Computer Model Calibration for Large-Scale Spatially Distributed Counts

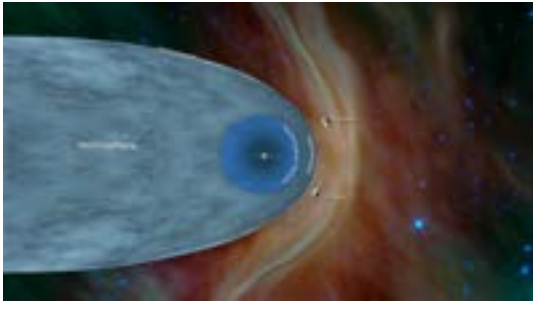
Steven D. Barnett[1,2], Robert B. Gramacy[1], Lauren J. Beesley[2], Dave Osthus[2], Yifan Huang[3], Fan Guo[3], Eric J. Zirnstein[4], Daniel B. Reisenfeld[5]

[1]Department of Statistics, Virginia Tech [2]Statistical Sciences Group, Los Alamos National Laboratory [3]Nuclear and Particle Physics, AstroPhysics and Cosmology, Los Alamos National Laboratory [4]Department of Astrophysical Sciences, Princeton University [5]Space Science and Applications Group, Los Alamos National Laboratory

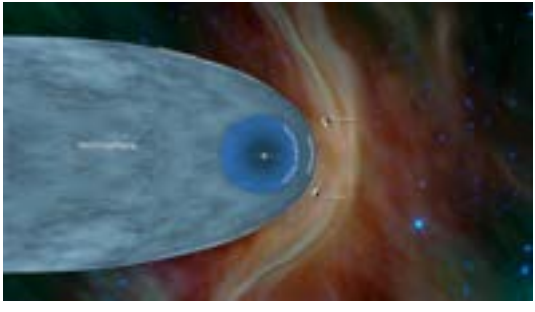
June 19, 2024

LA-UR-23-29368

The Heliosphere

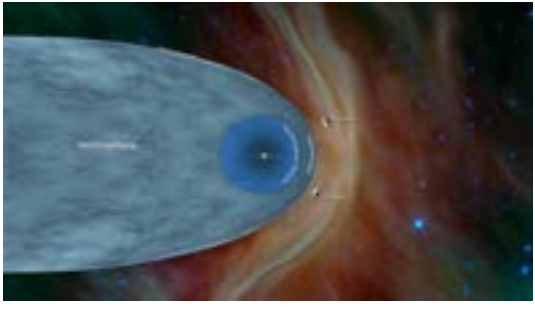


The Heliosphere



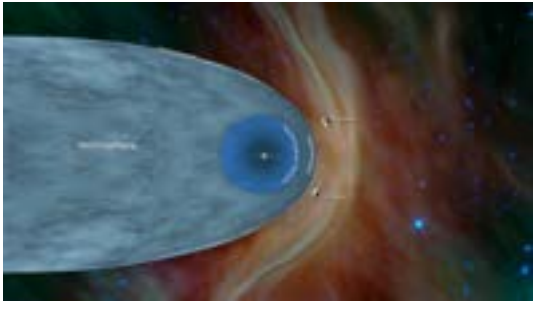
- Heliosphere: the bubble formed by the solar wind that encompasses the solar system

The Heliosphere



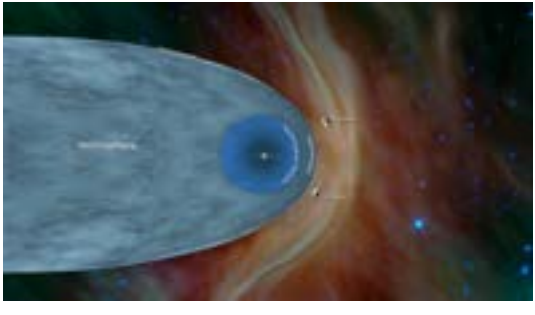
- Heliosphere: the bubble formed by the solar wind that encompasses the solar system
- Energetic Neutral Atoms (ENA): particles formed at the heliosheath with high energy and no charge

The Heliosphere



- Heliosphere: the bubble formed by the solar wind that encompasses the solar system
- Energetic Neutral Atoms (ENA): particles formed at the heliosheath with high energy and no charge
- Two different sources of ENAs: **globally distributed flux (GDF)** and the **ribbon**

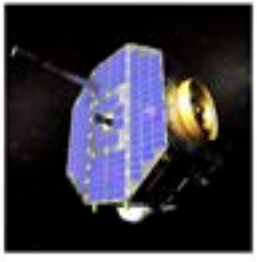
The Heliosphere

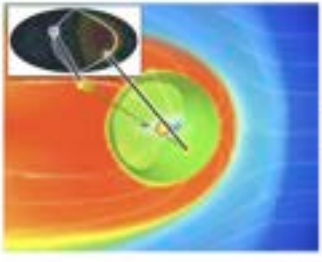


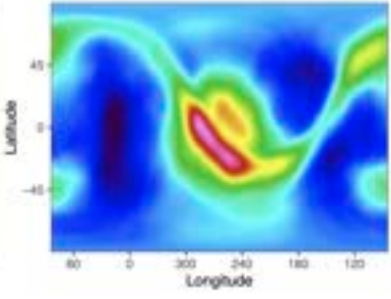
- Heliosphere: the bubble formed by the solar wind that encompasses the solar system
- Energetic Neutral Atoms (ENA): particles formed at the heliosheath with high energy and no charge
- Two different sources of ENAs: **globally distributed flux (GDF)** and the **ribbon**
- Goal: understand the structure and dynamics of the boundary between our solar system and interstellar medium

Interstellar Boundary Explorer (IBEX)

