STAT 341 Final Report

Levi Carr

Research Question

In cirrus clouds, how does the effective radius of the droplets change with ice composition?

The PACE Mission and the Cloud Dataset

Mission Overview

PACE (Plankton, Aerosol, Cloud, ocean Ecosystem) is a NASA science mission launched into orbit on February 8th, 2024 to extend and improve observations of global ocean biology, aerosols, and clouds. PACE has three science instruments aboard to examine the earth from above: the SPEXone polarimeter, the HARP2 polarimeter, and the Ocean Color Instrument (OCI). From NASA: "PACE's primary sensor, the Ocean Color Instrument (OCI), is a highly advanced optical spectrometer that measures properties of light over portions of the electromagnetic spectrum. It enables continuous measurement of light at finer wavelength resolution than previous NASA satellite sensors, extending key system ocean color data records for climate studies." This instrument yields atmospheric data products that I use for my analysis of cirrus cloud effective radii.

The satellite's orbital period is one solar day, so data is naturally binned by day. From the data, I selected a "training" day (where I looked directly at the data) and a "test" day which I fit the model to. These days were March 30th and April 2nd, respectively. The data files used are given in the "Data Availability" section.

Selecting Cirrus Clouds

From the cloud data products we also have variables such as the optical thickness τ (a measure of transmission) and cloud top pressure T_{top}. Heymsfield et al. (2017) state that cirrus clouds are observed to have cloud top temperatures greater than -40° C and optical thicknesses less than 3.1. Thus, I select only values from the cloud data set for $T_{top} > 233.15$ K and $\tau < 3.1$.

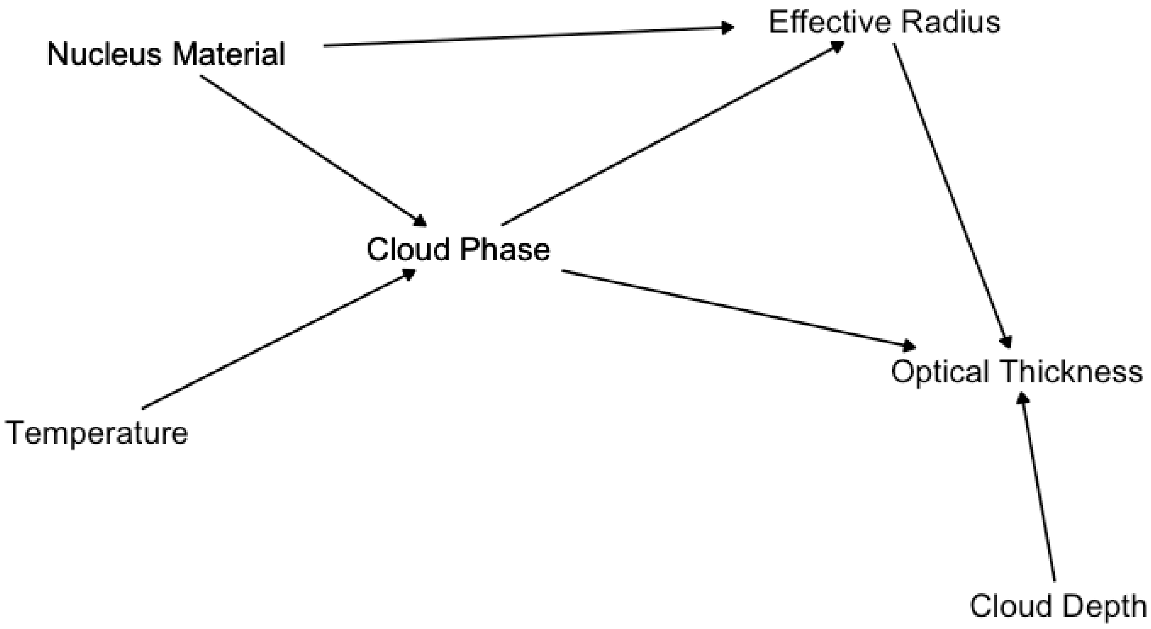
Making Categorical Predictors

Included in the PACE atmospheric data set is the `cf_ice` variable which describes how much of the cloud is ice. `cf_ice` ranges from 0 to 1 where 0 is an all liquid water cloud and 1 is an all ice cloud. By examining the distribution of `cf_ice` on March 30th, 2025, a pattern of two high density peaks around 0 and 1 with a few cases in between is predominant. I therefore separate the data into three groups: “no ice” (0.05 > `cf_ice`), “some ice” (0.05 < `cf_ice` < 0.95), and “all ice” (`cf_ice` > 0.95).

Model Description

My model will synthesize the cloud phase and effective radii sampled at the 1.6 μm band to determine how cloud phase and effective radii interact.

Causal Diagram



From the dataset, we are given temperature, effective radius, optical thickness, and cloud phase. Note that optical thickness is a collider, and so I do not use it in my model. However, as noted above, cirrus clouds as a population are known to reside within a certain range of optical thicknesses, so I only use it as a population selector. This does imply, however, that there is a priori knowledge of the effective radius, though implicitly so. As evidenced by the histogram of effective radii for the "training" data, there appears to be an upper limit on the effective radius occurring around 60 μm, even though we have no a priori knowledge of this limit. Likewise, temperature is used as a population selector based on the theoretical physics and nothing more. Cloud depth is the vertical thickness of the cloud. Nucleus material is the atomic/molecular composition of the droplet nucleus that formed the droplet. Different nuclei will lead to larger droplets and/or droplets that freeze faster, so this is a confounding variable in my model.

Mathematical Description

$$R_{eff} \sim \text{Gamma}(\alpha, \lambda)$$

$$\alpha = \frac{\mu_i^2}{\sigma^2}$$

$$\lambda = \frac{\mu_i}{\sigma^2}$$

The effective radius (R_eff) of droplets in clouds has been identified in the literature as one that is described by a gamma process (Hansen 1971). This likelihood distribution is used in almost all droplet size distribution models.

$$\ln(\mu_i) = \beta_0 + \beta_1 \,[\text{ice fraction}_i]$$

Since $\ln \mu$ is a linear relationship, we can say that R_{eff} is a GLM. Since R_{eff} is modeled as a gamma GLM, we use a logarithmic link function to assure that all inputs to the likelihood distribution are strictly positive. In particular, $\ln(\mu)$ is a function of β_0 and β_1 [ice fraction]. β_0 is as the global average R_{eff} for all cloud types. β_1 is a 3-vector with one component for each cirrus subset (no ice, some ice, and all ice). The components of β_1 are then the corrections from the global average R_{eff} value for each phase type.

$$\beta_0 \sim \text{Normal}(\ln(23), \ln(5))$$

It is very hard to track down the size of liquid water in cirrus clouds, so I hypothesize that the radius of liquid water droplets are smaller in clouds to the merit that liquid water expands when it freezes, often into shapes that are not minimal surfaces; this is to say that ice's effective radius should be larger. Since β_0 is the global average which includes the ice clouds, my hypothesis suggests that the radii are smaller than the average effective radius of cirrus clouds.

