

# Supplement for “Accounting for spatial variation in climatic factors predicts spatial variations in mosquito abundance in the desert southwest”

Oshinubi et al.

September 11, 2025

In this supplementary material, we present some additional information, plots, and tables from our analysis. In Section 1, we present some details about outlier ZCTAs. In Section 2, we present hierarchical model inference across clusters we developed, which helped us to estimate unique parameters for each cluster. In Section 3, we present the formulation of our likelihood function using negative binomial probability distribution. In Section 4, we provided the interpretation of our results of the outliers ZCTAs. In Section 5, we present the model evaluation for 2 clusters. In Section 6, we present supplementary tables. In Section 7, we present some additional figures generated from our analysis.

## 1 Outlier ZCTA

During initial data exploration, we noticed that a small number of the 115 ZCTAs contained much higher average mosquito abundances. To quantitatively define ‘outlier’ ZCTA in our data set, we divided the training data set into two temporal windows, the first half and second half of each year, which tend to show unique patterns in mosquito abundances. We then calculated the 95% quantiles of mosquito abundance across the 115 ZCTAs for the first half and second half of the year across the three-year data set. We considered ZCTAs as ‘outliers’ if their average abundance was higher than the 95% abundance quantile in either the first or second half of the year. From this, we identified 6 outlier ZCTAs, 5 of which had high average abundance in both time periods, whereas one outlier had high average abundance in the second half of the year (see Figure S3 in the supplementary file). In fact, across the three years of training data, these 6 ZCTAs account for a total of 42.5% of the observed mosquito abundance.

Importantly, these six outliers also contribute to unique patterns in the mosquito abundance time series not observed across the other 109 ZCTAs (Figure 1b in the main text). The outliers show large spikes in abundance in the spring, which do not seem to correlate with accumulated precipitation. We discuss hypotheses for this phenomenon in more detail later, but briefly, these abundance increases in outlier locations early in the year are likely due to standing water that has accumulated over the winter months, rather than real-time precipitation patterns. Therefore, in our analysis, we evaluated how the inclusion or exclusion of these outliers impacted our model fitting and out-of-sample prediction.

## 2 Hierarchical model inference across clusters

For the three clustered data partitions, we estimated some parameters across the clusters, while some parameters were estimated uniquely to each cluster. Specifically, it seems most biologically reasonable to assume that the effects of temperature and precipitation (i.e., how mosquito population growth rate responds to fluctuations in climate) do not vary spatially. However, the average abundance of mosquitoes does vary across space. Therefore, we allowed the baseline population growth rate,  $v$ , to vary across clusters (i.e., estimating  $v_k$  across  $k$  clusters), while holding all other model parameters constant across the clusters.

To do this efficiently, though, we first had to explore how baseline growth rate  $v_k$  was related to average



mosquito population size, because mathematically,  $v_k$  should scale approximately linearly with average abundance. Therefore, we performed some preliminary MCMC explorations to develop an appropriate hierarchical prior distribution for  $v_k$ . We used k-means clustering to divide the ZCTA into 4 clusters, approximately dividing the county into quadrants, each with quite different average mosquito abundances (see Figure S4 in the supplementary file). Then, we ran four independent MCMC analyses, one on each cluster, to estimate all the model parameters, including  $v$ , using very broad and vague priors. Finally, we conducted a regression of the posterior estimates of  $v$  in each cluster (i.e., 4 values of  $v$ ) against the average mosquito abundance in each respective cluster (see Figure S5 in the supplementary file).

In the hierarchical MCMC, we then used the estimated regression parameters (intercept, slope, residual variance, and their covariance matrix) to construct a multivariate normal prior. Specifically, at each MCMC iteration, we randomly drew from a multivariate normal to generate proposal parameters of the linear model that explains the relationship between  $v$  and average mosquito abundance. We then use these proposed linear model parameters to generate a proposal vector of  $v_k$ , based on the clusters' average mosquito abundances. This way, we could use the same multivariate normal prior regardless of how many clusters we used in the data partition. Additionally, this hierarchical inference method was much more efficient than estimating each  $v_k$  as independent parameters in the Metropolis-Hastings MCMC. All prior distributions for the model parameters are provided in Tables S1–S2 of the supplementary material. To evaluate the parameter estimation results, we include histograms (Supplementary Figures S11–S15) illustrating the posterior distributions of each parameter. Additionally, we present pairs plots (Supplementary Figures S16–S20) to explore relationships and potential correlations between parameters. Finally, a 1:1 plot (Supplementary Figure S21) is used to visually assess the agreement between observed and predicted values, providing insight into the predictive performance of the model.

### 3 Likelihood Formulation

The likelihood function is defined using a negative binomial probability distribution to accommodate over-dispersion in observed mosquito abundance data across multiple spatial clusters. The likelihood function is constructed to evaluate how well the simulated mosquito abundance matches observed values across clusters. The abundance dynamics are modeled by a temperature- and precipitation-driven ordinary differential equation (ODE), and the likelihood aggregates over all clusters.

Let  $X_i^{(k)}$  denote the observed mosquito abundance at time point  $i$  in cluster  $k$ , and let  $\gamma_i^{(k)}$  denote the corresponding model-predicted abundance obtained by numerically integrating the ODE climate-forced mathematical model in Section 2.2 in the main text for each cluster  $k$ . To allow for overdispersion, the observation model assumes that each  $X_i^{(k)}$  follows a negative binomial distribution:

$$X_i^{(k)} \sim \text{NegBin}(\gamma_i^{(k)}, r), \quad (1)$$

where  $r = 1/\text{inv\_disp}$  is the size (shape) parameter and  $\text{inv\_disp}$  is the estimated inverse-dispersion parameter. The negative binomial probability mass function (PMF) for this parameterization is:

$$\Pr(X = j) = \frac{\Gamma(j+r)}{j!\Gamma(r)} \left(\frac{r}{r+\gamma}\right)^r \left(\frac{\gamma}{r+\gamma}\right)^j, \quad (2)$$

for  $j \in \mathbb{N}$ ,  $\gamma > 0$ , and  $r > 0$ .

The log-likelihood contribution from the cluster  $k$  is

$$\mathcal{L}^{(k)} = \sum_{i \in \mathcal{T}_k} \log \left[ \text{NegBin}(X_i^{(k)} \mid \gamma_i^{(k)}, 1/\text{inv\_disp}) \right], \quad (3)$$

where  $\mathcal{T}_k$  is the set of observation times for cluster  $k$ .

The total log-likelihood across all clusters is



$$\mathcal{L}_{\text{total}} = \sum_{k=1}^{n_{\text{cluster}}} \mathcal{L}^{(k)}. \quad (4)$$

In implementation, a small positive value (e.g.,  $10^{-5}$ ) is substituted for any negative or zero  $\gamma_i^{(k)}$  values to ensure numerical stability when evaluating the log-likelihood.

The likelihood is evaluated using parameters grouped as follows:

- **Fixed parameters**  $(\alpha, \phi, T_{\min}, T_{\max}, \text{inv\_disp})$ ,
- **Cluster-specific parameters**  $v_1, \dots, v_{(n_{\text{cluster}})}$ .

The cluster-specific  $v_k$  values are drawn from a linear regression model with mean mosquito abundance as a predictor:

$$v_k = \beta_0 + \beta_1 \cdot A_k + \varepsilon_k, \quad \varepsilon_k \sim \mathcal{N}(0, \sigma^2), \quad (5)$$

where  $A_k$  is the mean observed abundance for cluster  $k$ . The regression parameters  $(\beta_0, \beta_1, \sigma)$  are estimated via maximum likelihood and used to generate  $v_k$  values in the MCMC proposal step.

This formulation allows the model to capture heterogeneous dynamics across spatial clusters while accommodating observational over-dispersion through the negative binomial framework.

## 4 Evaluating outlier zip codes

When we include the six outlier ZCTAs in the training data, our qualitative comparisons between county- and cluster-level models are unchanged, such that the county-level modeling is more parsimonious and the dynamical model fits well to the data in most parts of the year. However, these outlier ZCTAs reveal important areas for model improvement (supplementary Figure S2 and Figure S8 (first row)). For the outlier ZCTAs, the RMSE for the fitted data is 308.7358, and the WAIC value is 2017.8, while the predicted RMSE values for 2013 and 2017 are 298.3209 and 189.0854, respectively. When aggregating two clusters—comprising the county level without the 'outlier' ZCTAs and the 'outlier' ZCTAs—the RMSE for the fitted data is 434.7729. The corresponding RMSE values for the predicted data in 2013 and 2017 are 459.0103 and 484.355, respectively. What becomes obvious is that the model fails to predict large early-year increases in mosquito abundance in these six ZCTAs. These early-year abundances do not seem to be explained by the climate data that we provide the model. This begs the question as to why mosquito abundances reach very large numbers in the early season for these specific ZCTAs. This consistent underperformance during the early year suggests a potential "spring effect" in these locations. A closer examination of these ZCTAs reveals that they are located near agricultural lands, irrigation fields, and river runoff areas, suggesting that unique environmental factors in these regions may influence the mechanisms by which precipitation and/or temperature affect mosquito abundance during the spring season.

## 5 Evaluating 2 cluster (outliers ZCTAs and other ZCTAs)

In this section, we examine the clustering results when we identified two distinct clusters during our initial data exploration, namely, cluster 1 (outlier ZCTAs) and cluster 2 (other ZCTAs). For the 2-cluster subdivision, the predicted RMSE values for Cluster 1 and Cluster 2 are 313.807 and 254.352, respectively, while the WAIC for Cluster 1 and Cluster 2 are 2124.5 and 3722.6, respectively. For the fitted data in 2013, the RMSE values are 291.598 for Cluster 1 and 312.559 for Cluster 2. In 2017, the RMSE values for Cluster 1 and Cluster 2 were 169.330 and 505.083, respectively. In supplementary Figure S10, we present 2 cluster predictions.



## 6 Supplementary Tables

**Table 1:** Parameter estimated value and its prior for one cluster

Parameters	Prior				
$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \sigma \end{pmatrix}$	$\mathcal{N}(\begin{pmatrix} \beta_0 \\ \beta_1 \\ \sigma \end{pmatrix}, \begin{pmatrix} 0.05937776 \\ 0.002153349 \\ 0.09031871 \end{pmatrix}, \begin{pmatrix} 1.262493e^{-04} & -4.534549e^{-07} & -2.767479e^{-09} \\ -4.534549e^{-07} & 1.942469e^{-09} & 1.035168e^{-11} \\ -2.767479e^{-09} & 1.035168e^{-11} & 1.018966e^{-05} \end{pmatrix})$				
$(v_k)_{k=1\dots 9}$	Normal( $v_k, \beta_0 + \beta_1 * A_k, \sigma * \tau$ ), $\tau$ is 1 for $k > 2$ and 20 for $k = 1, 2$				
County without 'outliers' ZCTAs					
	Median	25.0%	97.5%	Prior	Rhat
$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \sigma \end{pmatrix}$	$\begin{pmatrix} 0.02976152 \\ 0.002246532 \\ 0.07624972 \end{pmatrix}$	$\begin{pmatrix} 0.02433612 \\ 0.002226002 \\ 0.07464232 \end{pmatrix}$	$\begin{pmatrix} 0.04552954 \\ 0.002314214 \\ 0.10095124 \end{pmatrix}$	-	$\begin{pmatrix} 1.000188 \\ 1.004348 \\ 1.005444 \end{pmatrix}$
$v_1$	1.719676	1.560024	2.402226	-	1.01624
$Tmin$	18.95258	18.56927	19.89583	Uniform( $Tmin$ , 15,23)	1.010856
$Tmax$	45.21634	44.42262	49.69285	Uniform( $Tmax$ , 40,55)	1.013312
$\alpha$	1.4823936	1.2838298	2.1524761	Uniform( $\alpha$ , 0.25,2.5)	1.019886
$\phi$	1.3716845	1.1376146	2.7305792	Uniform( $\phi$ , 0.9,10)	1.014801
$inv\_disp$	0.3554648	0.3283849	0.4851967	Uniform( $inv\_disp$ , 0.1,0.8)	1.003655
'outliers' ZCTAs					
$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \sigma \end{pmatrix}$	$\begin{pmatrix} 0.03976153 \\ 0.002246533 \\ 0.08624973 \end{pmatrix}$	$\begin{pmatrix} 0.03433613 \\ 0.002226003 \\ 0.08464233 \end{pmatrix}$	$\begin{pmatrix} 0.05552955 \\ 0.002314215 \\ 0.10095125 \end{pmatrix}$	-	$\begin{pmatrix} 1.000336 \\ 1.005444 \\ 1.004348 \end{pmatrix}$
$v_1$	1.675081	1.539098	2.214737	-	1.01692
$Tmin$	21.19682	20.86893	21.86433	Uniform( $Tmin$ , 15,23)	1.019605
$Tmax$	42.85949	42.38078	46.80677	Uniform( $Tmax$ , 40,55)	1.028547
$\alpha$	0.9169440	0.6672233	1.9362353	Uniform( $\alpha$ , 0.25,2.5)	1.040783
$\phi$	1.0858819	0.9859543	2.3321332	Uniform( $\phi$ , 0.9,10)	1.014189
$inv\_disp$	0.7014340	0.6563753	0.7951554	Uniform( $inv\_disp$ , 0.1,0.8)	1.004445