- IBEX satellite launched in 2008
- Contains IBEX-Hi ENA imager
- Records number of ENAs entering apparatus
- Maps the entire sky over a period of six months
- Background ENAs are possibly detected by the instrument
- Data is noisy and irregular







IBEX data

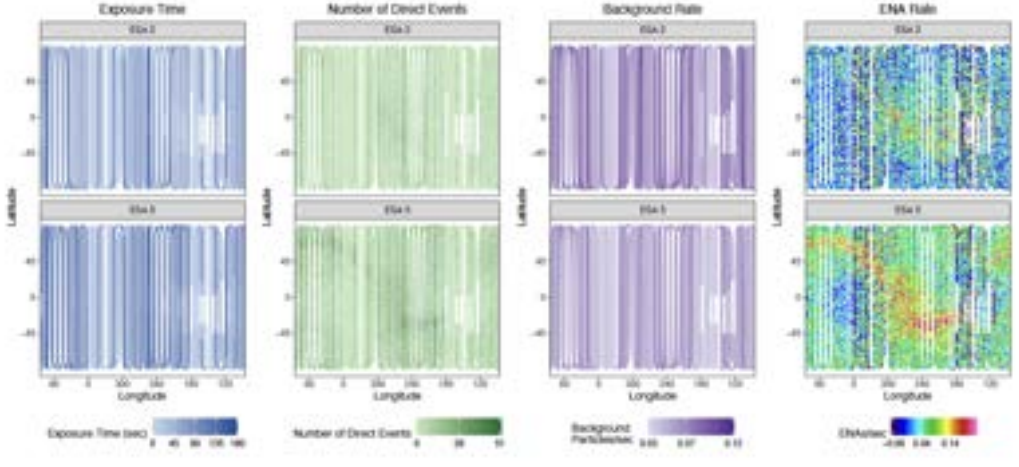
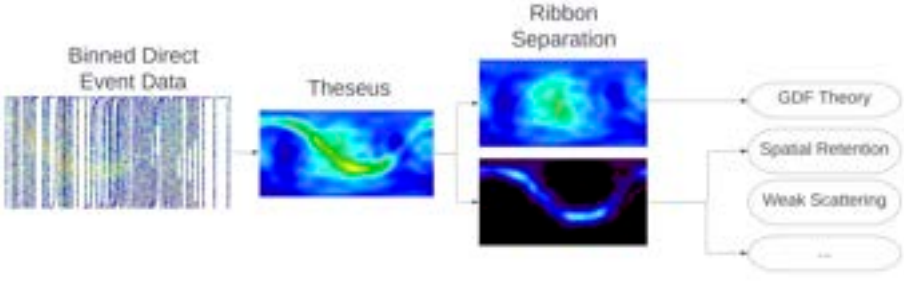
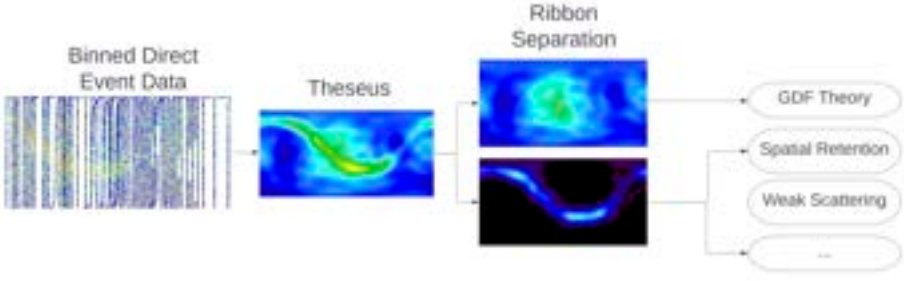


Figure taken from Osthus et al., (2022)

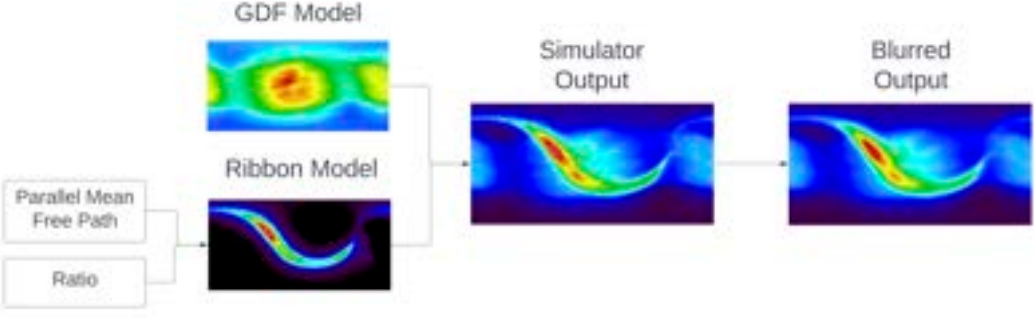
What is our goal?



What is our goal?



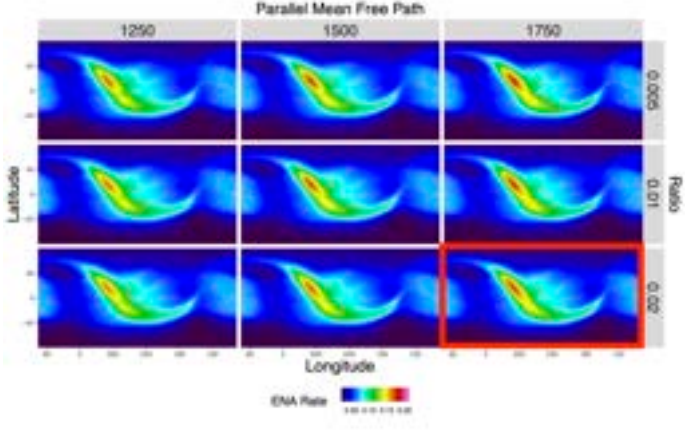
Simulated data generation



Computer model output

$$y(\boldsymbol{s}_i)|t(\boldsymbol{s}_i), \theta(\boldsymbol{s}_i), b(\boldsymbol{s}_i) \sim \text{Poisson}\left(t(\boldsymbol{s}_i)[\theta(\boldsymbol{s}_i) + b(\boldsymbol{s}_i)]\right)$$

