SOMETIMES HAVING A CONTINUOUS FORMULATION IS USEFUL. SOMETIMES IT ISN'T.

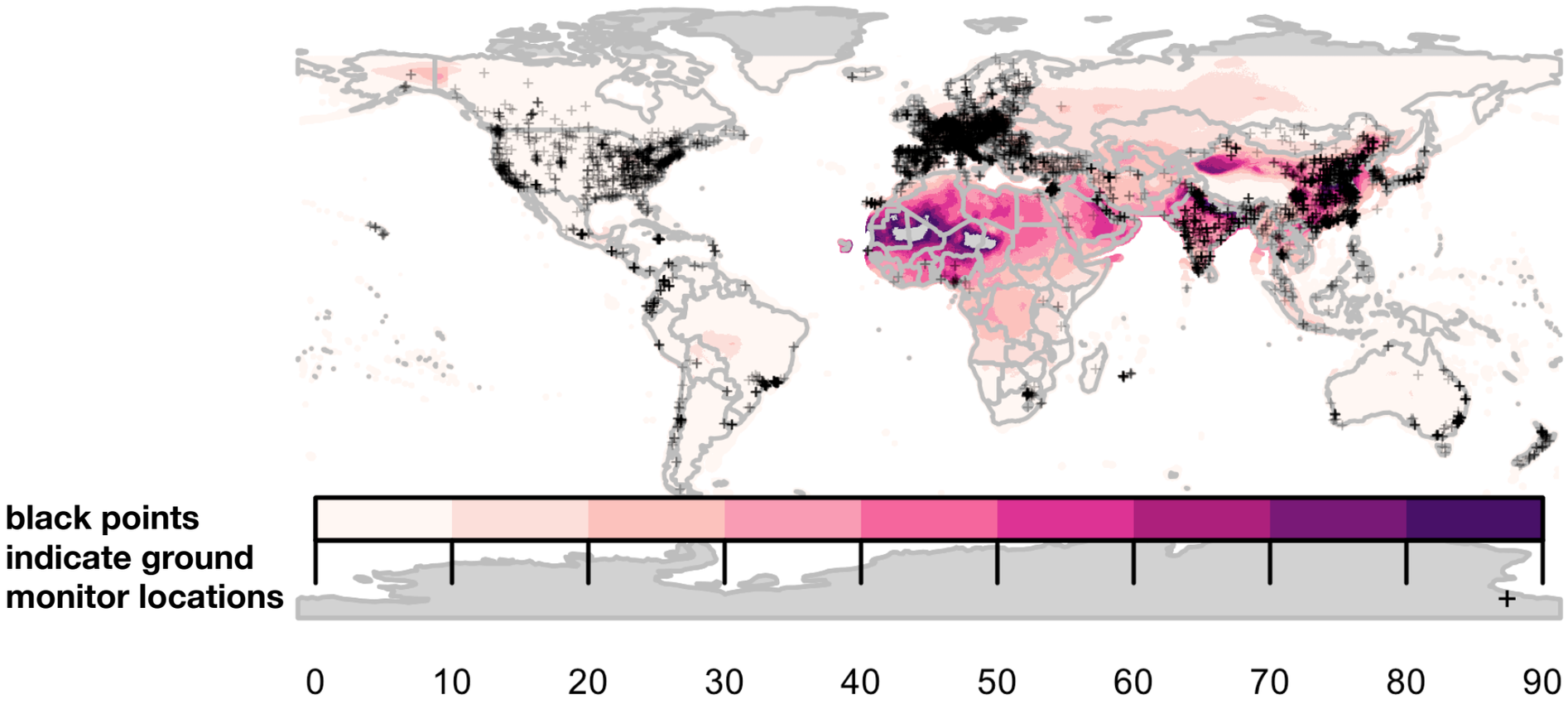
Daniel Simpson
University of Toronto
with Finn Lindgren
Gavin Shaddick, Matt Thomas, Geir-Arne Fuglstad, Håvard Rue, Lei Sun, Jonah Gabry, Andrew Gelman, Aki Vehtari

BREATHE WAS A REALLY GOOD KYLIE SONG, WASN'T IT

WHEN KYLIE SAID "BREATHE" THIS WASN'T WHAT SHE WANTED

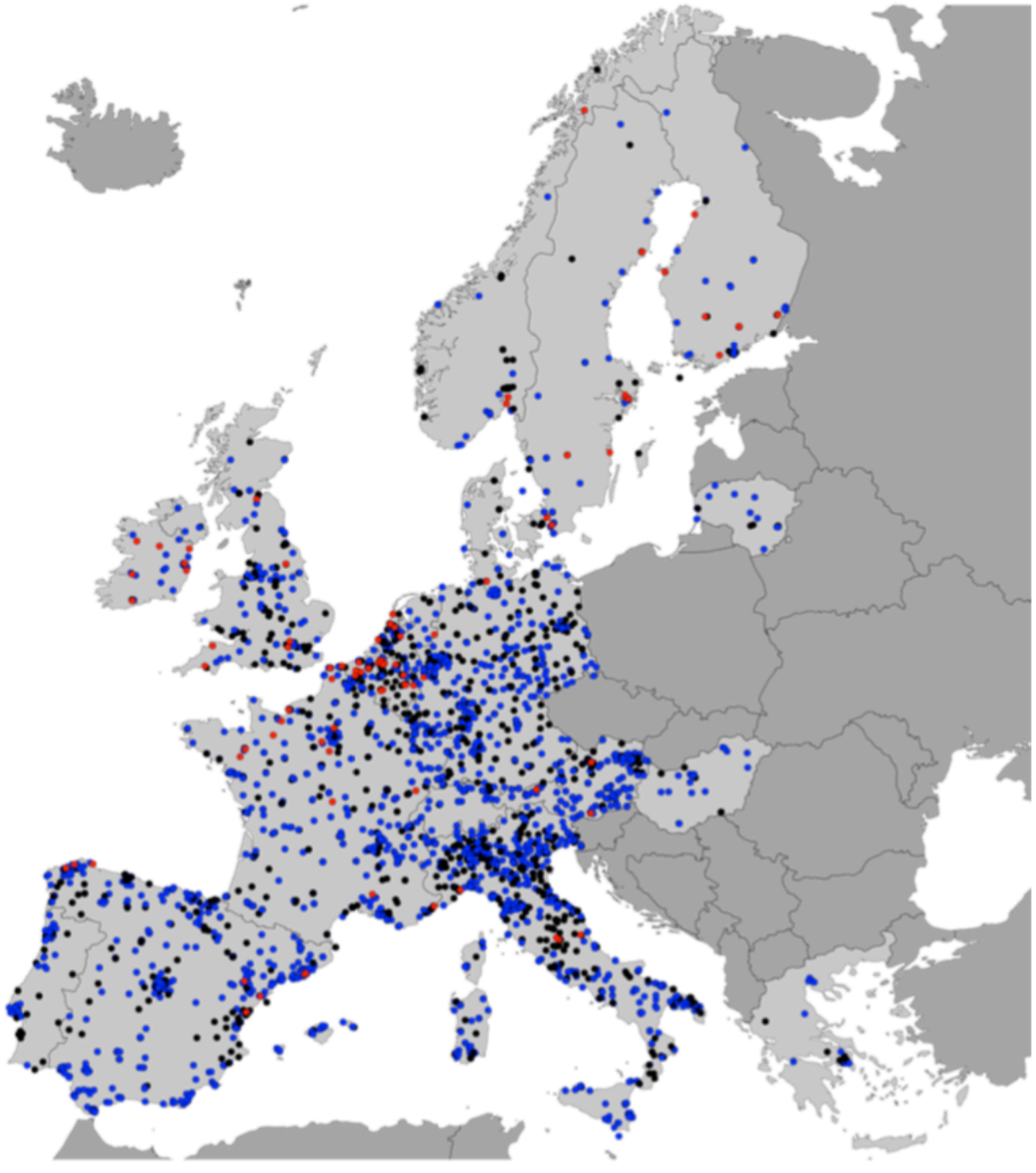
Goal Estimate global PM2.5 concentration

