



# Stan

Software Ecosystem for Modern Bayesian Inference

**[github.com/jgabry/nyr-stan-workshop-2023](https://github.com/jgabry/nyr-stan-workshop-2023)**

NY R Conference  
Stan + Bayes Workshop  
July 2023

# Instructors

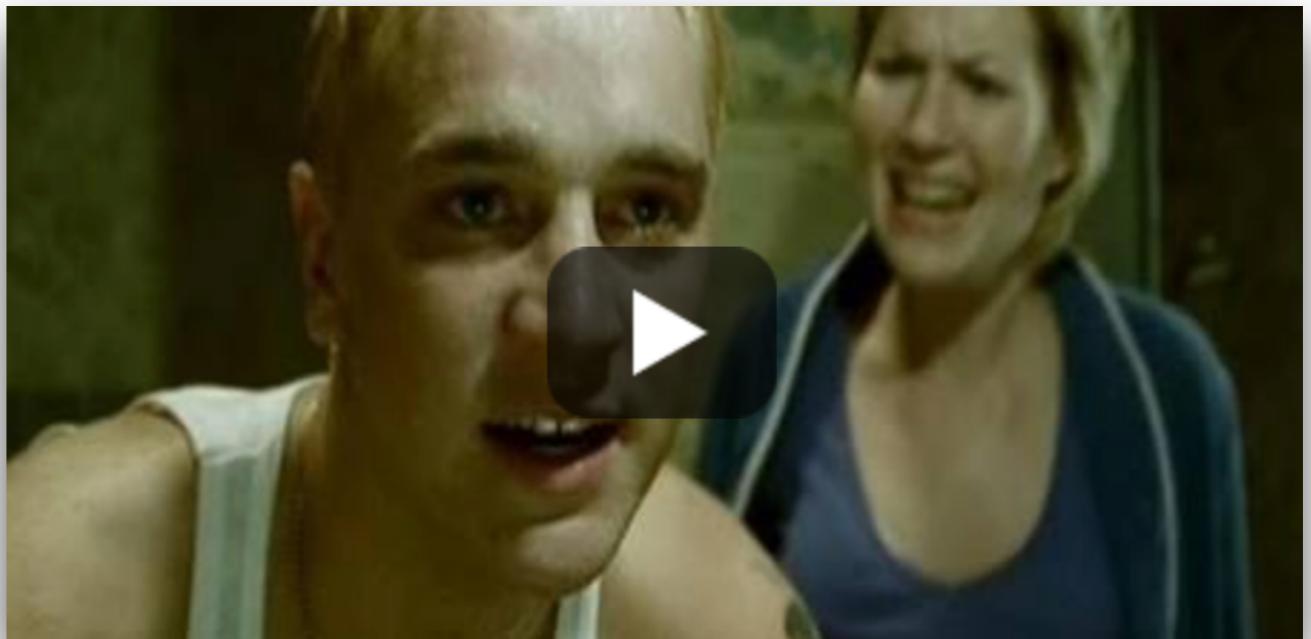
**Jonah Gabry**  
Columbia University

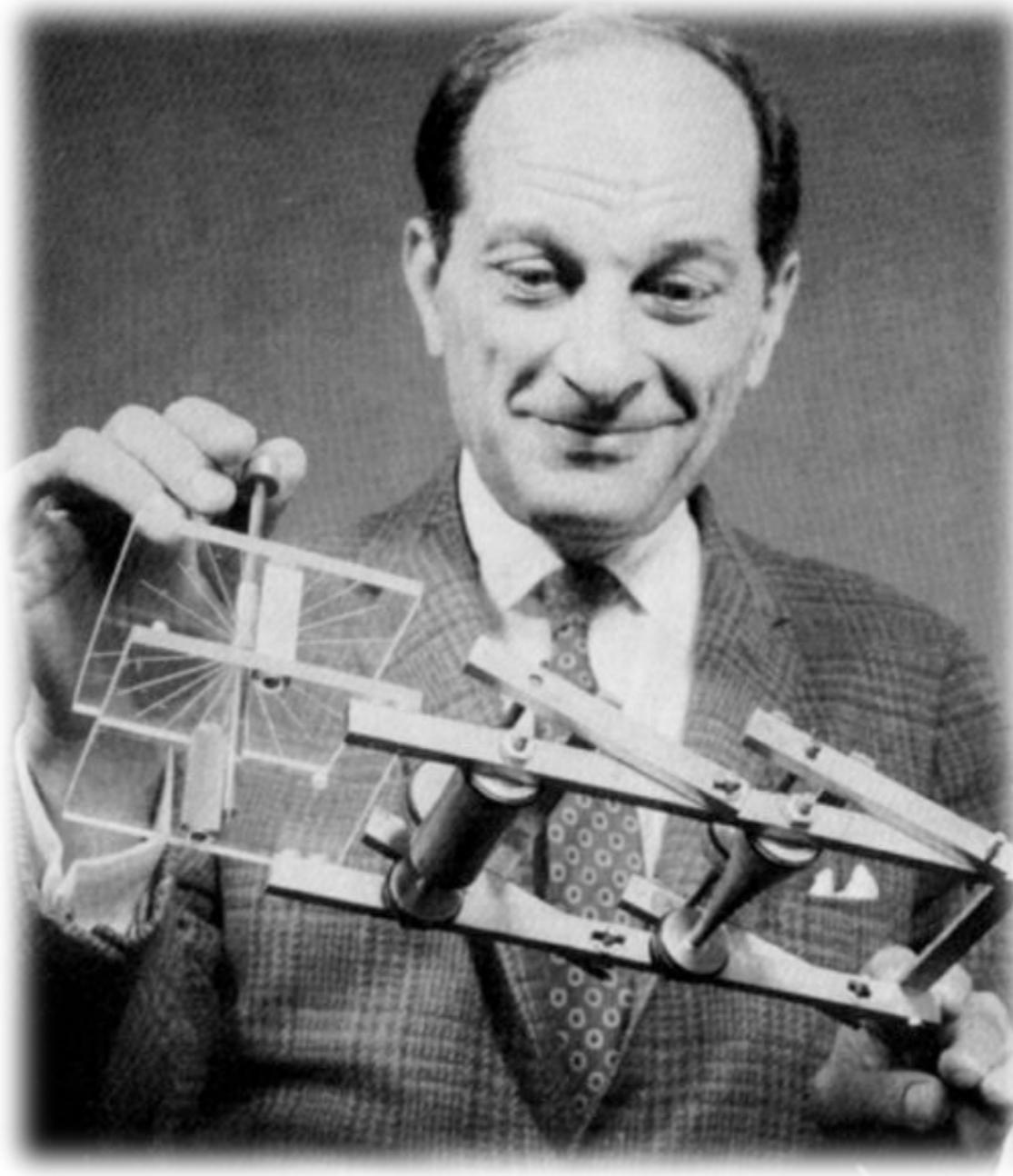
[linkedin](#)  
[website](#)

**Scott Spencer**  
Columbia University

[linkedin](#)  
[Columbia faculty page](#)

# Why “Stan”? suboptimal SEO





Stanislaw Ulam  
(1909–1984)

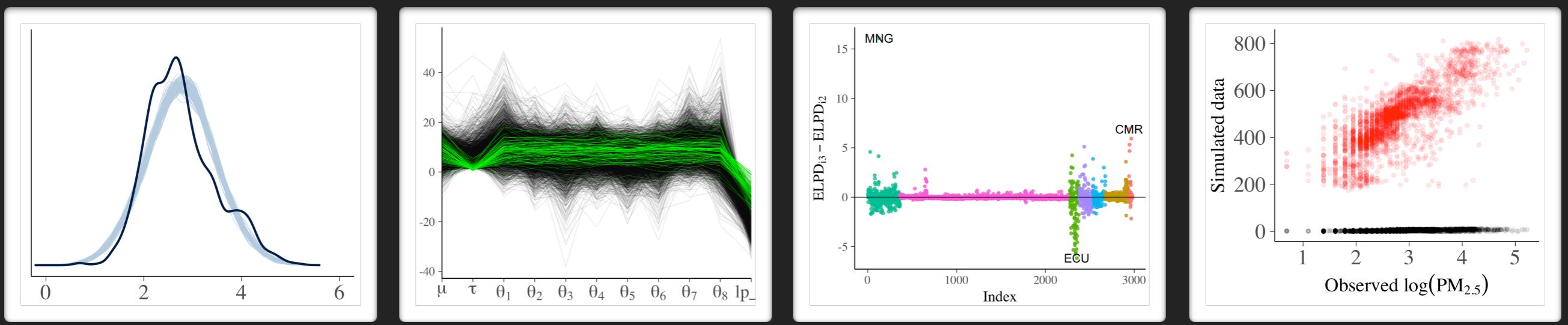
Monte Carlo  
Method

H-Bomb

# What is Stan?

- Open source probabilistic **programming language, inference algorithms**
- Stan **program**
  - declares data and (constrained) parameter variables
  - defines log posterior (or penalized likelihood)
- Stan **inference**
  - MCMC for full Bayes
  - VB for approximate Bayes
  - Optimization for (penalized) MLE
- Stan **ecosystem**
  - language, math library (C++)
  - interfaces and tools (R, Python, Julia, many more)
  - documentation (reference manual and user guide, case studies, R package vignettes, and more)
  - online community (Stan Forums on Discourse)

# Visualization in Bayesian workflow



**Jonah Gabry**

Columbia University  
Stan Development Team

# A Bayesian modeler commits to an *a priori joint distribution*

$$p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathbf{y})p(\mathbf{y})$$

*Likelihood x Prior*

*Posterior x Marginal Likelihood*

The diagram illustrates the decomposition of a joint probability. On the left, the joint probability  $p(\mathbf{y}, \boldsymbol{\theta})$  is shown as a product of two terms:  $p(\mathbf{y} \mid \boldsymbol{\theta})$  and  $p(\boldsymbol{\theta})$ . This is labeled "Likelihood x Prior". Arrows point from the labels "Data (observed)" and "Parameters (unobserved)" to the terms  $p(\mathbf{y} \mid \boldsymbol{\theta})$  and  $p(\boldsymbol{\theta})$  respectively. On the right, the joint probability is shown as a product of two terms:  $p(\boldsymbol{\theta} \mid \mathbf{y})$  and  $p(\mathbf{y})$ . This is labeled "Posterior x Marginal Likelihood". Arrows point from the labels "Parameters (unobserved)" and "Data (observed)" to the terms  $p(\boldsymbol{\theta} \mid \mathbf{y})$  and  $p(\mathbf{y})$  respectively.

**Data (observed)**    **Parameters (unobserved)**

# Workflow

## Bayesian data analysis

- Exploratory data analysis
- *Prior* predictive checking
- Model fitting and algorithm diagnostics
- *Posterior* predictive checking
- Model comparison (e.g., via cross-validation)

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019).

### **Visualization in Bayesian workflow.**

*Journal of the Royal Statistical Society Series A*

Journal version: [rss.onlinelibrary.wiley.com/doi/full/10.1111/rssc.12378](https://rss.onlinelibrary.wiley.com/doi/full/10.1111/rssc.12378)

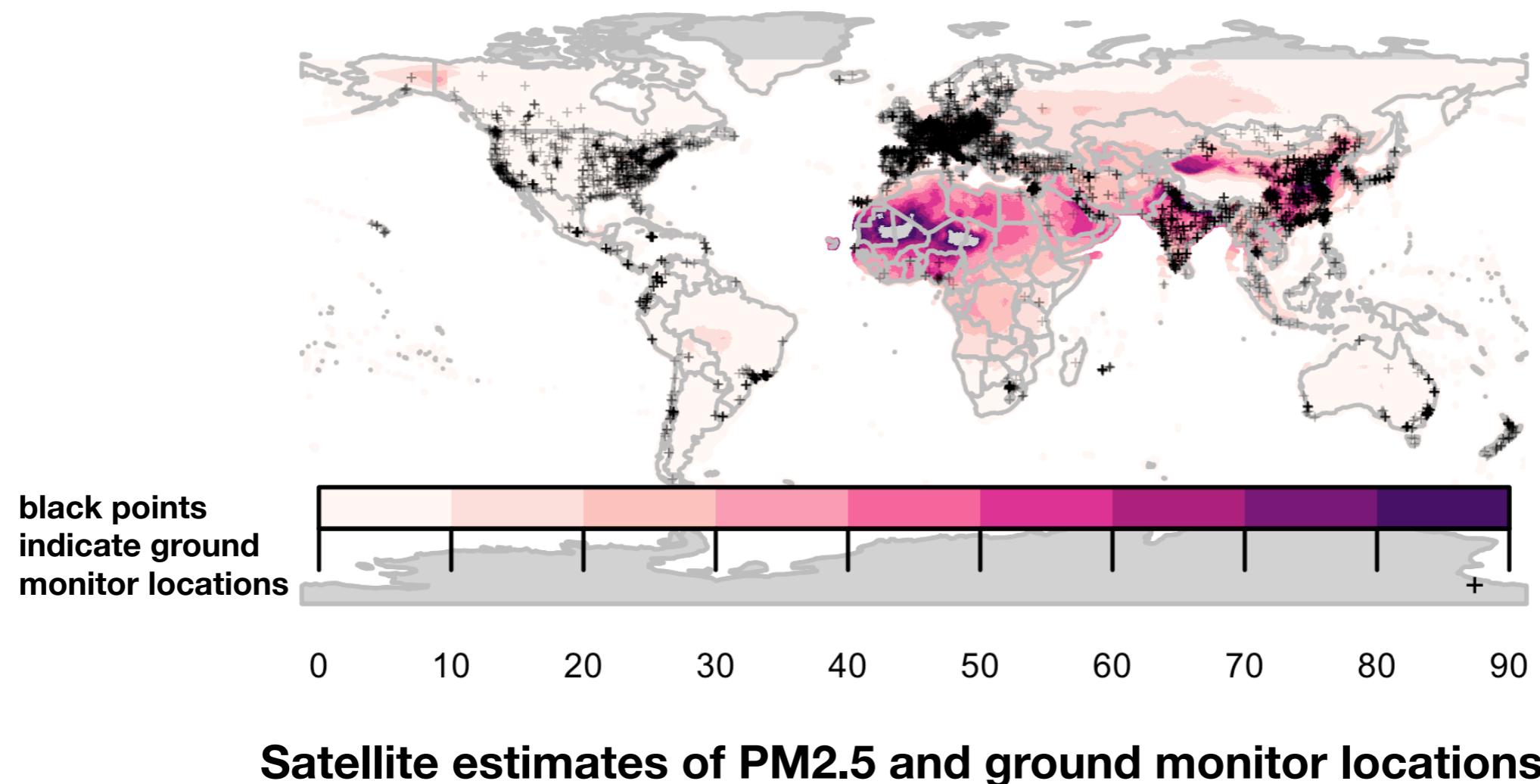
arXiv preprint: [arxiv.org/abs/1709.01449](https://arxiv.org/abs/1709.01449)