Assume that the hypothesis is true and that all ice and no ice clouds are found in equal proportions across the globe. It should be the case then that the global average is the average of the means of the all ice and no ice cloud droplet distributions. Since water expands by ~10%, the mean effective radius for no ice clouds should be smaller than the global average by ~5% or ~1μm. Likewise, the mean effective radius for all ice clouds should be ~5% larger than the global average, or ~1μm. How do we get this to work with the link function? Start by writing

$$\ln(\vec{\mu}_j) = \vec{\beta}_0 + \vec{\beta}_{1,j}$$

where $j = 1, 2, 3$ are the no ice, some ice, and all ice populations, respectively, represented as components of $\vec{\beta}_1$.

$$\vec{\mu}_j = e^{\vec{\beta}_0 + \vec{\beta}_{1,j}} = e^{\vec{\beta}_0} e^{\vec{\beta}_{1,j}} = e^{\vec{\beta}_0} \left[e^{\beta_{1,1}}, e^{\beta_{1,2}}, e^{\beta_{1,3}} \right]^\top$$

But $e^{\vec{\beta}_0}$ is just the global average R_{eff} and $\vec{\mu}$ is a vector of our expected values for each phase of matter. So $e^{\vec{\beta}_0} = [23, 23, 23]^\top$, $\mu = [22, 23, 24]$, and

$$\vec{\mu} \cdot e^{-\vec{\beta}_0} = \frac{1}{23}[22, 23, 24]^\top = \left[e^{\beta_{1,1}}, e^{\beta_{1,2}}, e^{\beta_{1,3}} \right]^\top = e^{\vec{\beta}_{1,j}}$$

therefore, the prior means, $\vec{\beta}_{1,j}$, are

$$\vec{\beta}_{1,j} = \ln\left(\frac{1}{23}[22, 23, 24]^\top \right) = [-0.04445, 0.00000, 0.04256]^\top$$

As a result, I suppose

$$\beta_1[\text{no ice}] \sim \text{Student}(\nu = 2,\ \text{mean} = -0.04445,\ \text{sigma} = 0.2)$$

I could not track down a value for the variance of the liquid phase's effective radius, but I expect it to be fairly exact since it is based on a physical basis. I chose a standard deviation of 0.2 μm. Even so, I regularize the prior with a student t-distribution so that the data may still drive the posterior.

$$\beta_1[\text{some ice}] \sim \text{Student}(\nu = 2,\ \text{mean} = 0.00000,\ \text{sigma} = 0.4)$$

There are a small number of mixed-phase cirrus clouds, but they range from 5-95% ice by my artificial distinction. I expect the average radii of droplets in these clouds to be close to the global average, but with a larger variation than their ice and liquid counterparts given the wide range of possible phase concentrations. A standard deviation of 0.4 μm was chosen via the prior predictive distribution.

$$\beta_1[\text{all ice}] \sim \text{Student}(\nu = 2,\ \text{mean} = 0.04256,\ \text{sigma} = 0.2)$$

By my hypothesis, the freezing of liquid water droplets should yield larger effective radii, so the correction to the average should be positive. Like the liquid droplet case, I could not find much for variance information, but I still expect it to be fairly constrained since the cloud is composed only of one phase of matter. I chose a standard deviation of 0.2 μm. Even so, I regularize the prior with a student t-distribution so that the data may still drive the posterior.

By analyzing one day of data, I found that most effective radii were above 10 μm. Therefore, I adopt a lognormal distribution (strictly positive outputs, faster probability drop than an exponential) to describe the standard deviation in effective radius. Knowing that many of the R_{eff} distributions have an effective lower bound at 10 μm and an effective upper bound ~80μm , I chose to adopt the following as my standard deviation:

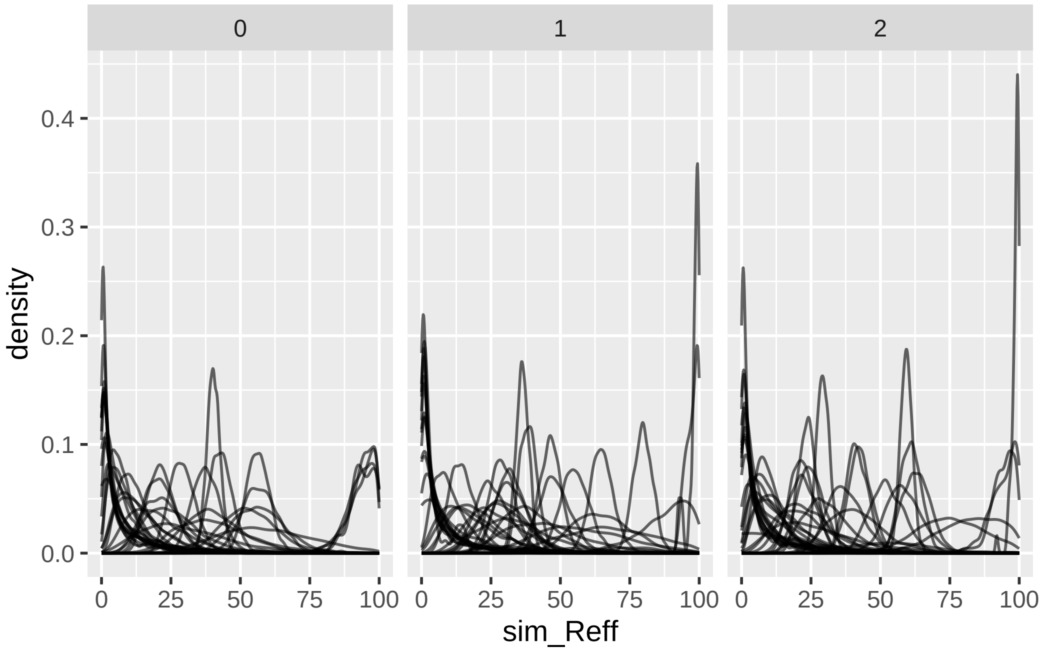
$$\sigma \sim \text{Lognormal}(\ln(8), 0.5)$$

Prior Predictive Distribution

Using the prior information above, I construct a prior predictive distribution for the effective radii in each population.