**Table 2:** Parameter estimated value and its prior for more than one cluster

Parameters	2 clusters				
	Median	25.0%	97.5%	Prior	Rhat
$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \sigma \end{pmatrix}$	$\begin{pmatrix} 0.04976155 \\ 0.002246531 \\ 0.09624977 \end{pmatrix}$	$\begin{pmatrix} 0.04433612 \\ 0.002226002 \\ 0.09464232 \end{pmatrix}$	$\begin{pmatrix} 0.06552957 \\ 0.002314210 \\ 0.10095125 \end{pmatrix}$	-	$\begin{pmatrix} 1.003371 \\ 1.005444 \\ 1.004348 \end{pmatrix}$
$v_1$	1.1475184	0.9767011	1.9081644	-	1.320659
$v_2$	1.6285085	1.4167489	2.7619224	-	1.297385
$Tmin$	19.78218	19.50840	20.51363	Uniform( $Tmin$ , 15,23)	1.088686
$Tmax$	44.47620	43.91322	47.81152	Uniform( $Tmax$ , 40,55)	1.093162
$\alpha$	0.9694519	0.7066512	1.8108229	Uniform( $\alpha$ , 0.25,2.5)	1.190052
$\phi$	1.0910817	0.9882769	1.8744850	Uniform( $\phi$ , 0.9,6)	1.05833
$inv\_disp$	0.5408785	0.5129148	0.6304361	Uniform( $inv\_disp$ , 0.1,0.8)	1.001621
5 clusters					
$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \sigma \end{pmatrix}$	$\begin{pmatrix} 0.06341531 \\ 0.002133807 \\ 0.08550928 \end{pmatrix}$	$\begin{pmatrix} 0.05826092 \\ 0.002114197 \\ 0.08413671 \end{pmatrix}$	$\begin{pmatrix} 0.07795327 \\ 0.002191405 \\ 0.08968036 \end{pmatrix}$	-	$\begin{pmatrix} 1.000575 \\ 1.000489 \\ 1.000326 \end{pmatrix}$
$v_1$	0.3014930	0.2818043	0.3627244	-	1.032145
$v_2$	0.4382690	0.4125237	0.5201658	-	1.026701
$v_3$	0.2591353	0.2430725	0.3131752	-	1.035533
$v_4$	0.1311577	0.1214776	0.1587472	-	1.033824
$v_5$	0.1712375	0.1593046	0.2086824	-	1.037363
$Tmin$	16.39452	16.21301	17.36529	Uniform( $Tmin$ , 15,23)	1.109313
$Tmax$	49.57801	48.67387	52.79773	Uniform( $Tmax$ , 40,55)	1.136697
$\alpha$	1.865516	1.748551	2.145747	Uniform( $\alpha$ , 0.25, 2.5)	1.222095
$\phi$	1.2131562	1.1118972	1.5750627	Uniform( $\phi$ , 0.9,6)	1.047742
$inv\_disp$	0.4930550	0.4751286	0.5470297	Uniform( $inv\_disp$ , 0.1,0.8)	1.002397
9 clusters					
$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \sigma \end{pmatrix}$	$\begin{pmatrix} 0.05890547 \\ 0.002162998 \\ 0.08881471 \end{pmatrix}$	$\begin{pmatrix} 0.05377620 \\ 0.002142091 \\ 0.08730433 \end{pmatrix}$	$\begin{pmatrix} 0.07409803 \\ 0.002223244 \\ 0.09321634 \end{pmatrix}$	-	$\begin{pmatrix} 1.000442 \\ 1.000309 \\ 1.000327 \end{pmatrix}$
$v_1$	0.2294917	0.2140709	0.2751476	-	1.071692
$v_2$	0.05781203	0.05320818	0.07176764	-	1.068338
$v_3$	0.06271807	0.05767068	0.07784730	-	1.077563
$v_4$	0.09522864	0.08782413	0.11563258	-	1.078966
$v_5$	0.1962258	0.1825582	0.2371645	-	1.089494
$v_6$	0.3926066	0.3672039	0.4624060	-	1.07284
$v_7$	0.07094355	0.06506929	0.08601582	-	1.096995
$v_8$	0.1296059	0.1199495	0.1573306	-	1.077736
$v_9$	0.09760678	0.09056399	0.11853778	-	1.079821
$Tmin$	12.13756	11.86880	13.32317	Uniform( $Tmin$ , 10.0, 17.0)	1.019512
$Tmax$	54.49941	54.10153	54.97587	Uniform( $Tmax$ , 40.0, 55.0)	1.023635
$\alpha$	2.432134	2.353598	2.496935	Uniform( $\alpha$ , 0.25, 2.5)	1.09014
$\phi$	0.8115820	0.7652920	0.9560820	Uniform( $\phi$ , 0.5, 2.0)	1.016278
$inv\_disp$	0.6950292	0.6776585	0.7510000	Uniform( $inv\_disp$ , 0.1, 0.8)	1.003514



**Table 3:** Model evaluation at cluster level for five cluster

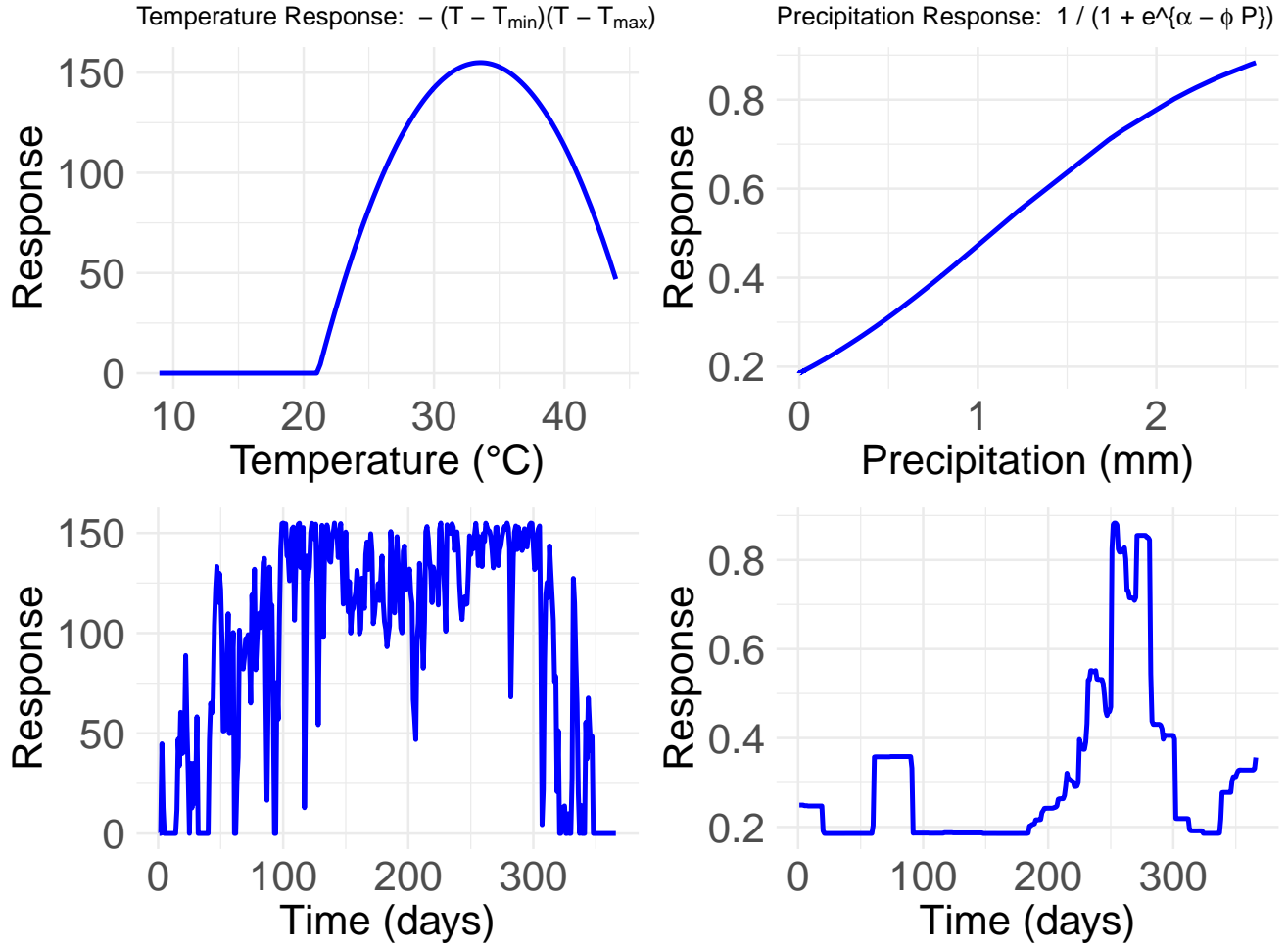
Model	Cluster1 (RMSE)	Cluster2 (RMSE)	Cluster3 (RMSE)	Cluster4 (RMSE)	Cluster5 (RMSE)
5 Cluster (2013)	53.980	116.430	61.840	259.500	66.280
5 Cluster (2014-2016)	80.240	104.150	64.860	35.210	40.620
5 Cluster (2017)	120.970	191.655	142.421	68.479	83.322

**Table 4:** Model evaluation at cluster level for nine cluster

Model / RMSE	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7	Cluster8	Cluster9
9 Cluster (2013)	49.301	17.698	17.199	35.118	46.547	292.185	18.659	48.447	52.276
9 Cluster (2014-2016)	64.342	16.063	22.407	23.297	54.770	96.401	30.620	43.495	29.474
9 Cluster (2017)	69.260	21.112	18.949	62.809	135.524	161.591	60.480	42.767	84.735