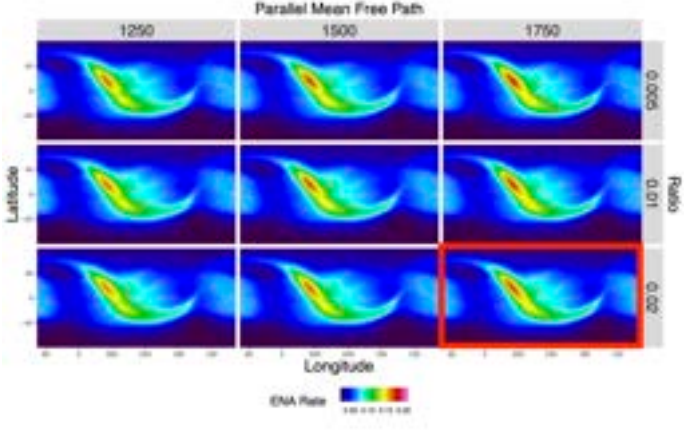
$$\boldsymbol{s}_i = (\text{lon}_i, \text{lat}_i)$$

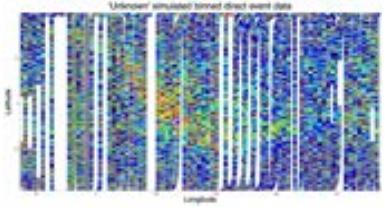


Computer model output

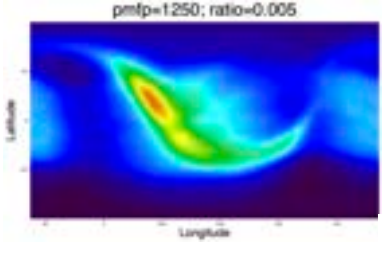
$$y(\boldsymbol{s}_i)|t(\boldsymbol{s}_i), \theta(\boldsymbol{s}_i), b(\boldsymbol{s}_i) \sim \text{Poisson}\left(t(\boldsymbol{s}_i)[\theta(\boldsymbol{s}_i) + b(\boldsymbol{s}_i)]\right)$$

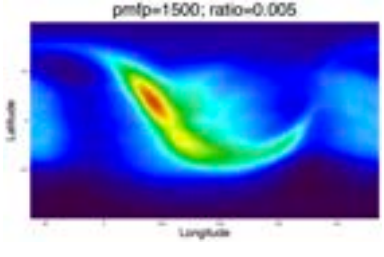
$$\boldsymbol{s}_i = (\text{lon}_i, \text{lat}_i)$$



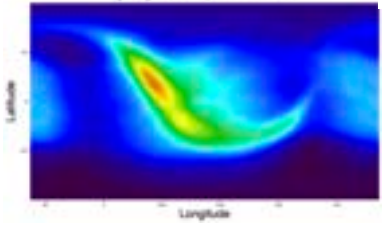


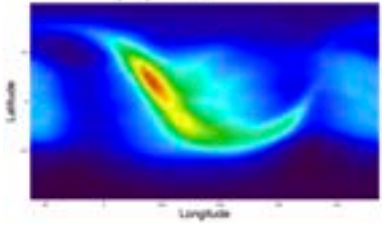
An emulator is needed





?



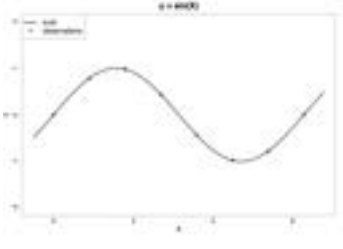


Gaussian process emulators

- Gaussian process: any finite set of observations that is modeled as following a multivariate normal distribution (Gramacy 2020)
- Can be completely specified by its mean $\mu(x)$ and covariance $\Sigma(x, x')$ functions
- $Y(x) \sim MVN(\mu(x), \Sigma(x, x'))$
- A prior over random functions or a posterior over functions given observed values
- Common tool in spatial statistics and computer experiments
- Naturally good at quantifying uncertainty

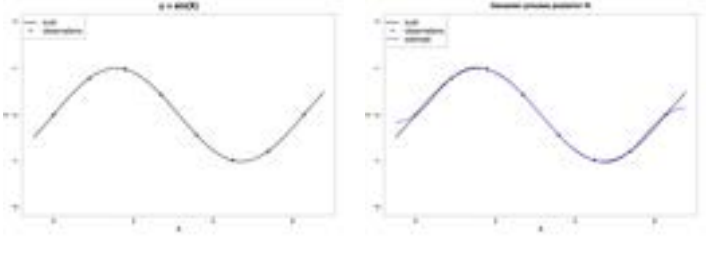
Gaussian process emulators

- Gaussian process: any finite set of observations that is modeled as following a multivariate normal distribution (Gramacy 2020)
- Can be completely specified by its mean $\mu(x)$ and covariance $\Sigma(x, x')$ functions
- $Y(x) \sim MVN(\mu(x), \Sigma(x, x'))$
- A prior over random functions or a posterior over functions given observed values
- Common tool in spatial statistics and computer experiments
- Naturally good at quantifying uncertainty