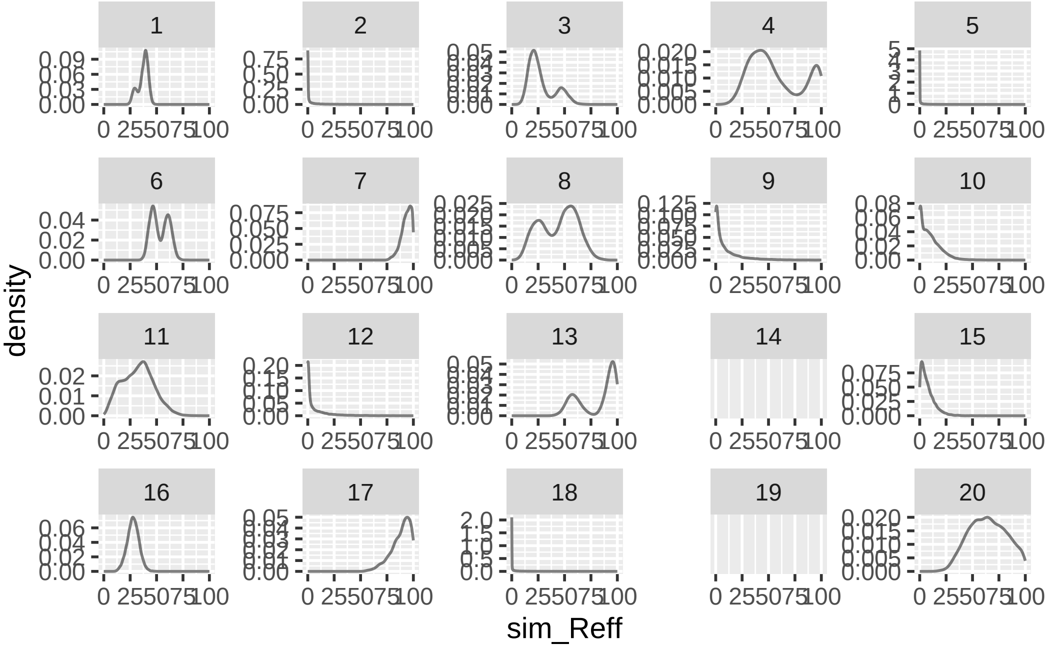
```r
# Wrangle Data
data <- nc_open("PACE_OCI.20250402.L3m.DAY.CLOUD.V3_0.1deg.NRT.nc")

lat <- ncvar_get(data, "lat") # columns of data matrices
lon <- ncvar_get(data, "lon") # rows of data matrices
lat_matrix <- matrix(lat, nrow = length(lon), ncol = length(lat), byrow = TRUE)
lon_matrix <- matrix(lon, nrow = length(lon), ncol = length(lat), byrow = FALSE)

cf_ice <- ncvar_get(data, "ice_cloud_fraction")
cot <- ncvar_get(data, "cth_cot")
ctt <- ncvar_get(data, "ctt")
cer_16 <- ncvar_get(data, "cer_16")

cloud_data <- data.frame(
  lat = as.vector(lat_matrix),
  lon = as.vector(lon_matrix),
  cf_ice = as.vector(cf_ice),
  cot = as.vector(cot),
  ctt = as.vector(ctt),
  cer16 = as.vector(cer_16)
  )|>
   drop_na(cf_ice, cot, ctt, cer16) |>   # optional, there are a lot of NAs
   filter(cot < 3.1, ctt > 233.15) |>
   # Select cirrus range. Rationale: https://journals.ametsoc.org/view/journals/amsm/58/1/ams
   mutate( ice_idx = ifelse(cf_ice < 0.05, 0, ifelse(cf_ice < 0.95, 1, 2)) )
   # No ice: 0, Some ice: 1, All ice: 2

# Simulate Priors and Posteriors
n_sim <- 50 # number of simulated datasets

prior_pred_dist <- tibble(sim_id = c(1:n_sim)) |>
   mutate(b0 = rnorm(n_sim, log(23), log(5)),
          b1_0 = rstudent(n_sim, nu=2, mu=log(22/23), sigma=0.2),
          # I expect the effective radius to decrease from average for
          # (supercooled) non precipitating liquid clouds
          b1_1 = rstudent(n_sim, nu=2, mu=log(1), sigma=0.4),
          # I expect the effective radius to stay around average for
          # intermediate liquid/ice states. I expect a larger variance here
          b1_2 = rstudent(n_sim, nu=2, mu=log(24/23), sigma=0.2),
          # I expect the effective radius to increase for ice clouds.
          # Ice crystals tend to be bigger in non precipitating clouds
          sigma = rlnorm(n_sim, log(8), 0.5)
          ) |>
   rowwise() |>
   mutate(mu = list(exp( b0 +
                         b1_0 * ifelse(cloud_data$ice_idx == 0, 1, 0) +
                         b1_1 * ifelse(cloud_data$ice_idx == 1, 1, 0) +
                         b1_2 * ifelse(cloud_data$ice_idx == 2, 1, 0) )),
          ice_idx = list(cloud_data$ice_idx),
          cf_ice = list(cloud_data$cf_ice),
          ) |>
   unnest(cols =  c(mu, ice_idx, cf_ice)) |>
   ungroup() |>
   mutate(alpha = mu^2 / sigma^2,
          lambda = mu / sigma^2
   ) |>
   rowwise() |>
   mutate( sim_Reff = rgamma(1, shape = alpha, rate = lambda) ) |>
   ungroup()
```

```
gf_dens(~sim_Reff | factor(ice_idx), group = ~sim_id,
        data = prior_pred_dist |> filter(sim_Reff > 0.1),
         alpha = 0.6) |>
   gf_lims(x = c(0,100))
```



