
Let Y_i denote the Secchi depth for lake i ($i = 1, \ldots, 134$), with corresponding CHL value x_i. Because each lake appears only once, it is reasonable to treat observations as independent. Specifying a conditional model for $Y_i \mid x_i$ induces a joint distribution for the full collection Y_i.

The modeling strategy mirrors Part I: we first identify an appropriate overall model for Secchi as a function

of CHL, then fit the same model structure separately to the Plains and Ozarks. Under this approach, any differences in intercept, slope, curvature, or dispersion reflect genuine regional differences in the CHL–Secchi relationship rather than artifacts of using different model families or transformations. This ensures that regional comparisons are ecologically interpretable and supports direct comparison of parameters, confidence bands, and other quantities of interest.

Henceforth, for brevity, I omit explicit "as in Part I" references, though the underlying modeling logic is the same.

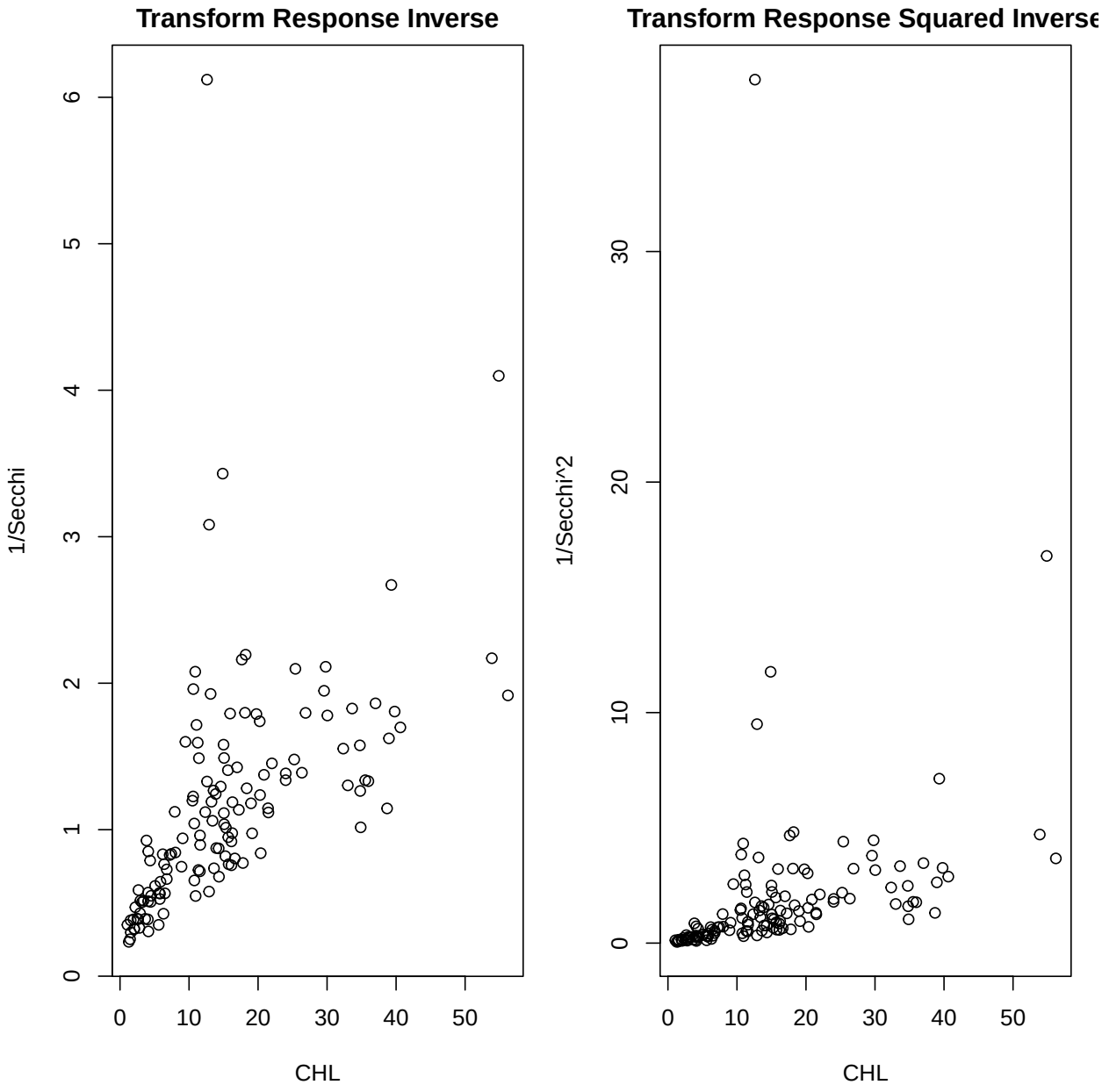
Generalized Linear Model



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

We begin by considering generalized linear models for Secchi as a function of CHL. Using the `boxcoxfctns` routines, I estimated Box–Cox mean–variance slopes; across both equal-count and equal-width binning schemes, the slopes ranged from 1.6 to 2.1. This indicates that $\mathrm{Var}(Y \mid x)$ grows roughly like μ^2–μ^3, which is consistent with Gamma or Inverse Gaussian random components.

With the random component thus narrowed to Gamma or IG, the next step is to select an appropriate link function for modeling the CHL–Secchi mean relationship.