Code: [github.com/jgabry/bayes-vis-paper](https://github.com/jgabry/bayes-vis-paper)

# Example

**Goal** Estimate global PM2.5 concentration

**Problem** Most data from noisy satellite measurements (ground monitor network provides sparse, heterogeneous coverage)

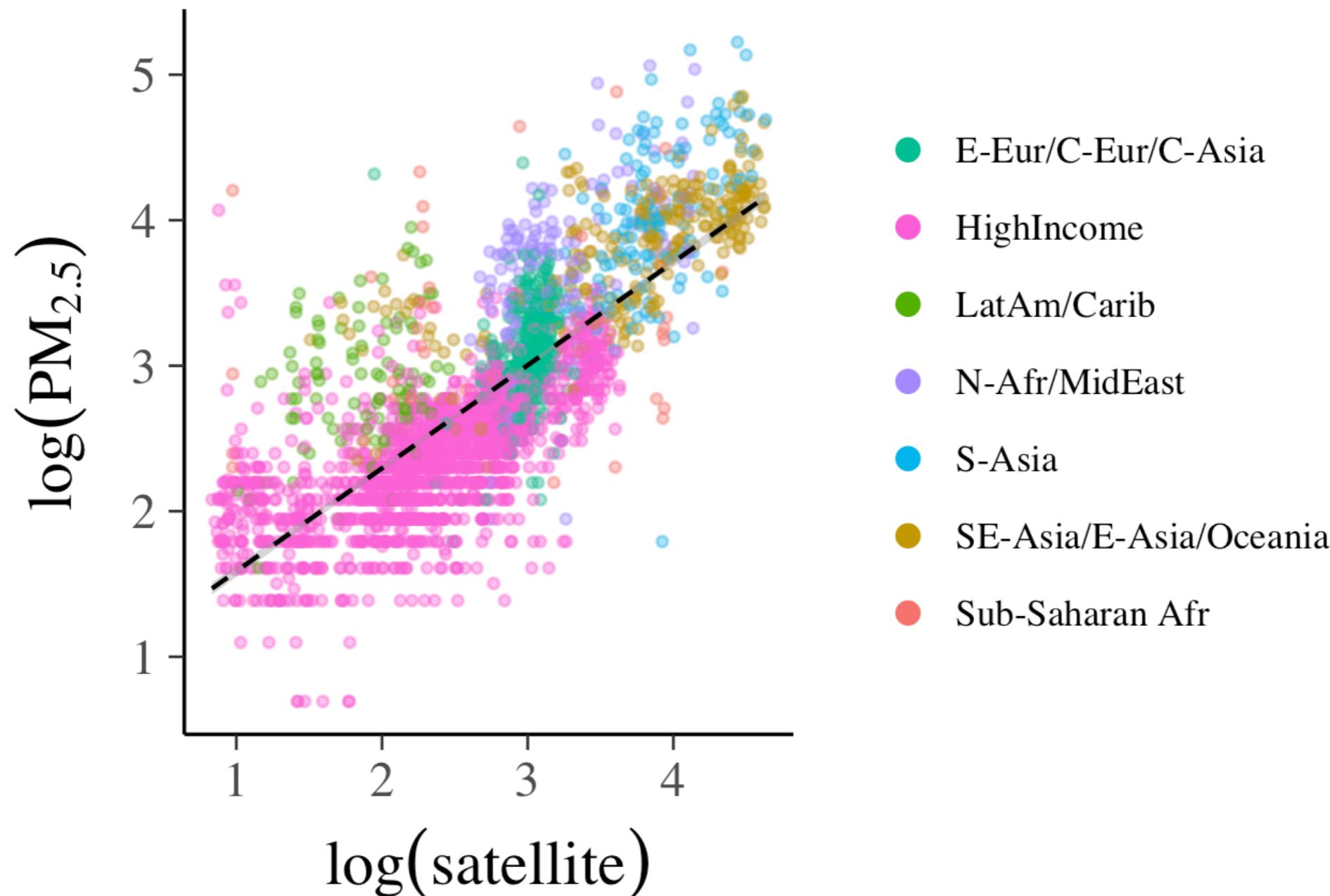


# Exploratory Data Analysis

*Building a network of models*

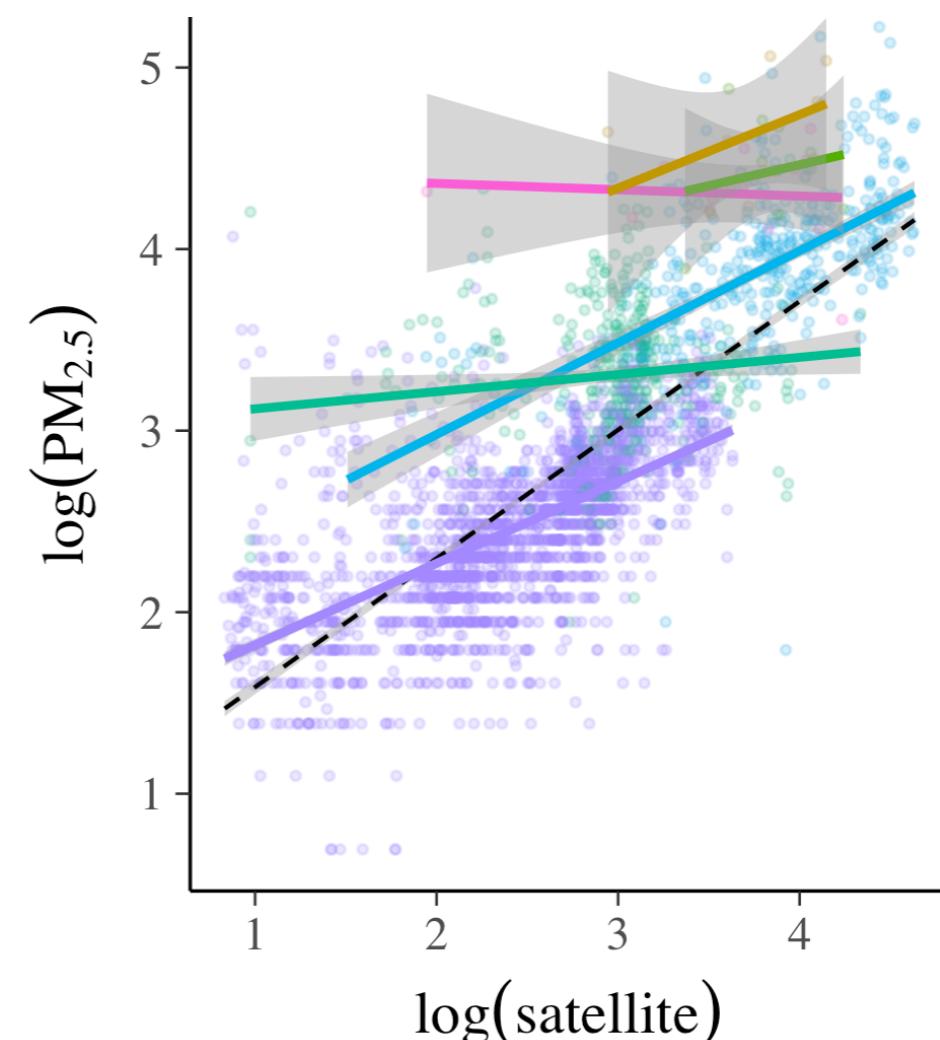
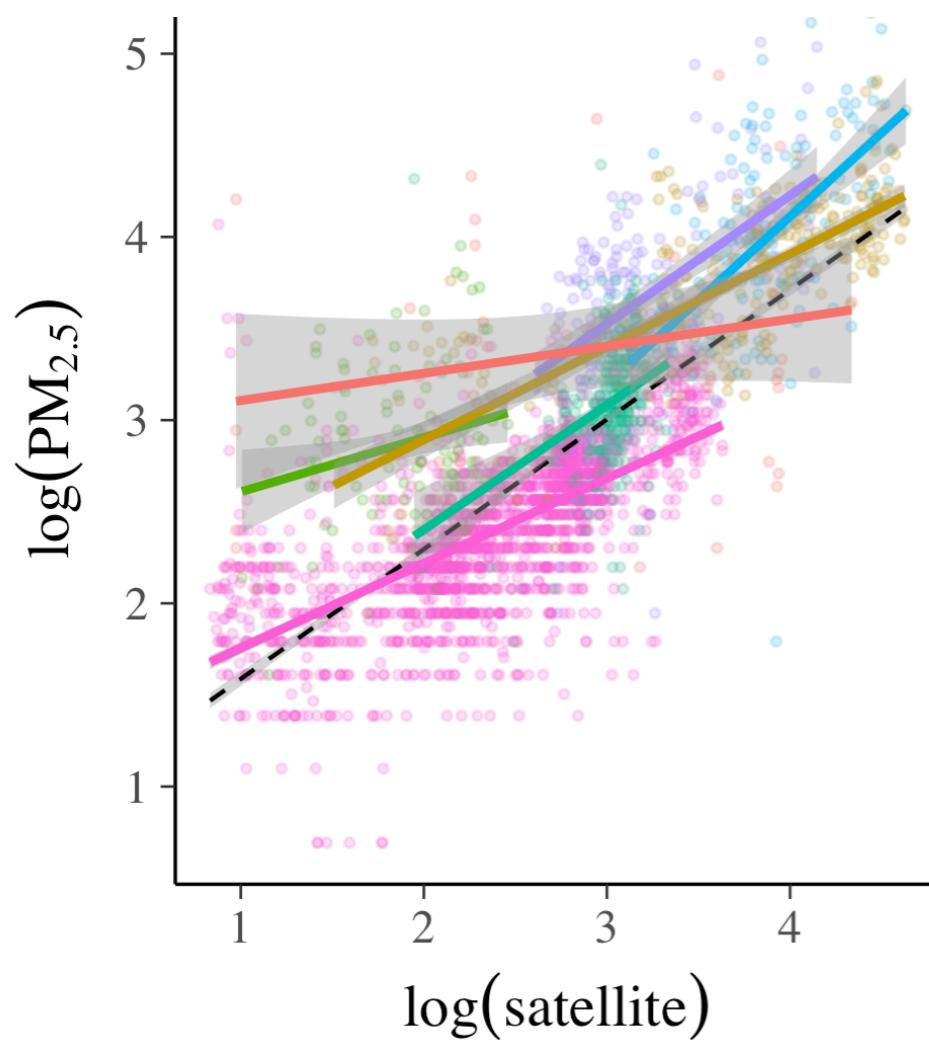
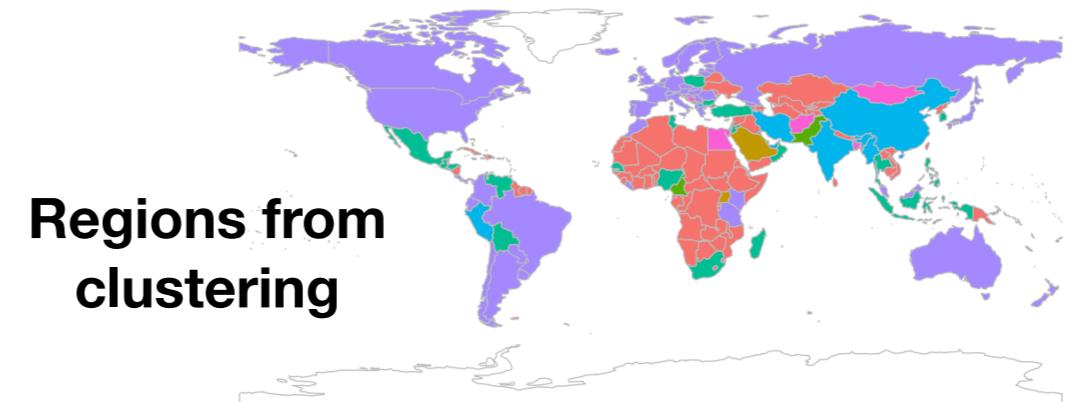
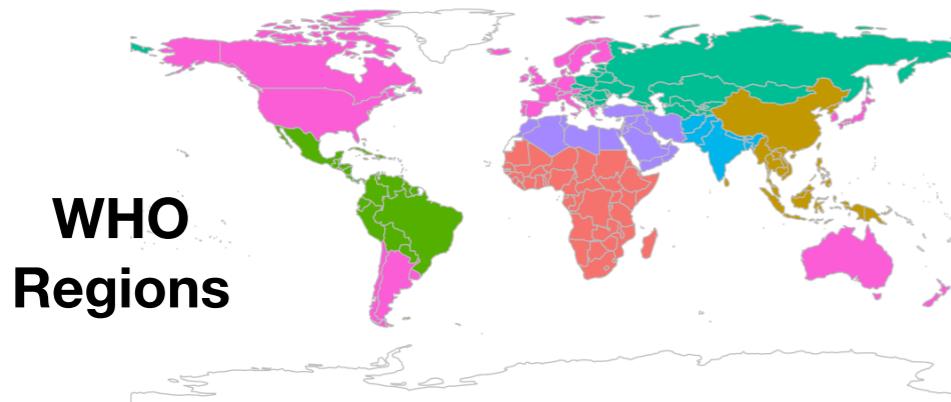
# Exploratory data analysis