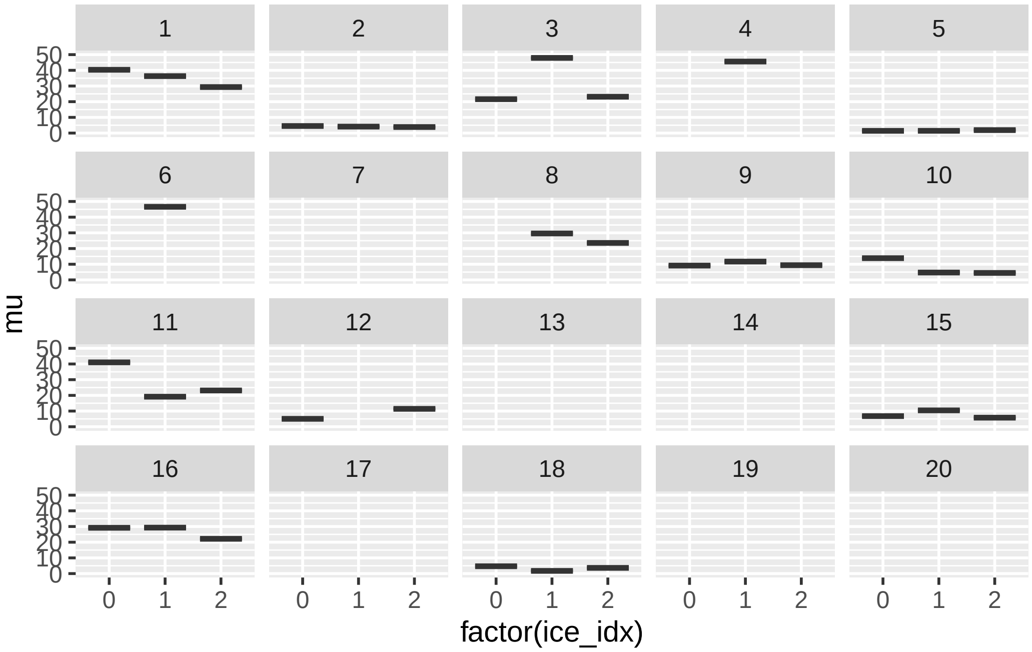
The prior predictive distribution for each population of clouds, grouped by phase of matter. No ice is on the left, some ice is in the center, all ice is on the right. As seen from the prior predictive distribution(s), the three populations are nearly identical. Furthermore, there may be an unaccounted for effect driving many of the effective radii towards zero. I do not have the theoretical background to know what physical mechanism this might be.

```
gf_dens(~sim_Reff, group = ~sim_id,
        data = prior_pred_dist |>
           filter(sim_id <21)) |>
   gf_facet_wrap(~sim_id, scales = 'free') |>
   gf_lims( x = c(0,1E2))
```



The first 20 prior predictive distributions for effective radius. Note that the distributions not with mean ~ 0 follow the expected/observed distribution for real cirrus clouds.

```
gf_boxplot(mu ~ factor(ice_idx) | sim_id,
         alpha = 0.1,
         data = prior_pred_dist |>
          filter(sim_id < 21) ) |>
   gf_lims( y = c(0, 50))
```



Boxplots of the first 20 simulated values of *mu* with respect to ice content. No ice is seen on the left, some ice in the center, and all ice on the right of each pane

Fitted Model

This model was fit by the MCMC package cmdstan. Four chains were computed with 2000 total iterations each using `adapt_delta = 0.8`.

Diagnostics for Fit Convergence

I analyzed two fits of convergence readily available in stan: neff and Rhat. This model uses 1000 sampling iterations. The neff for the parameters associated with the categorical predictors is ~500 for all, indicating that the sampling process was mildly inefficient. Rhat for all parameters associated with the predictors was larger than 1.00 which likely reflected challenges for my model to account for the large variability in cloud data. I present the results of the mostly-converged data here with the knowledge that a better model should be fit in the future.

```
summary(cirrus_fit_w_mixed_phase, pars = c("b0", "b1[1]", "b1[2]", "b1[3]", "sigma"))$summar
```

```
             mean     se_mean         sd       2.5%         25%         50%
b0     3.20241304 0.029819588 0.66477228  1.8200695  2.78595174  3.20327656
b1[1] -0.26278675 0.029799305 0.66462480 -1.5389417 -0.68930081 -0.26527911
b1[2]  0.02305593 0.029800529 0.66454111 -1.2511190 -0.40133190  0.02027291
```

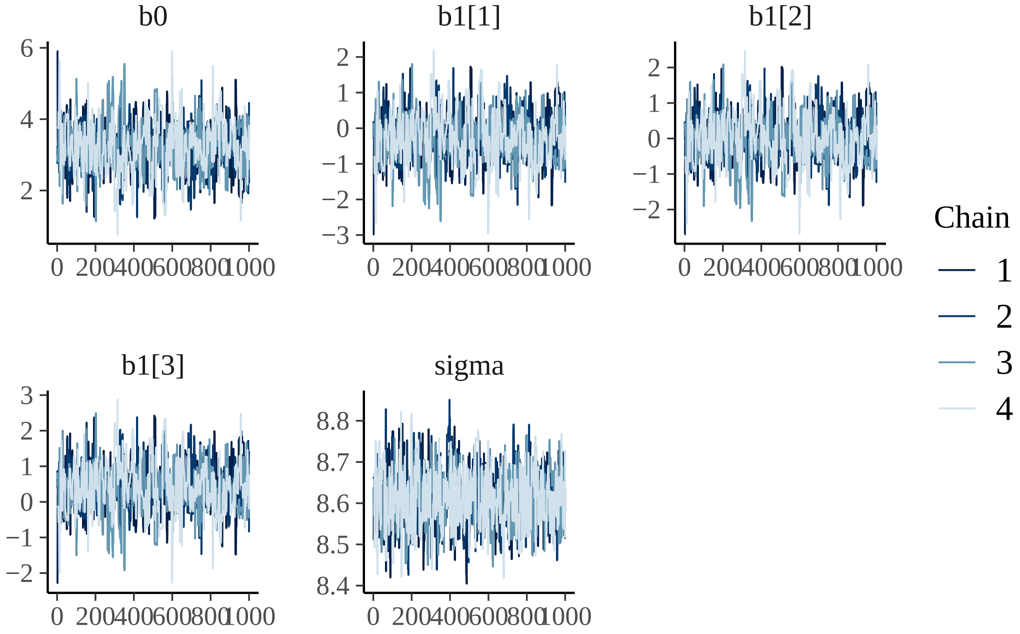
```
b1[3]  0.42822075 0.029821605 0.66467433 -0.8458085  0.00374446  0.42735057
sigma  8.61387789 0.002381398 0.06549791  8.4867594  8.57019519  8.61294701
            75%    97.5%    n_eff      Rhat
b0    3.6290987 4.476855 496.9841 1.006558
b1[1] 0.1555302 1.117513 497.4401 1.006548
b1[2] 0.4404912 1.407335 497.2740 1.006555
b1[3] 0.8425490 1.813610 496.7705 1.006590
sigma 8.6584744 8.748995 756.4683 1.000838
```

```
summary(cirrus_fit_wo_mixed_phase, pars = c("b0", "b1[1]", "b1[2]", "sigma"))$summary
```

```
            mean      se_mean          sd      2.5%         25%         50%
b0     3.2739387 0.030823333 0.75271040  1.797611  2.8065300  3.2635264
b1[1] -0.2884790 0.030831262 0.75263663 -1.857392 -0.7371462 -0.2786606
b1[2]  0.2507371 0.030841719 0.75268622 -1.315981 -0.1985443  0.2605742
sigma  9.0826140 0.002354852 0.06780853  8.950939  9.0380194  9.0836584
            75%    97.5%    n_eff      Rhat
b0    3.7221908 4.838281 596.3437 1.009928
b1[1] 0.1818931 1.190937 595.9202 1.009933
b1[2] 0.7196899 1.731436 595.5946 1.009947
sigma 9.1278103 9.212311 829.1660 1.011213
```