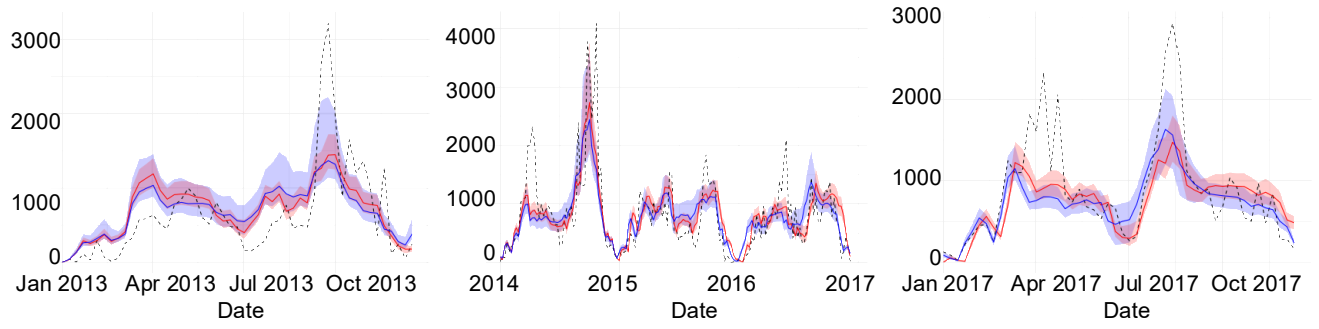


## 7 Supplementary Figures

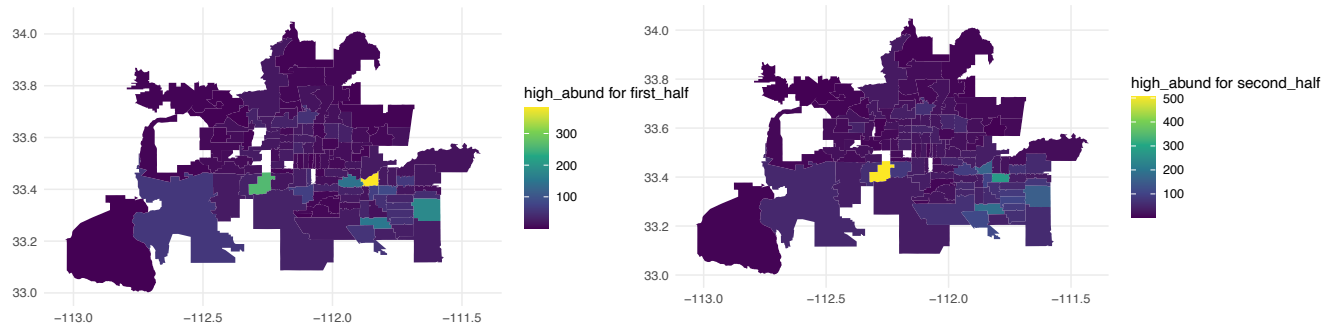


**Figure 1:** Temperature and precipitation response functions used in the model. We used 2014 temperature and precipitation data from PRISM. **(Top left)** Temperature response function defined as  $-(T - T_{\min})(T - T_{\max})$ , illustrating the parabolic relationship between temperature and suitability, with zero response below the minimum threshold  $T_{\min}$  and above the maximum threshold  $T_{\max}$ , and maximum response at intermediate temperatures. **(Top right)** Precipitation response function defined as  $\frac{1}{1 + \exp(\alpha - \phi P)}$ , showing the sigmoidal increase in suitability with increasing precipitation, saturating at higher rainfall. **(Bottom left)** Time series of the temperature response across one year, showing variability in response driven by daily fluctuations in observed temperature. **(Bottom right)** Time series of the precipitation response across the same period, reflecting the nonlinear effects of rainfall pulses and dry periods on response values.



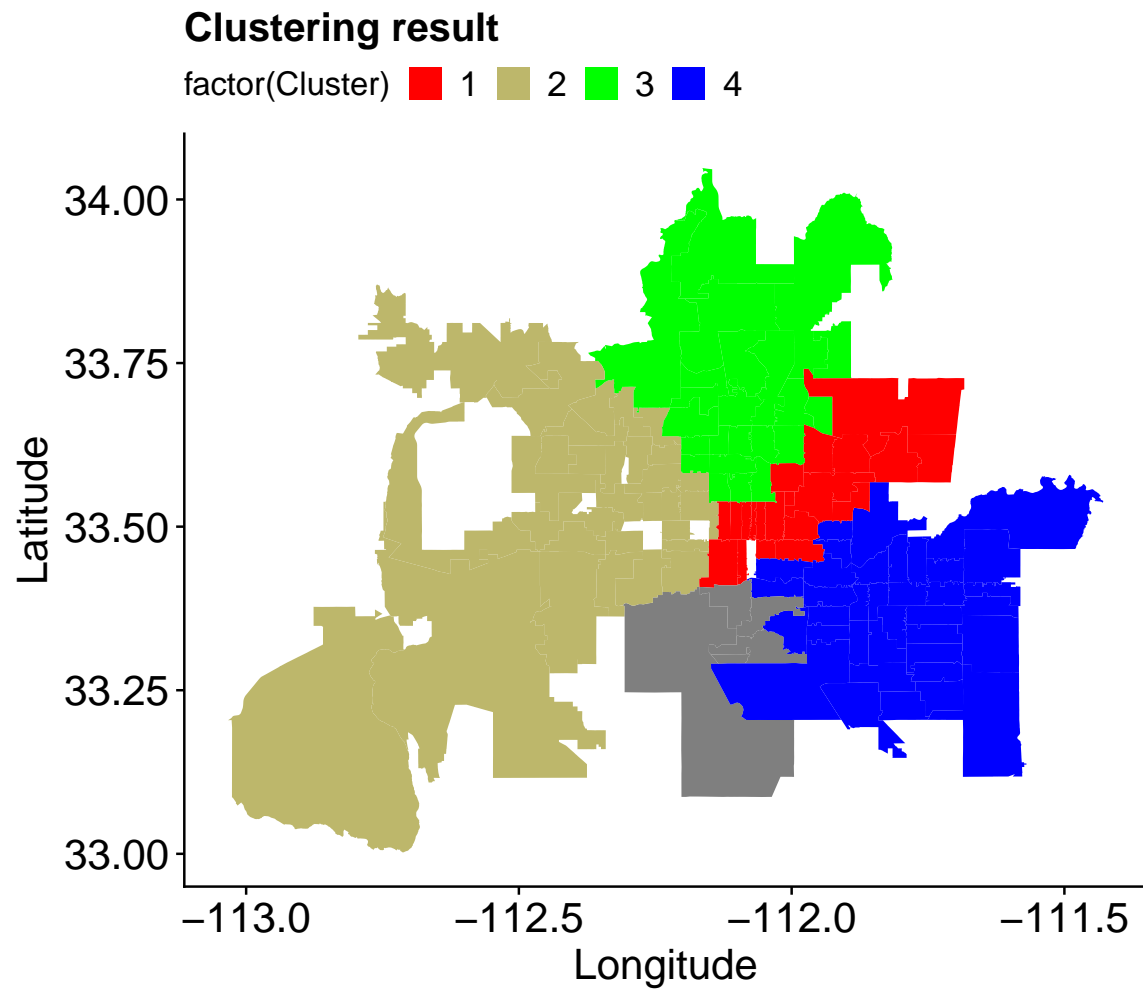


**Figure 2:** Comparison of clustering prediction, real data, and county prediction for 2 clusters. Shaded ribbons in blue indicate 95% credible intervals for county prediction, while shaded ribbons in red indicate 95% credible intervals for cluster sum prediction. The black dotted lines represent observed mosquito abundance data over time.



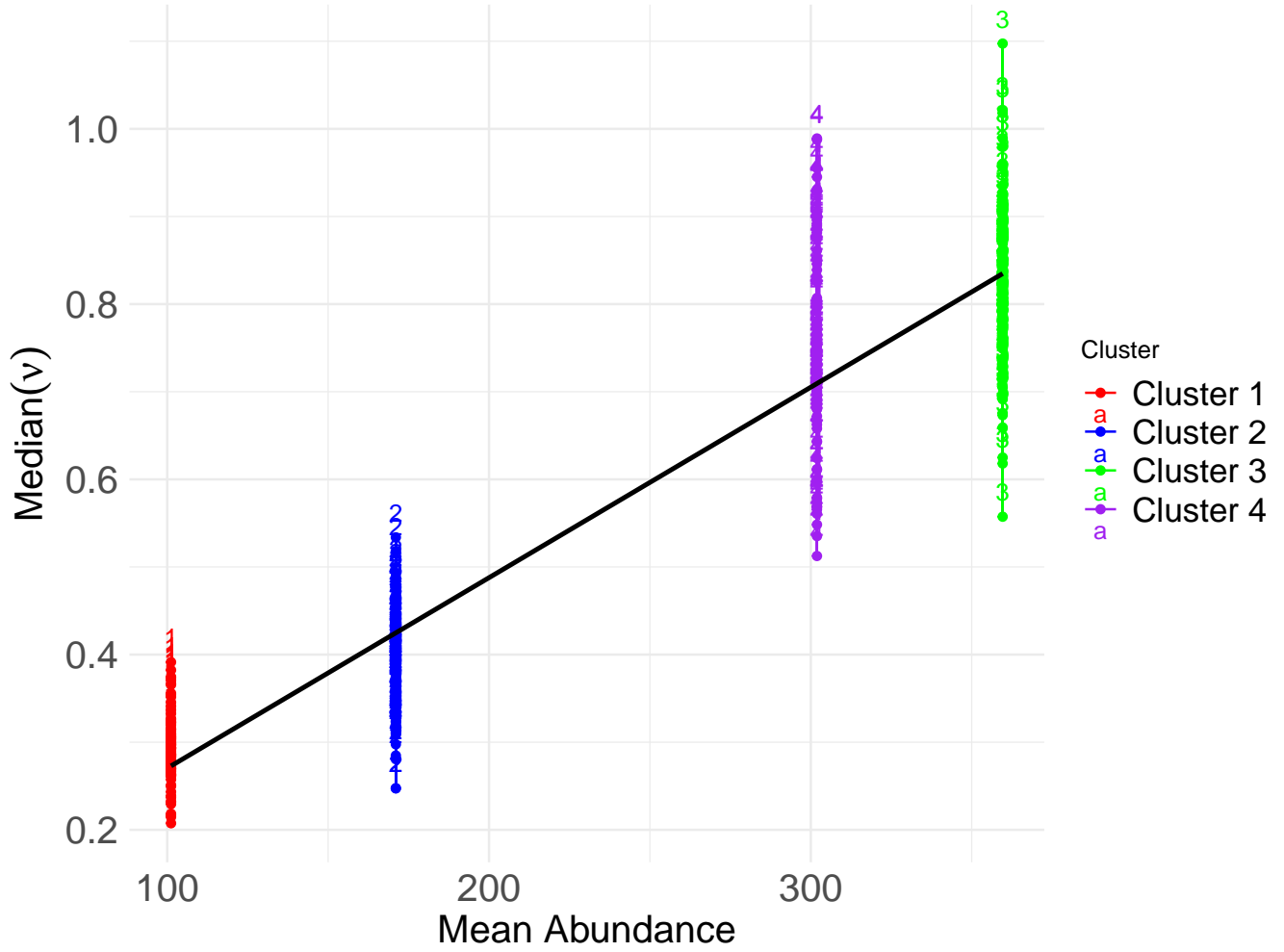
**Figure 3:** Left: A spatio-temporal plot of mosquito abundance per ZCTA for the first half of the year, with high abundance defined as the 97.5<sup>th</sup> percentile of total mosquitoes sampled. Right: A spatio-temporal plot of mosquito abundance per ZCTA for the second half of the year, with high abundance defined as the 97.5<sup>th</sup> percentile of total mosquitoes sampled.