## building a network of models



# Exploratory data analysis

## building a network of models



# Exploratory data analysis

## building a network of models

For measurements  $n = 1, \dots, N$

and regions  $j = 1, \dots, J$

### Model 1

$$\log(\text{PM}_{2.5,nj}) \sim N(\alpha + \beta \log(\text{sat}_{nj}), \sigma)$$

# Exploratory data analysis

building a network of models

For measurements  $n = 1, \dots, N$

and regions  $j = 1, \dots, J$

## Models 2 and 3

$$\log(\text{PM}_{2.5,nj}) \sim N(\mu_{nj}, \sigma)$$

$$\mu_{nj} = \boxed{\alpha_0 + \alpha_j} + \boxed{(\beta_0 + \beta_j)} \log(\text{sat}_{nj})$$

$$\alpha_j \sim N(0, \tau_\alpha) \quad \beta_j \sim N(0, \tau_\beta)$$

# Prior predictive checks

*Fake data can be almost as valuable as real data*

# Prior predictive checking: fake data is almost as useful as real data

- Used to understand the role of the prior in the data generating process
- Even people who don't have strong feelings about a specific prior know what their observed data shouldn't look like (i.e., pollution levels aren't so thick that we can't move, so the model should reflect that)
  1. Use either observed x or generate reasonable values for x
  2. Sample from your prior (assuming a proper prior)
  3. Calculate the predicted y for that x and that particular random draw from the prior
  4. Repeat 2 & 3 many times
  5. Compute summaries / visualize

# Prior predictive checking: fake data is almost as useful as real data

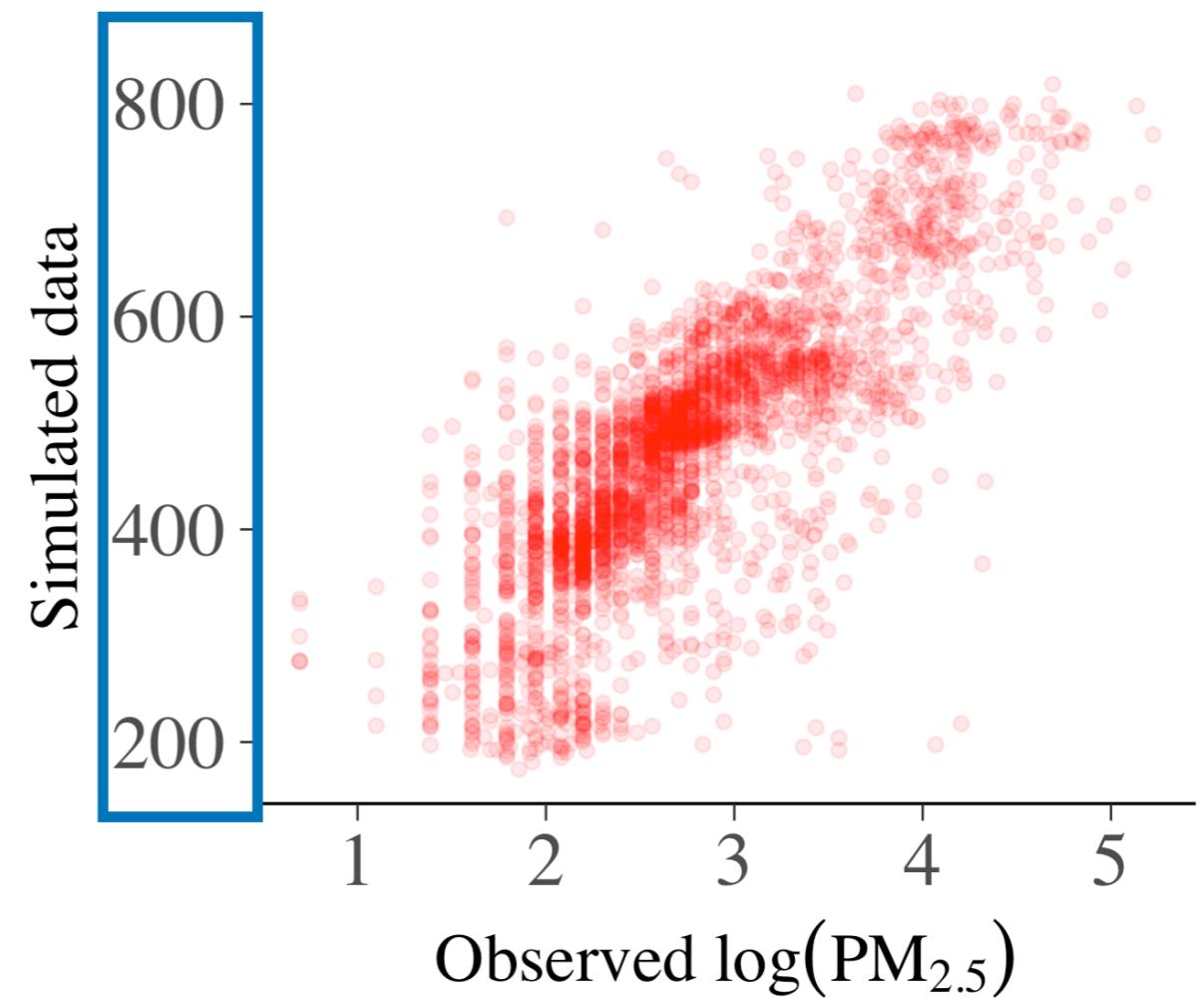
*What do vague/non-informative priors imply  
about the data our model can generate?*

$$\alpha_0 \sim N(0, 100)$$

$$\beta_0 \sim N(0, 100)$$

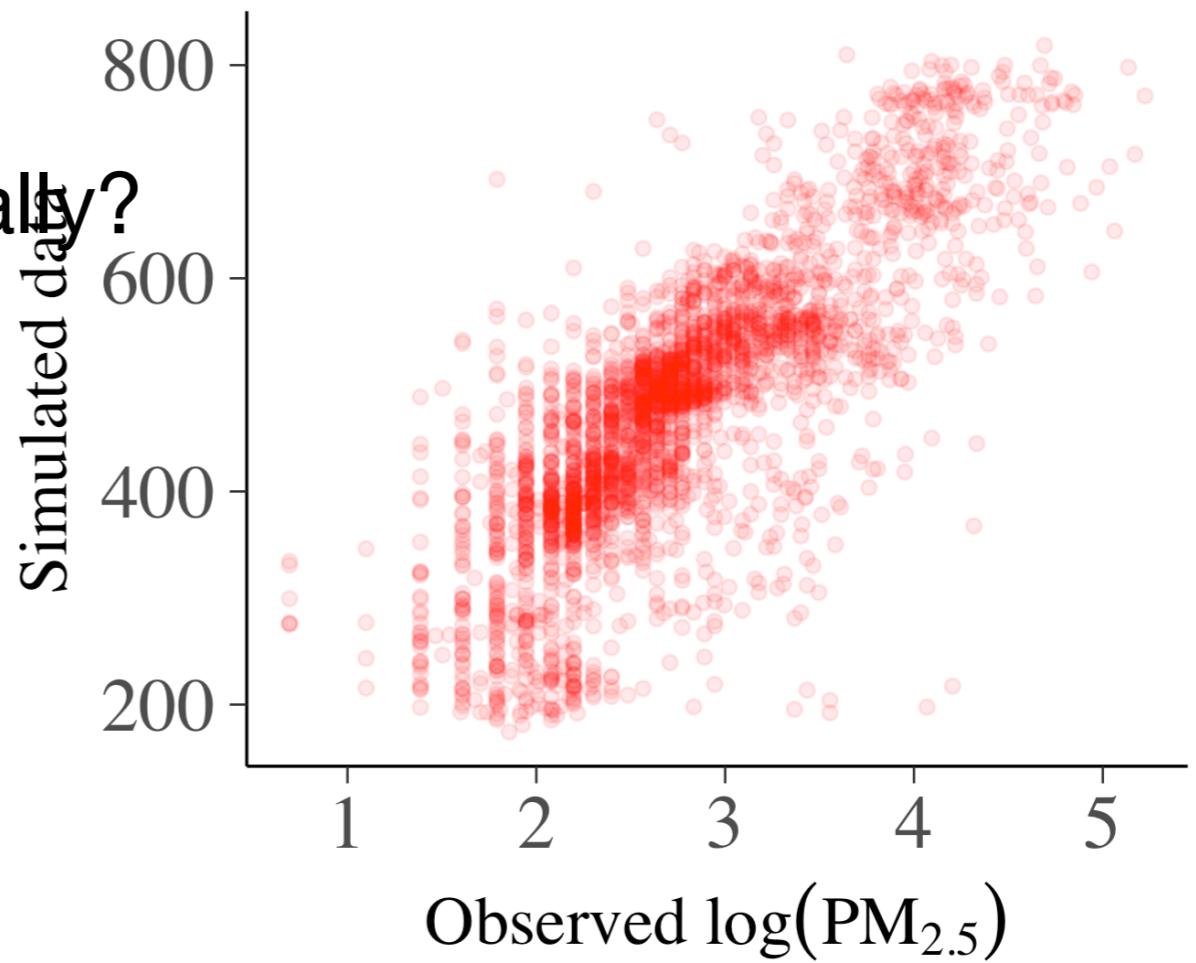
$$\tau_\alpha^2 \sim \text{InvGamma}(1, 100)$$

$$\tau_\beta^2 \sim \text{InvGamma}(1, 100)$$



# Prior predictive checking: fake data is almost as useful as real data

- The prior model is **two orders of magnitude** off the real data
- Two orders of magnitude **on the log scale!**
- What does this mean practically?
- The data will have to overcome the prior...



# Prior predictive checking: fake data is almost as useful as real data

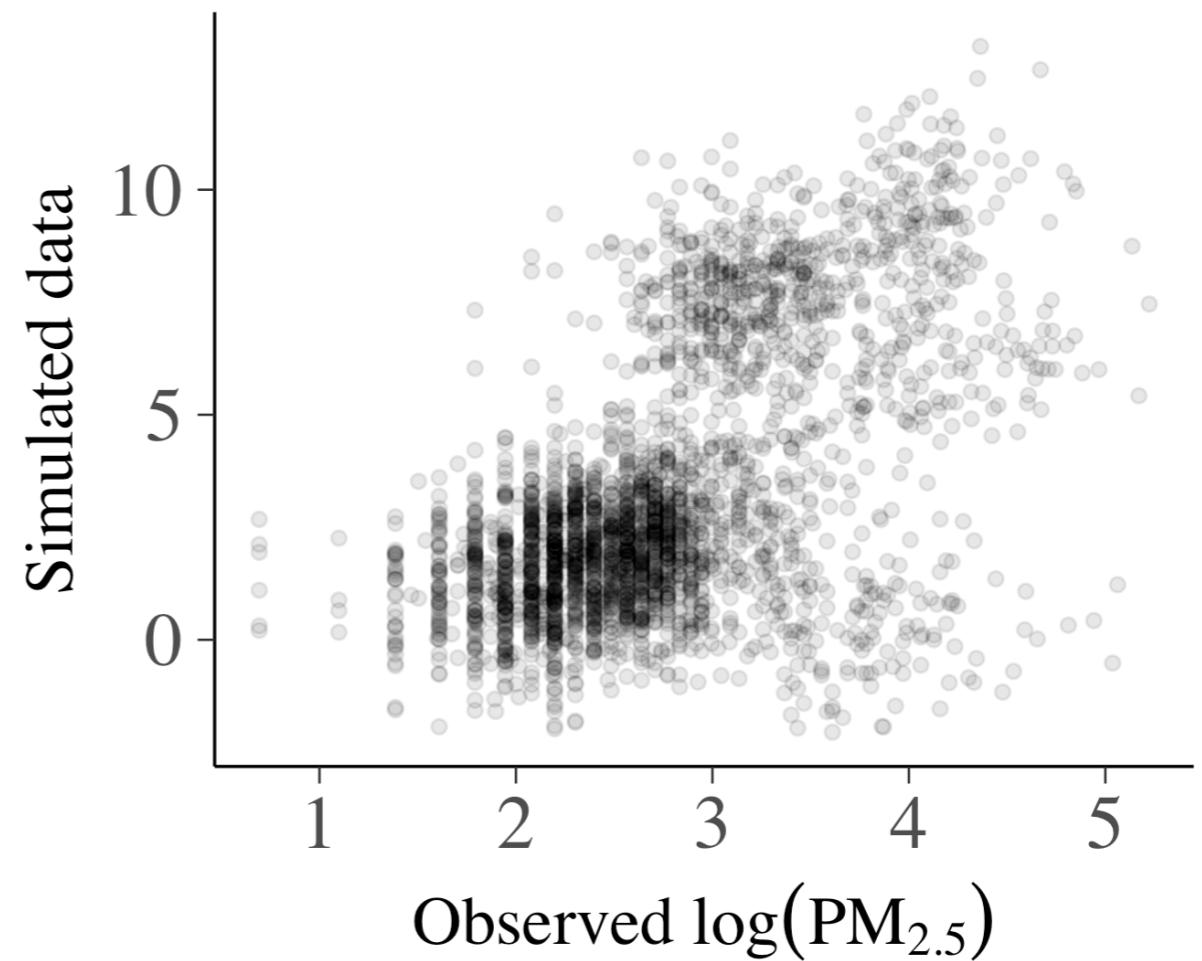
*What are better priors for the global intercept and slope  
and the hierarchical scale parameters?*