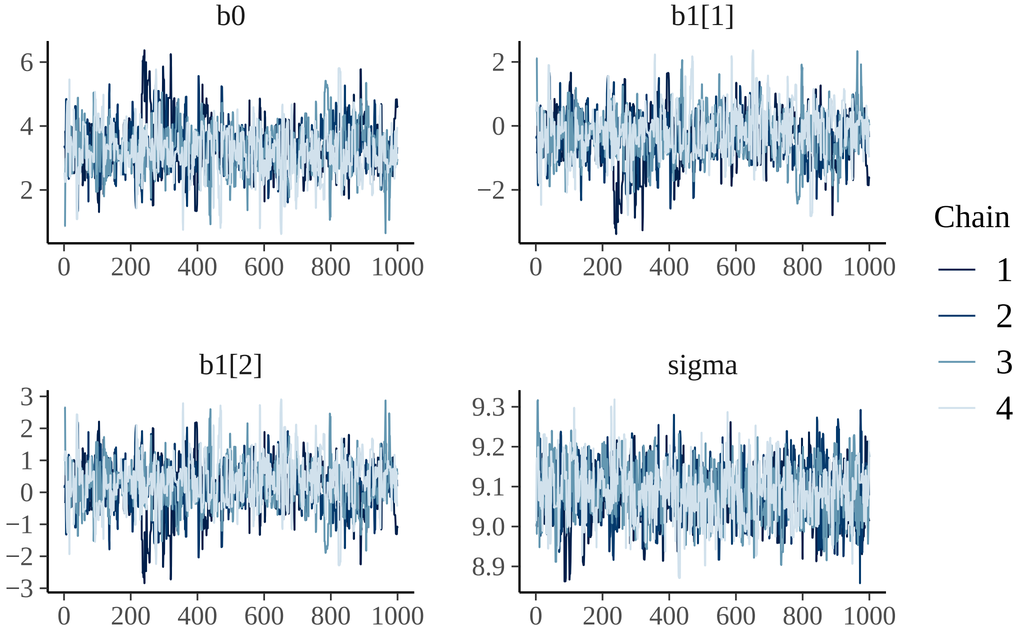
As seen, neff values are large enough that they should be sufficient for the model to converge, but the Rhat values near one suggest that the model is not fully converged.I plot the MCMC trace for all four chains to check for visual convergence.

```
mcmc_trace(cirrus_fit_w_mixed_phase, pars = c("b0", "b1[1]", "b1[2]", "b1[3]", "sigma"))
```



The MCMC trace for the cirrus model that includes mixed phase clouds.

```r
mcmc_trace(cirrus_fit_wo_mixed_phase, pars = c("b0", "b1[1]", "b1[2]", "sigma"))
```

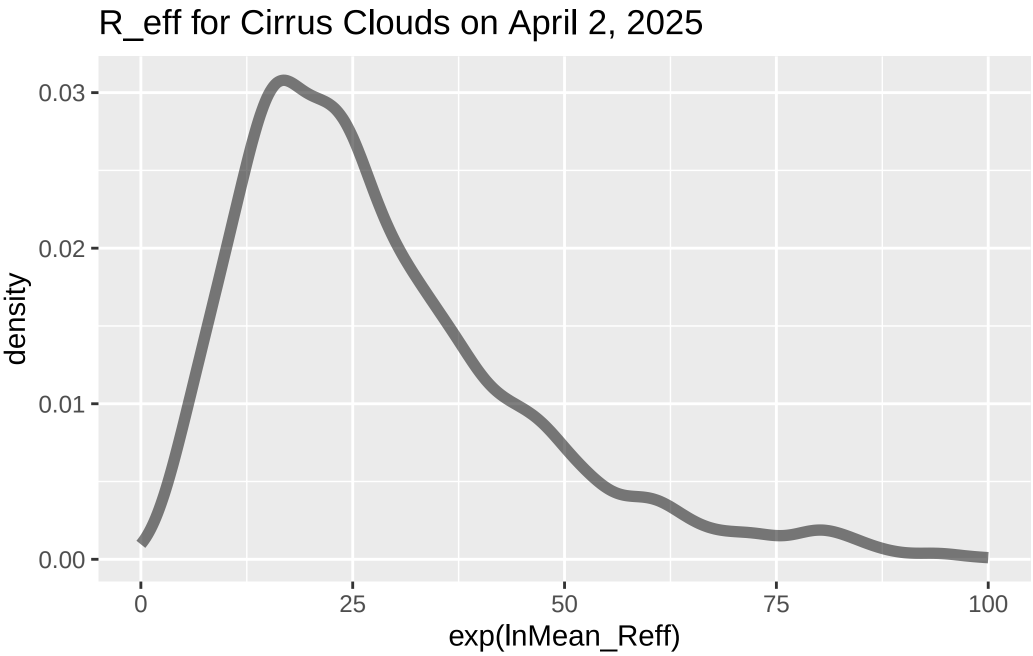


The chains are mixed, but not well-mixed. Similarly, they appear to be mostly stationary. Future work on this model should restrict what values the priors can take.