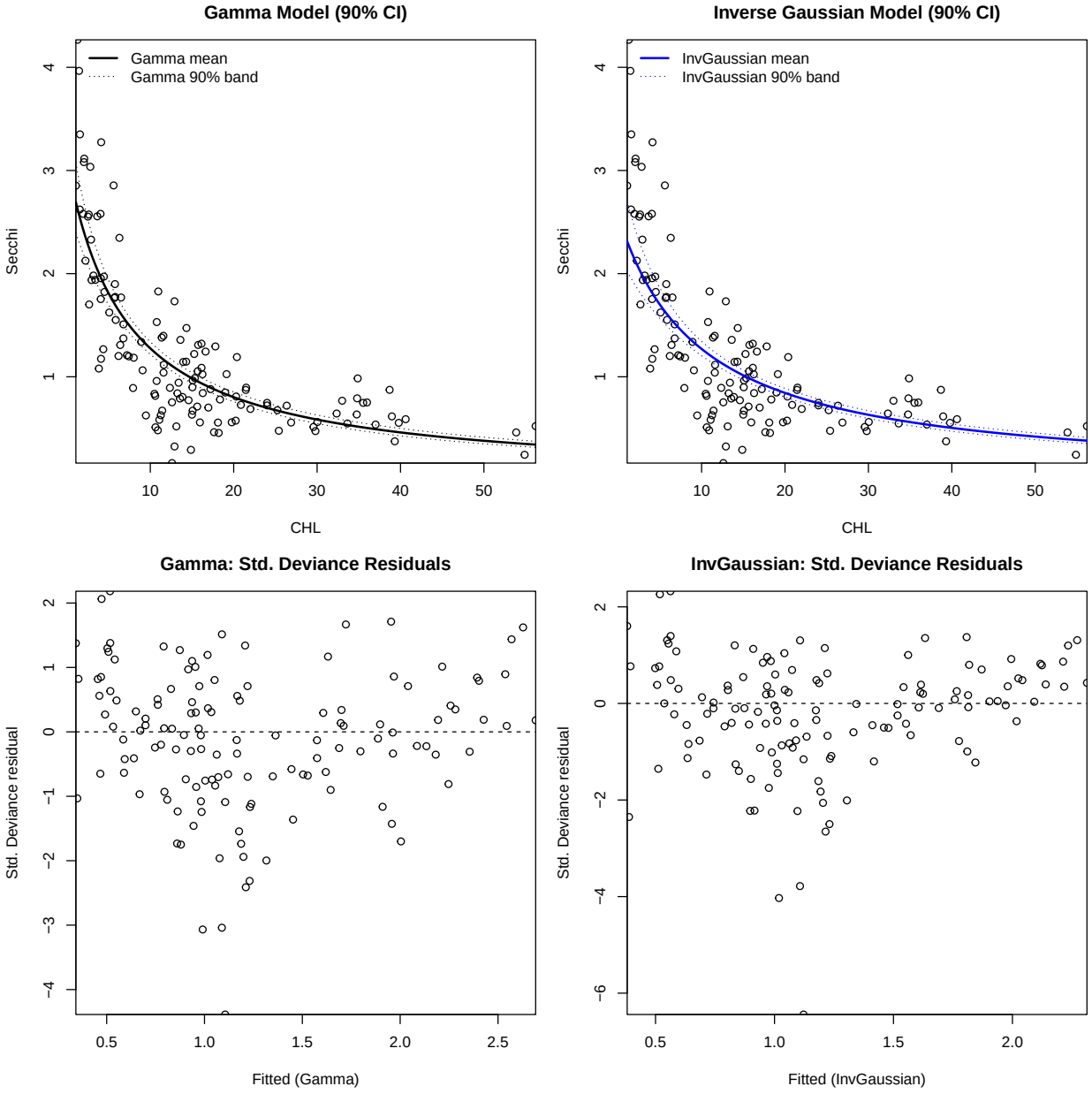


To identify a suitable link function, we seek a transformation of the mean that both linearizes the Secchi–CHL relationship and preserves the expected monotone decreasing pattern (higher CHL $\rightarrow$ lower Secchi). Exploratory transformation plots indicate that an inverse transformation of the mean yields an approximately linear trend, with the square-inverse transformation performing similarly well. These transformations are used only to guide link selection; they do *not* imply transforming Secchi directly.

Combined with the earlier mean–variance assessment, this motivates fitting Gamma and Inverse Gaussian GLMs ($\theta = 2, 3$) paired with inverse and square-inverse links. These candidate models are then compared.

I attempted to fit several GLMs, including the Inverse Gaussian model with its canonical inverse-squared link. Even after generating reasonable starting values from a preliminary `glm()` fit and restricting initial linear predictors to be positive, the `basic.glm` routine failed to converge because iterations quickly produced invalid η values. Due to this numerical instability, only a subset of models could be reliably fitted.

The Gamma GLM with inverse link and the Inverse Gaussian GLM with inverse link both converged cleanly and exhibited stable behavior. These models were therefore carried forward for comparison. I evaluate them using fitted-curve overlays on the scatterplot and by examining their deviance residuals.