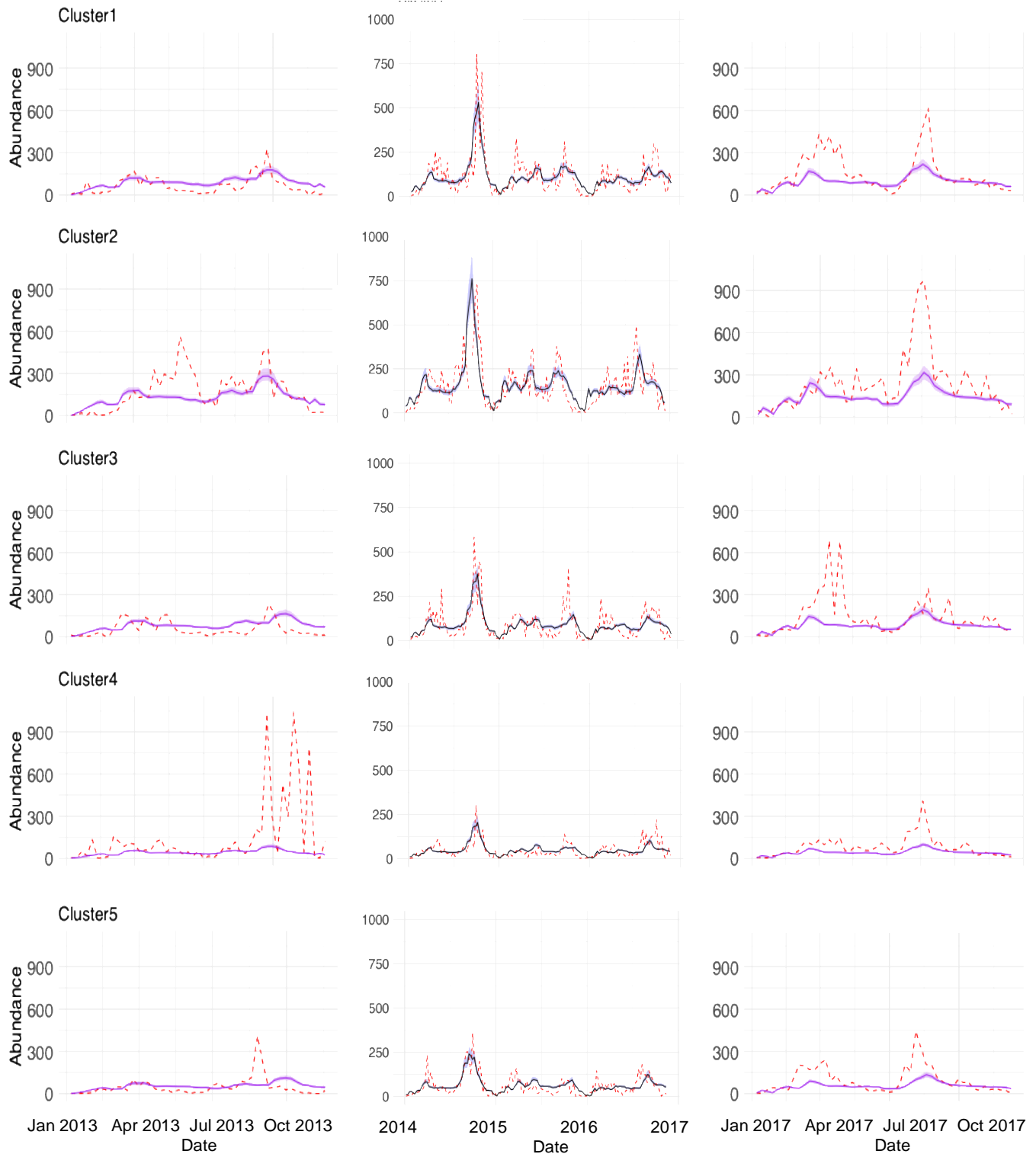
**Figure 4:** 4 cluster result





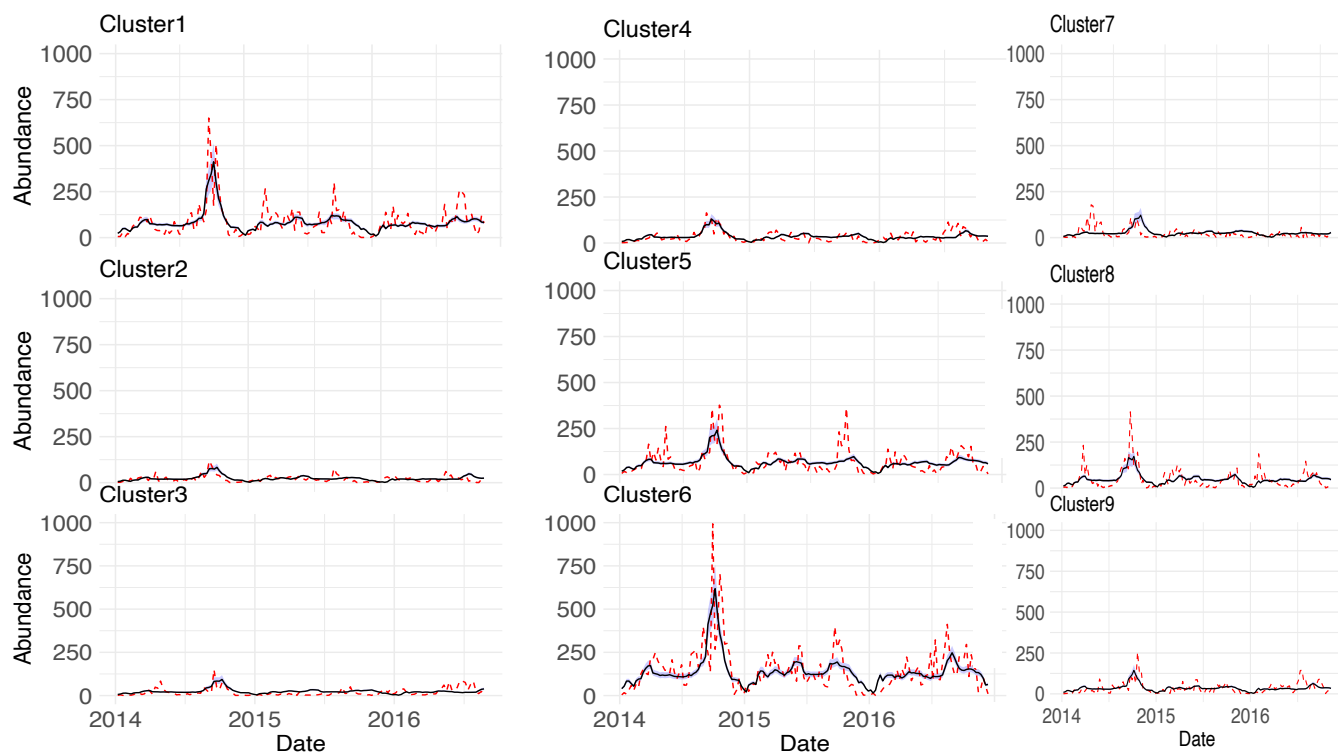
**Figure 5:** Relationship between estimated baseline mosquito population growth rate ( $v_k$ ) and average mosquito abundance across clusters. Each point represents the posterior mean estimate of  $v$  for one of the four spatial clusters obtained from independent MCMC analyses using vague priors. A linear regression line is shown, summarizing the trend used to define a hierarchical prior for  $v_k$  in the main hierarchical MCMC analysis. This relationship supports the assumption that  $v_k$  scales approximately linearly with average mosquito abundance across clusters.





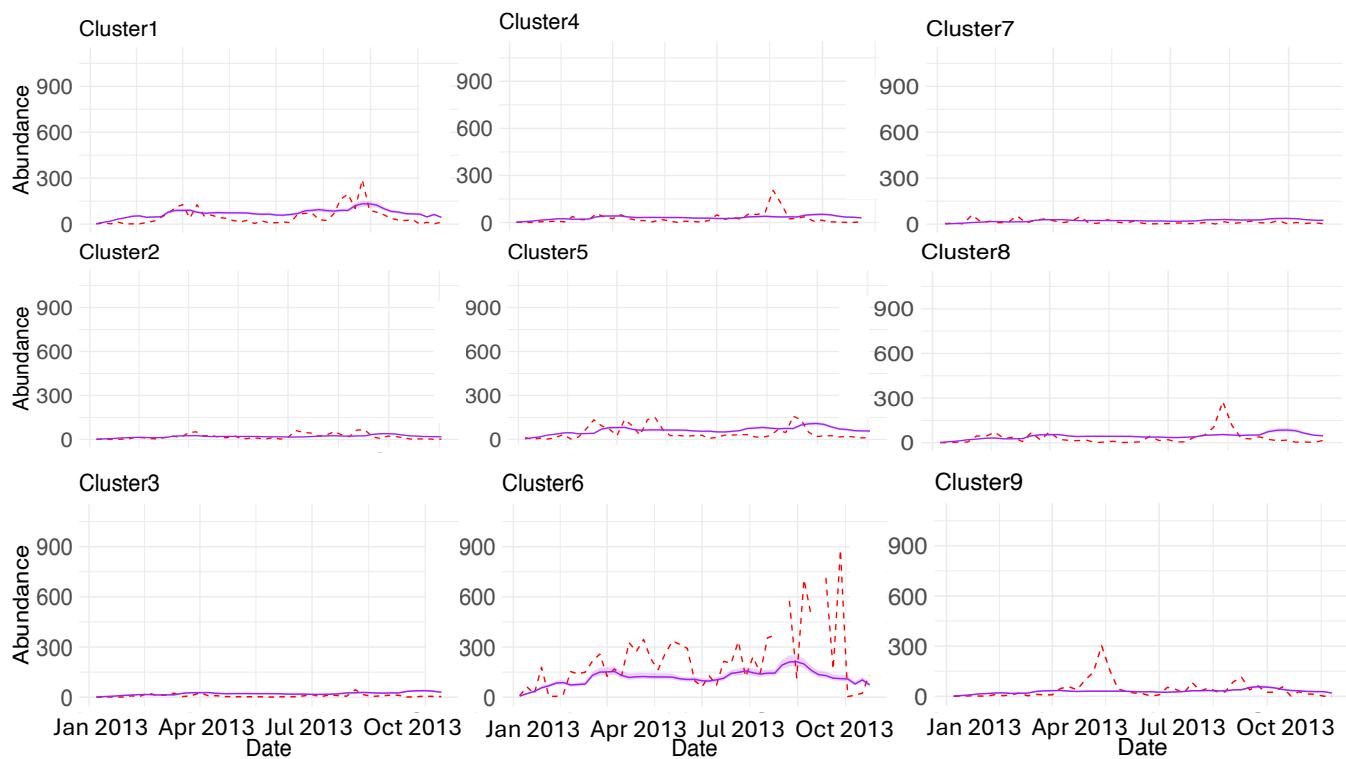
**Figure 6:** 5 cluster prediction. Shaded ribbons indicate 95% credible intervals from the county fit. Red dotted lines are the real data, while the black line is the model fit for within-sample fits, and the orange line is for out-of-sample fits.