Posteriors

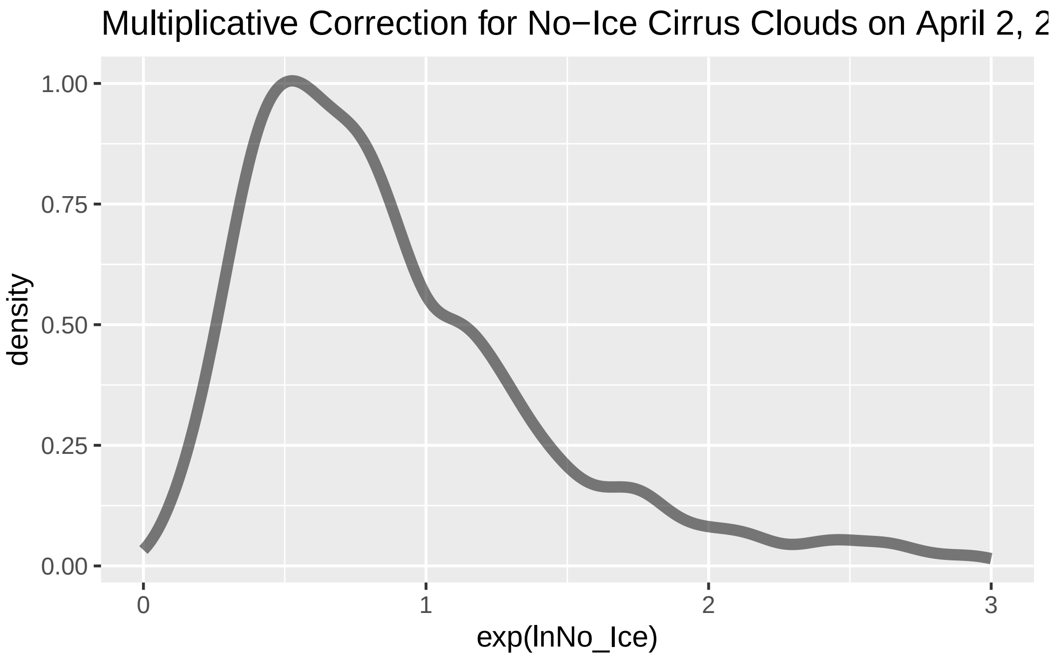
You may include any graphs or summary statistics of your choice that you think are useful in understanding the posterior. I present the posteriors for the effective radius intercept and "correction" parameters. Because of the link function, the values are the natural log of the actual effective radius.

```
gf_dens(~exp(lnMean_Reff), data=cirrus_fit_df, linewidth = 2, color = 'black') |>
  gf_labs(title='R_eff for Cirrus Clouds on April 2, 2025', ylab=('Posterior Density')) |> g
```



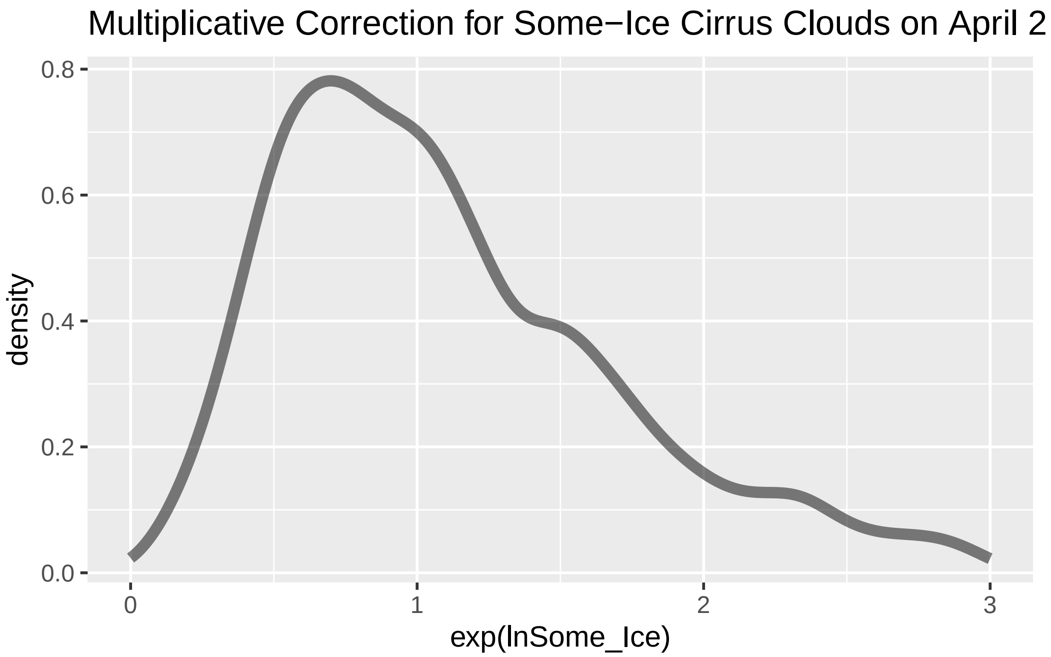
The effective radius intercept posterior. This represents the global average of cirrus cloud droplet effective radius.

```
gf_dens(~exp(lnNo_Ice), data=cirrus_fit_df, linewidth = 2, color = 'black') |>
  gf_labs(title='Multiplicative Correction for No-Ice Cirrus Clouds on April 2, 2025', ylab=
```



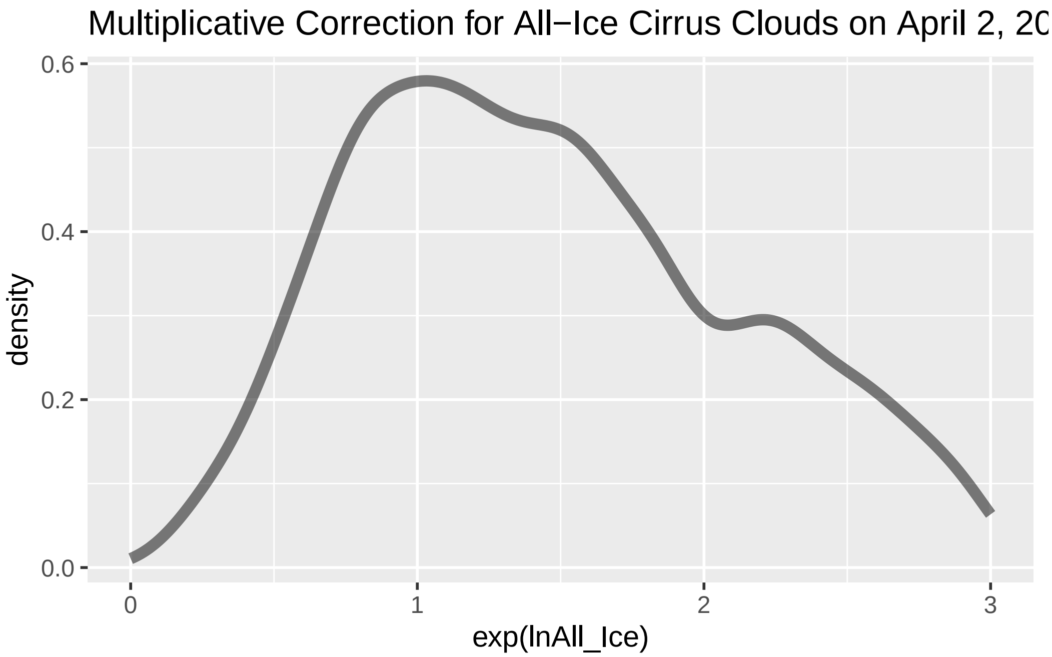
The effective radius “no ice” correction posterior. This posterior suggests that the average effective radius of no-ice cirrus clouds is smaller than the global average, with most no-ice cirrus clouds having about half the global average effective radius.

```
gf_dens(~exp(lnSome_Ice), data=cirrus_fit_df, linewidth = 2, color='black') |>
  gf_labs(title='Multiplicative Correction for Some-Ice Cirrus Clouds on April 2, 2025', ylab
```