The fitted curves from the Gamma and Inverse Gaussian GLMs are visually similar, and the scatterplot alone does not clearly favor one model. Although the Inverse Gaussian model appears to reflect the increased variability at low CHL values (CHL < 10), its fitted mean curve misses most observations in this range, indicating a poorer fit to the central trend.

Residual diagnostics provide stronger separation. The Inverse Gaussian deviance residuals show more pronounced heteroscedasticity and noticeably weaker symmetry compared to the Gamma model. In contrast, the Gamma GLM with an inverse link produces well-centered, more homogeneous residuals. These patterns persist when using studentized residuals, which ultimately drove the model choice.

On this basis, the Gamma GLM with inverse link is selected as the preferred model. With this GLM established, I next assess several additive error models for completeness before identifying a single working

model for the region-specific comparisons.

Other Models Considered – Additive Error Models

Transform Both Sides

We begin by fitting a simple linear regression of Secchi on CHL. Although the linear trend is directionally correct, the studentized residuals show strong heteroscedasticity, motivating consideration of a transform-both-sides (TBS) additive error model.

Several variance–stabilizing transformations were evaluated, with the inverse and cube-root transformations appearing most promising based on exploratory plots. The purpose of these transformations is not to model Secchi on a transformed scale per se, but to obtain an additive-error model with approximately constant variance.