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



Satellite estimates of PM2.5 and ground monitor locations

THE PROBLEM ALSO EXISTS ON A SMALLER SCALE



AND ON A MUCH SMALLER SCALE

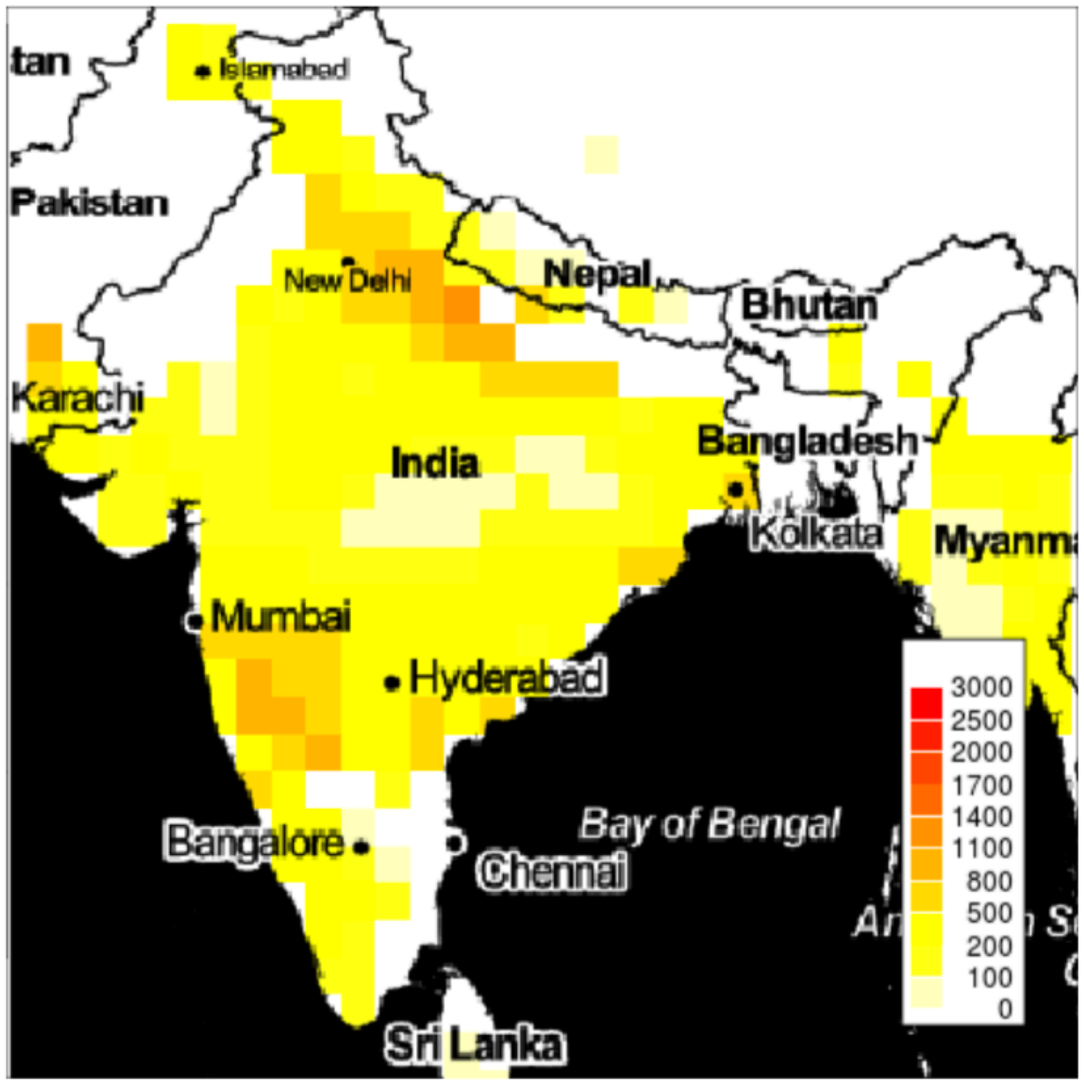


Figure 1: MODIS AOD Jan. 5

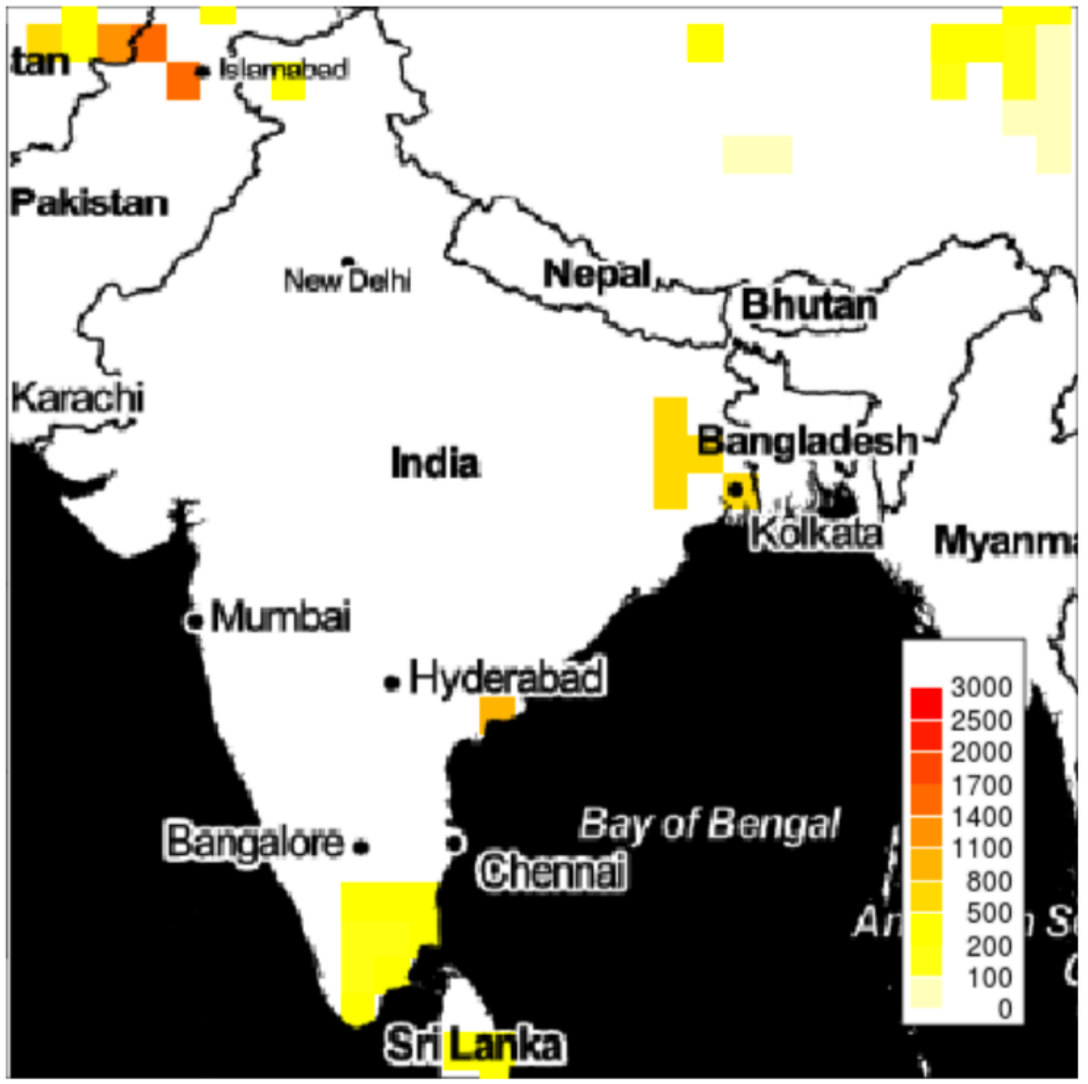


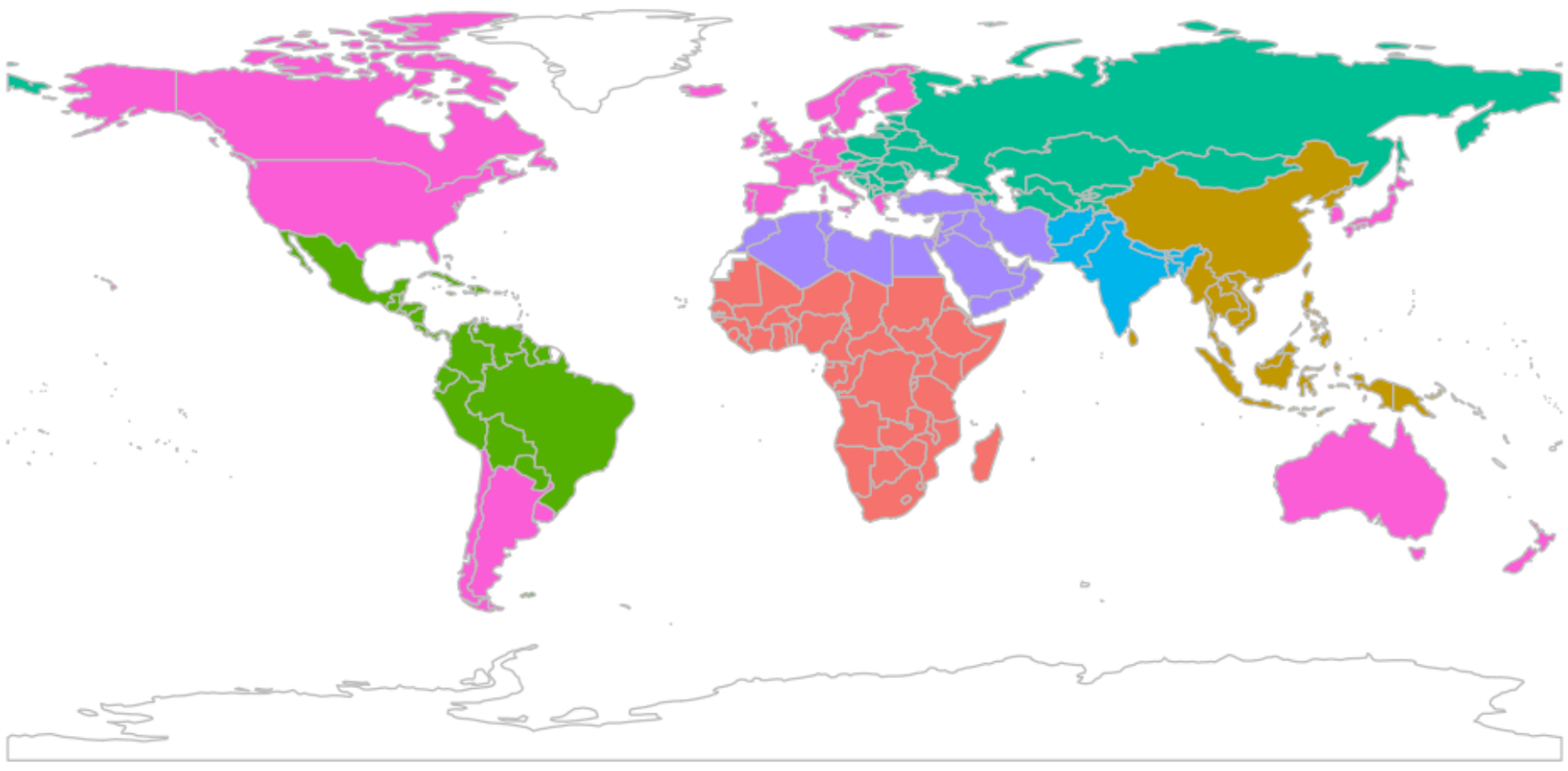
Figure 2: MODIS AOD Jul. 31

JUST A SIMPLE SPONGE

- These are fairly standard spatial statistics problems
- There is nothing particularly breathtaking about the size of the data
- These models can be fit with standard software
- So I'm not talking about extending the real of what is possible, so much as about building up good practice for what is already possible.

SO HOW DO WE FORMULATE THESE PROBLEMS

- We have data y_{ij} observed at location s_i that is in "group" j
- At location s_i we have covariates x_i (possibly at an area level
- What is group in this context?



WHY IS THINKING ABOUT MACROSTRUCTURE IMPORTANT?

- Spatial statistics is all about working across multiple scales.
- A lot of good work is being done on multi-resolution work (Hi Andrew!)
- Getting the large-scale structure correct can make your life a whole lot easier.
- It also stops the model leaning too heavily on some convenient–but–wrong assumptions like stationarity and isotropy.
- Getting this right is probably as important as getting the mean-structure correct.

SO WHAT DOES THIS LOOK LIKE

- In general, we will have some covariates that have a fixed relationship with the response, and some where the relationship varies by group and spatial location.
- In maths, this is $y_{ij} = \sum_{k=1}^{L} \beta_k x_{ki}^{(F)} + \sum_{\ell=1}^{L} \beta_{j\ell}(s_i) x_{\ell i} + \epsilon_{ij}$
- For simplicity, we're just considering Gaussian observation models, but there are lots of cases where we need to be more general, so it's a good idea not to lean too much on that assumption

FIRST YOU'RE ANOTHER SLOE-EYED VAMP,
THEN SOMEONE'S MOTHER,
THEN YOU'RE CAMP

BUT HOW DO WE MAKE THOSE FUNCTIONS?