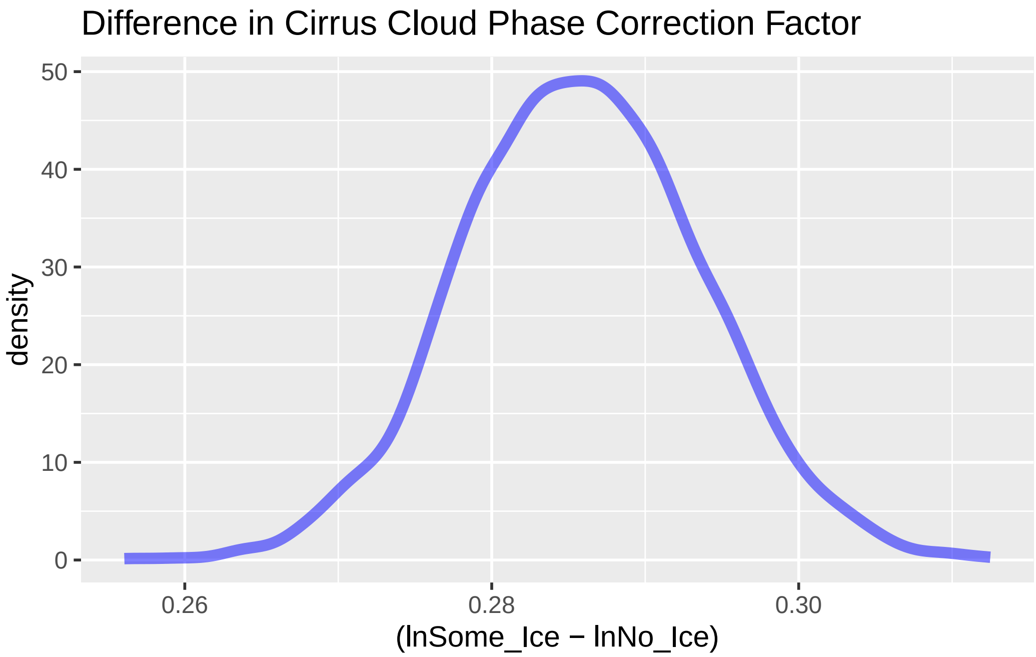
The effective radius “some ice” correction posterior. This posterior suggests that the mean effective radius of some-ice cirrus clouds is well-represented by the global average effective radius, with most some-ice cirrus clouds having a smaller effective radius than the global average.

```
gf_dens(~exp(lnAll_Ice), data=cirrus_fit_df, linewidth = 2, color='black') |>
  gf_labs(title='Multiplicative Correction for All-Ice Cirrus Clouds on April 2, 2025', ylab=
```



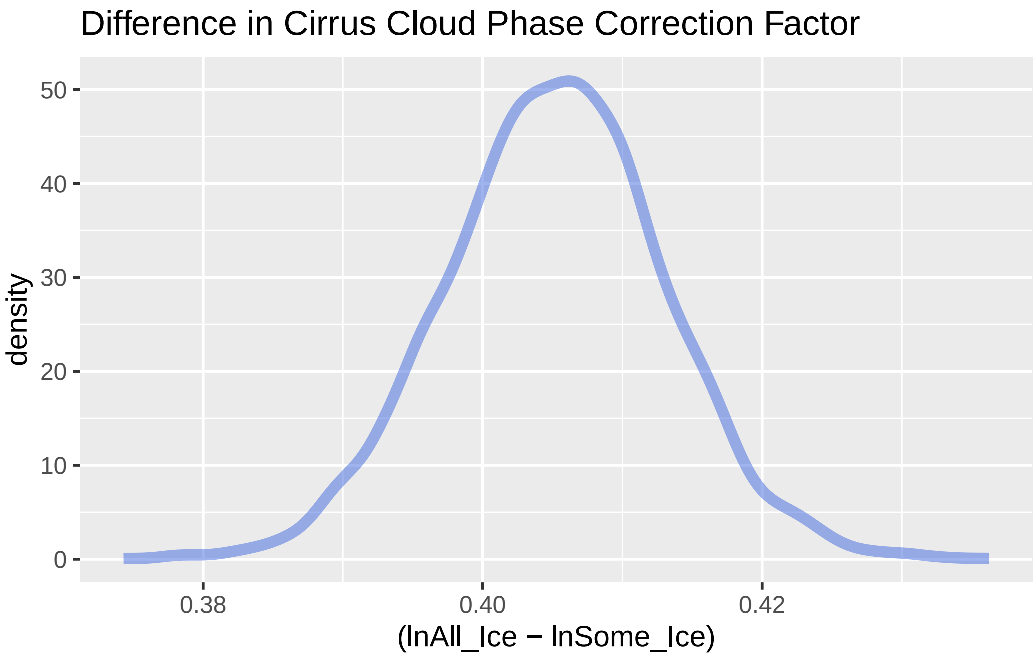
The effective radius “all ice” multiplicative correction posterior. Given the mean around 1 and a thick upper tail, this suggests that the effective radius of the all-ice cirrus clouds is larger than the global average, but with most all-ice clouds having the global average effective radius.

```
gf_dens(~(lnSome_Ice-lnNo_Ice), data=cirrus_fit_df, linewidth=2, color='blue')|> gf_labs(titl
                                                      ylab("Posterior Density"))
```