The resulting TBS fits, however, indicate that this approach is not suitable for the Secchi–CHL relationship. Even under the best-performing transformations, the fitted curves fail to capture the clear curvature in the data, and the residuals retain noticeable patterns and heteroscedasticity. The cube-root transformation performs better than the inverse transformation—avoiding instability near very small Secchi values and offering some variance stabilization—but substantial funneling remains, especially at low CHL.

Overall, the TBS additive error models provide a poorer fit than the GLMs and do not adequately account for the nonlinear mean–variance structure. For this reason, they are not competitive as working models, though they offer some guidance for selecting transformations within power-of-the-mean models considered next.

Power of the Mean

We next consider power-of-the-mean (POM) additive error models as an alternative to the GLM. The Box–Cox mean–standard-deviation patterns for Secchi versus CHL suggested

$$\mathrm{Var}(Y \mid x) \propto \mu(x)^{\theta}, \qquad \theta \approx 1.5\text{–}3,$$

which motivates variance weights of the form

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

where larger values of δ attempt stronger variance stabilization. Based on the Box–Cox summaries, we examined POM models with

$$\delta = 1.5, 2, 2.5, 3.$$

Across these specifications, the fitted curves and diagnostics revealed serious deficiencies. For many choices of δ, the fitted Secchi–CHL curve becomes nearly flat—or even slightly increasing—across much of the CHL range, with an abrupt "upturn'' only for CHL > 40. This behavior contradicts the strongly decreasing relationship evident in the raw data and in the GLM fits.

Residual diagnostics reinforce this conclusion: even after weighting by $\mu(x)^{-2\delta}$, substantial heteroscedasticity and structure remain. Large residuals persist at both low and moderate CHL, and the spread does not stabilize across fitted means, indicating that the assumed PoM variance form does not capture the true mean–variance relationship.

Thus, although the Box–Cox diagnostics initially suggested exploring PoM additive models, the resulting fits are not adequate. They fail to reproduce the observed curvature and do not resolve the heteroscedasticity present in the linear and TBS fits. The Gamma GLM with inverse link remains far superior.

Notably, the PoM models above were all formulated directly with CHL as the predictor. Because Secchi depth reflects light penetration through a water column, it may be more natural to re-express the model in terms of a simple geometric argument linking depth to the amount of material per unit volume. This motivates the "cylinder" formulation considered next.

A Possibly Better Approach? Disk & Volume

The motivation for an additive model based on 1/CHL follows from a simple geometric argument. If the visible water column above the Secchi disk is approximated as a **cylinder** of radius r and height h (the Secchi depth), then its volume is

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

Chlorophyll concentration (CHL) is measured as mass per unit volume. If visibility is lost once a roughly fixed mass of chlorophyll is present above the disk, then

$$\text{CHL} \times V \approx \text{constant}, \qquad V \propto \frac{1}{\text{CHL}},$$

and therefore

$$h \propto \frac{1}{\text{CHL}}.$$

This suggests a reciprocal relationship between Secchi depth and CHL, motivating an additive model in the transformed predictor $\phi = 1/\text{CHL}$:

$$Y_i = \beta_0 + \beta_1 \phi_i + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid \phi_i] = 0.$$

Expressed in the original CHL scale,

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

a reciprocal curve that is steep for low CHL and gradually flattens as CHL increases. This parallels the geometric reasoning in the textbook *tree-trunk* and *walleye* examples, where physical arguments yield linearity after transforming the predictor.

Residual diagnostics for this reciprocal model still show heteroscedasticity. Following the cube-root approach used in the fish-weight example, we apply a transform-both-sides model:

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

and fit

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

Back-transforming yields the approximate mean function

$$m(x) \approx \left(\alpha_0 + \alpha_1 x^{-1/3}\right)^3,$$

a smooth, strictly decreasing curve analogous to the cube-root linearization in the fish-weight model.