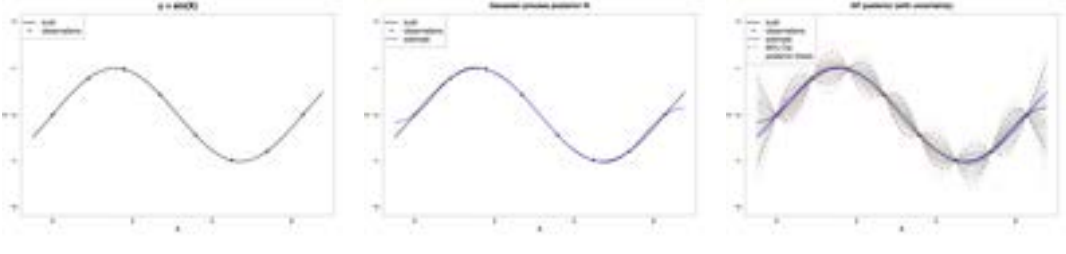
Gaussian process emulators

- Gaussian process: any finite set of observations that is modeled as following a multivariate normal distribution (Gramacy 2020)
- Can be completely specified by its mean $\mu(x)$ and covariance $\Sigma(x, x')$ functions
- $Y(x) \sim MVN(\mu(x), \Sigma(x, x'))$
- A prior over random functions or a posterior over functions given observed values
- Common tool in spatial statistics and computer experiments
- Naturally good at quantifying uncertainty



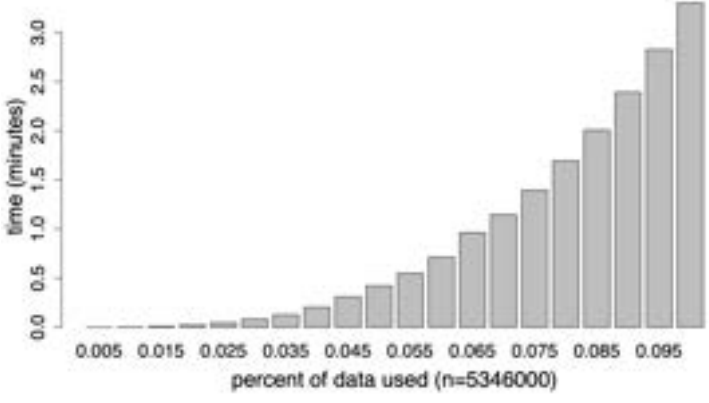
Gaussian process emulators

- Gaussian process: any finite set of observations that is modeled as following a multivariate normal distribution (Gramacy 2020)
- Can be completely specified by its mean $\mu(x)$ and covariance $\Sigma(x, x')$ functions
- $Y(x) \sim MVN(\mu(x), \Sigma(x, x'))$
- A prior over random functions or a posterior over functions given observed values
- Common tool in spatial statistics and computer experiments
- Naturally good at quantifying uncertainty



Gaussian processes and large-scale data

- $L(\boldsymbol{y}|\boldsymbol{X}) \propto |\Sigma(\boldsymbol{X})|^{-1/2} \exp\{\boldsymbol{y}^T \Sigma(\boldsymbol{X})^{-1} \boldsymbol{y}\}$
- Computationally intractable to continue increasing design size
- Inverting a covariance matrix of size $n \times n$ requires computation of order $O(n^3)$



Vecchia approximation

- $L(Y) = \prod_{i=1}^{n} L(Y_i|Y_1, Y_2, \ldots, Y_{i-1}) = \prod_{i=1}^{n} L(Y_i|Y_{g(i)})$
- $g(1) = \emptyset$; $g(i) = \{1, 2, \ldots, i-1\}$

Vecchia approximation

- $L(Y) = \prod_{i=1}^{n} L(Y_i|Y_1, Y_2, \ldots, Y_{i-1}) = \prod_{i=1}^{n} L(Y_i|Y_{g(i)})$
- $g(1) = \emptyset$; $g(i) = \{1, 2, \ldots, i-1\}$

Vecchia:

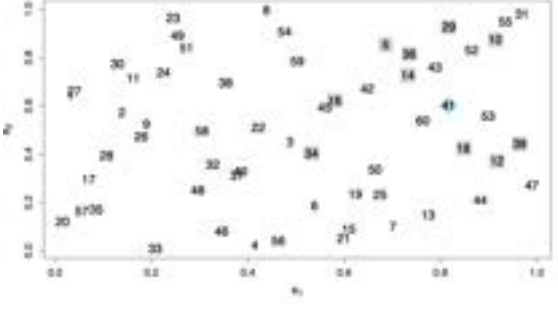
- $L(Y) \approx \prod_{i=1}^{n} L(Y_i|Y_{h(i)})$
- $h(i) \subset \{1, 2, \ldots, i-1\}$; $|h(i)| = m \ll n$

Vecchia approximation

- $L(Y) = \prod_{i=1}^{n} L(Y_i|Y_1, Y_2, \ldots, Y_{i-1}) = \prod_{i=1}^{n} L(Y_i|Y_{g(i)})$
- $g(1) = \emptyset$; $g(i) = \{1, 2, \ldots, i-1\}$

Vecchia:

- $L(Y) \approx \prod_{i=1}^{n} L(Y_i|Y_{h(i)})$
- $h(i) \subset \{1, 2, \ldots, i-1\}$; $|h(i)| = m \ll n$

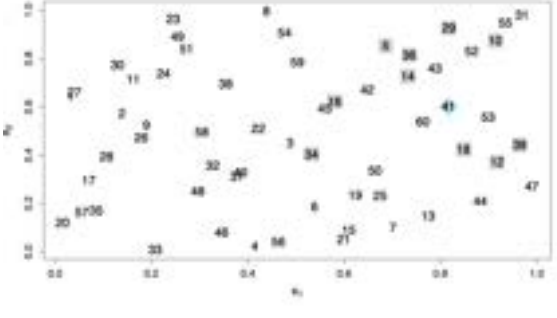


Vecchia approximation

- $L(Y) = \prod_{i=1}^{n} L(Y_i|Y_1, Y_2, \ldots, Y_{i-1}) = \prod_{i=1}^{n} L(Y_i|Y_{g(i)})$
- $g(1) = \emptyset$; $g(i) = \{1, 2, \ldots, i-1\}$

Vecchia:

- $L(Y) \approx \prod_{i=1}^{n} L(Y_i|Y_{h(i)})$
- $h(i) \subset \{1, 2, \ldots, i-1\}$; $|h(i)| = m \ll n$



How does this help?