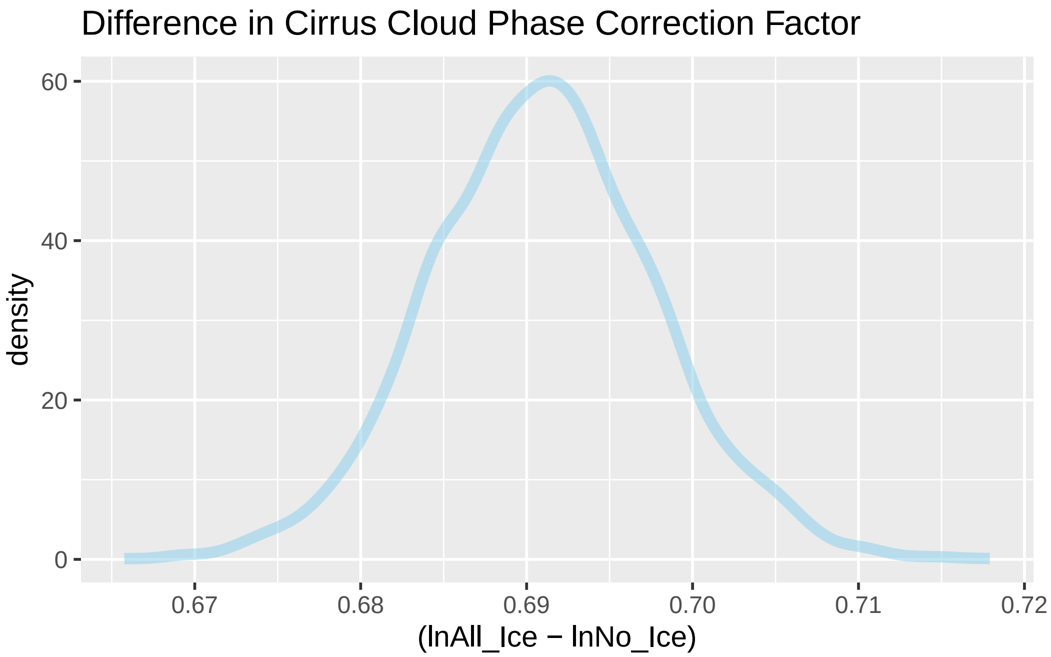
The difference between the natural log of the correction parameters in “some ice” and “no ice” clouds. A difference of 0.285 on this scale corresponds to a multiplicative difference of 1.33 μm. This suggests that the no_ice cloud correction is greater than the some_ice cloud correction.

```
gf_dens(~(lnAll_Ice-lnSome_Ice), data=cirrus_fit_df, linewidth=2, color='royalblue')|> gf_lab
                                                                ylab("Posterior Density"))
```



The difference between the natural log of the correction parameters in "some ice" and "no ice" clouds. A difference of 0.405 on this scale corresponds to a multiplicative difference of 1.50 μm. This suggests that the some ice cloud correction is greater than the some_ice cloud correction.

```
gf_dens(~(lnAll_Ice-lnNo_Ice), data=cirrus_fit_df, linewidth=2, color='skyblue')|> gf_labs(ti
                                                              ylab("Posterior Density"))
```



The difference between the natural log of the correction parameters in “all ice” and “no ice” clouds. A natural log difference of 0.692 on this scale corresponds to a difference of 2.00 μm. This suggests that the no_ice correction is greater than the all ice correction.

Model Comparison

Another model was fit with where the mixed phase clouds were divided between the “no ice” and “all ice” populations. This is a test to probe whether modeling the mixed phase clouds can be done with a simpler model since the mixed phase clouds are not common. In data prep, the `cf_ice` number was divided at 0.5; all observations with `cf_ice < 0.5` went to “no ice” and the remainder went to “all ice”. A modification to the stan code was made to reflect the elimination of a prior, but no other features were changed. The model was analyzed for convergence as above and was found to be convergent. I use the WAIC function from the `Rethinking` package by Richard McElreath to compare the two models. The results are shown below.

```r
compare(cirrus_fit_w_mixed_phase, cirrus_fit_wo_mixed_phase, func=WAIC)
```

```
Error in extract(object, pars = log_lik) :
  Arguments in `...` must be used.
x Problematic argument:
* pars = log_lik
i Did you misspell an argument name?
```

```
                              WAIC       SE   dWAIC      dSE    pWAIC
cirrus_fit_w_mixed_phase  75303.69 182.5909    0.00       NA 6.349071
cirrus_fit_wo_mixed_phase 76328.47 177.3938 1024.78 125.6189 4.498500
                                 weight
cirrus_fit_w_mixed_phase   1.000000e+00
cirrus_fit_wo_mixed_phase 2.964257e-223
```

It is clear from the WAIC scores that the cirrus model that accounts for the mixed phase is significantly better at describing the data than the model that does not account for the mixed phase.

Conclusion

Based on my analysis of the PACI OCI atmospheric characteristics for April 2nd, 2025, I conclude that cirrus clouds with no ice tend to have smaller effective radii than the global average, while cirrus clouds composed entirely of ice tend to have similar or larger radii than the global average. In between, cirrus clouds with a mix of ice and water tend to have similar or smaller effective radii than the global average. Future work on this project might look at performing an analysis on the polar, mid-latitude, and tropical cirrus clouds and compare the no ice, some ice, and all ice populations between the latitudes to determine if one type of cloud is more common at one latitude than another.

Data Availability

All my files are available upon request through GitHub. Files can be found at https://github.com/carr-levi-a/calvinbayes25_atmsci.

I used data from the NASA PACE mission accessed through https://oceandata.sci.gsfc.nasa.gov/directdataaccess In particular, I use Level 3 Mapped data product from the OCI instrument with a resolution of 0.1 deg or ~11 km. The dates were March 30th, 2025, and April 2nd, 2025. The names of these files are "PACE_OCI.20250330.L3m.DAY.CLOUD.V3_0.1deg.NRT.nc" and "PACE_OCI.20250402.L3m.DAY.CLOUD.V3_0.1deg.NRT.nc", respectively.

References:

Hansen, J.E., Travis, L.D. Light scattering in planetary atmospheres. Space Sci Rev 16, 527–610 (1974). https://doi.org/10.1007/BF00168069

Heymsfield, A. J., et al. Cirrus Clouds. Meteor. Monogr. 58, 2.1–2.26 (2017). https://doi.org/10.1175/AMSMONOGRAPHS-D-16-0010.1

Ocean Color Instrument. Accessed 05/01/2025. https://pace.oceansciences.org/oci.htm