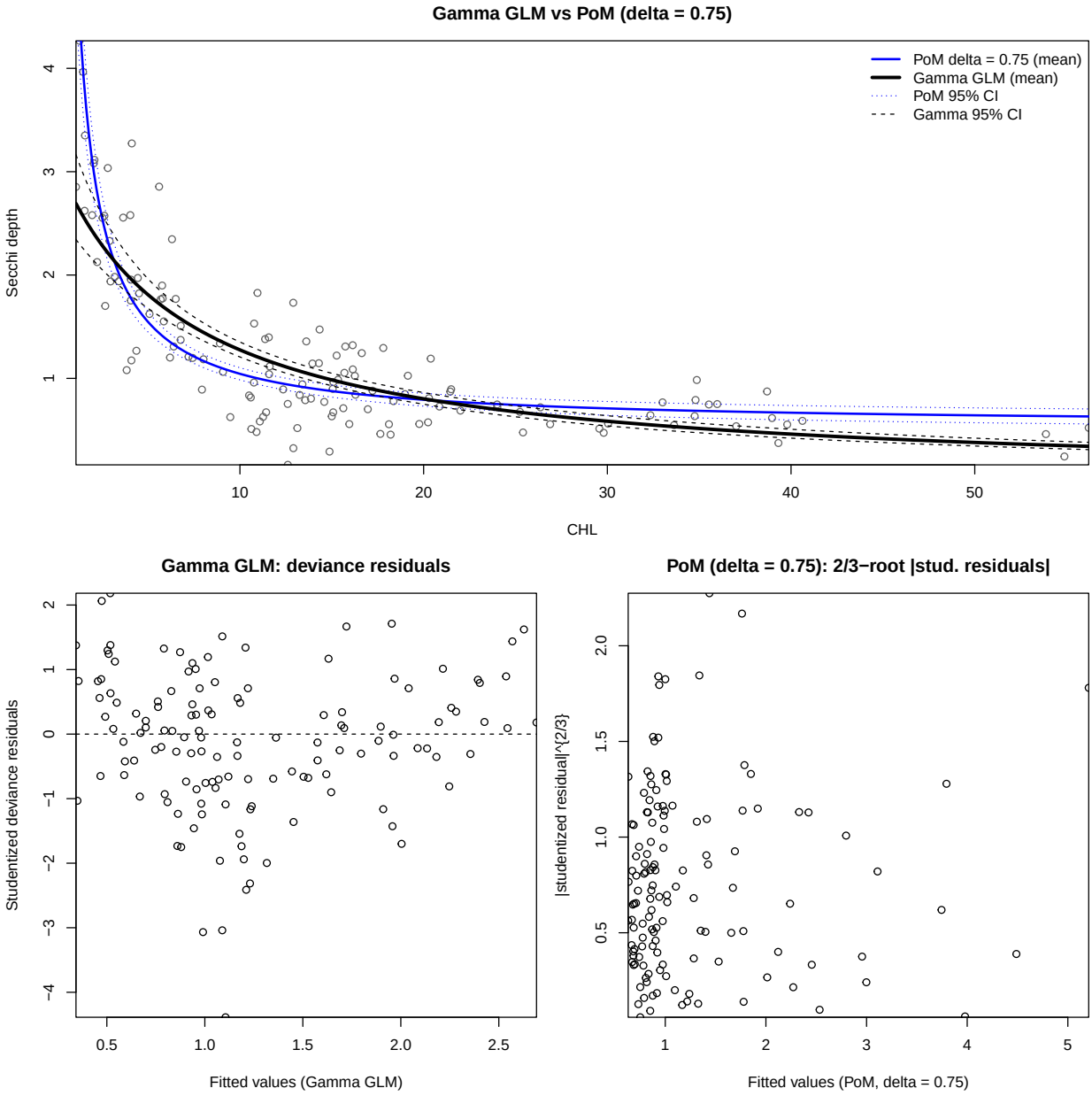
Although both the reciprocal additive model and its cube-root transform-both-sides (TBS) version capture the general decreasing trend, neither provides satisfactory variance stabilization. The power-of-the-mean (PoM) extensions with $\delta = 0.5$ and $\delta = 0.75$ perform the best among the additive models considered, producing smoother reciprocal curves and partially stabilizing the variance. However, even these weighted fits retain visible structure in the residuals and fall short of the diagnostic quality achieved by the Gamma GLM with inverse link.