- $\Sigma(X)^{-1}$ is now a sparse matrix
- (i, j)th element is 0 if y_i, y_j are conditionally independent
- Further, Cholesky decomposition $\Sigma(X)^{-1} = UU^T$ is even more sparse

Scaled Vecchia approximation (Katzfuss et al., 2022)

- Distance between inputs depends on some choice of the scaling of inputs
- Inputs vary in the magnitude of their effect on the response

Scaled Vecchia approximation (Katzfuss et al., 2022)

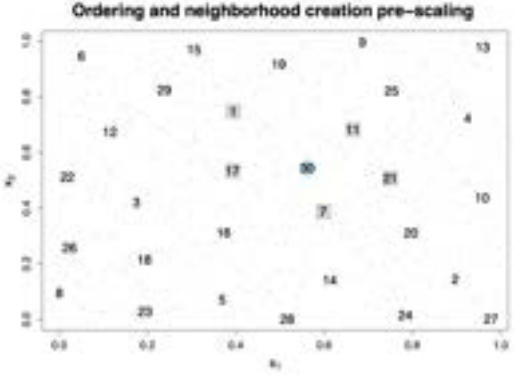
- Distance between inputs depends on some choice of the scaling of inputs
- Inputs vary in the magnitude of their effect on the response
- **Scaled Vecchia:** Pre-scale inputs before determining neighborhoods

Scaled Vecchia approximation (Katzfuss et al., 2022)

- Distance between inputs depends on some choice of the scaling of inputs
- Inputs vary in the magnitude of their effect on the response
- **Scaled Vecchia:** Pre-scale inputs before determining neighborhoods
- Isotropic covariance function can be used when fitting the Gaussian process

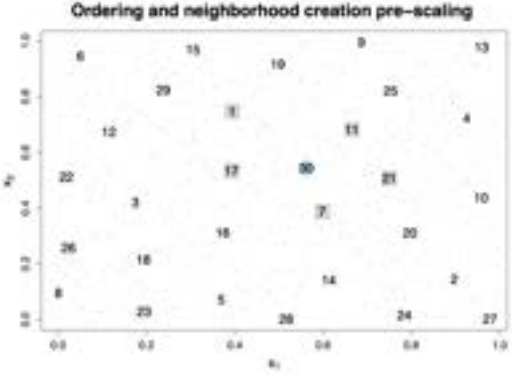
Scaled Vecchia approximation (Katzfuss et al., 2022)

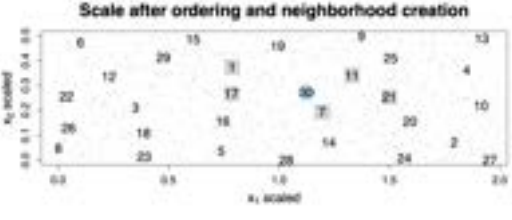
- Distance between inputs depends on some choice of the scaling of inputs
- Inputs vary in the magnitude of their effect on the response
- **Scaled Vecchia:** Pre-scale inputs before determining neighborhoods
- Isotropic covariance function can be used when fitting the Gaussian process



Scaled Vecchia approximation (Katzfuss et al., 2022)

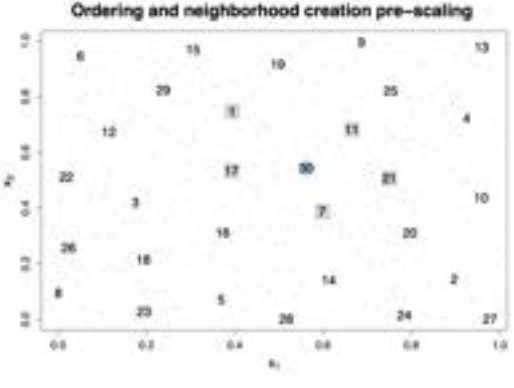
- Distance between inputs depends on some choice of the scaling of inputs
- Inputs vary in the magnitude of their effect on the response
- **Scaled Vecchia:** Pre-scale inputs before determining neighborhoods
- Isotropic covariance function can be used when fitting the Gaussian process

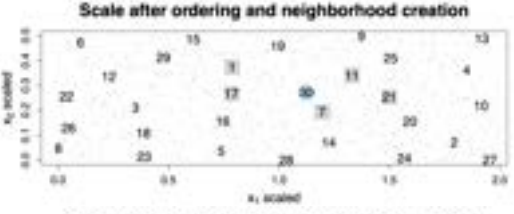


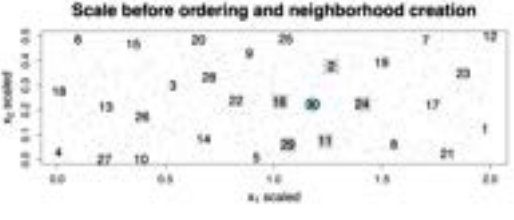


Scaled Vecchia approximation (Katzfuss et al., 2022)

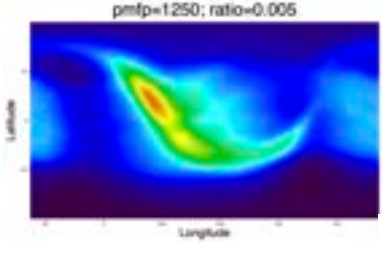
- Distance between inputs depends on some choice of the scaling of inputs
- Inputs vary in the magnitude of their effect on the response
- **Scaled Vecchia:** Pre-scale inputs before determining neighborhoods
- Isotropic covariance function can be used when fitting the Gaussian process