- What do we need from a random function
  1. We need to be able to evaluate it at any finite set of locations and get a joint distribution
  2. (We probably will want area averages at some point)
  3. The order of evaluation shouldn't matter
  4. We should be able to add and remove points from the evaluation set consistently
- On top of this, we want things to be mathematically and computationally tractable

GAUSSIAN RANDOM FIELDS

- This almost inevitably leads to the idea of a Gaussian random field (GRF) $u(\cdot)$, which is a random function with the property that

$$[u(s_1), \ldots, u(s_n)] \sim N(0, \Sigma_{s_1,\ldots,s_n})$$

- The entries of the covariance matrix are usually specified through a **covariance function** $\Sigma_{ij} = c(s_i, s_j)$
- There are lots of parameterized families of covariance functions that we can use to do our inference

A SMALL PROBLEM

- The dimension of Σ depends on the number of observations, which means that it's pretty easy for it to be big
- This is a problem: we cannot evaluate the density of a GRF or sample from it if the matrix is too big
- So we need to do something clever

MANY PEOPLE HAVE DONE CLEVER THINGS

- Two main schools of work:
  - Finite dimensional approximations (Kernel methods, Fixed Rank Kriging, Predictive Processes, SPDEs, Multiresolution Approximations, etc)
  - Local approximations (Covariance tapering, composite likelihood, Vecchia approximations)
- There are plusses and minuses to both
- But I'm going to focus on the first class

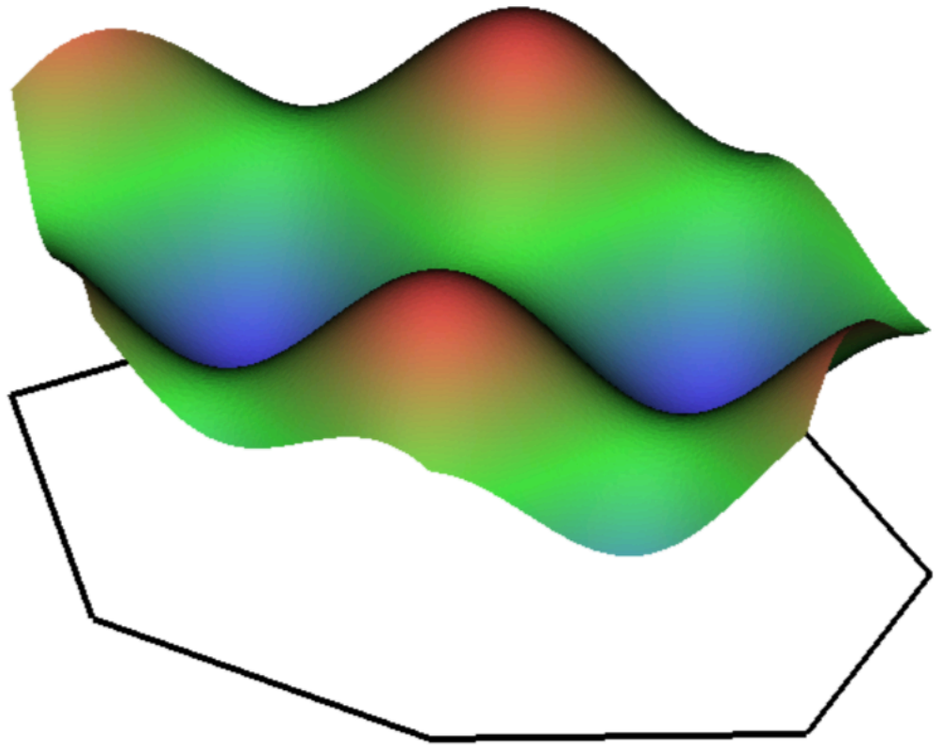
FINITE DIMENSIONAL GAUSSIAN RANDOM FIELD

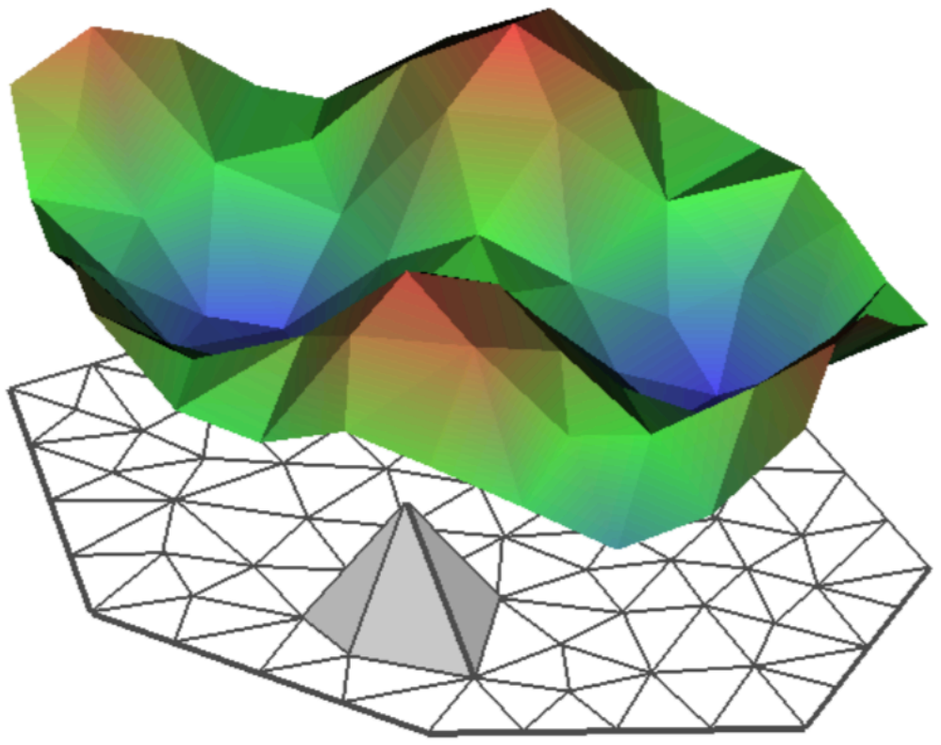
- Given n deterministic basis functions $\phi_i(s)$, a finite dimensional GRF has the form

$$u_n(s) = \sum_{i=1}^{n} w_i \phi_i(s)$$

- The weights w_i are random and they jointly have a multivariate Gaussian distribution $w \sim N(0,\Sigma)$
- This works pretty well but
  - Sensitive to the choice of basis function
  - More basis functions are better

VIDEO GAMES





Note the compact support! It makes these basis functions cheap to evaluate!

WHAT DOES THEORY TELL US ABOUT BASIS FUNCTIONS?

- Theory tells us that if $u_n(\cdot) = R_n u(\cdot)$, then the error in posterior functionals looks like

$$\| R_n \|_{V->H} = \sup_{\|v\|_V=1} \|v - R_n v\|_H$$

- Ewwwwwwww.
- What this says is that the error depends on how well sums of basis functions can approximate sample paths from the "true" GRF.
- But this requires a true GRF.
- Turns out, this gives us a way to specify the weight distribution

ENTER WHITTLE AND MATÉRN

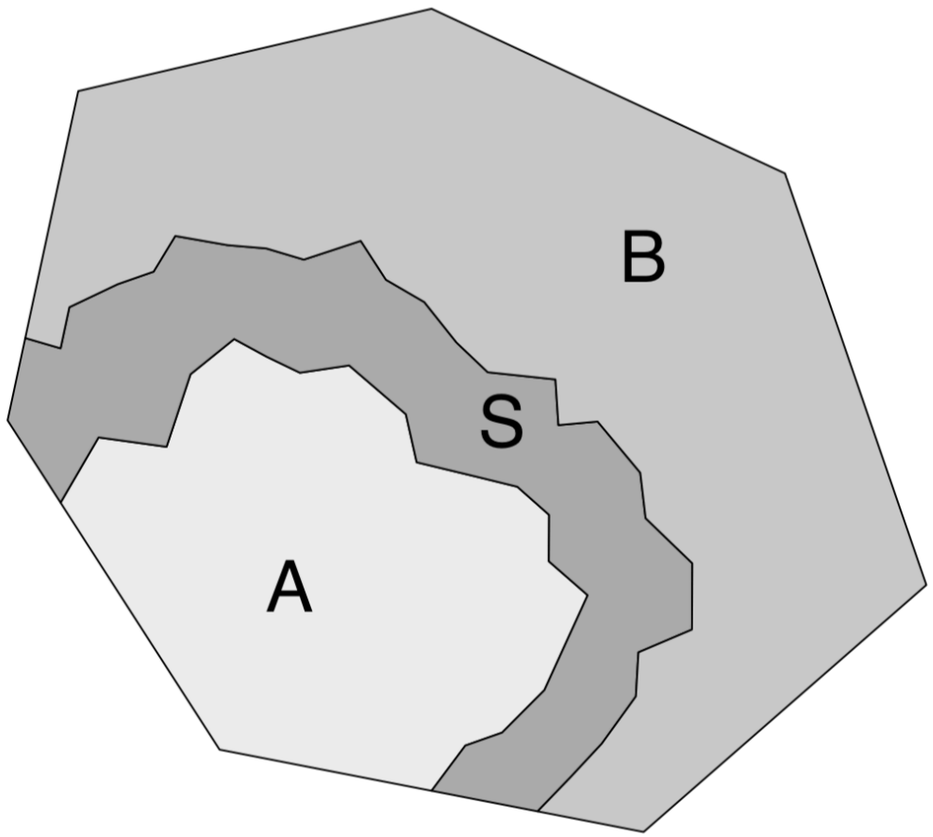
- A common class of covariance function is the Matérn covariance function

$$c(s_1, s_2) \propto \sigma^2 \left(\kappa \|s_1 - s_2\|^2\right)^{\nu} K_{\nu} \left(\kappa \|s_1 - s_2\|^2\right)$$

- Here ν is a smoothness parameter that we will fix
- κ controls the bandwidth
- σ^2 is a scale parameter
- (K_{ν} is the modified Bessel function of the second kind)
- **So what prior should we put on κ and σ**

WHY IS THIS RELEVANT?

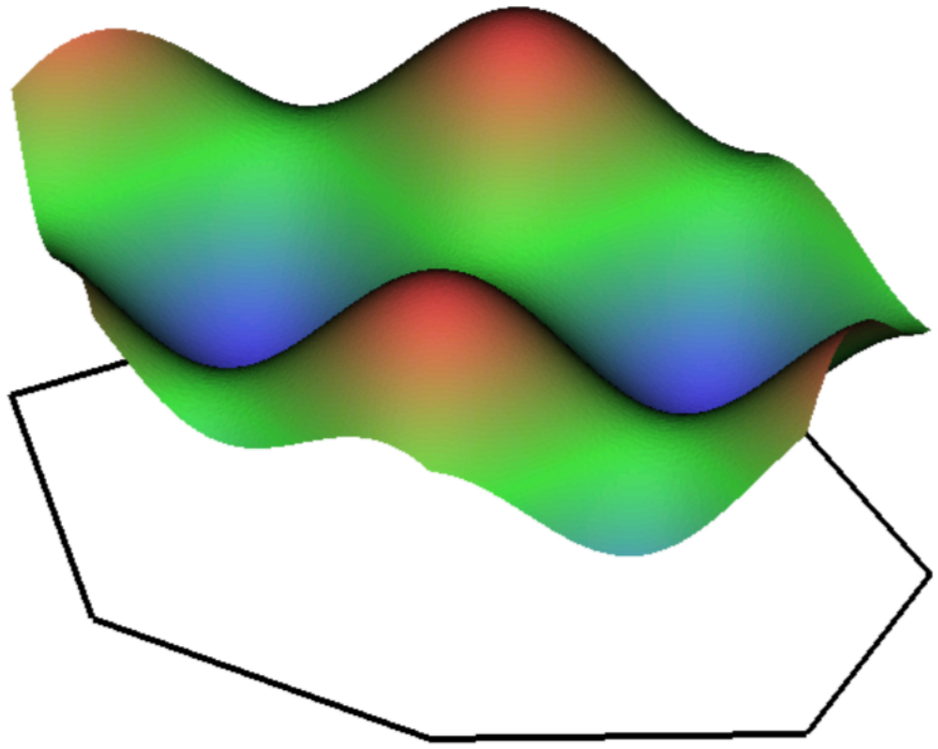
- Well in the mid 1950s, Whittle showed that GRFs with Matérn covariance functions are the stationary solutions to a particular class of stochastic partial differential equations (SPDEs)
- In the 1970s, Rozanov showed that SPDEs have a special connection to the continuous space Markov property