Consequently, the only additive models that might be viewed as "adequate" are the reciprocal TBS model and the PoM models with $\delta = 0.5$ or $\delta = 0.75$, but these remain supportive rather than competitive alternatives. They help confirm the reciprocal relationship suggested by the geometric argument, but they do not match the interpretability, stability, or overall fit of the selected Gamma GLM.

Between the two PoM specifications, $\delta = 0.75$ provided slightly better performance at high CHL values: the fitted curve tracked the observed decline in Secchi depth more closely in the upper tail than the corresponding $\delta = 0.5$ model.

Note: The "best" additive model is not presented in full here, but its fitted curve and representative residual diagnostics appear in the figures that follow.

Comparing Models



The two models that performed adequately were the Gamma GLM with an inverse link and the power-of-the-mean additive model with $\delta = 0.75$ combined with a cube-root TBS transformation. Although both produce broadly similar decreasing curves, the Gamma GLM remains the preferred model.

First, the Gamma GLM naturally encodes the variance structure

$$\mathrm{Var}(Y \mid x) = \phi, \mu(x)^2,$$

which closely matches the heteroscedasticity observed in Secchi depth. The PoM–TBS model, by contrast, imposes variance stabilization through estimated weights, making it more sensitive to instability in the fitted means and more prone to irregularities at the extremes of CHL.

Second, the fitted curve and confidence intervals from the Gamma GLM are smoother and more stable across the entire CHL range. The PoM–TBS curve tends to flatten at higher CHL values, and its confidence bands

widen noticeably at both ends due to back-transformation and weight variability.

Residual diagnostics also favor the Gamma GLM: the deviance and studentized residuals are centered, approximately symmetric, and exhibit no major patterns. The PoM–TBS model retains curvature and uneven spread, indicating that its variance adjustments remain incomplete.

Finally, the Gamma GLM is simpler to interpret. The mean function

$$\mu(x) = \frac{1}{\beta_0 + \beta_1 x}$$

is directly interpretable on the original scale. The PoM–TBS model requires transforming both sides and then back-transforming

$$m(x) \approx (\alpha_0 + \alpha_1 x^{-1/3})^3,$$

which complicates inference, especially for quantities beyond the mean.

Taken together, these considerations make the Gamma GLM with inverse link the most coherent and reliable modeling choice for the region-specific comparisons that follow.