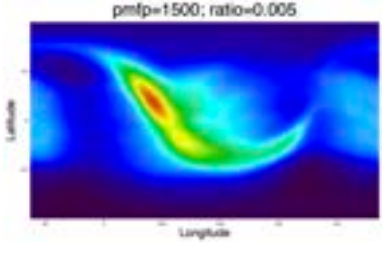




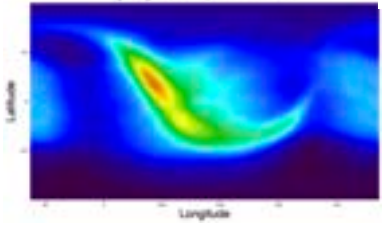


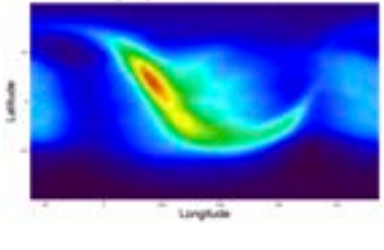
Emulator output



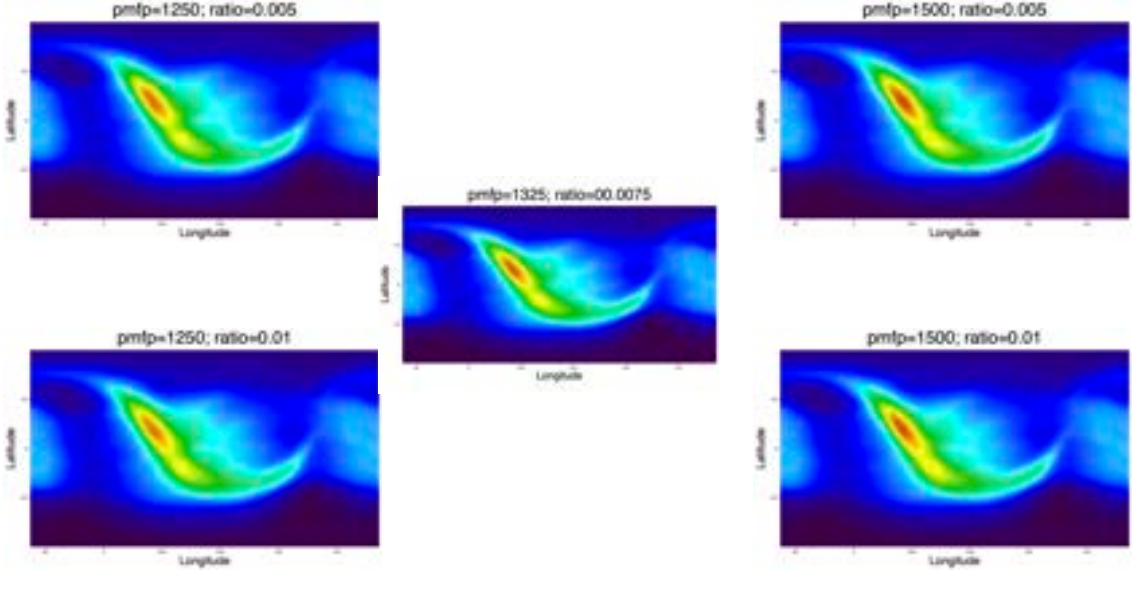


?

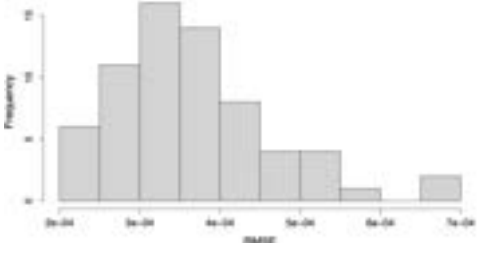


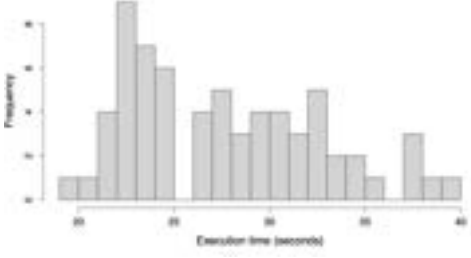


Emulator output



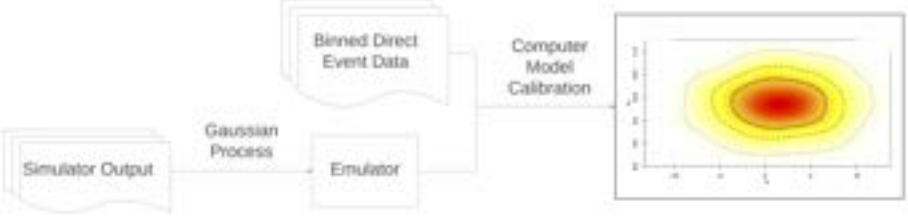
Quantitative emulator results



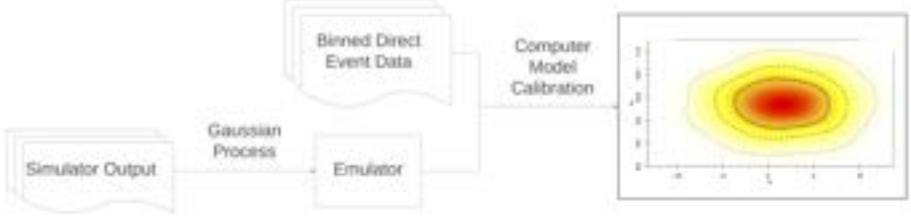


- Takeaway: the GP emulator is doing a good job both visually and quantitatively, and there is room for improvement!