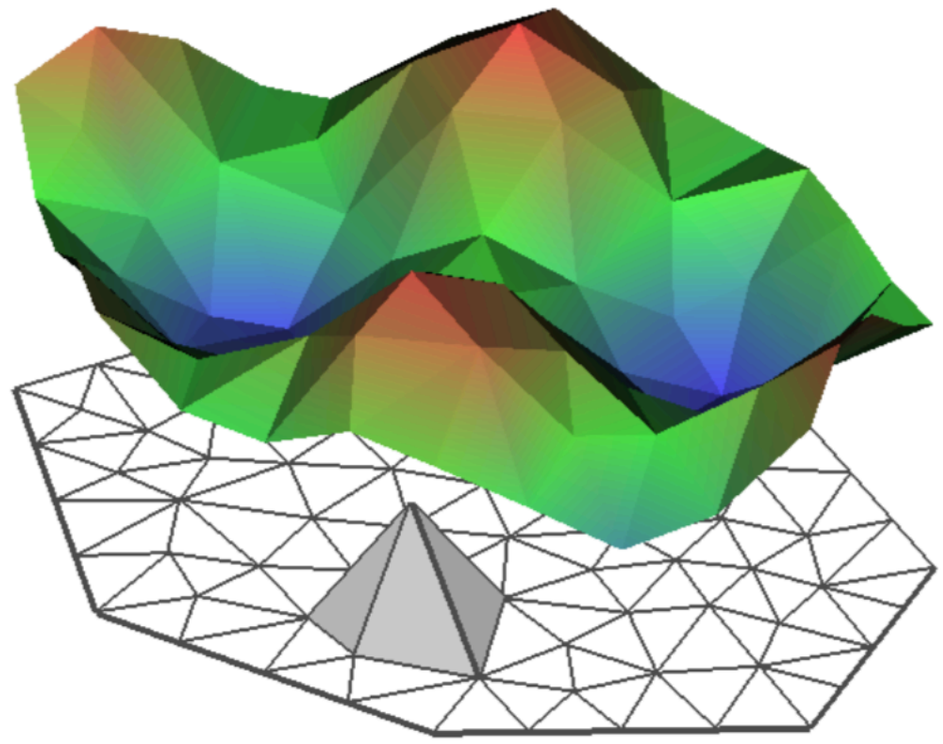


LINDGREN, RUE, AND LINDSTRÖM (2011)

- Finn Lindgren, Håvard Rue, and Johan Lindström put this all together into a cohesive method they called the SPDE method
- It allows for a fairly computationally inexpensive approximation to Matérn-type GRFs
- Built around triangulations of the domain (gives away small-scale features)
- Works nicely as a single piece in a more complex model
- Implemented in the INLA, which has an R interface (and more recently in the mofre friendly INLABru package)

THE SPDE METHOD USES TRIANGLES





Note the compact support! It makes these basis functions cheap to evaluate!

AND WITH THAT WE SOLVED SPATIAL STATISTICS

- Well, not really.
- It did well in a competition (Heaton *et al.*), but that was (too) easy
- Gotta do something about this sub-grid error!
- Hard to scale beyond a certain limit
- **We still have challenges with model specification**

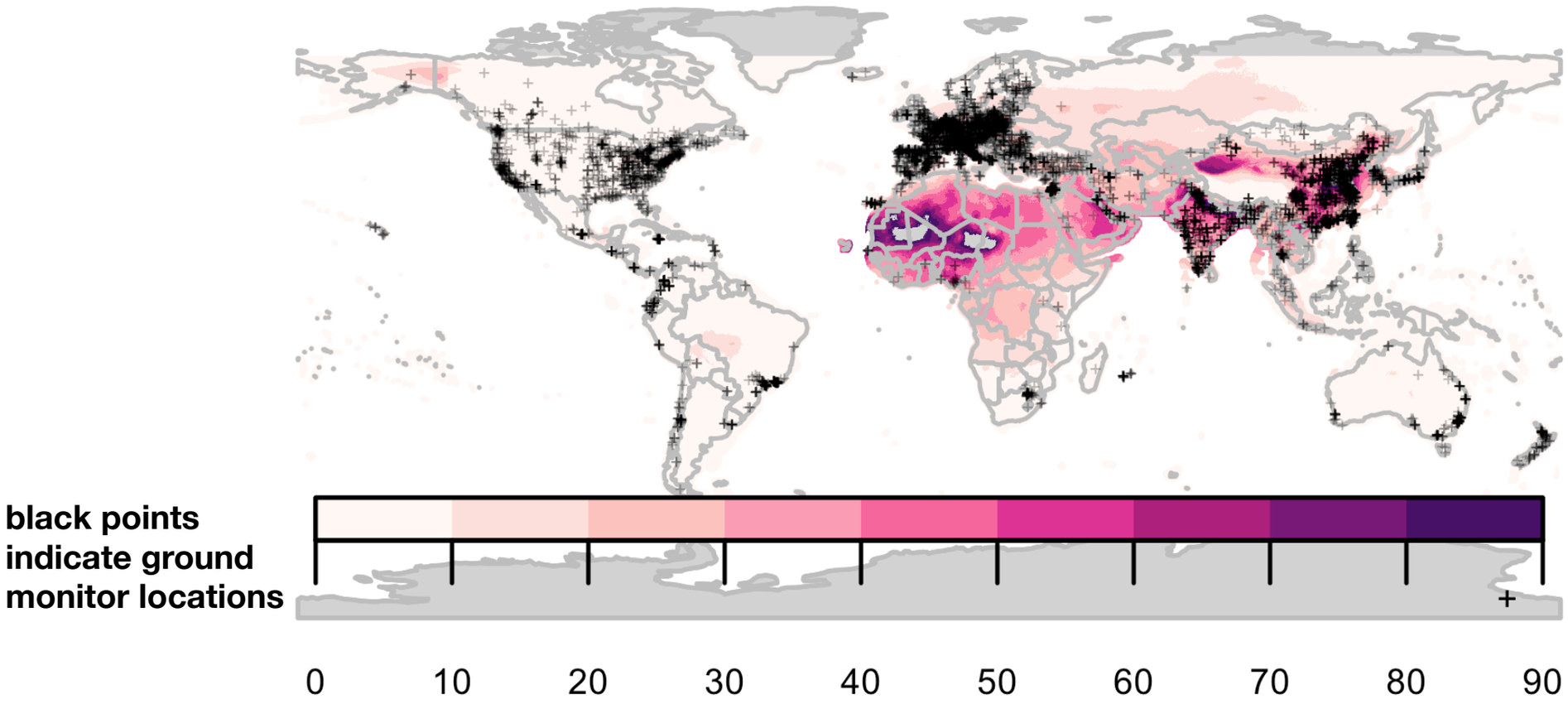
SOME ASPECTS OF MODELLING

AS ALWAYS, BRITNEY SPEARS WAS AHEAD OF THE GAME

WHEN KYLIE SAID "BREATHE" THIS WASN'T WHAT SHE WANTED

Goal Estimate global PM2.5 concentration

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



Satellite estimates of PM2.5 and ground monitor locations

ARIANISM WAS A HERESY FOR A REASON

- Many are taught that the likelihood is the fundamental building block of a Bayesian model and the prior is a secondary object
- This is a very limiting view.
- In reality, we build a **joint distribution** for the data and the likelihood
- People who don't do this (like people who use reference priors) are making some heavy assumptions
- (and, in this analogy, are heretics but don't worry so much about that)

Gelman, A., Simpson, D., and Betancourt, M. (2017).
The prior can often only be understood in the context of the likelihood.
arXiv preprint: arxiv.org/abs/1708.07487

THE MAJESTY OF GENERATIVE MODELS

- If we disallow improper priors, then Bayesian modelling is generative.
- In particular, we have a simple way to simulate from $p(y)$:
  - Simulate $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$
  - Simulate $\mathbf{y}^* \sim p(\mathbf{y} \mid \boldsymbol{\theta}^*)$
  - (Repeat for each sample)

PRIOR PREDICTIVE CHECKING

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

$$\log(\text{PM}_{2.5})_i = \alpha_i + \beta_i \log(\text{sat}_i) + \epsilon_i$$

$$\alpha_j \sim N(\alpha_0, \tau_\alpha^2)$$

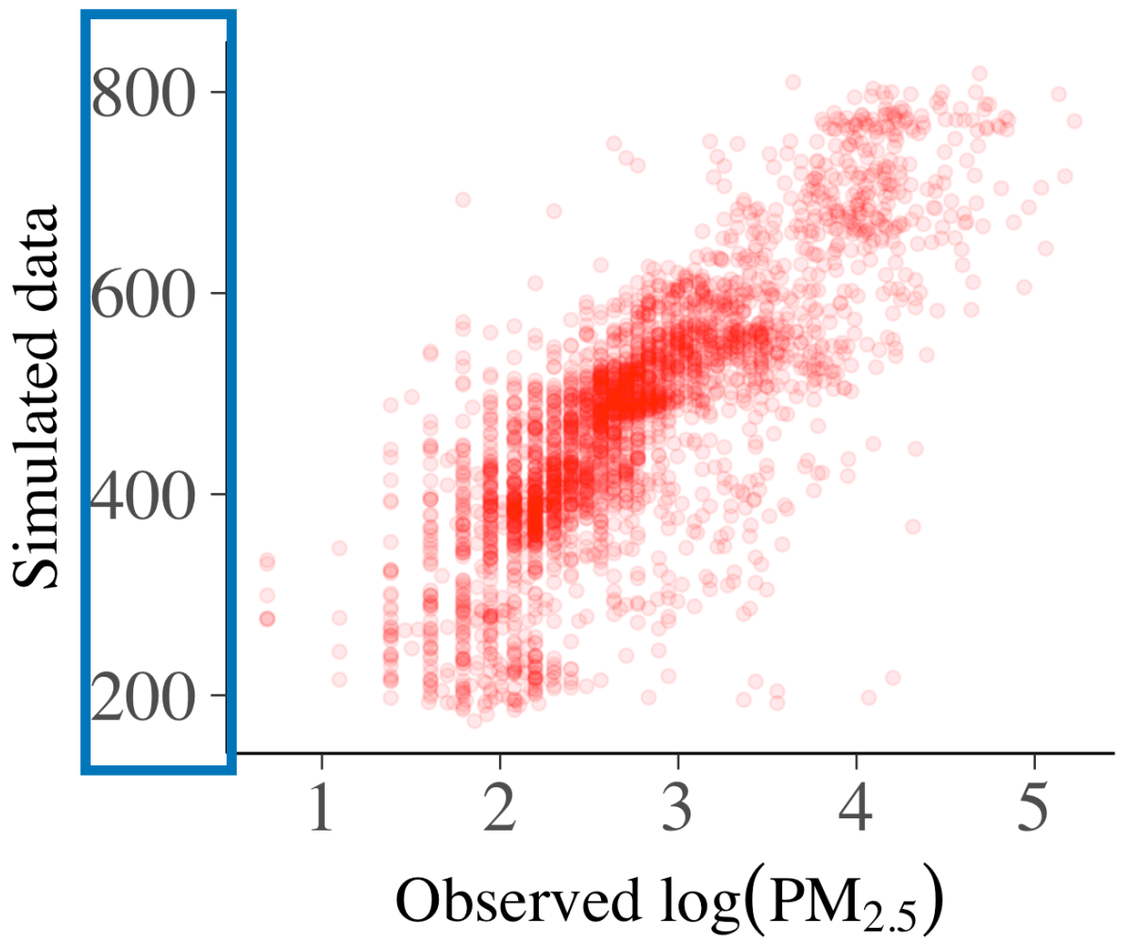
$$\beta_j \sim N(\beta_0, \tau_\beta^2)$$