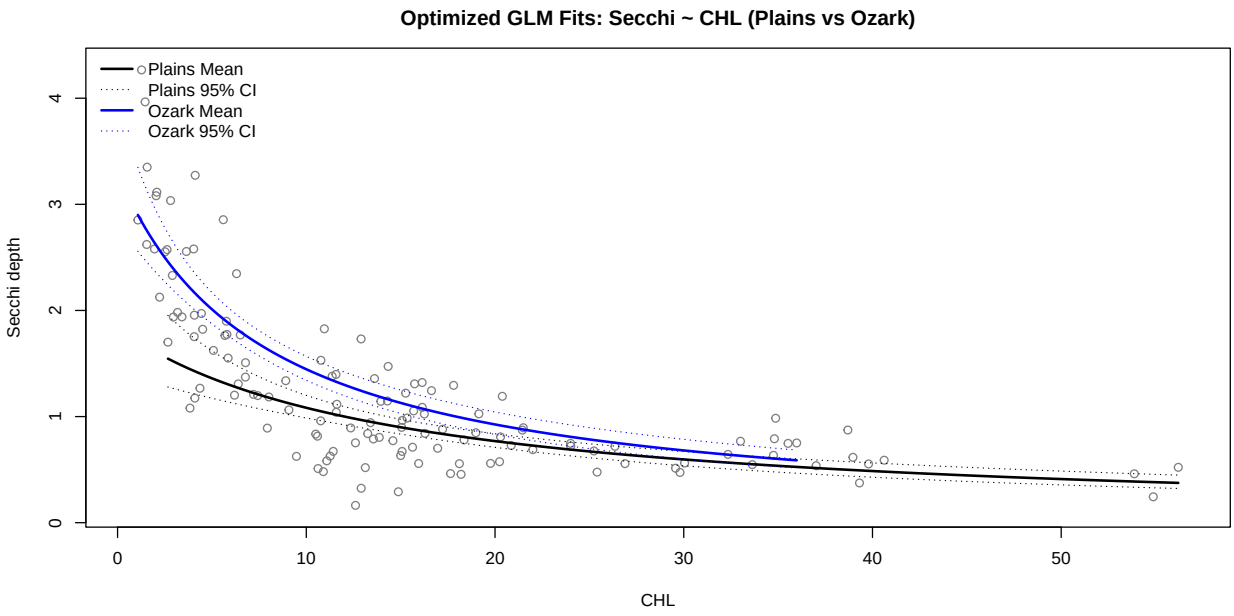
Applying to Regions

Table 6: 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 7: 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



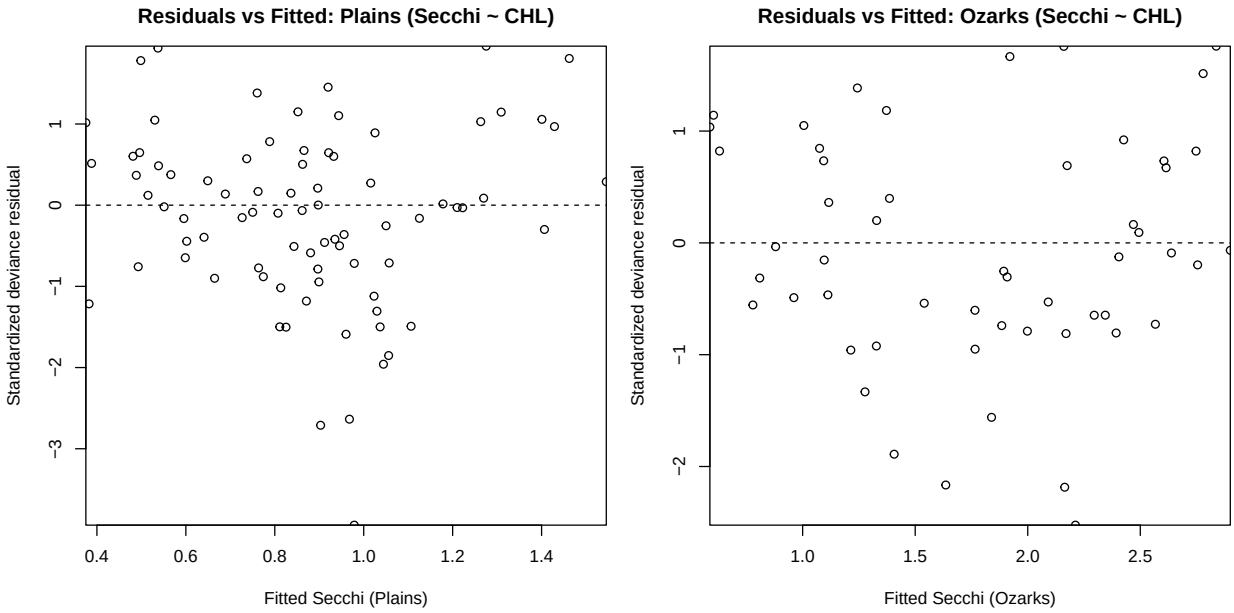


Table 8: Scaled deviance by region for Secchi ~ CHL Gamma GLMs

Region	Scaled_Deviance
Plains	93.27254
Ozarks	53.66591

Interpreting the Model & Results

Across both regions, Secchi depth exhibits the same basic negative and nonlinear relationship with chlorophyll: Secchi decreases as CHL increases, and the Gamma GLM with inverse link provides a coherent 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–CHL curve is essentially identical across regions, indicating that

the underlying light–attenuation mechanism operates similarly throughout Missouri lakes.