$$\alpha_0 \sim N(0, 1)$$

$$\beta_0 \sim N(1, 1)$$

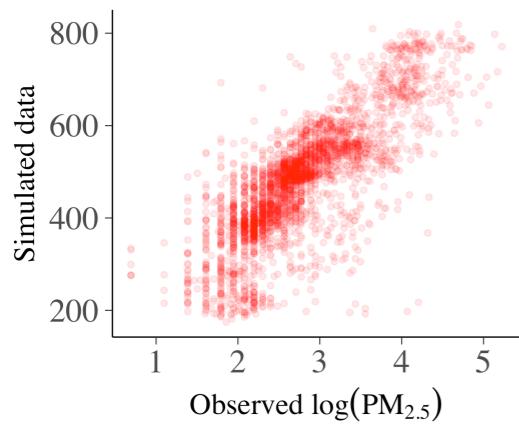
$$\tau_\alpha \sim N_+(0, 1)$$

$$\tau_\beta \sim N_+(0, 1)$$

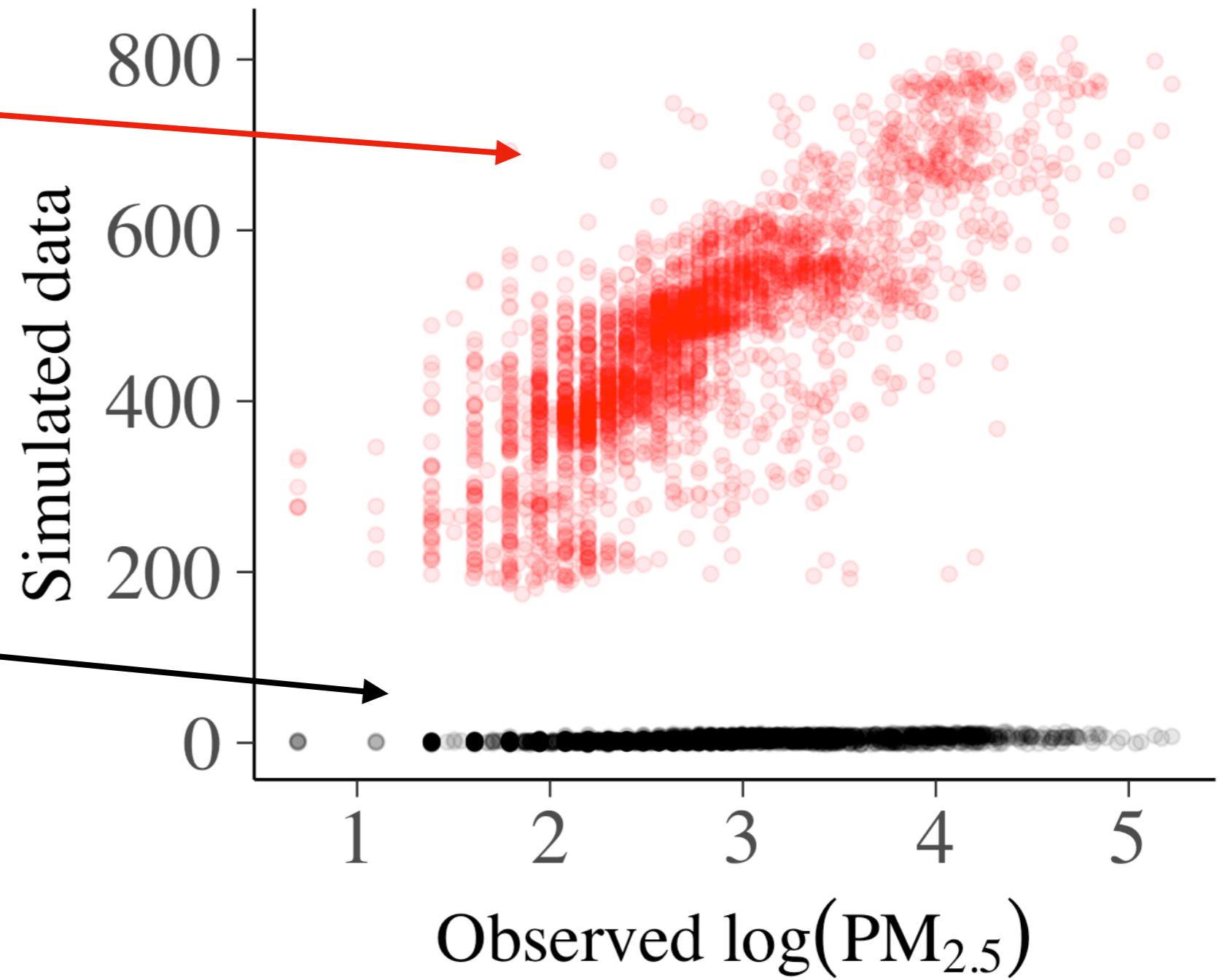
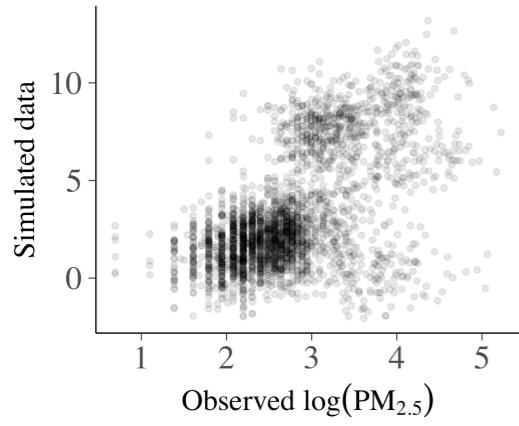


# Prior predictive checking: fake data is almost as useful as real data

**Non-informative**



**Weakly informative**



# What is the problem with “vague” priors?

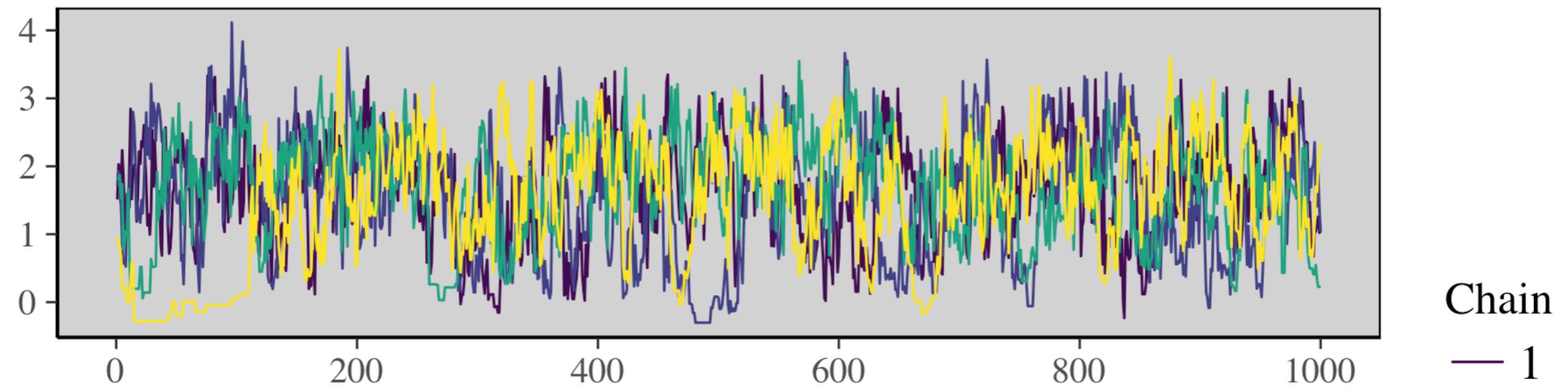
- If we use an *improper* prior, then we do not specify a joint model for our data and parameters
- More importantly, we do not specify a data generating mechanism  $p(\mathbf{y})$
- By construction, these priors do not regularize inferences, which is quite often a bad idea
- Proper but diffuse is better than improper but is still often problematic