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

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

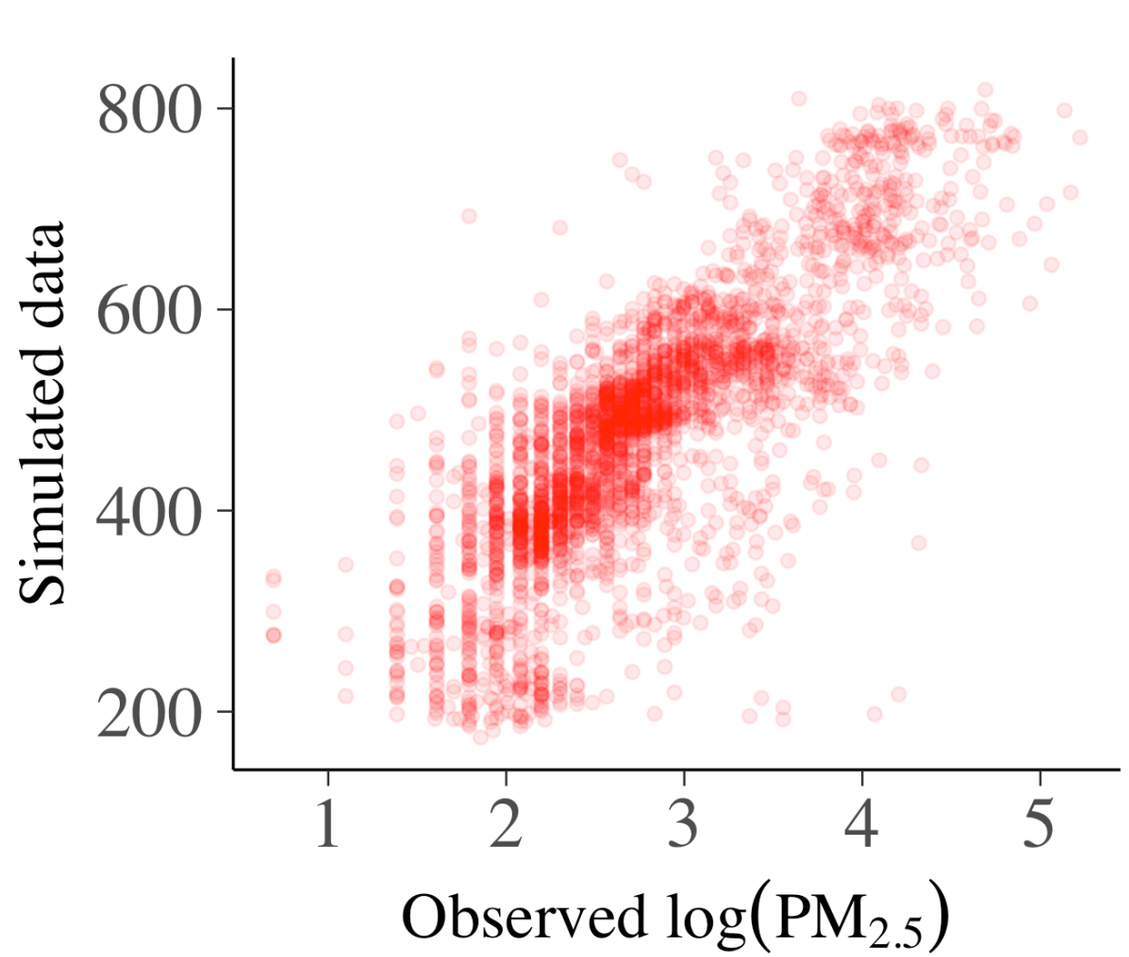
$$\tau_\alpha^2 \sim \text{InvGamma}(1, 100)$$

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



WAIT! WHAT?

- The prior model is two orders of magnitude off the real data
- Two orders of magnitude on the log scale!
- Log density of neutron star only 60 μgm^{-3}!!
- What does this mean practically?
- The data will have to overcome the prior…



IT CAN GUIDE YOUR CHOICE OF PRIOR

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

$$\log(\mathrm{PM}_{2.5})_i = \alpha_i + \beta_i \log(\mathrm{sat}_i) + \epsilon_i$$

$$\alpha_j \sim N(\alpha_0, \tau_\alpha^2)$$

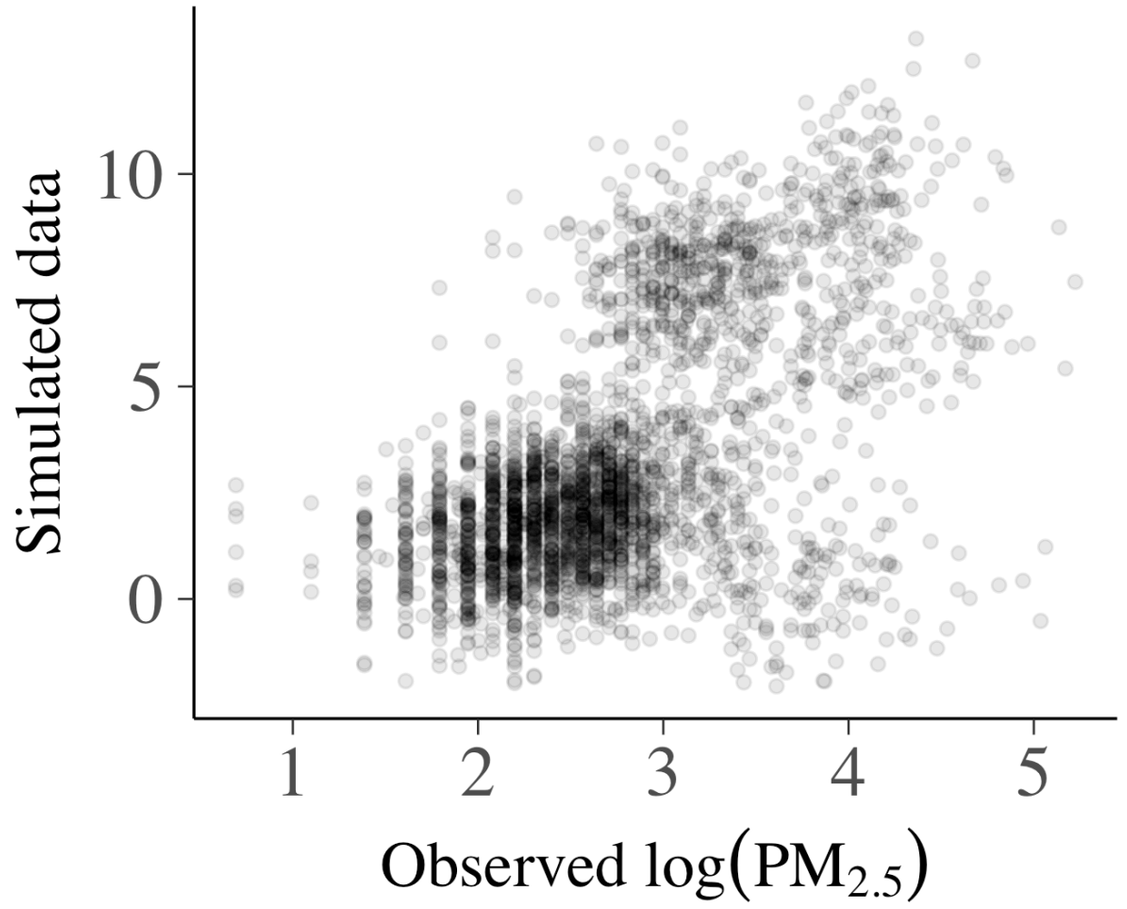
$$\beta_j \sim N(\beta_0, \tau_\beta^2)$$