The *magnitude* of the response, however, differs in a consistent way. For any fixed CHL value, the Plains exhibit higher Secchi depth than the Ozarks. This vertical separation appears even at low CHL values (Plains Secchi roughly 1.7–2.5 m vs. 1.3–2.0 m in the Ozarks) and becomes more pronounced as CHL increases. By CHL values near 20–30, the fitted curves are clearly distinct and their 95% confidence bands show minimal overlap.

The coefficient estimates reflect the same pattern. The two regions have nearly identical slopes on the inverse–link scale, implying similar rates at which Secchi declines with CHL. The intercepts differ, however: the Ozarks have a smaller fitted intercept, translating to uniformly lower Secchi depth across the entire CHL range. This suggests greater light attenuation in Ozark lakes for the same chlorophyll concentration.

Residual diagnostics indicate that the Gamma GLM fits both regions well, with no evidence of differing functional forms or dispersion patterns. Thus, the observed separation in Secchi depth represents a genuine regional difference rather than a modeling artifact.

In summary, the Plains and Ozarks share the same functional Secchi–CHL relationship, but differ in overall water clarity: Ozark lakes exhibit consistently lower Secchi depth than Plains lakes at matched CHL levels.

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:

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

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:

$$\begin{aligned} \eta_{Pi} &= \beta_{P0} + \beta_{P1} x_{Pi}, & \eta_{Oi} &= \beta_{O0} + \beta_{O1} x_{Oi}, \\ \mu_{Pi} &= \frac{1}{\beta_{P0} + \beta_{P1} x_{Pi}}, & \mu_{Oi} &= \frac{1}{\beta_{O0} + \beta_{O1} x_{Oi}}. \end{aligned}$$

The parameter vector is

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

The full log–likelihood is

$$\begin{aligned} \ell_1(\theta_1) = &\sum_{i=1}^{n_P} \Big[(\alpha - 1)\log y_{Pi} - \alpha y_{Pi}(\beta_{P0} + \beta_{P1} x_{Pi}) + \alpha \log(\beta_{P0} + \beta_{P1} x_{Pi}) + \alpha \log \alpha - \log \Gamma(\alpha)\Big] \\ &+ \sum_{i=1}^{n_O} \Big[(\alpha - 1)\log y_{Oi} - \alpha y_{Oi}(\beta_{O0} + \beta_{O1} x_{Oi}) + \alpha \log(\beta_{O0} + \beta_{O1} x_{Oi}) + \alpha \log \alpha - \log \Gamma(\alpha)\Big] \end{aligned}$$

The LRT compares

$$H_0 : \text{reduced model (no regional difference)} \qquad \text{vs.} \qquad H_1 : \text{full model (regions differ).}$$

Let $\hat{\theta}_0 = (\hat{\beta}_0, \hat{\beta}_1, \hat{\alpha})$ be the MLE under H_0 and $\hat{\theta}_1$ the MLE under H_1. The likelihood ratio statistic is

Table 9: 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

$$\Lambda = -2\big[\ell_0(\hat{\theta}_0) - \ell_1(\hat{\theta}_1)\big]$$

Under H_0,

$$\Lambda \overset{approx}{\sim} \chi^2_2,$$

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

We reject H_0 when

$$\Lambda > \chi^2_{2,\,1-\alpha}$$

Under regularity conditions:

- 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!