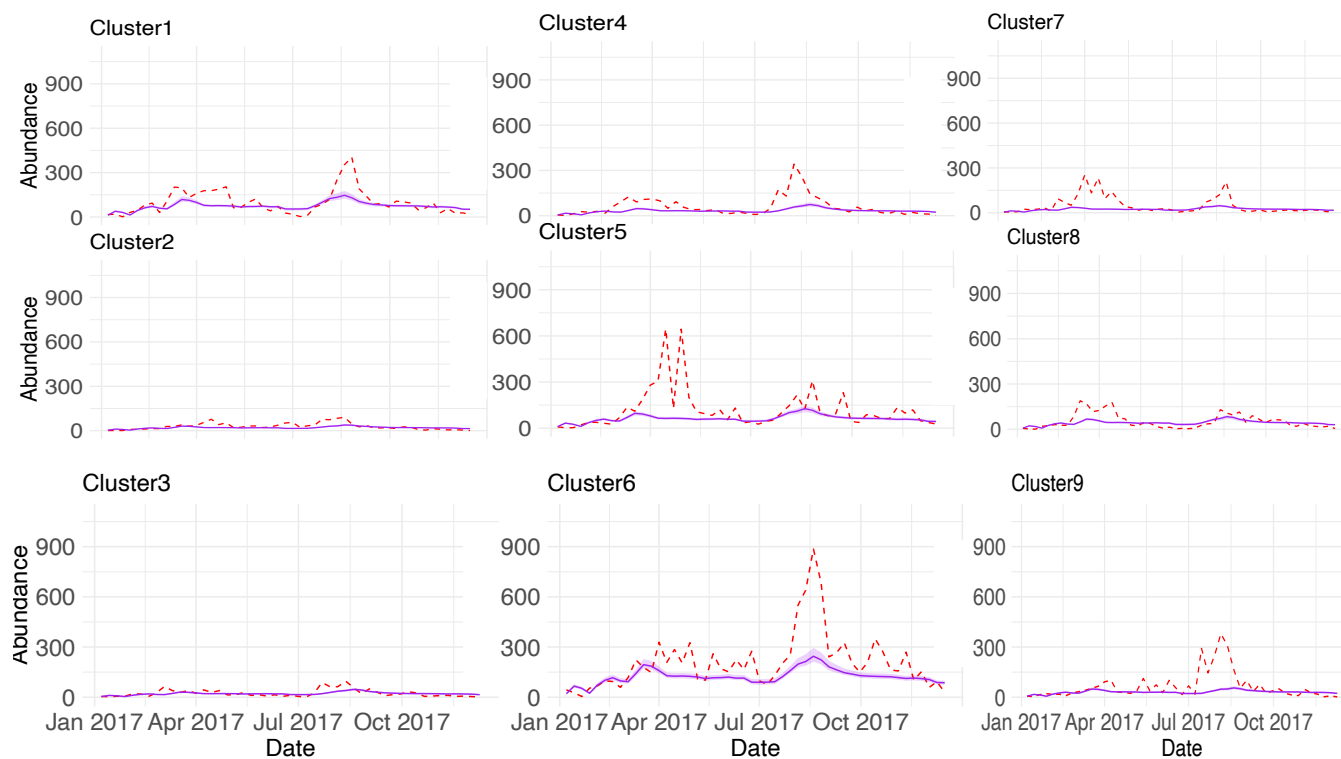
**Figure 7:** 9 cluster prediction. Shaded ribbons indicate 95% credible intervals from the county fit. Red dotted lines are the real data, while the black line is the model fit for within-sample fits.





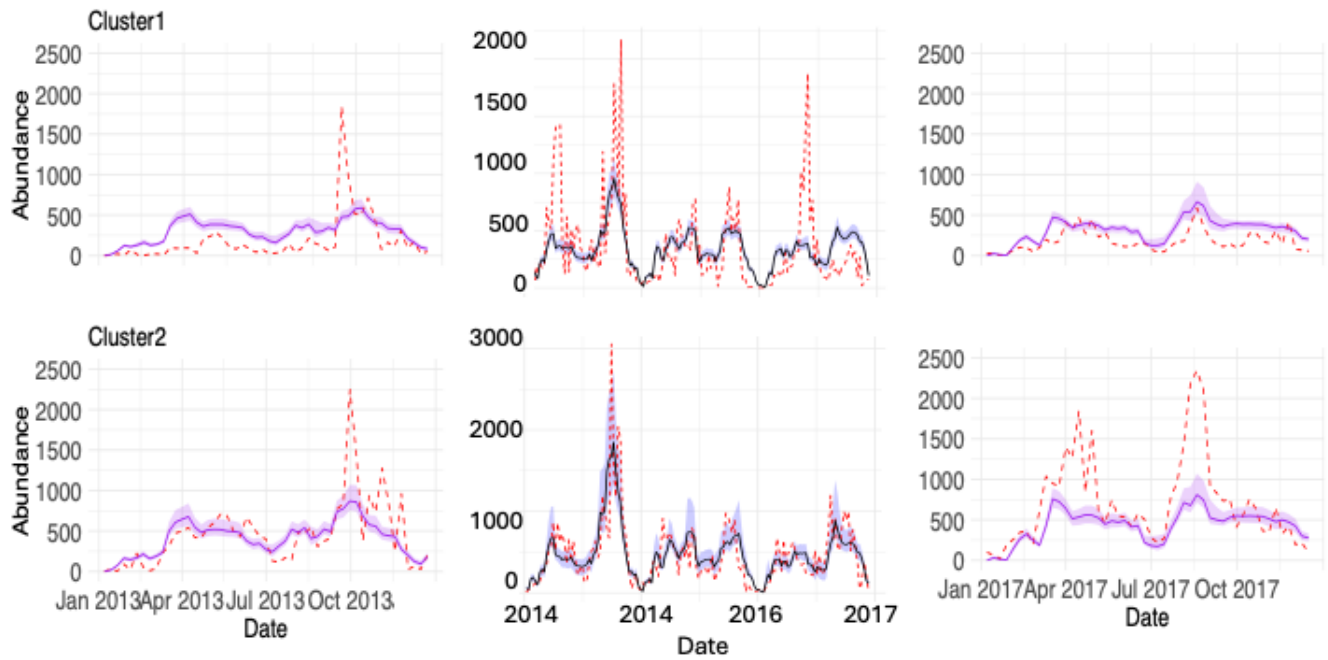
**Figure 8:** 9 cluster 2013 prediction. Shaded ribbons indicate 95% credible intervals from the county fit. Red dotted lines are the real data, while the orange line is for out-of-sample fits for 2013.





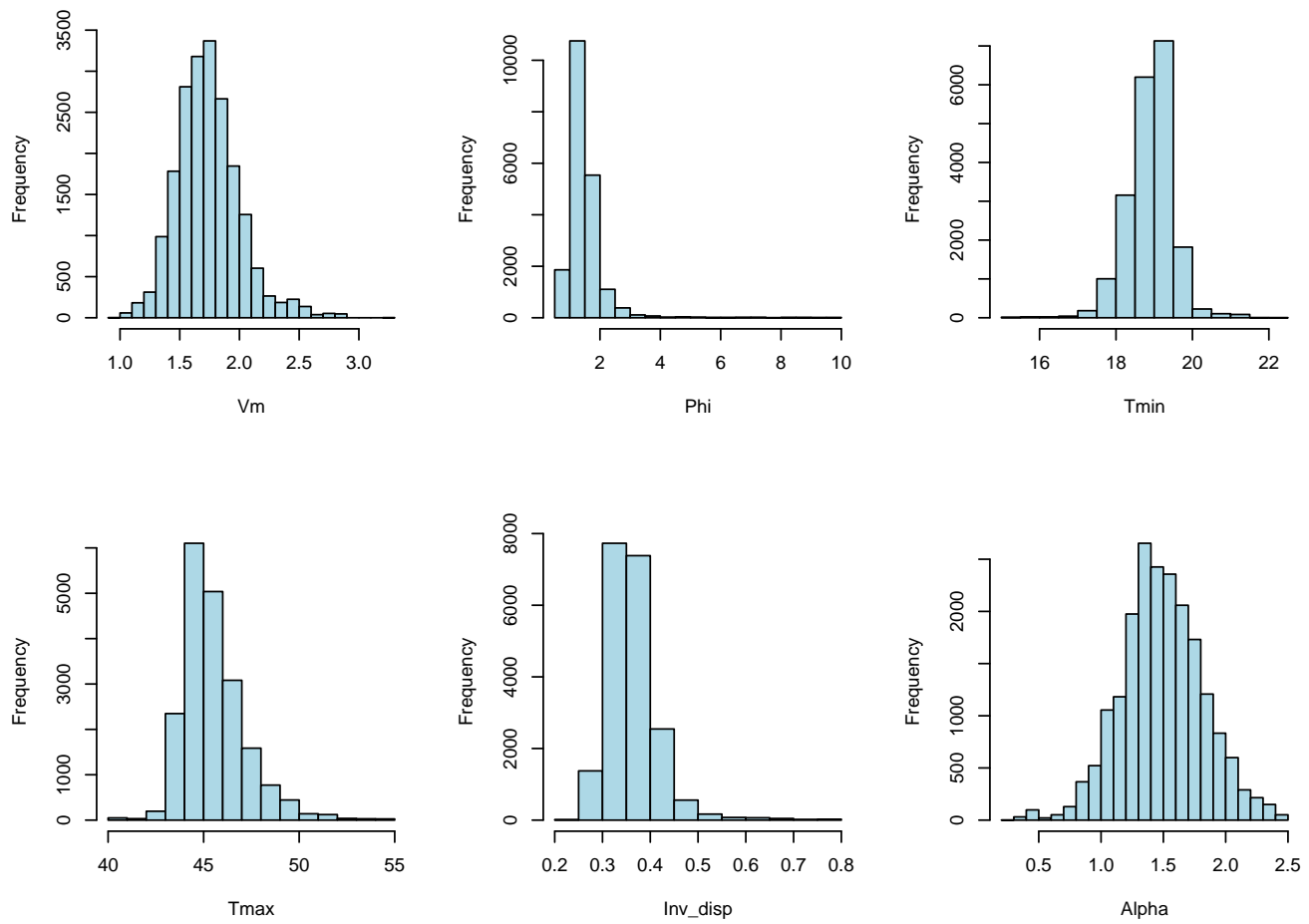
**Figure 9:** 9 cluster 2017 prediction. Shaded ribbons indicate 95% credible intervals from the county fit. Red dotted lines are the real data, while the orange line is for out-of-sample fits for 2017.





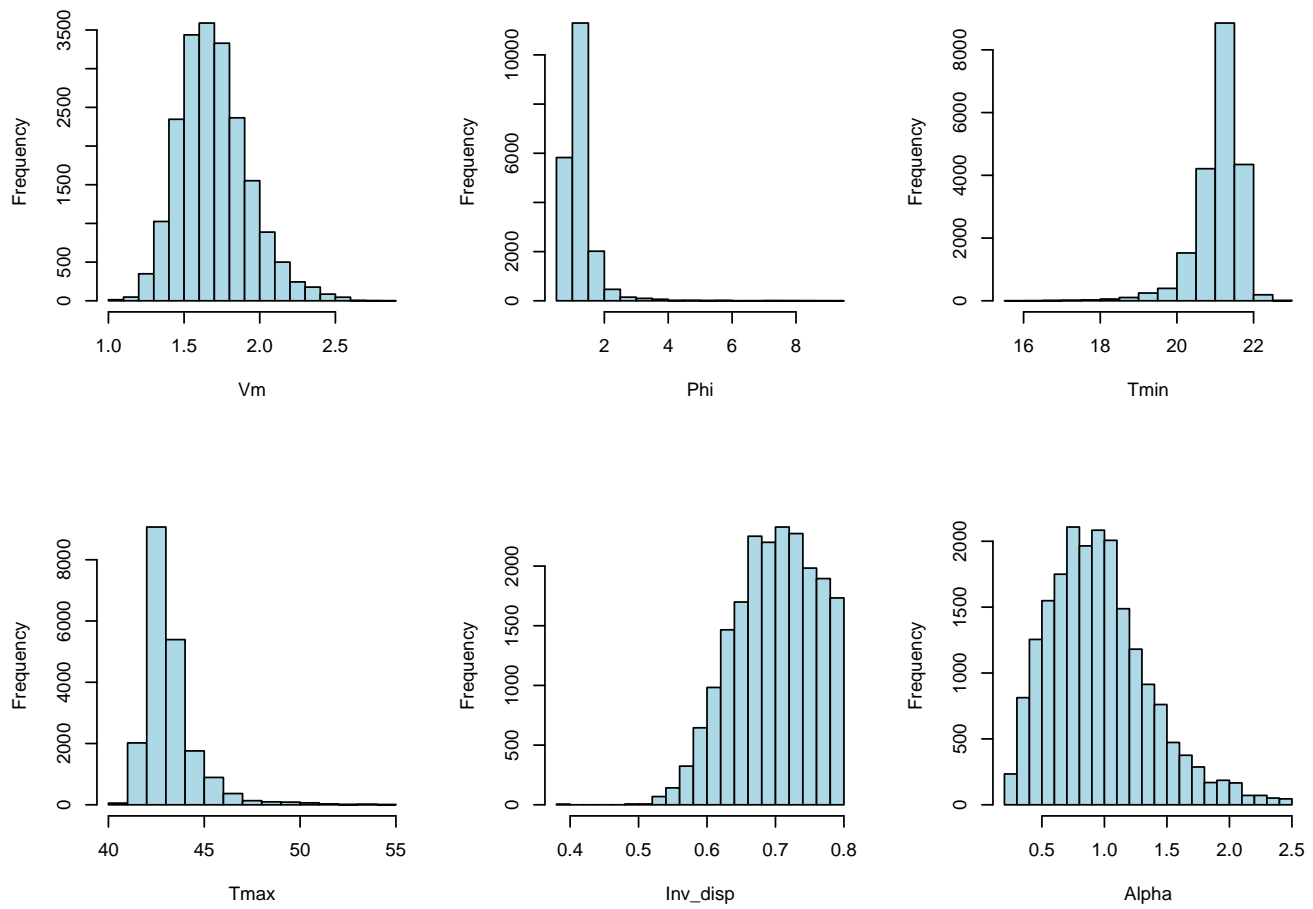
**Figure 10:** 2 cluster prediction. Shaded ribbons indicate 95% credible intervals from the county fit. Red dotted lines are the real data, while the black line is the model fit for within-sample fits, and the orange line is for out-of-sample fits.