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

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

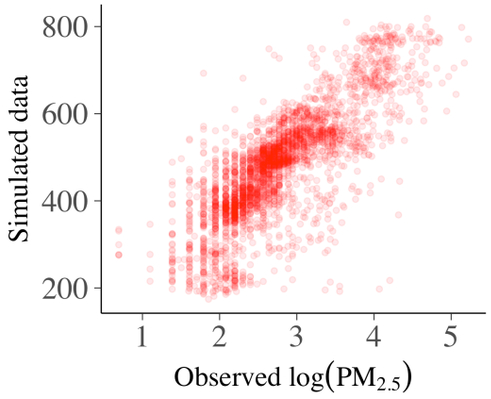
$$\tau_\alpha \sim N_+(0, 1)$$

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

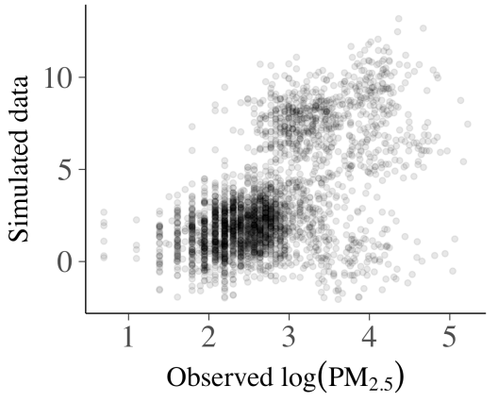


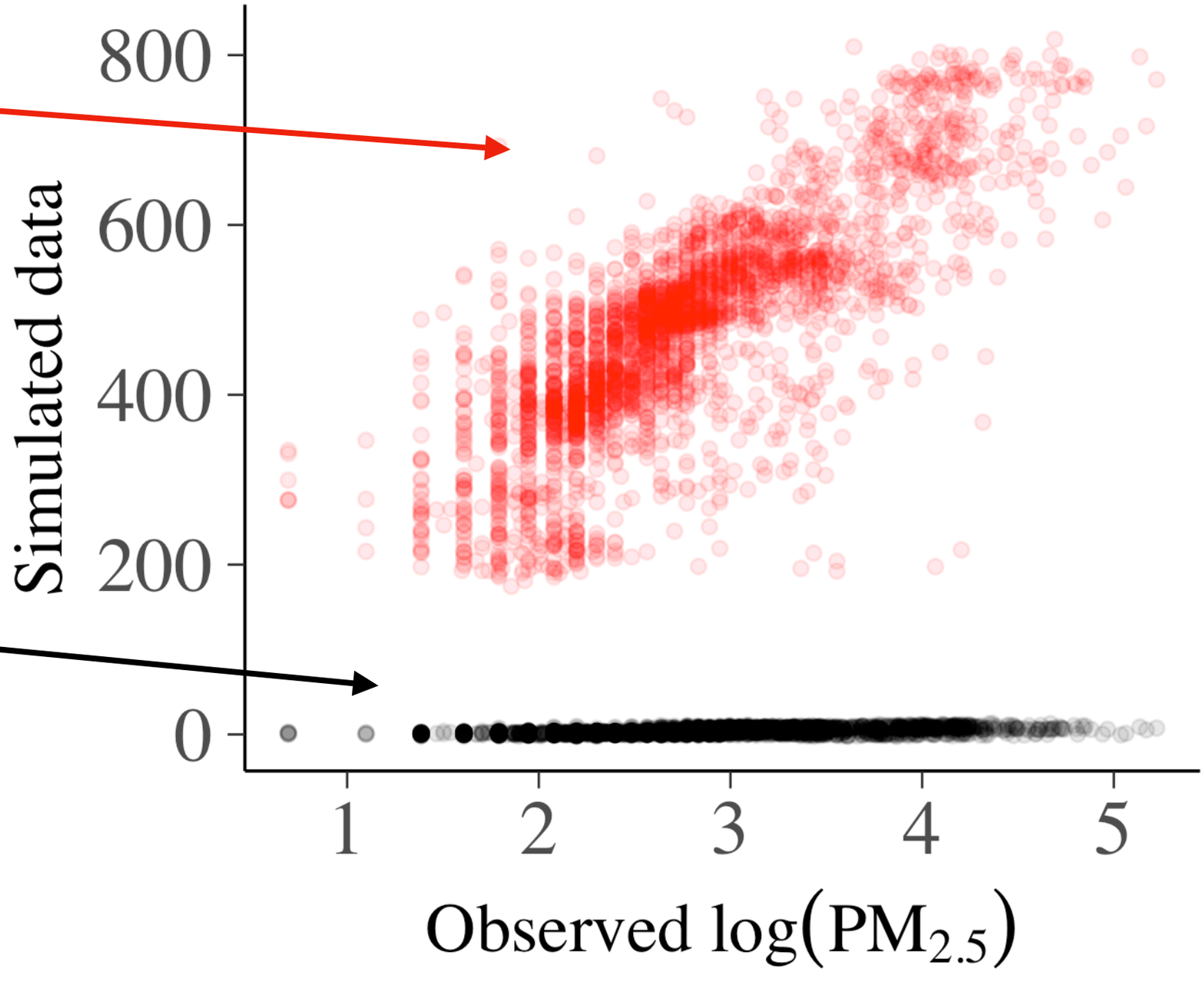
AND MAKE IT EASIER TO DEFEND YOUR MODELLING CHOICES

Non-informative

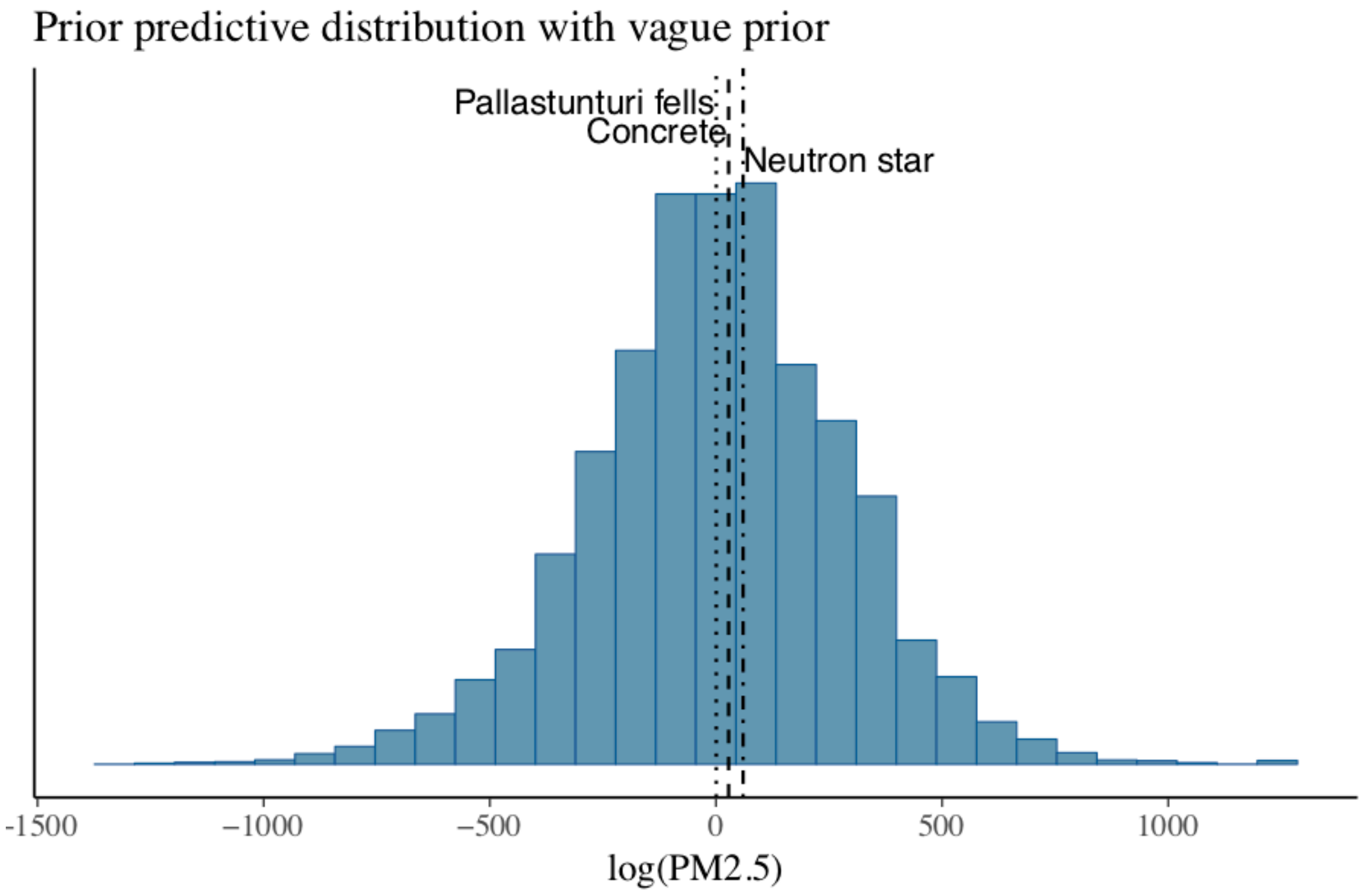


Weakly informative

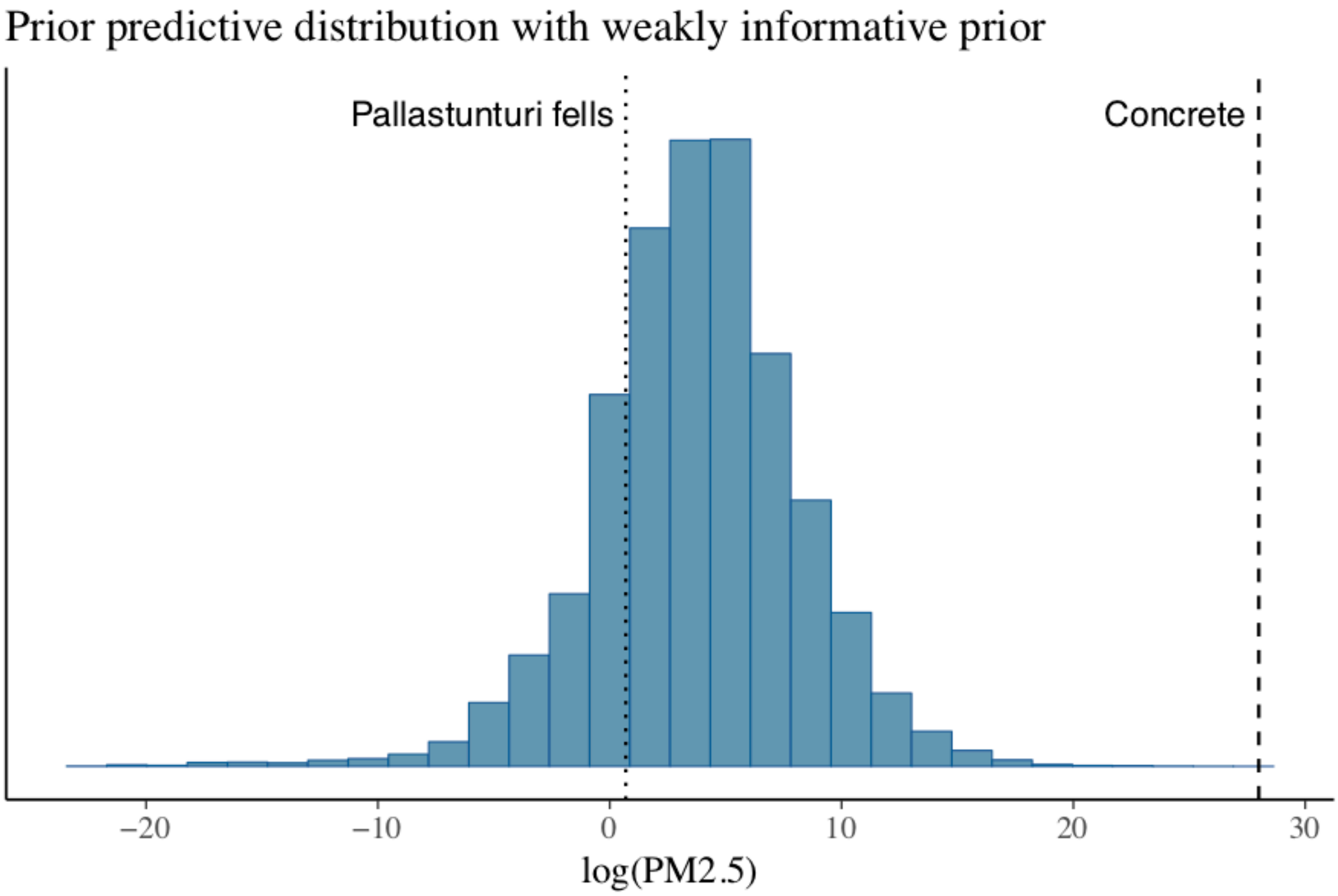




AND NOW, WITHOUT THE DATA



MORE REASONABLE PRIORS



SOME THOUGHTS

- We are very bad at reasoning about logarithms. Always check the natural scale!
- This is a GLM, so the natural summary of the problem that we can reason about is the observation
- For more complex models, a lot more substantive knowledge is needed
- Wang, Nott, Drovndi, Mengersen, Evans (2018) use a numerical summary of the predictive distribution as a way to choose priors ("history matching").