# MCMC diagnostics

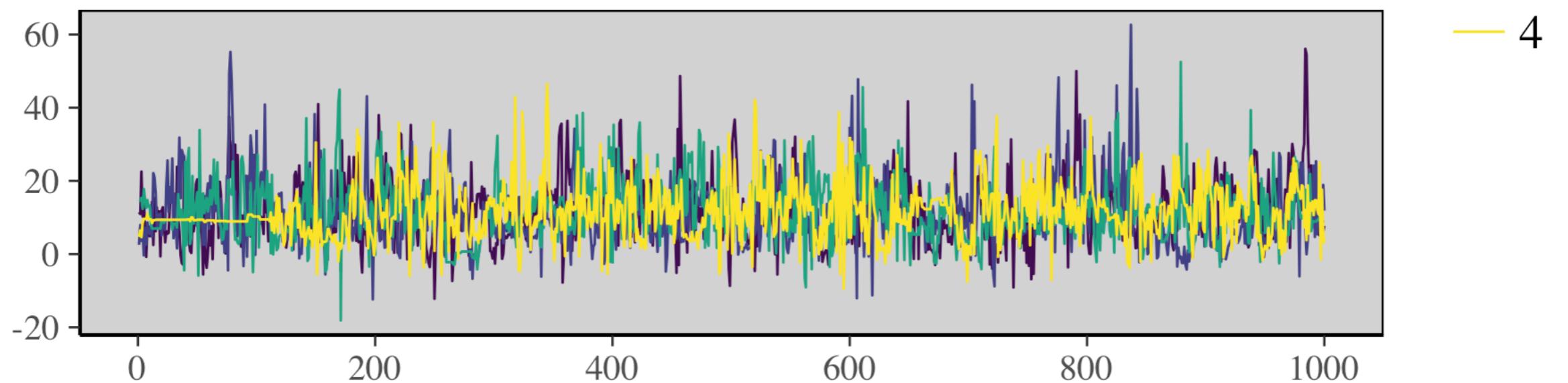
*Beyond trace plots*

<https://chi-feng.github.io/mcmc-demo/>

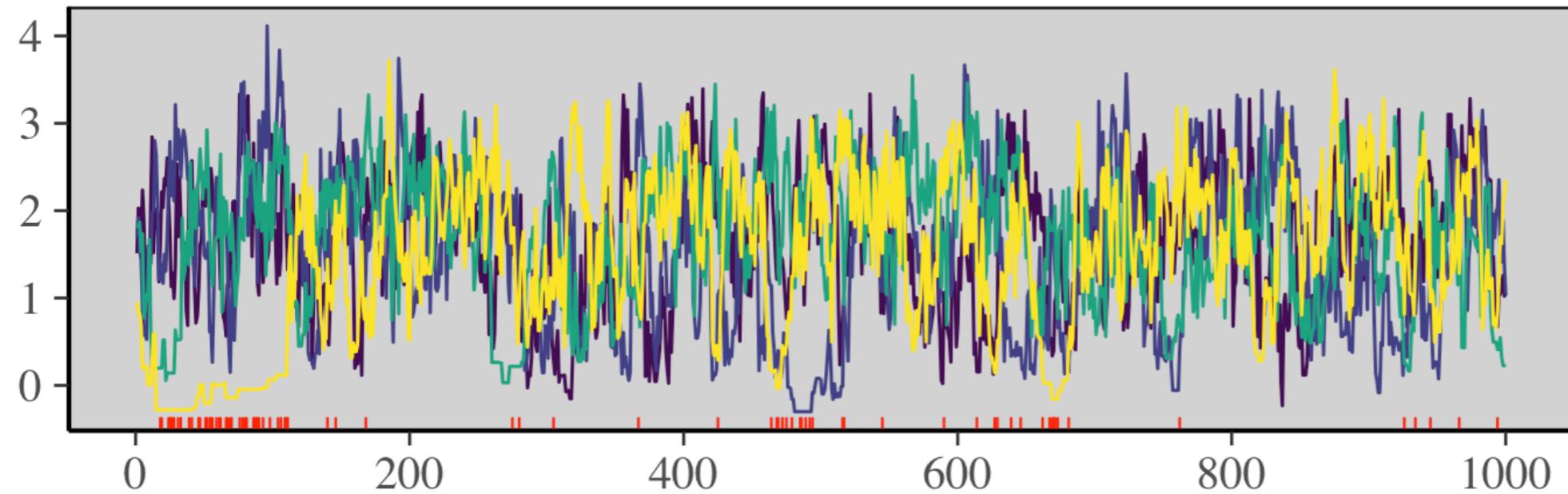
$\log(\tau)$



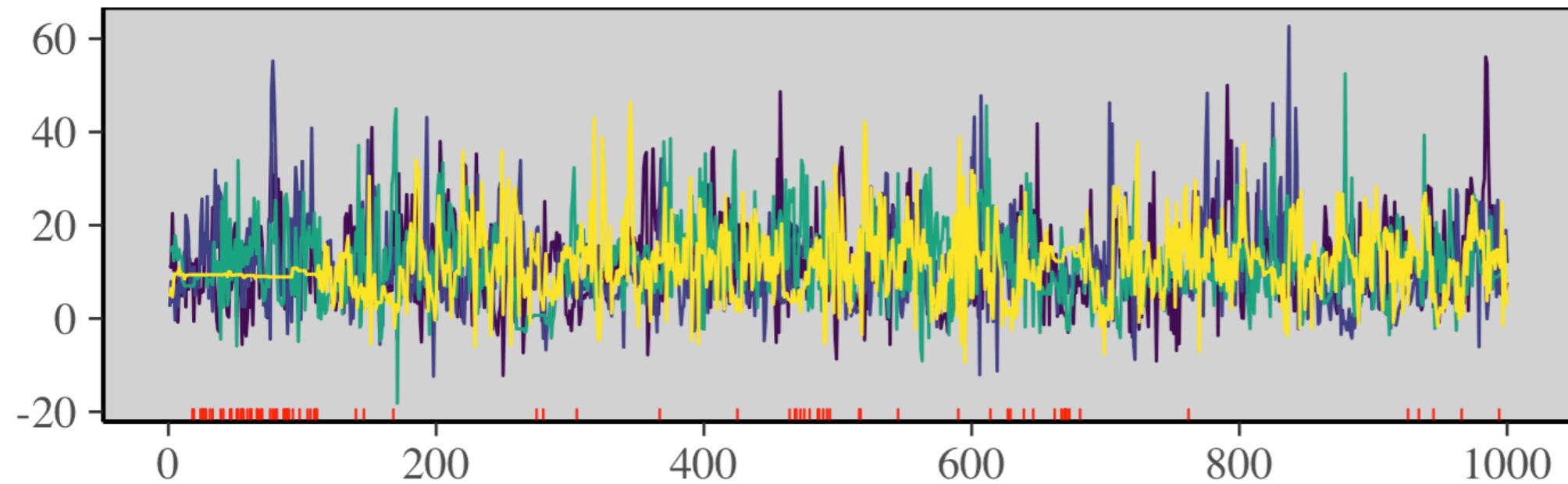
$\theta[1]$



$\log(\tau)$



$\theta[1]$



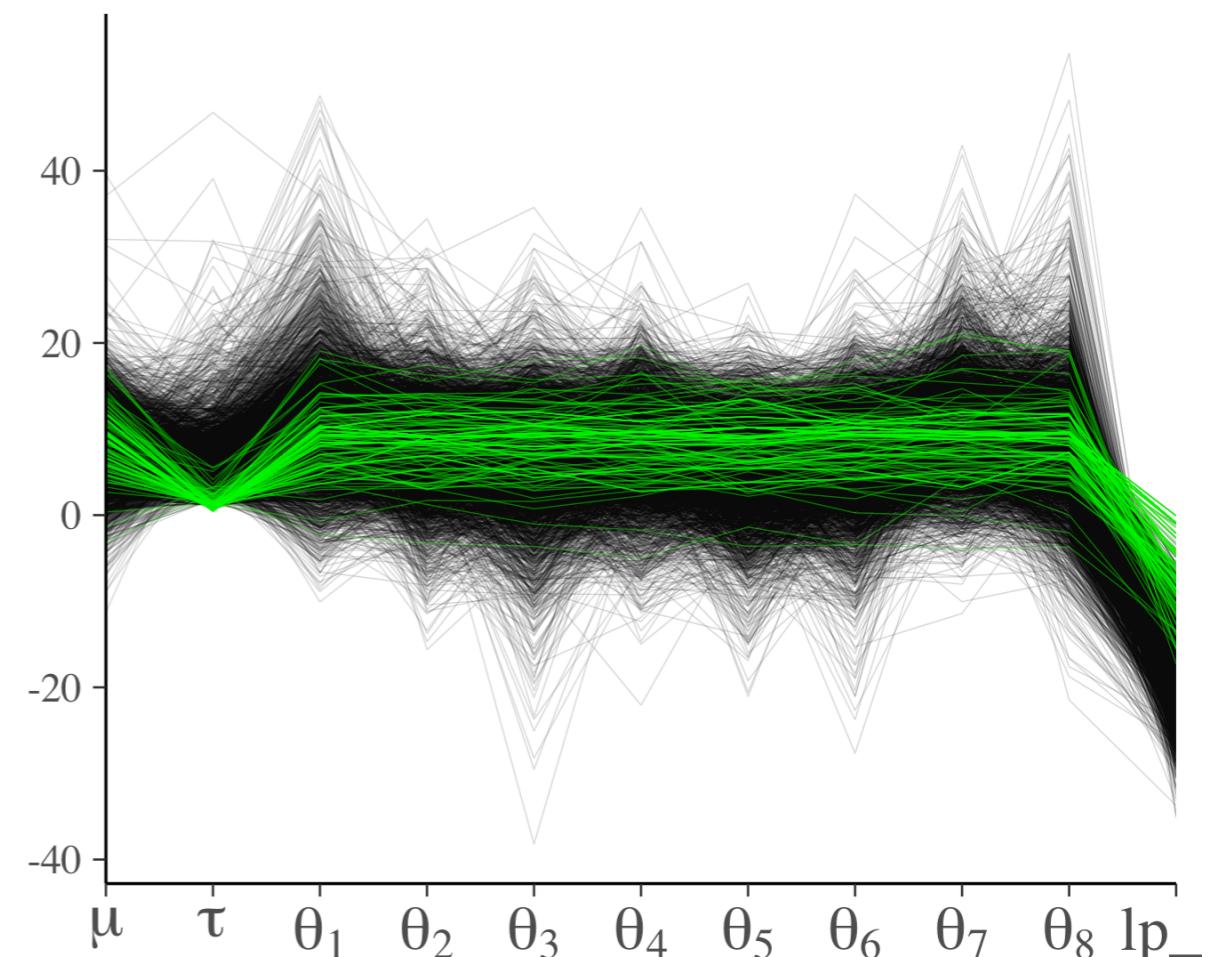
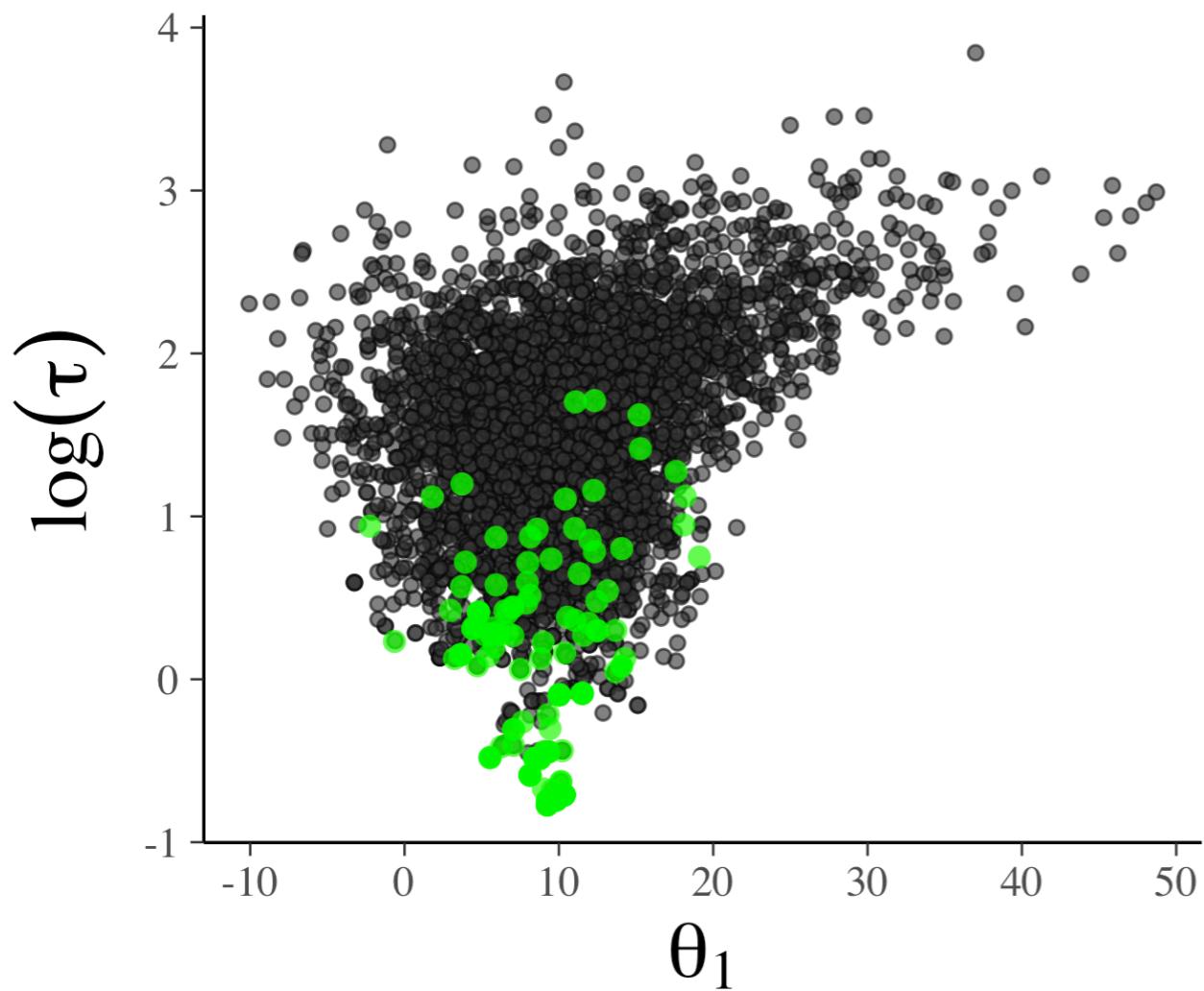
Chain

- 1
- 2
- 3
- 4

— Divergence

# MCMC diagnostics

## beyond trace plots

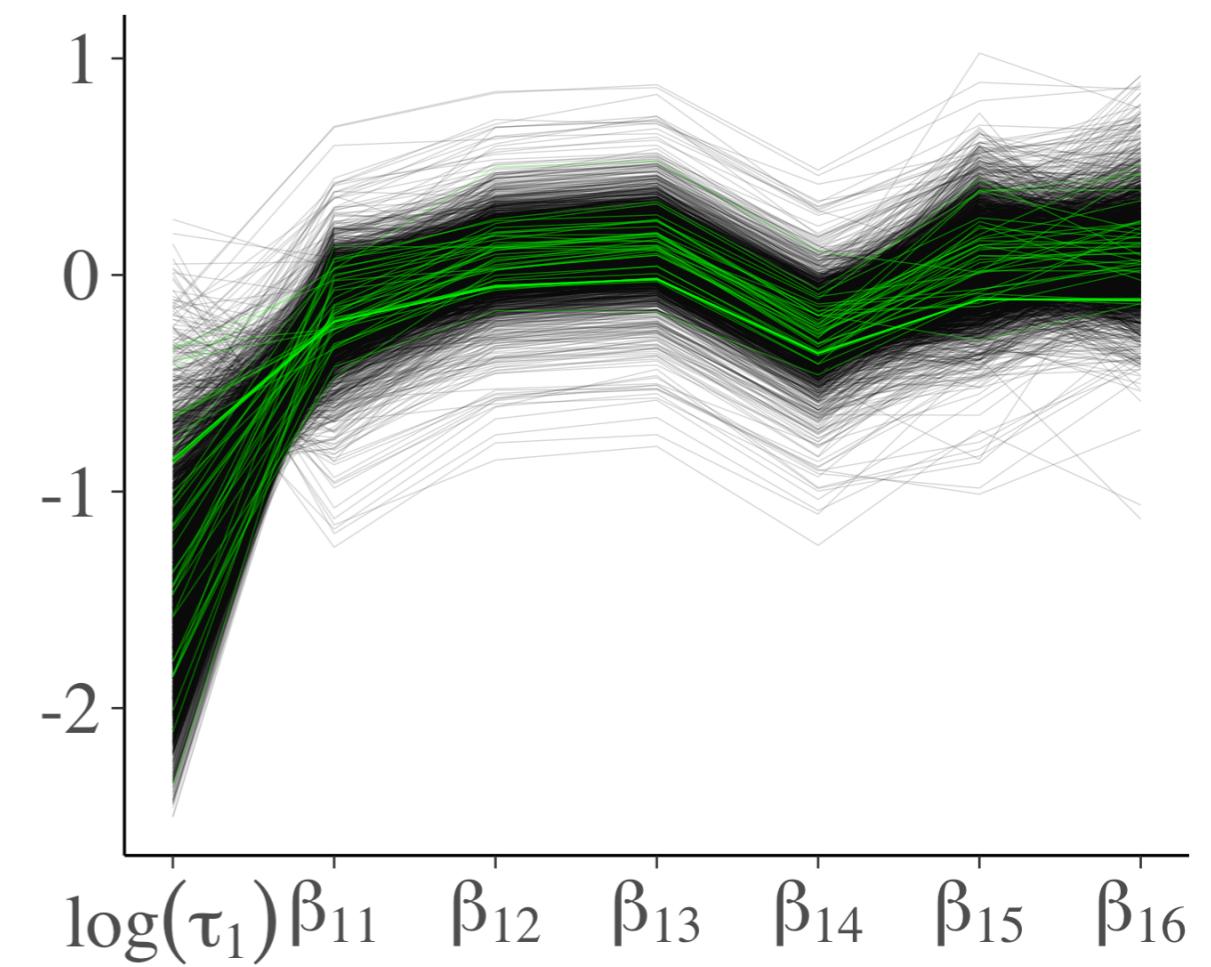
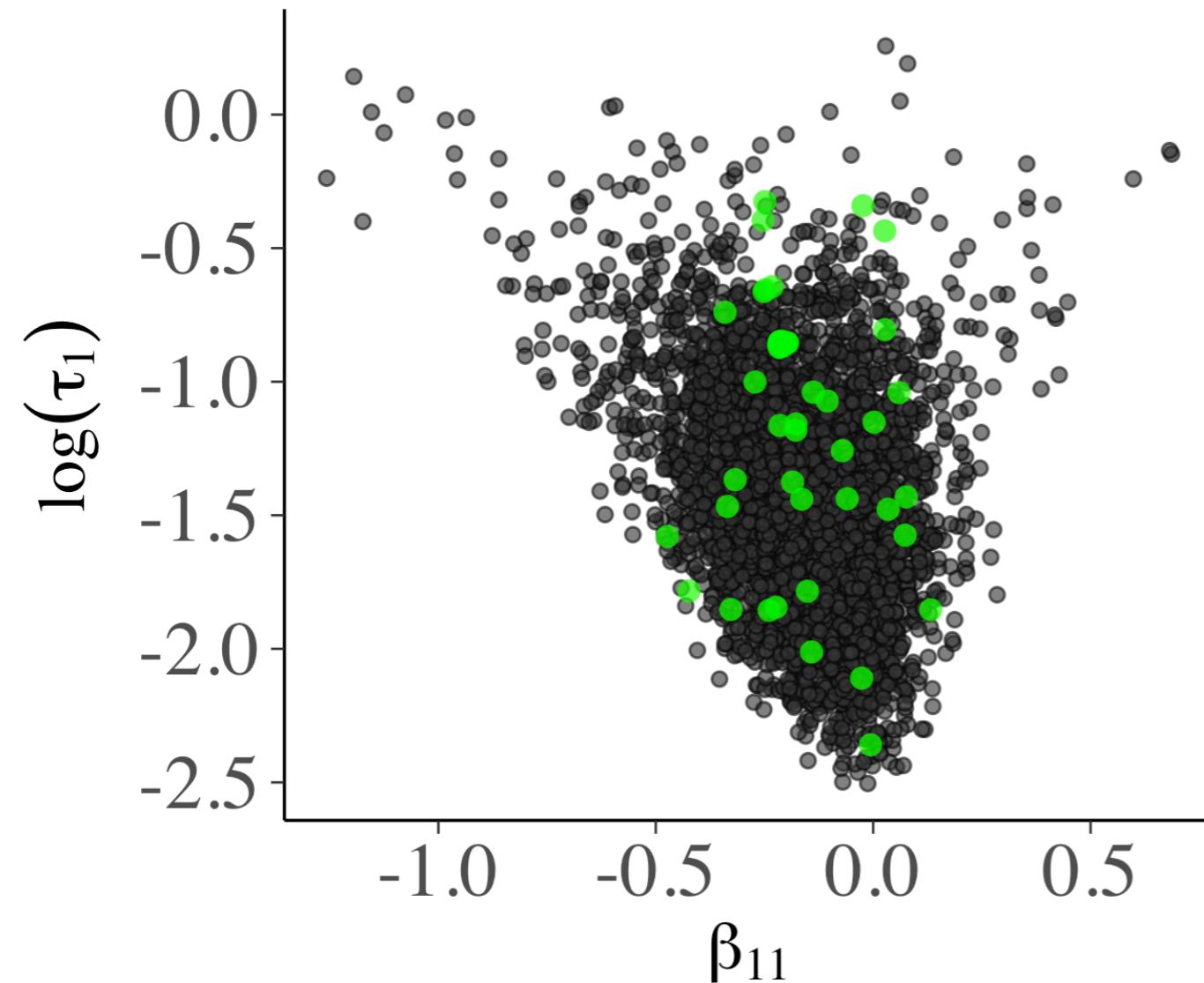


Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2018).  
**Visualization in Bayesian workflow.**  
*Journal of the Royal Statistical Society Series A*, accepted for publication.  
[arxiv.org/abs/1709.01449](https://arxiv.org/abs/1709.01449) | [github.com/jgabry/bayes-vis-paper](https://github.com/jgabry/bayes-vis-paper)

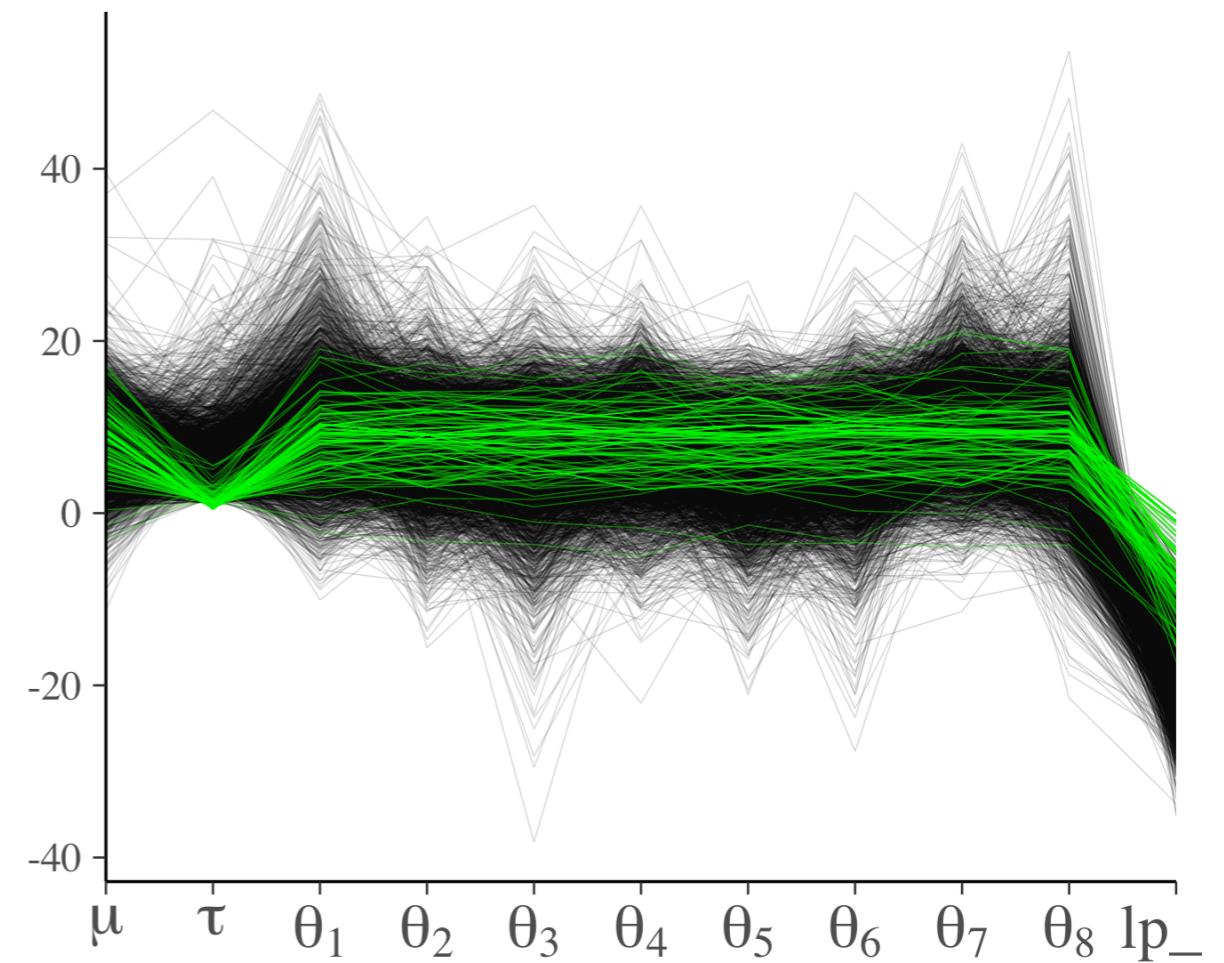
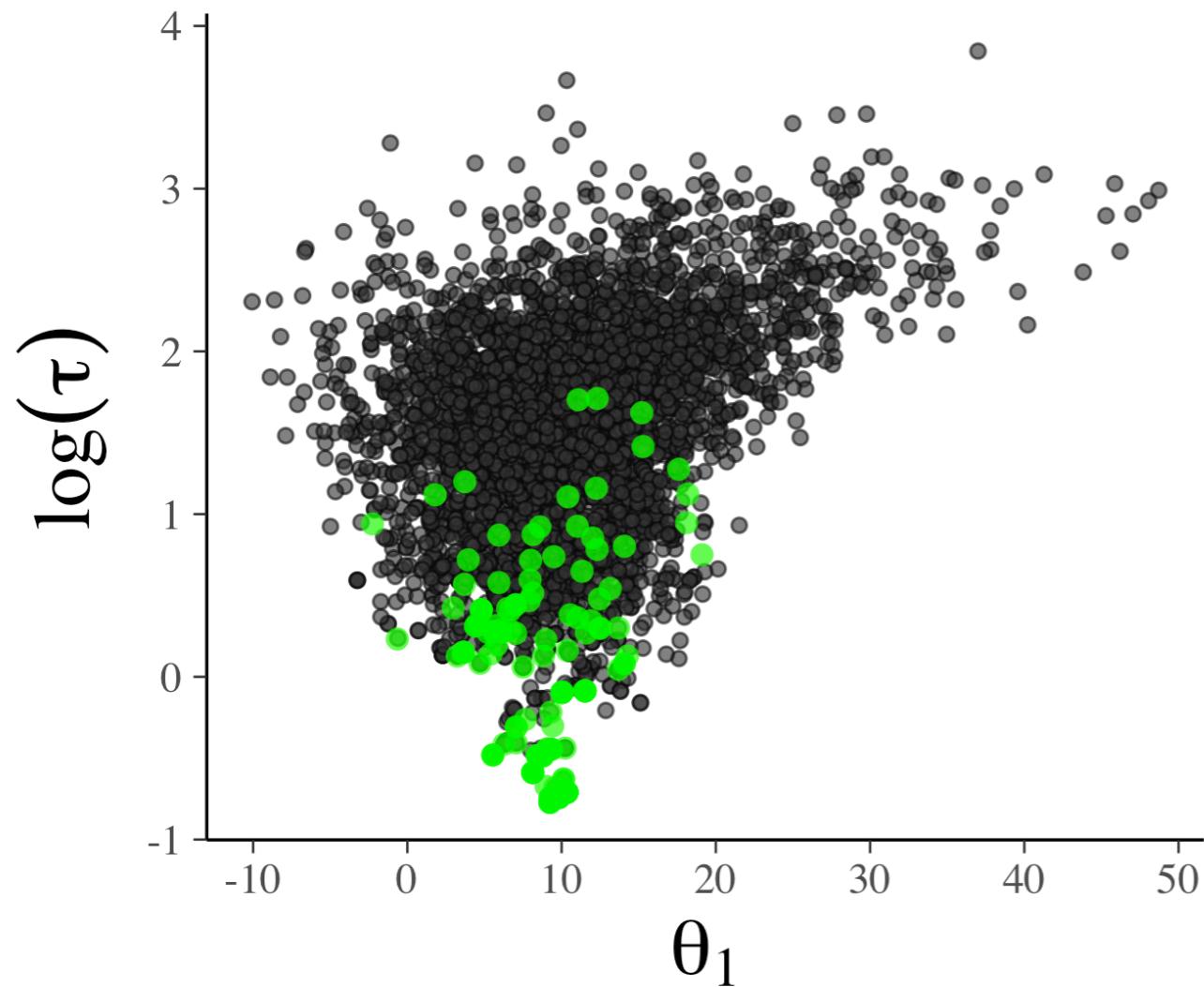
Betancourt, M. (2017).  
**A conceptual introduction to Hamiltonian Monte Carlo.**  
arXiv preprint:  
[arxiv.org/abs/1701.02434](https://arxiv.org/abs/1701.02434)