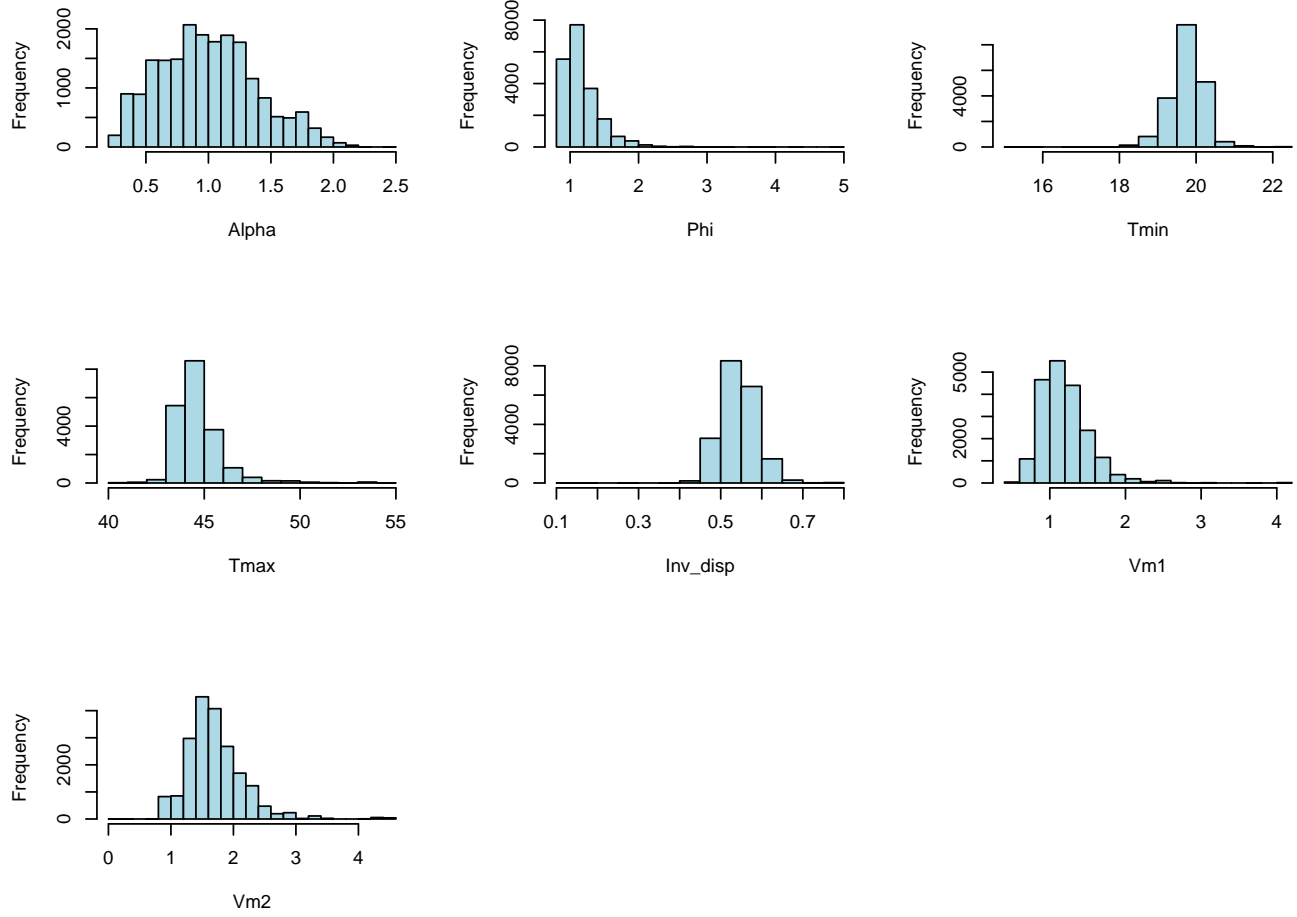
**Figure 11:** Histograms of marginal posterior draws of parameters in the county-level model-fitting without outlier ZTCA.





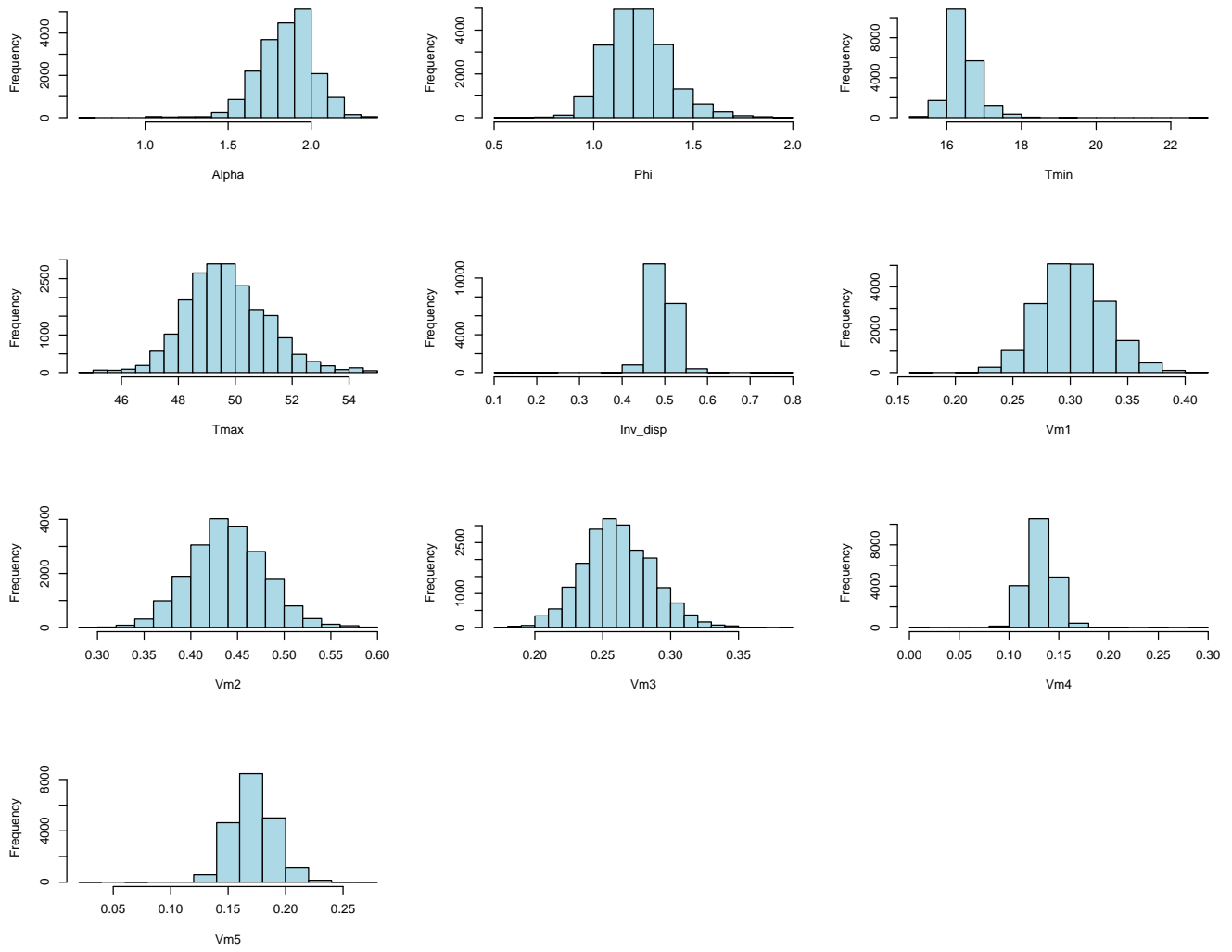
**Figure 12:** Histograms of marginal posterior draws of parameters for outlier ZTCA model fitting.





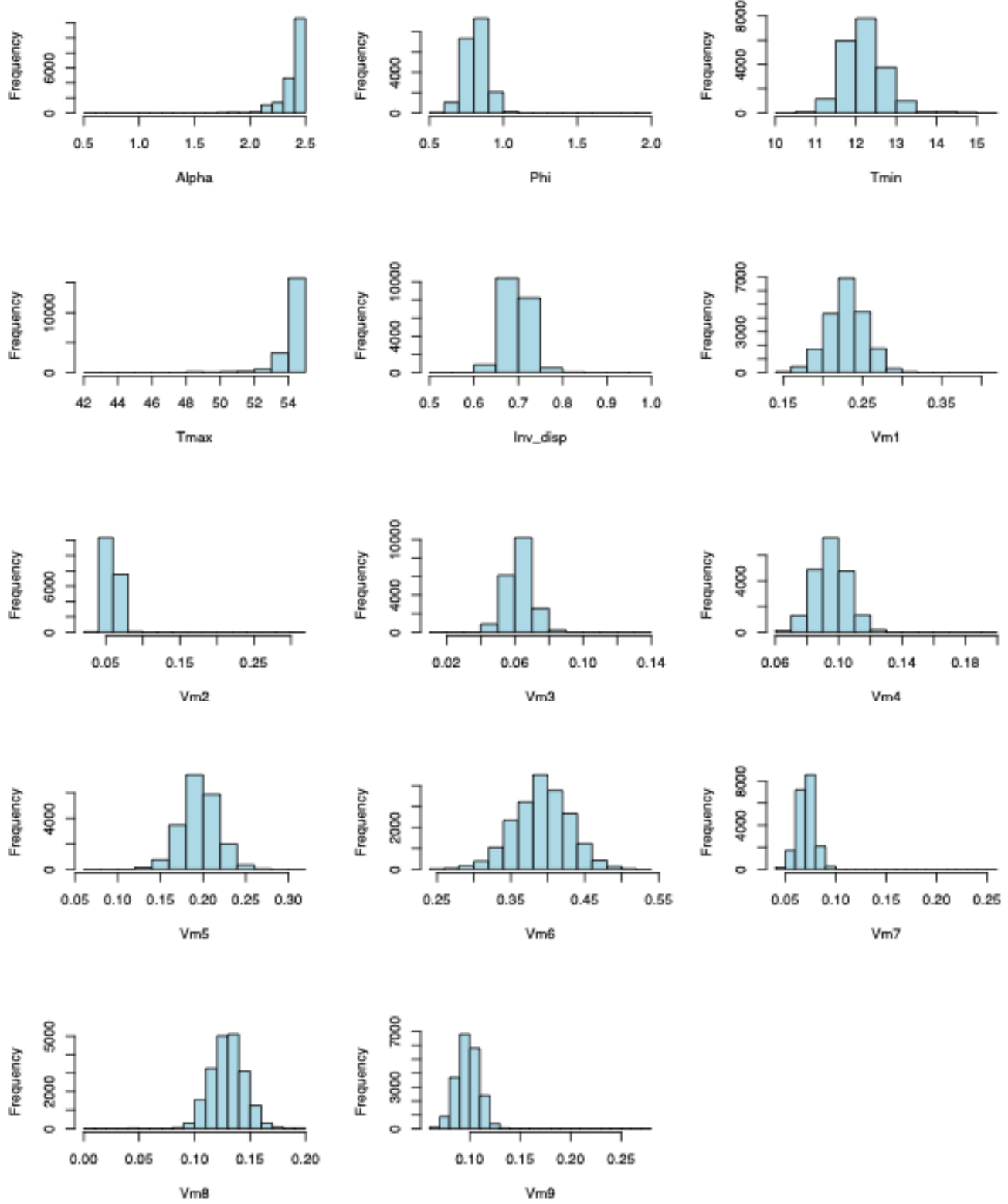
**Figure 13:** Histogram of marginal posterior draws of parameters for 2-cluster model fitting.