BUT MAYBE MATCHING TO A MATÉRN DIDN'T MAKE OUR LIFE EASIER

THE SIMPLEST MODEL

- Let's look at the simplest possible model

$$y_i = u(s_i) + \epsilon_i$$

- If $u(\cdot)$ has a Matérn prior, we know that it has two unknown parameters: The bandwidth and the marginal variance
- It turns out that even with infinite data, we can't disentangle these.
- So our priors are going to be important

EVERYTHING IS PROBABLY NOT GOING TO BE OK

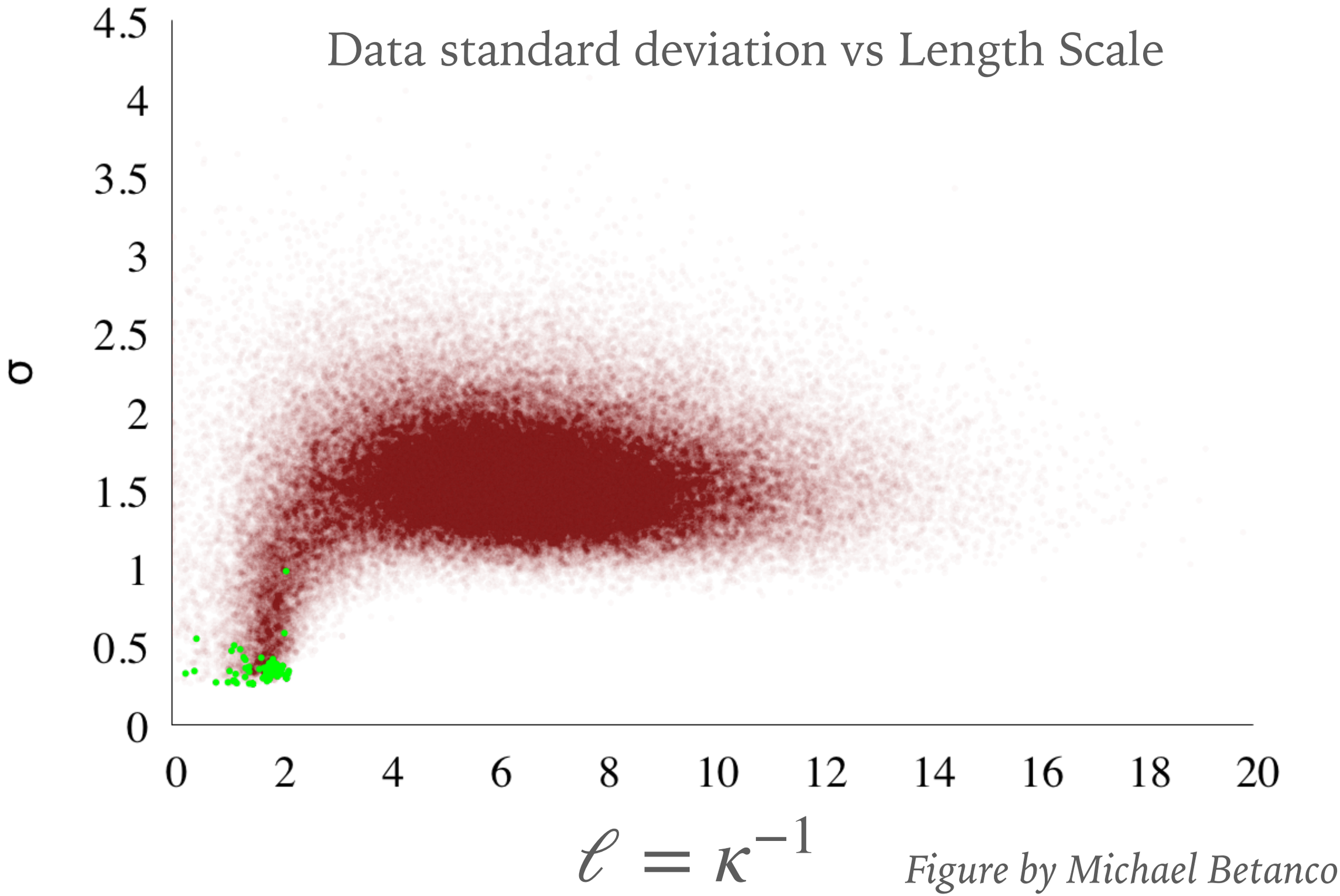


Figure by Michael Betancourt

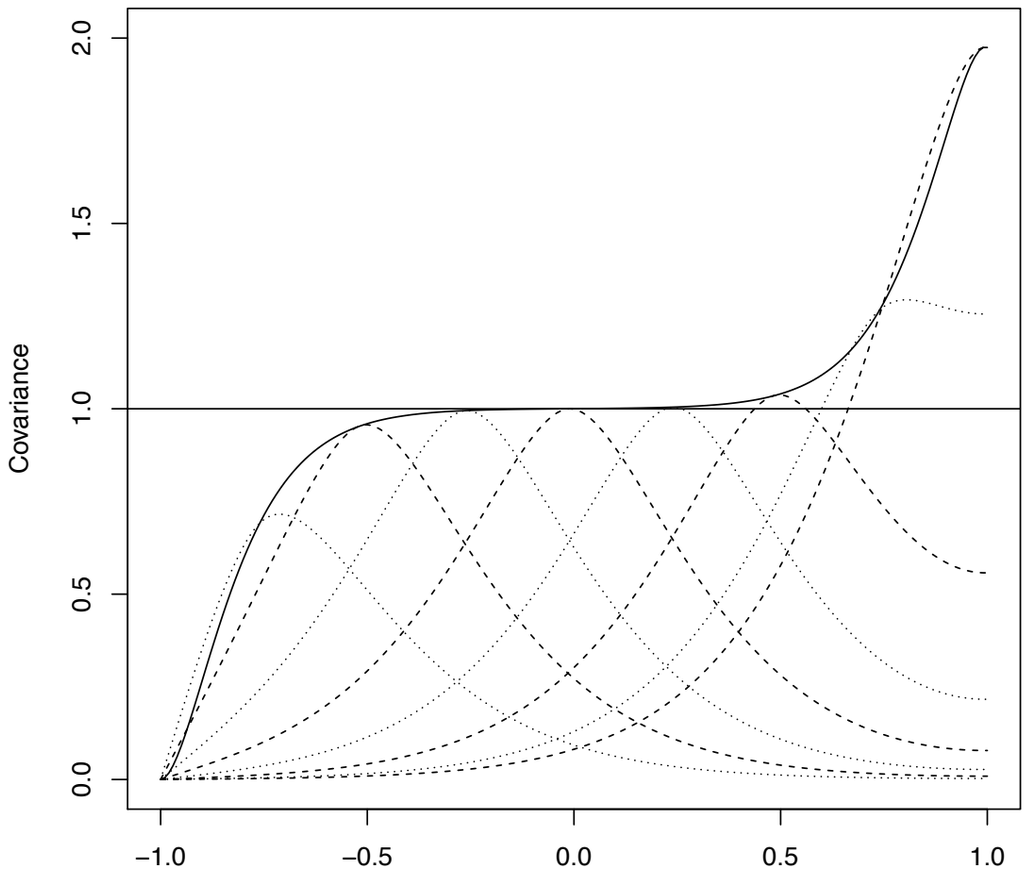
WHAT ARE WE OBSERVING

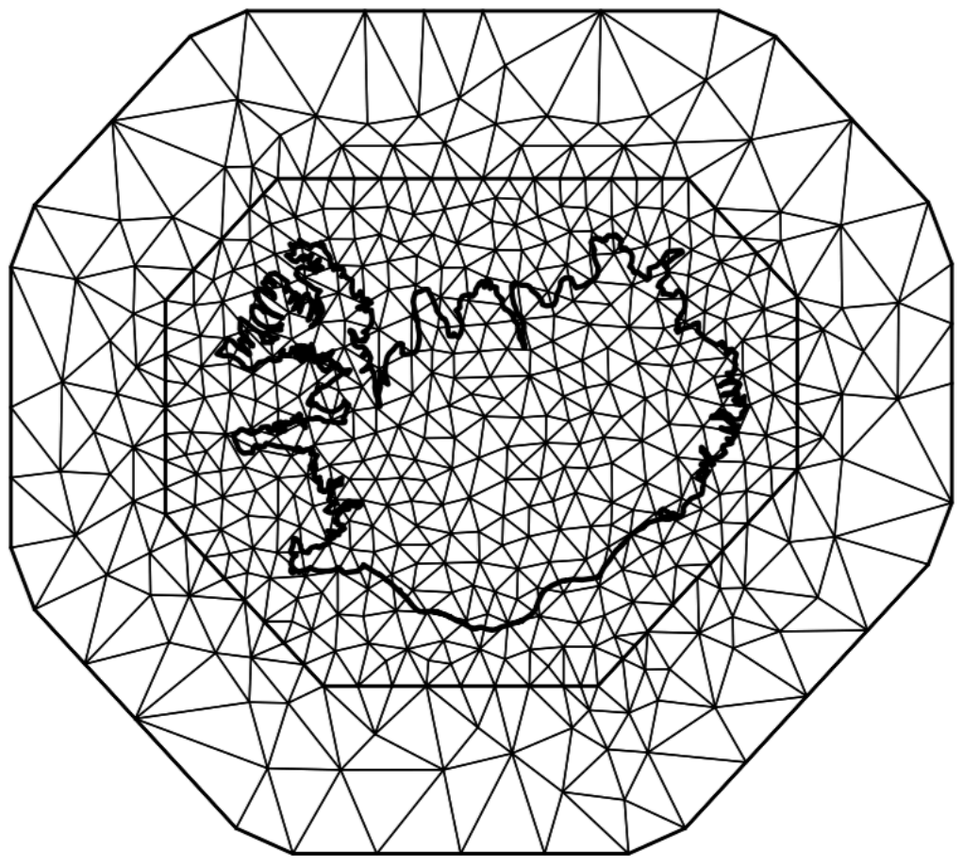
- Our problem is that we've only defined the GRF through what happens when you observe it.
- So the data isn't saying much about the shorter length scales.
- This will all be prior dependent.
- IDEA: Make a prior that doesn't put much mass on extremely short length scales
- The PC prior framework (Simpson *et al.*, 2017, Fuglstad *et al.*, 2018) gives a way to make this rigorous.
- Long story short: in 2D, put an inverse Gamma on the length scale (or a gamma on κ)

BUT SURELY THAT'S TOO STRAIGHTFORWARD

ONE SERIOUS DRAWBACK

- The SPDE method gives a Markovian approximation
- It exists **only** on the computational domain, whereas the true GRF usually exists beyond that
- This tension leads to **boundary effects (aka the field is different near the boundary compared to the centre)**





WHY ISN'T EXTENDING ENOUGH

- It's expensive (even though we can have bigger triangles in the extension)
- How far should you extend??
- (The answer depends on the unknown range of the GRF)
- It's inelegant. The full SPDE structure is a really lovely way to take a continuous object and approximate it. Extending the boundary is a cheap hack.

WHY IS THIS HARD TO FIX?