# MCMC diagnostics

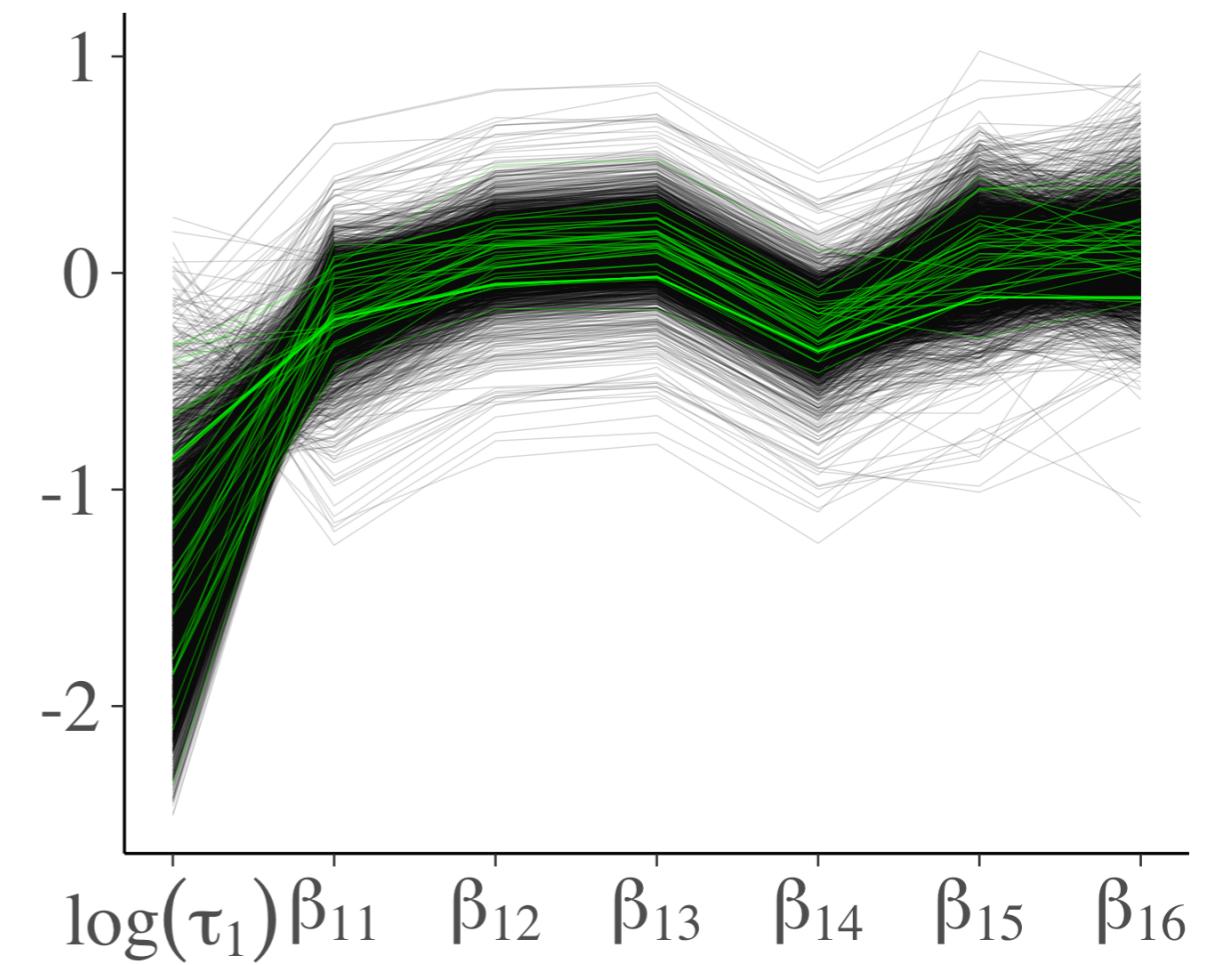
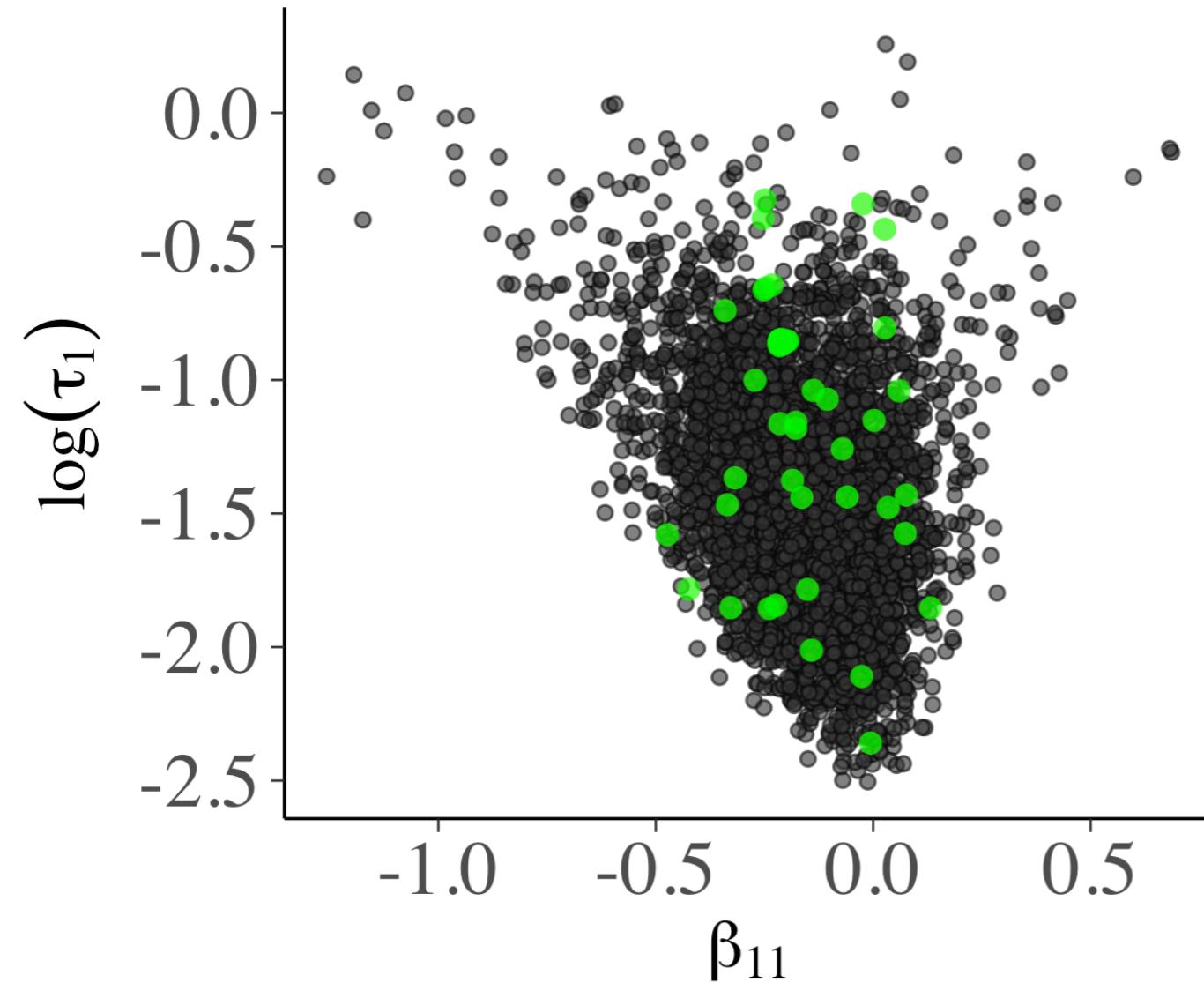
## beyond trace plots



# Pathological geometry



# “False positives”



# Posterior predictive checks

*Visual model evaluation*

# **Posterior predictive checking**

## visual model evaluation

- Misfitting and overfitting both manifest as tension between measurements and predictive distributions
- Graphical posterior predictive checks visually compare the observed data to the predictive distribution

# **Posterior predictive checking**

## visual model evaluation

The *posterior predictive distribution* is the average data generation process over the entire model

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta) p(\theta|y) d\theta$$

# Posterior predictive checking

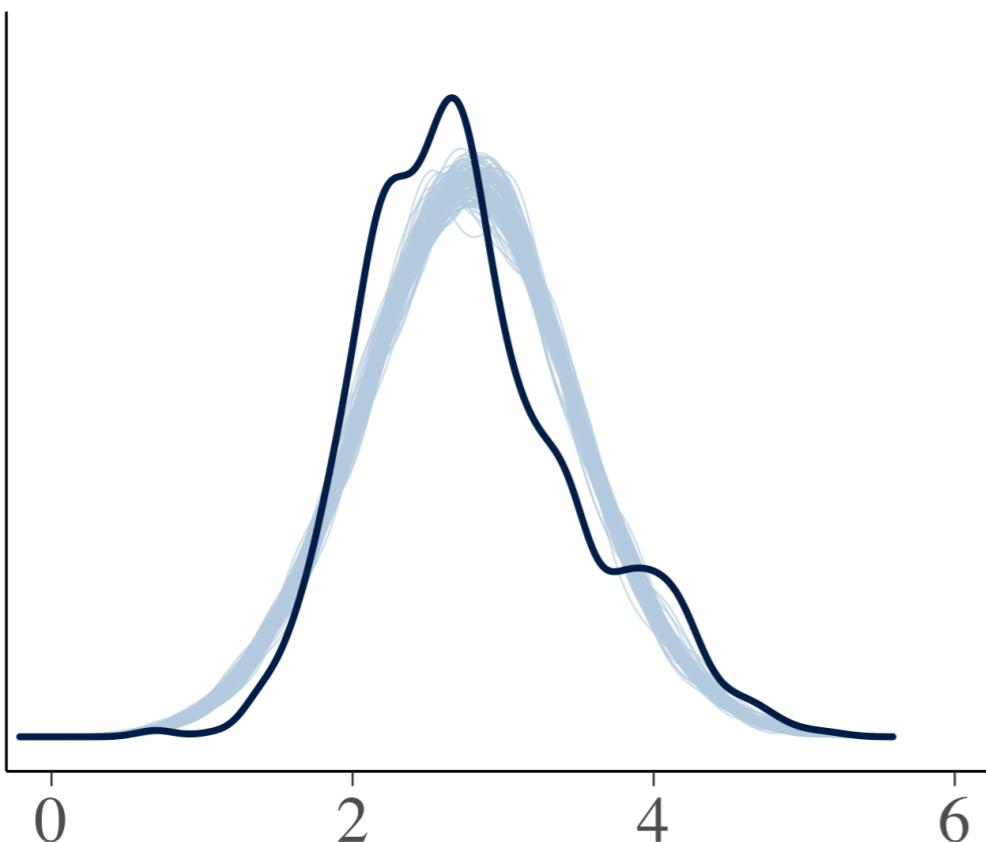
- Used to understand the fit of the model
- Procedure:
  1. Sample from your posterior
  2. Use observed  $x$
  3. Calculate the predicted  $y$  for that  $x$  and that particular random draw from the posterior
  4. Repeat 2 & 3 many times
  5. Compute summaries / visualize