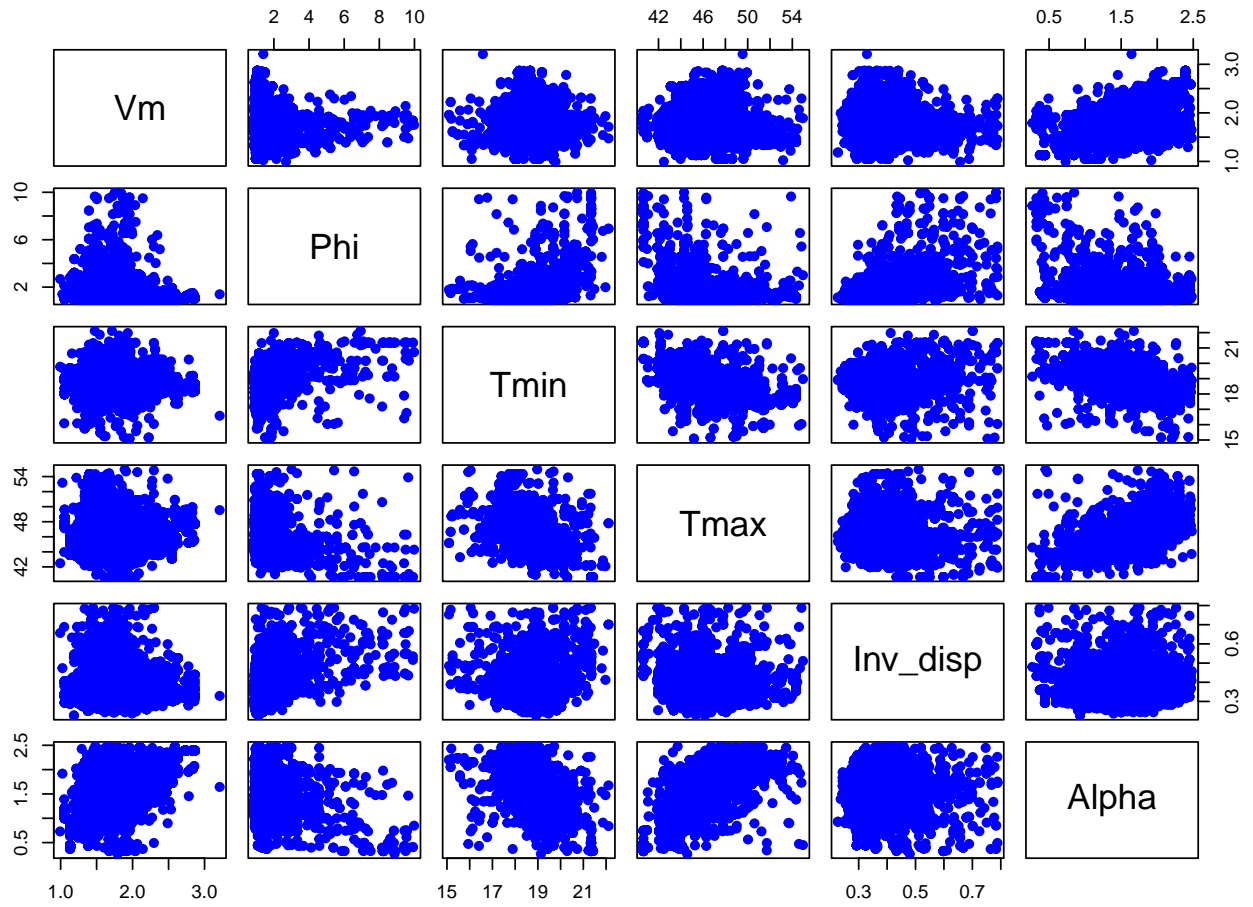
**Figure 14:** Histogram of marginal posterior draws of parameters for 5-cluster model fitting.





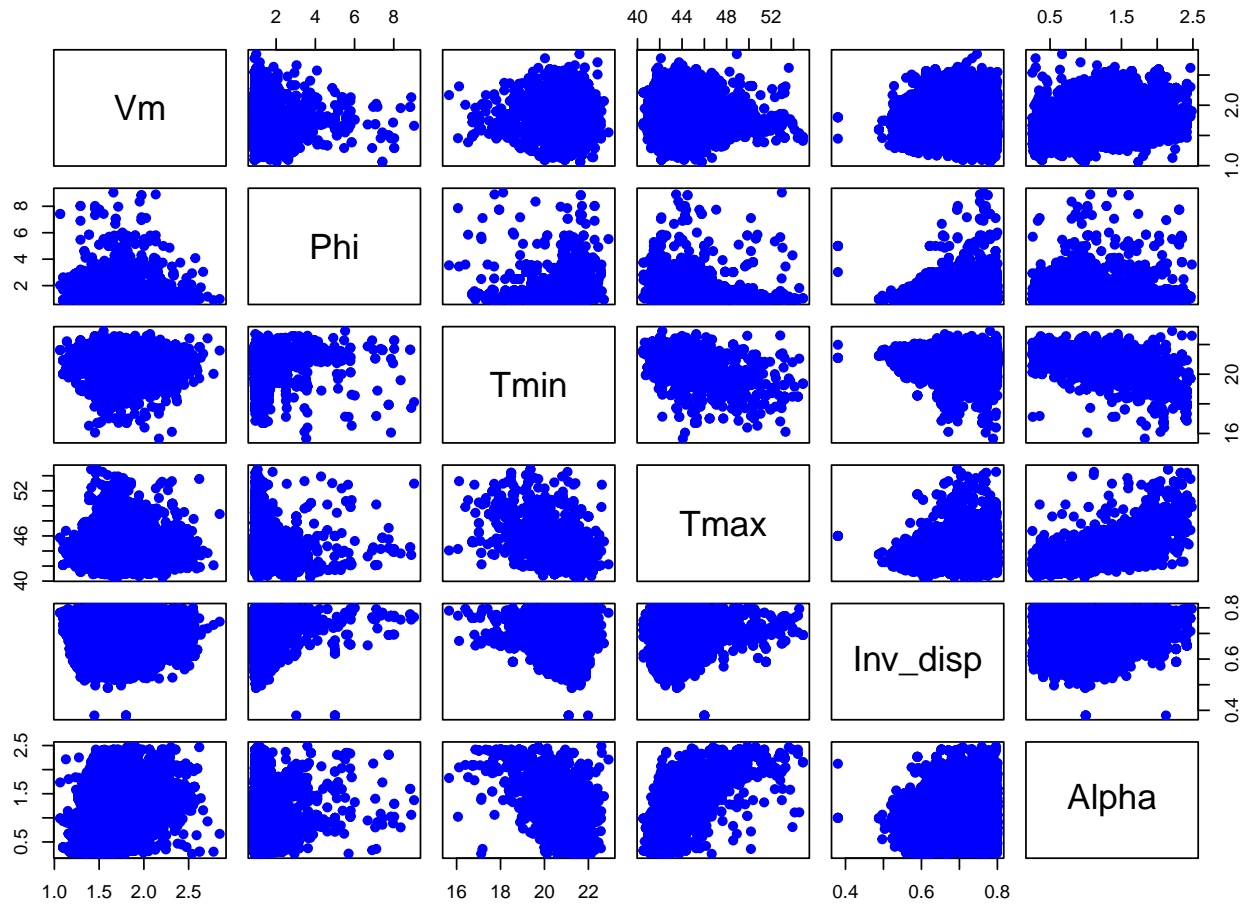
**Figure 15:** Histogram of marginal posterior draws of parameters for 9-cluster model fitting.





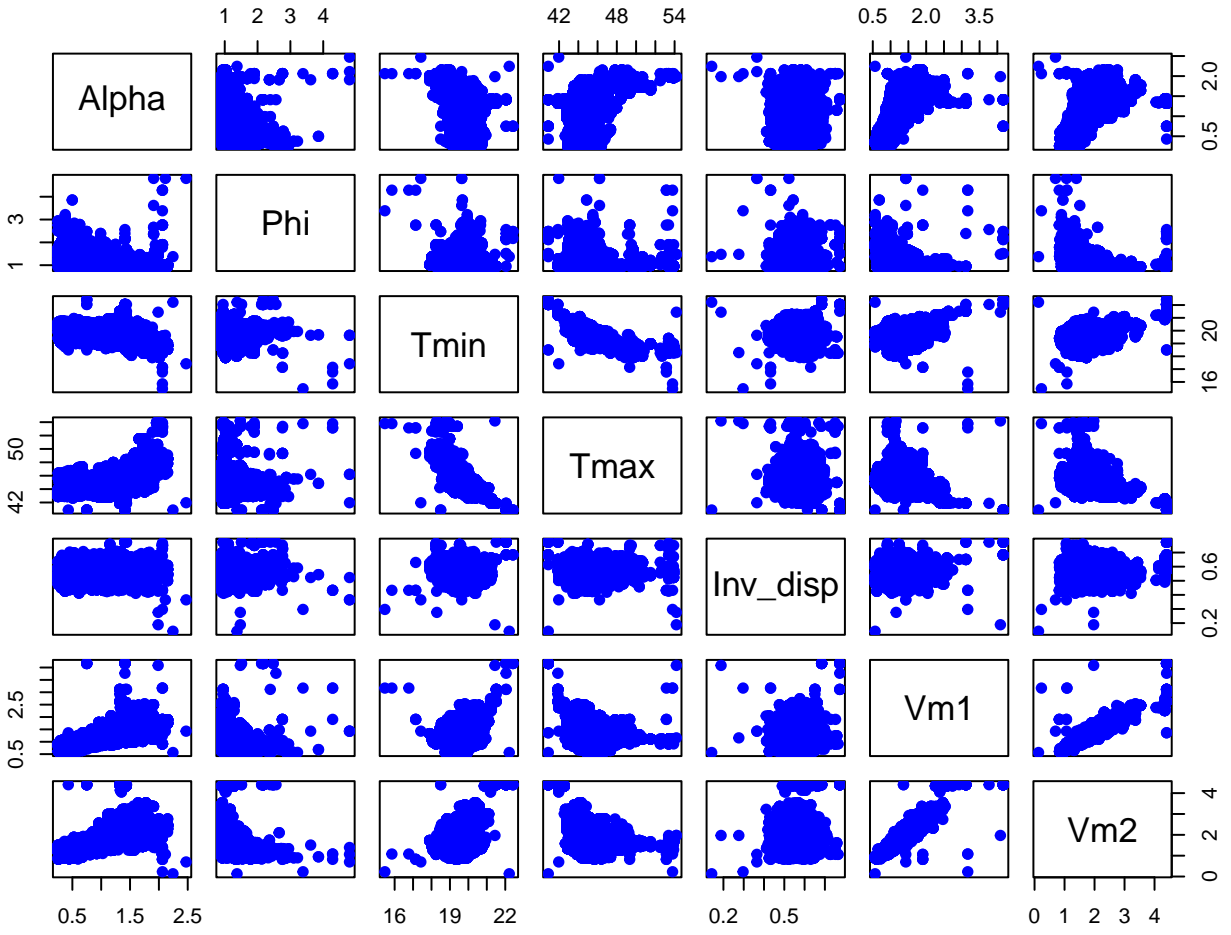
**Figure 16:** Pairs plot displaying joint posterior draws of parameters from the county-level model without outliers.