- So far I have breezed past the technicalities of the SPDE method.
- For the most part, they're not really necessary to understand what it does.
- But here they become very relevant

IT'S ALL ABOUT AN INNER PRODUCT

- It turns out that there are a bunch of equivalent ways to define the covariance structure of a Gaussian random field.
- Rather than use the covariance structure, the SPDE method defines an **inner product** for the Cameron-Martin (or Reproducing Kernel Hilbert) Space

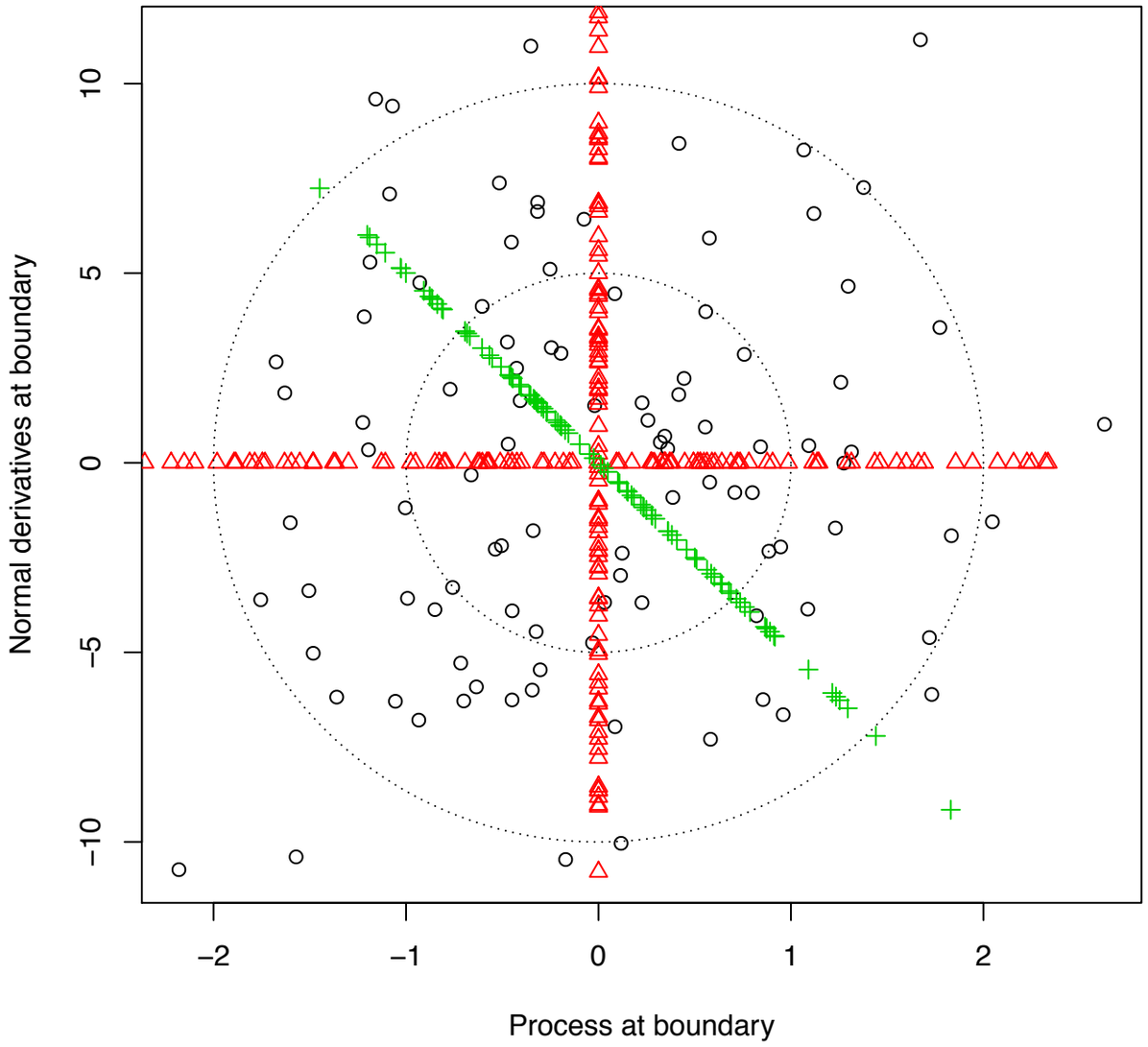
$$H_{\mathbb{R}^d}(f, g) = \int_{\mathbb{R}^d} \hat{f}(\omega)\overline{\hat{g}(\omega)}f(\omega)\, d\omega$$

- Here $f(\omega)$ is the **power spectrum** of $u(\,\cdot\,)$
- The problem turns out to be that the SPDE method approximates this inner product poorly near the boundary of the domain

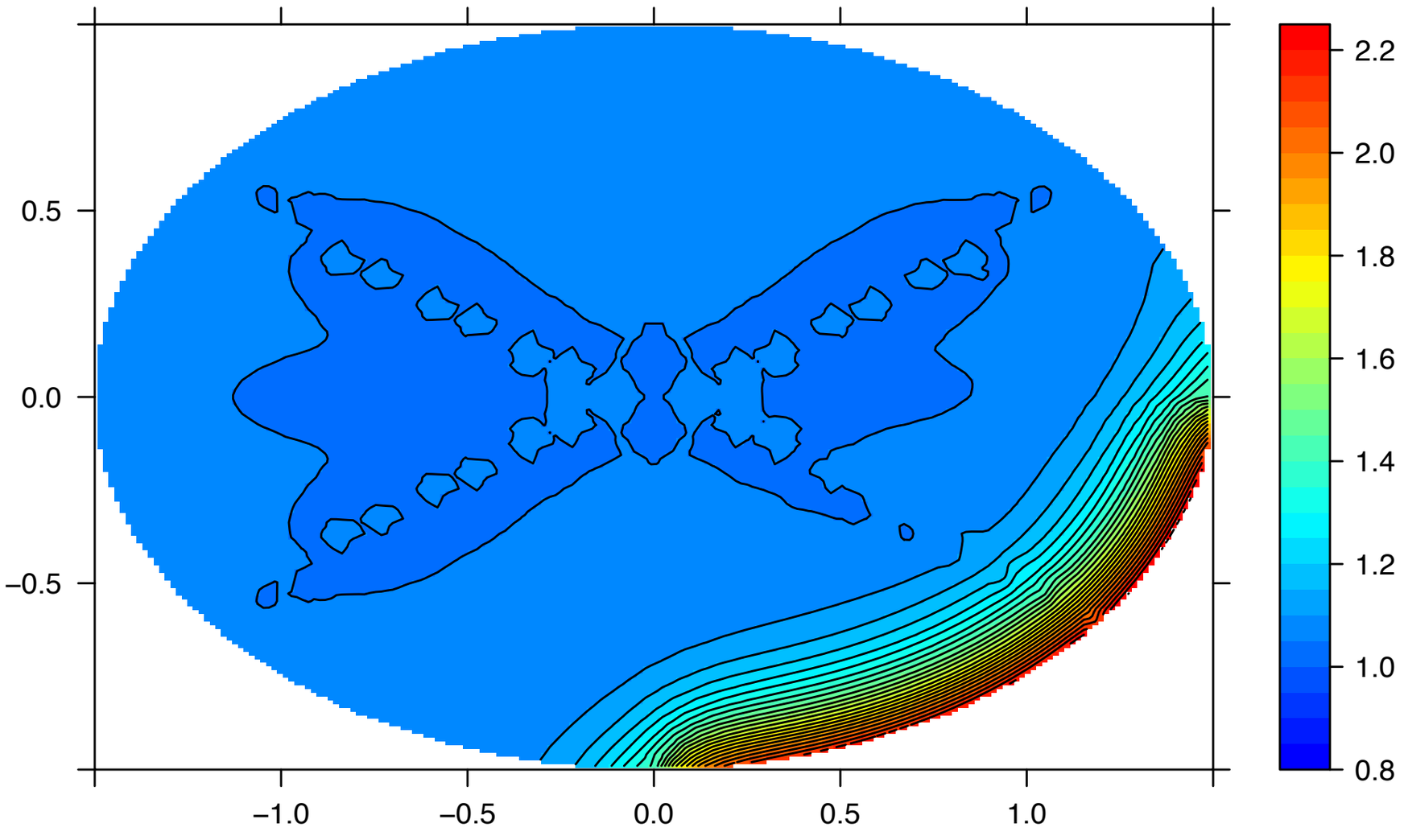
WE'VE BEEN TRYING TO FIX THIS FOR A WHILE…

- Here are some things we know.
- The GRFs approximated by the SPDE method are **Markov**, which means we don't need to know the field everywhere to make a correction.
- We only need the value of the field and its normal derivatives on the boundary.
- But how do we put them in?
- We'll probably need their joint distribution…

FIRST THING WE LEARN IS THAT THE BOUNDARIES HAVE TO BE RANDOM



THERE IS SOME INDICATION THAT IT'S POSSIBLE



BUT HOW?

- We need to mimic a very specific property of the correct inner product: the **reproducing kernel property.**
- If $c(s, s')$ is the covariance function associated with $H(\,\cdot\,,\,\cdot\,)_{\mathbb{R}^d}$ then

$$H(\psi(\,\cdot\,), c(s,\,\cdot\,))_{\mathbb{R}^d} = \psi(s), \quad s \in \mathbb{R}^d$$

- That is, the covariance function interacts with this inner product to evaluate functions.
- (Given the inner product or the covariance function, this uniquely defines the remaining one)

A RESTRICTED REPRODUCING KERNEL PROPERTY

- The solution appears to be to modify the inner product the SPDE method gives us, which we'll call $H_\Omega(\,\cdot\,,\cdot\,)$
- The most important quantity is

$$\psi^*(s) = \mathbb{E}\left(u(s) \mid u(s) = \psi(s), \partial_n^k u(s) = \partial_n^k \psi(s), s \in \partial\Omega\right)$$

- This is a **very nice, smooth** function that would be our Kriging estimate on the interior of the domain we only knew the value of the process and its derivatives on the boundaries
- The modified inner product is

$$H^*(f,g) = H_\Omega(f - f^*, g - g^*) + H_{\partial\Omega}(f,g)$$

- Here $H_{\partial\Omega}$ is the correct inner product for the boundary process(in state space form)

AND THAT'S WHERE WE ARE

- This modified inner product has the reproducing Kernel property, so it's the correct continuous formulation of the problem
- Our current challenge is trying to approximate this in a reasonable way.
- We've got some very rough initial results, but we aren't quite there yet

SOME CLOSING THOUGHTS

TAKING THE CONTINUOUS MODEL SERIOUSLY . . .

- … has a cost.
- We need to make the maths work, and the maths is fairly intense stochastic processes stuff.
- But it's also worth it because we don't get as many potential surprises, which makes it possible for non-experts to use
- It also means we can set priors in a sensible way
- Taking the model seriously as a global approximation to the prior means we get good "large scale" properties.
- The cost of this is small scale features.

BUT THERE REALLY IS A LOT MORE TO DO

- I've not talked about time here
- The big challenge with time is that even with all our unguents and potions, the matrices still end up too big to be useful
- So we need another layer of approximation
- I suspect that a more formal analysis of windowed estimators and domain decomposition might lead to something that links the continuous formulation with decomposable models
- But for now, we just do *ad hoc* stuff.

REFERENCES

- Haakon Bakka, Håvard Rue, Geir-Arne Fuglstad, Andrea Riebler, David Bolin, Elias Krainski, Daniel Simpson, and Finn Lindgren (2018). Spatial modelling with R-INLA: A review. WIRE Computational Statistics. Volume 10(6).
- Geir-Arne Fuglstad, Daniel Simpson, Finn Lindgren, and Håvard Rue. (2018). Constructing Priors that Penalize the Complexity of Gaussian Random Fields. Journal of the American Statistical Association. Volume 114(525), pp. 445–452.
- Thomas, Matthew L., Gavin Shaddick, Daniel Simpson, Kees de Hoogh, and James V. Zidek. "Data integration for high-resolution, continental-scale estimation of air pollution concentrations." arXiv preprint arXiv: 1907.00093 (2019).
- Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman (2018). Visualization in Bayesian workflow (with Discussion). Journal of the Royal Statistical Society Series A. Volume 182(2), pp. 389–402.