# Posterior predictive checking

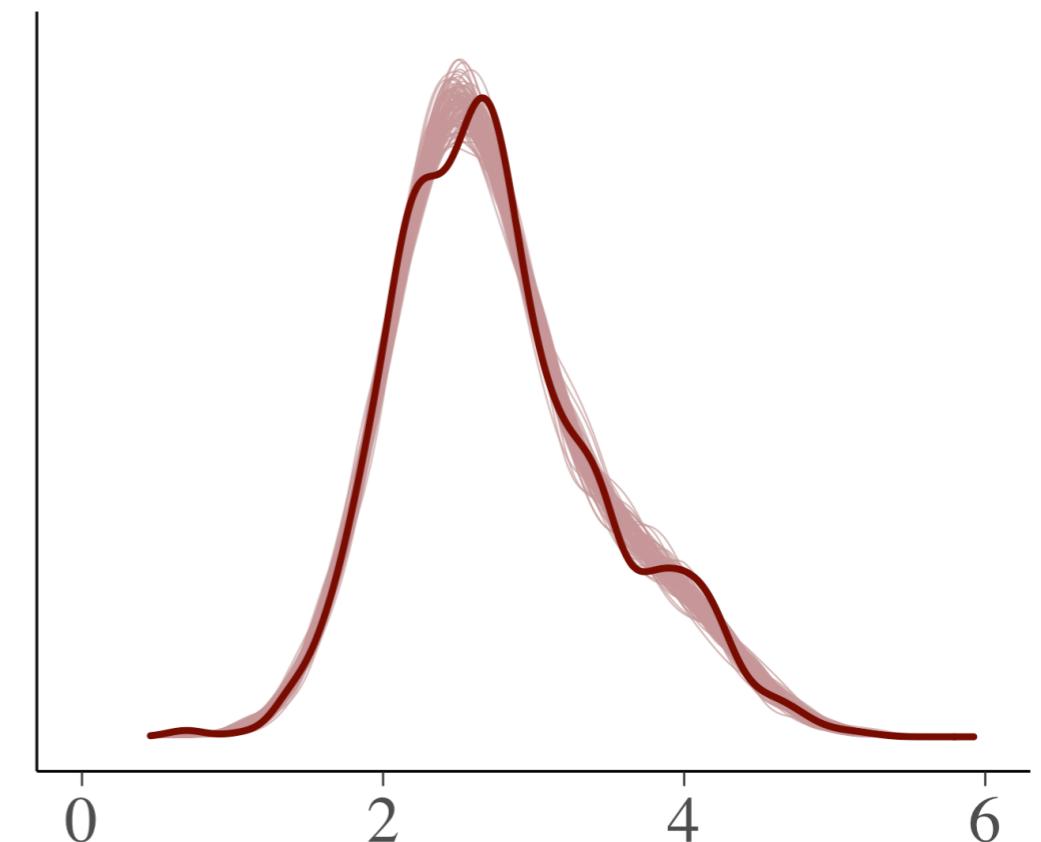
## visual model evaluation

Observed data vs posterior predictive simulations

**Model 1 (single level)**



**Model 3 (multilevel)**

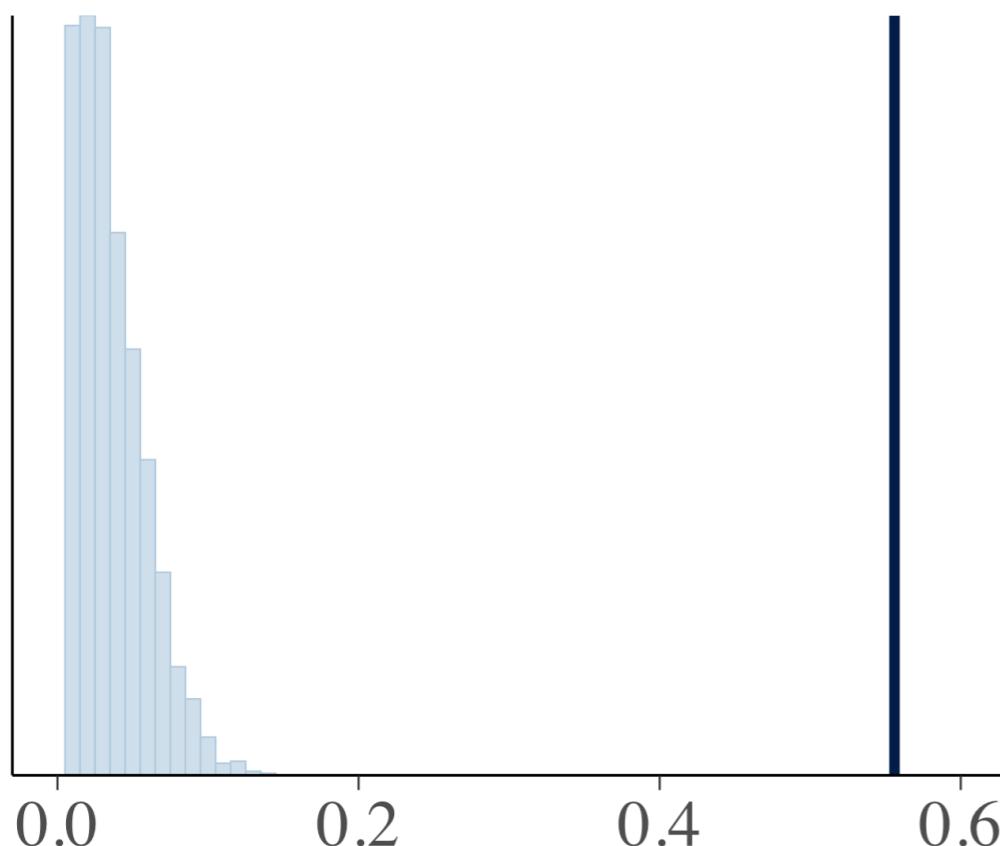


# Posterior predictive checking

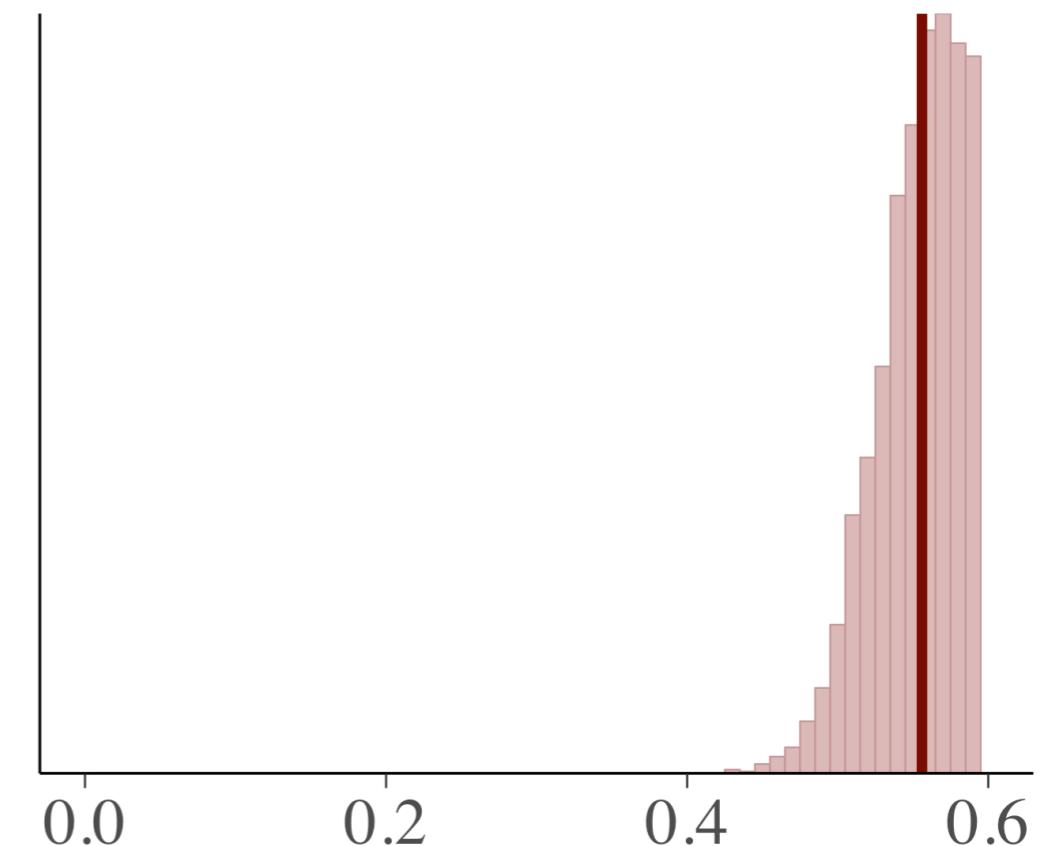
## visual model evaluation

Observed statistics vs posterior predictive statistics

**Model 1 (single level)**

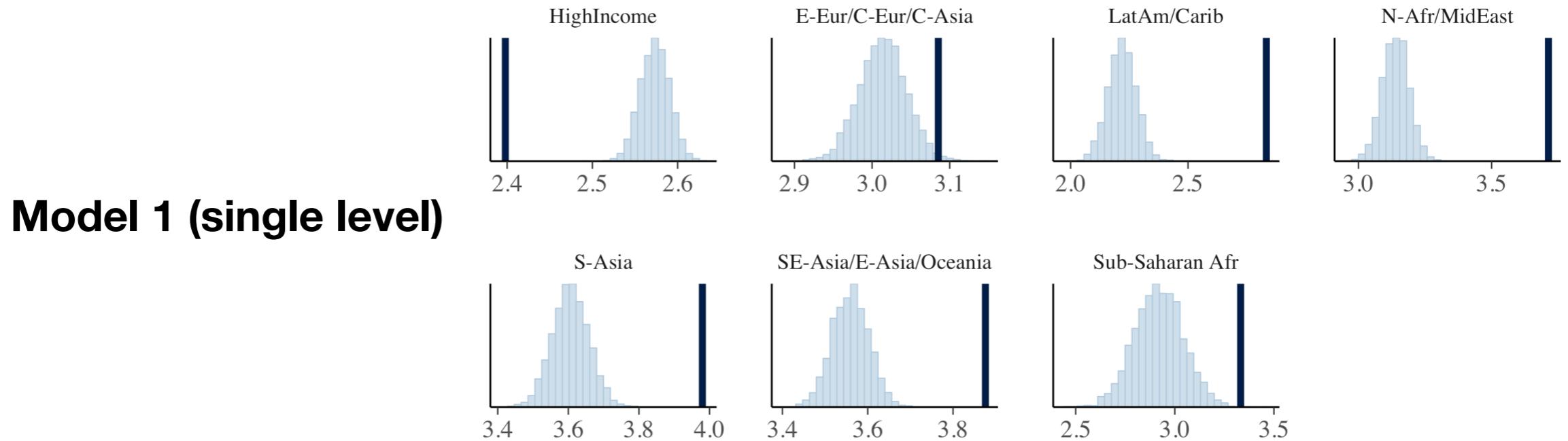


**Model 3 (multilevel)**

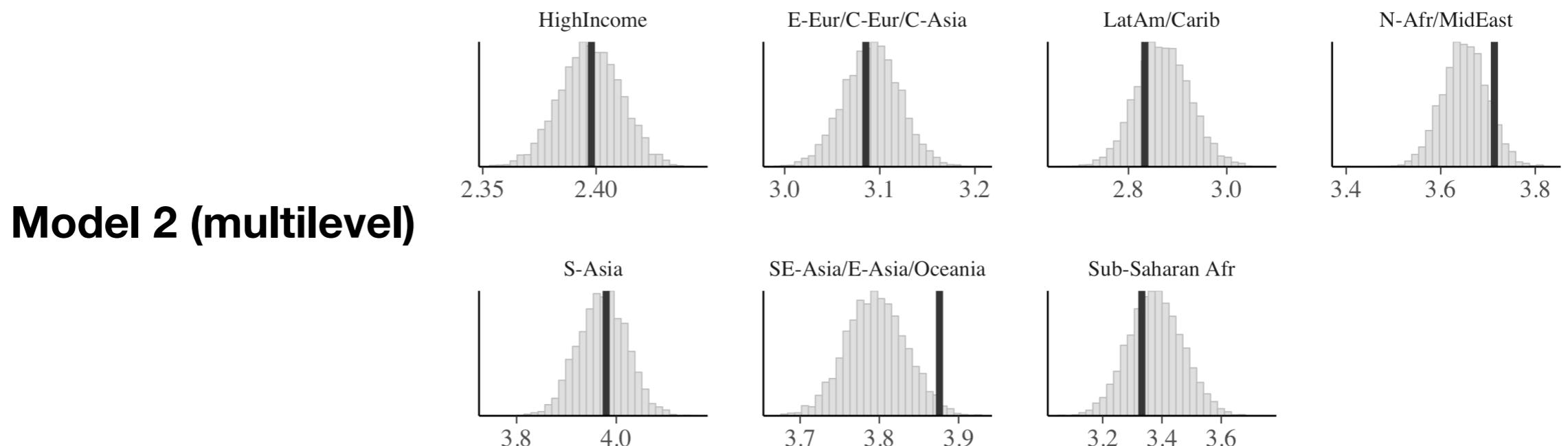


$$T(y) = \text{skew}(y)$$

# Posterior predictive checking: visual model evaluation



$$T(y) = \text{med}(y|\text{region})$$



# Model comparison

*Pointwise predictive comparisons & LOO-CV*

# Model comparison

## pointwise predictive comparisons & LOO-CV

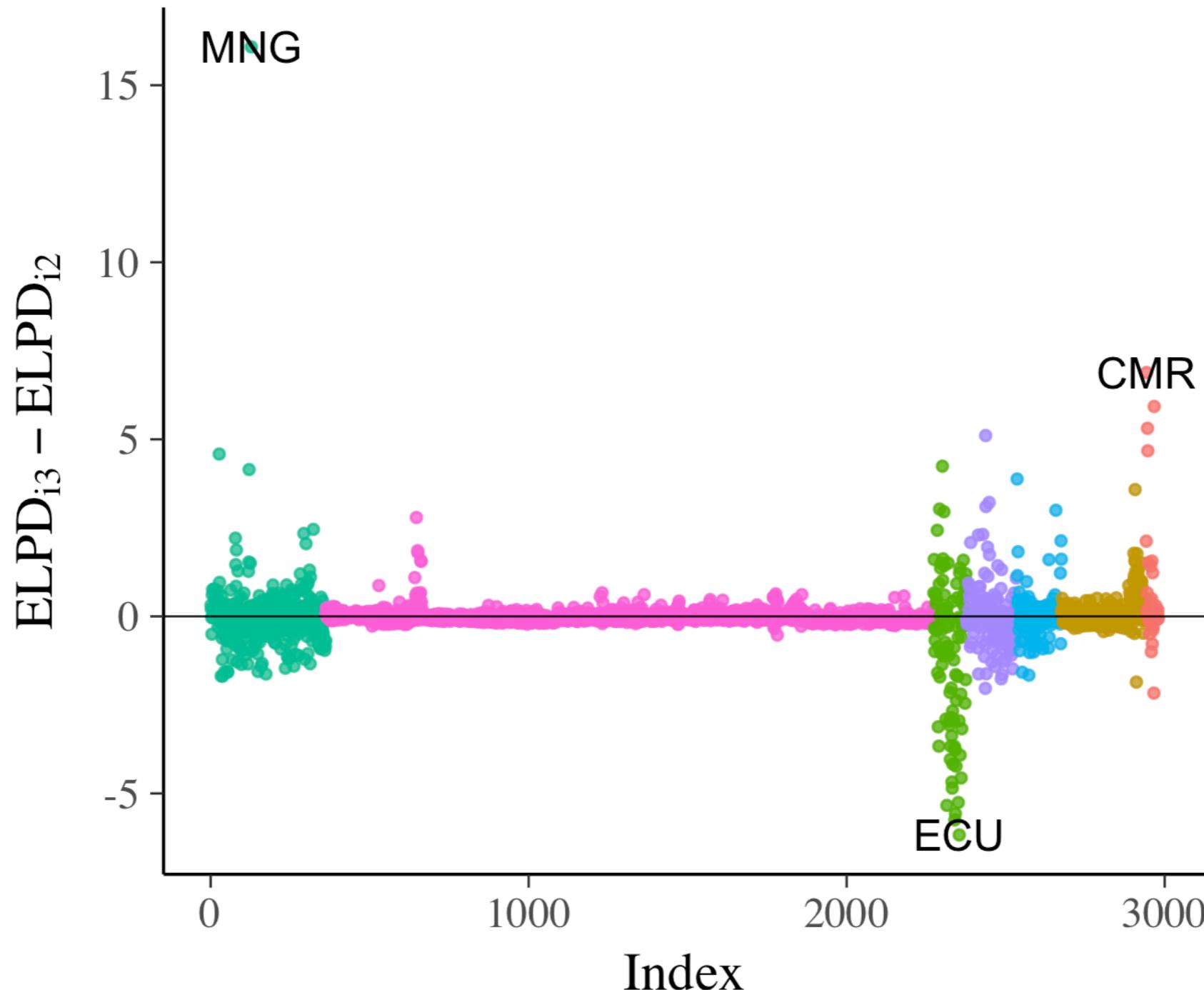
- Visual PPCs can also identify unusual/influential (outliers, high leverage) data points
- We like using cross-validated leave-one-out predictive distributions

$$p(y_i | y_{-i})$$

- Which model best predicts each of the data points that is left out?

# Model comparison

## pointwise predictive comparisons & LOO-CV



# Model comparison

## Efficient approximate LOO-CV

- How do we compute LOO-CV without fitting the model  $N$  times?
- Fit once, then use Pareto smoothed importance sampling (PSIS-LOO)
- Has finite variance property of truncated IS
- And less bias (replace largest weights with order stats of generalized Pareto)
- Assumes posterior not highly sensitive to leaving out single observations
- Asymptotically equivalent to WAIC
- Advantage: PSIS-LOO CV more robust + has diagnostics (check assumptions)

---

Vehtari, A., Gelman, A., and Gabry, J. (2017).

**Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.**

*Statistics and Computing*. 27(5), 1413–1432.

doi: [10.1007/s11222-016-9696-4](https://doi.org/10.1007/s11222-016-9696-4)

# Diagnostics

Shape parameter of generalized Pareto distribution & influential observations