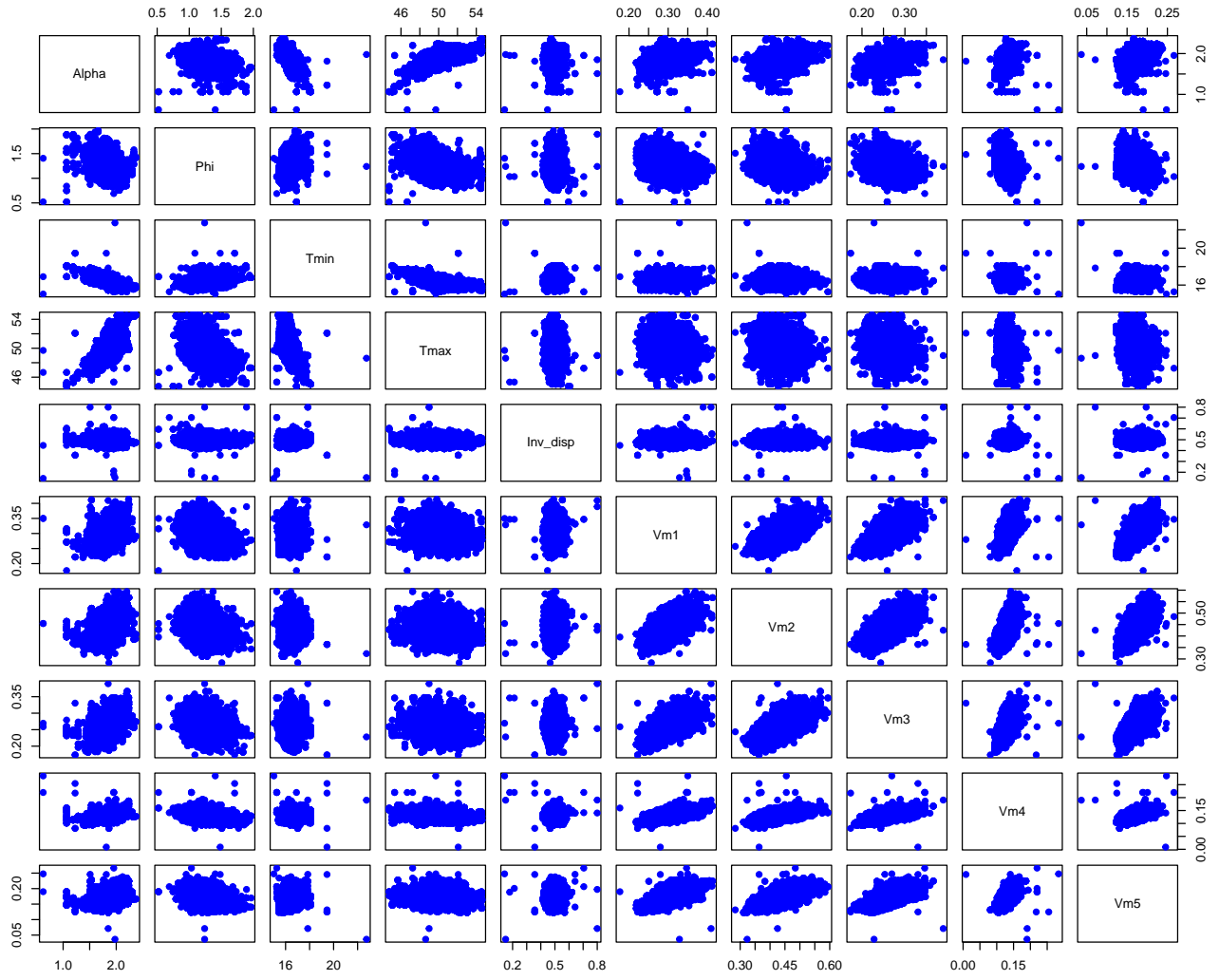
**Figure 17:** Pairs plot displaying joint posterior draws of parameters from the outlier ZCTA model.





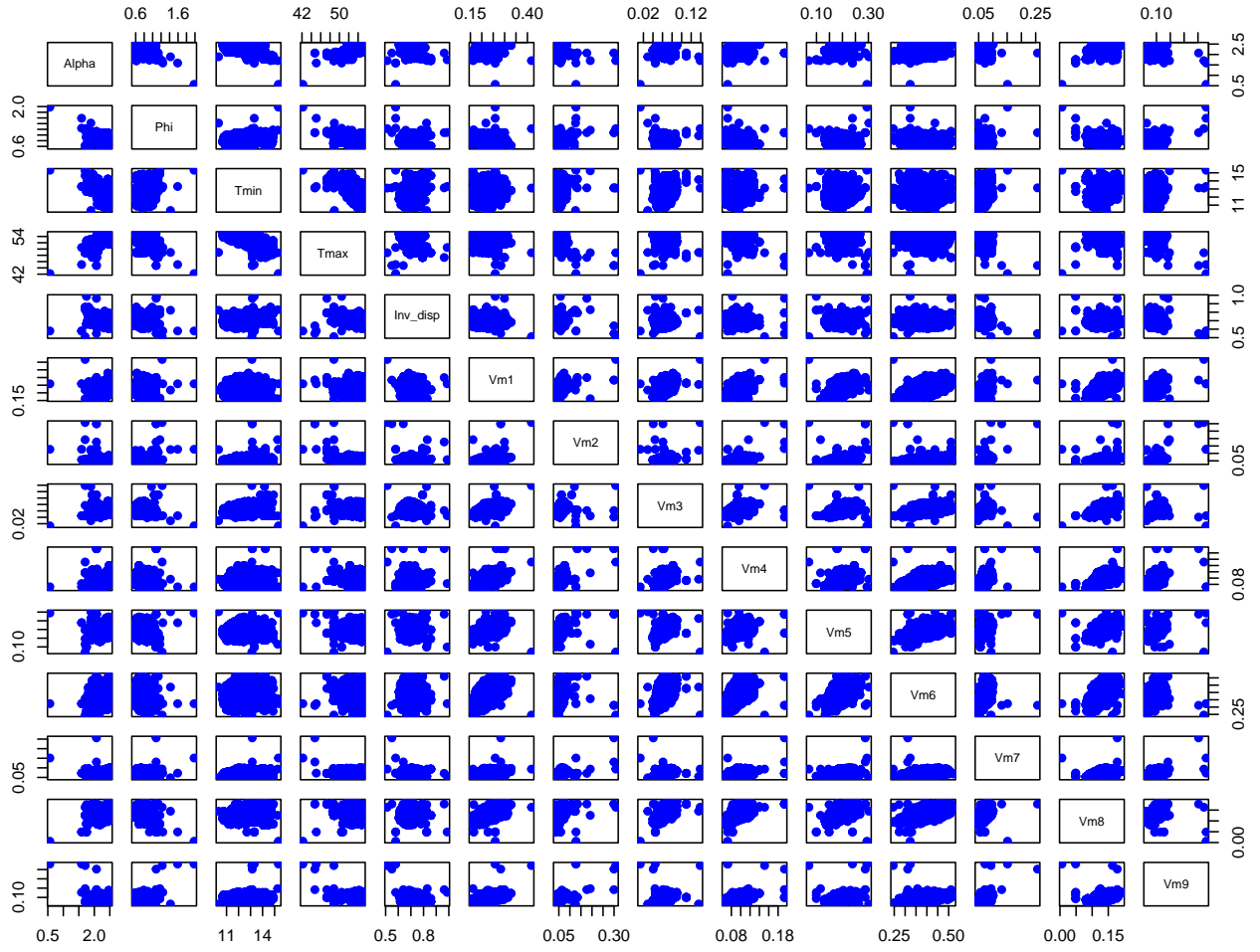
**Figure 18:** Pairs plot displaying joint posterior draws of parameters from 2 cluster model.





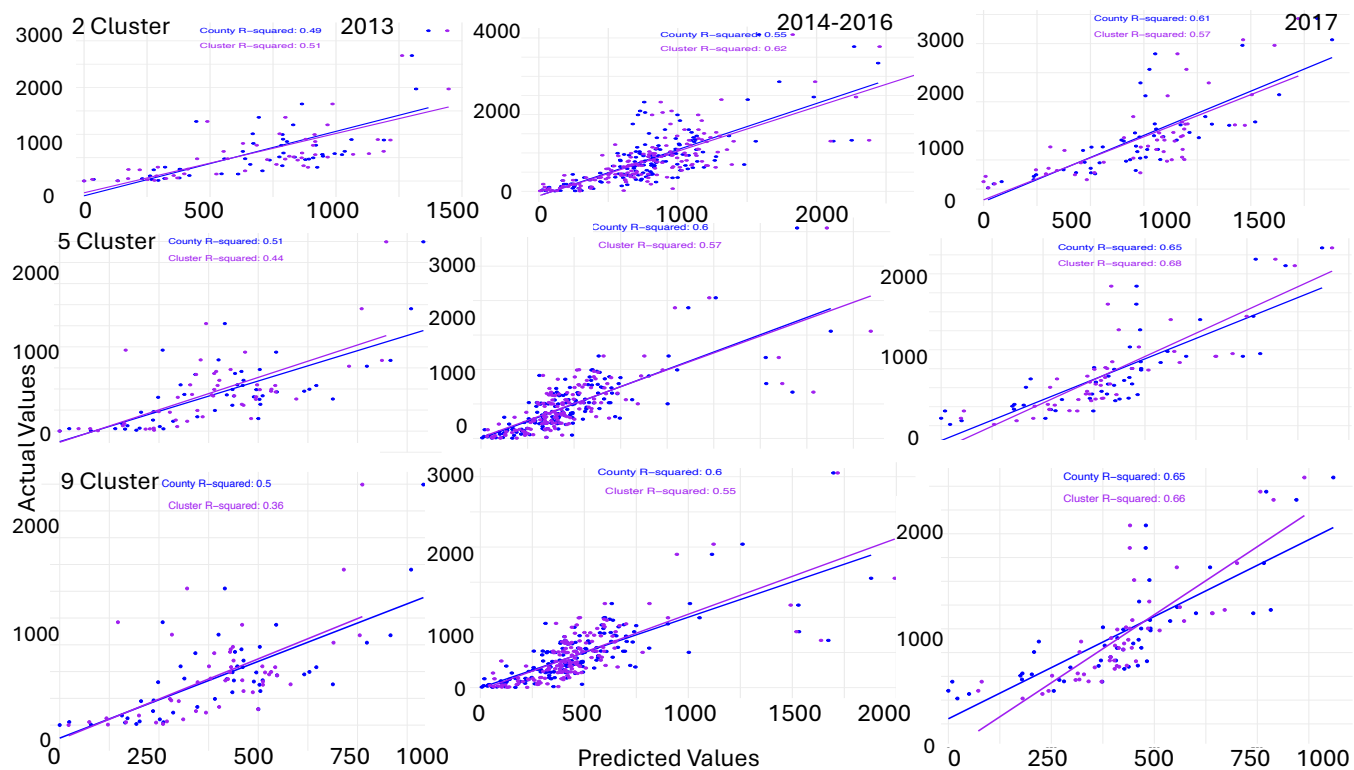
**Figure 19:** Pairs plot displaying joint posterior draws of parameters from 5 cluster model.





**Figure 20:** Pairs plot displaying joint posterior draws of parameters from 9 cluster model.





**Figure 21:** Predicted vs Actual for all cluster