Computer model calibration



Computer model calibration



Binned Direct Event Data = Simulator(**pmfp**, **ratio**, lat, lon, esa) + error

Metropolis-Hastings

Kennedy + O'Hagan:

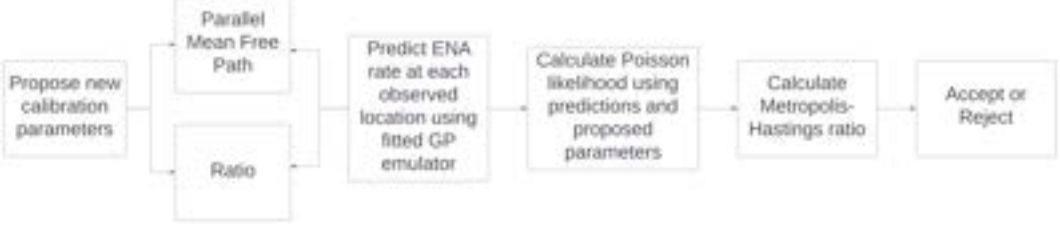
- $y^R(x) = y^M(x, u^*) + b(x)$
- $Y^F(x) = y^M(x, u^*) + b(x) + \epsilon$
- $\begin{bmatrix} Y_{n_M} \\ Y_{n_F} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{n_M} & \Sigma_{n_M}(X_{n_F}, u) \\ \Sigma_{n_M}(X_{n_F}, u)^T & \Sigma_{n_F}(u) + \Sigma^b_{n_F} \end{bmatrix} \right)$

Metropolis-Hastings

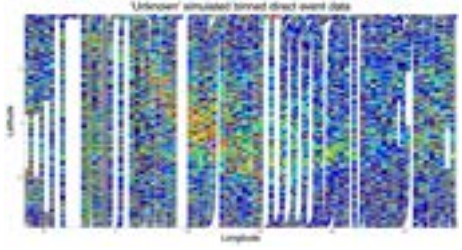
Kennedy + O'Hagan:

- $y^R(x) = y^M(x, u^*) + b(x)$
- $Y^F(x) = y^M(x, u^*) + b(x) + \epsilon$
- $\begin{bmatrix} Y_{n_M} \\ Y_{n_F} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{n_M} & \Sigma_{n_M}(X_{n_F}, u) \\ \Sigma_{n_M}(X_{n_F}, u)^T & \Sigma_{n_F}(u) + \Sigma^b_{n_F} \end{bmatrix} \right)$

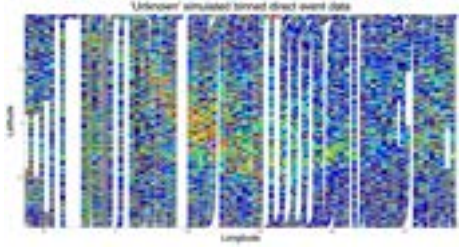
Proposed McMC framework with Poisson response:

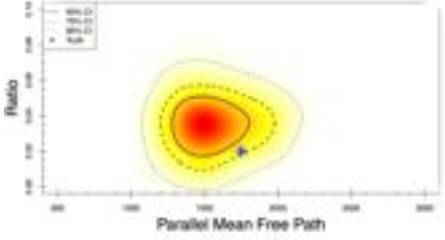


Computer model parameter estimation

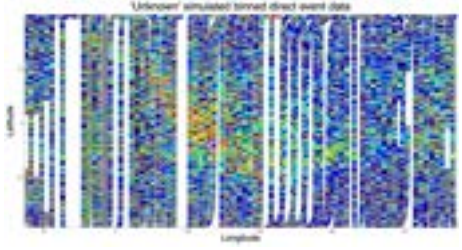


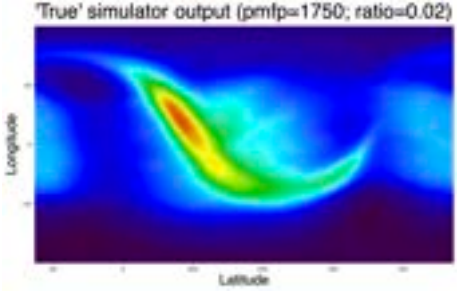
Computer model parameter estimation

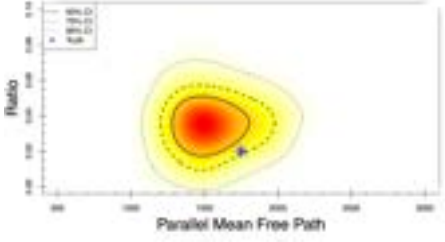




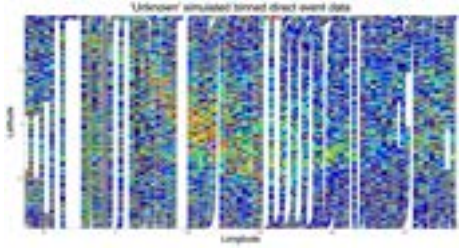
Computer model parameter estimation

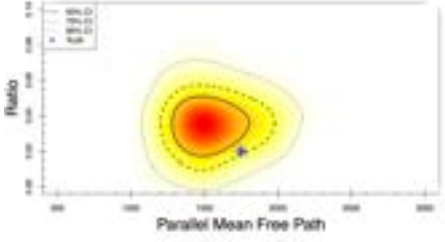


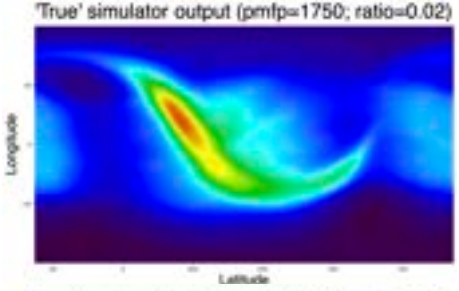


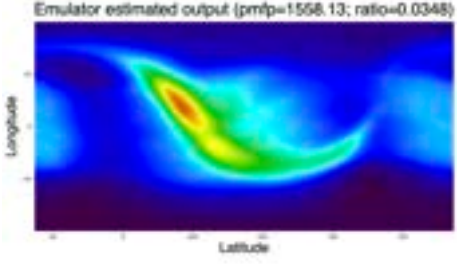


Computer model parameter estimation

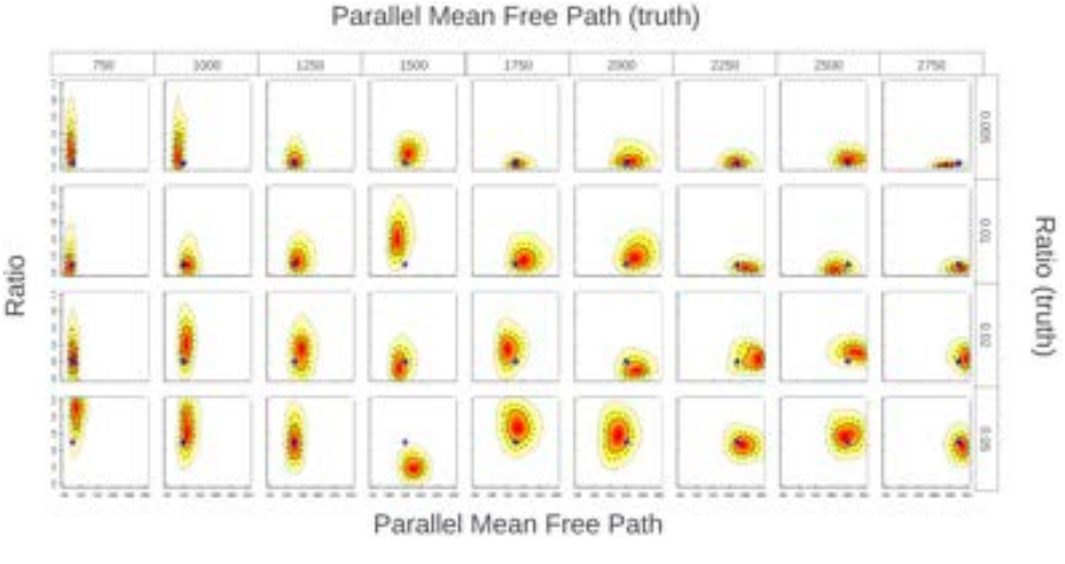




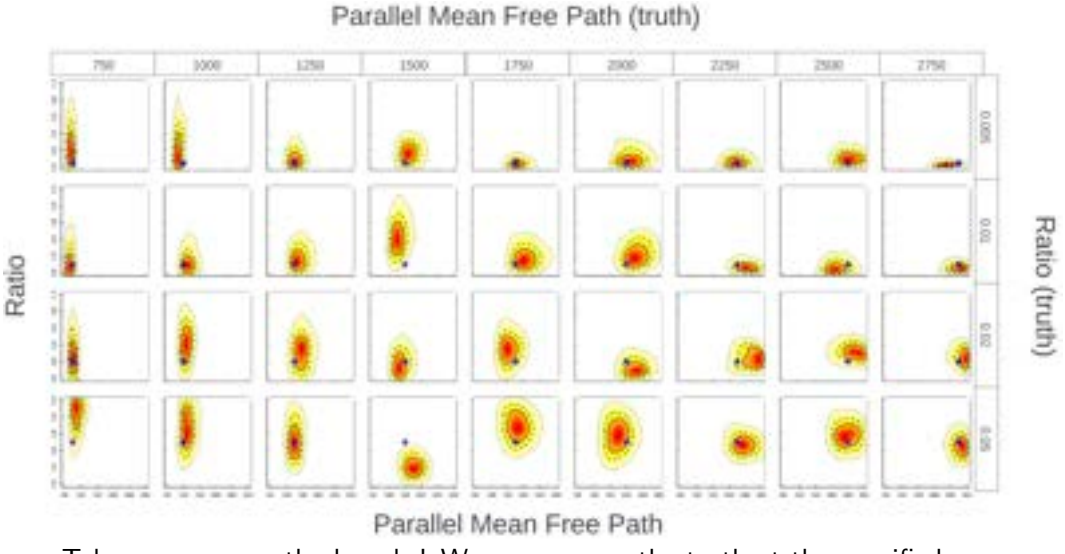




Computer model parameter estimation



Computer model parameter estimation



- Takeaway: our method works! We can recover the truth at the specified confidence level!

Conclusions

- Gaussian process surrogate models can expand the current set of sky maps to any combination of parameters with high accuracy

Conclusions

- Gaussian process surrogate models can expand the current set of sky maps to any combination of parameters with high accuracy
- Statistical computer model calibration can recover the true computer model parameters from simulated binned direct event data

Conclusions

- Gaussian process surrogate models can expand the current set of sky maps to any combination of parameters with high accuracy
- Statistical computer model calibration can recover the true computer model parameters from simulated binned direct event data
- Computer model has been shown to be incomplete, indicating that our process needs additional work to account for it

Next Steps

- Model discrepancy using a Gaussian process or deep Gaussian process

Next Steps

- Model discrepancy using a Gaussian process or deep Gaussian process
- Replace scaled Vecchia approximation for full uncertainty quantification

Next Steps

- Model discrepancy using a Gaussian process or deep Gaussian process
- Replace scaled Vecchia approximation for full uncertainty quantification
- Work with theoretical physicists to inform where the computer model needs improvement

Next Steps

- Model discrepancy using a Gaussian process or deep Gaussian process
- Replace scaled Vecchia approximation for full uncertainty quantification
- Work with theoretical physicists to inform where the computer model needs improvement
- Consider other inputs such as ESA (energy level) and time (sun cycle)