SAND2023-05100C

Sandia National Laboratories

Data Fusion with Uncertainty Quantification for Observational Data

Audrey McCombs

Justin Li, Mauricio Campos, Lyndsay Shand, J. Gabriel Huerta

SIAM GS23

Bergen, Norway

Wednesday, June 21, 2023

U.S. DEPARTMENT OF ENERGY NNSA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

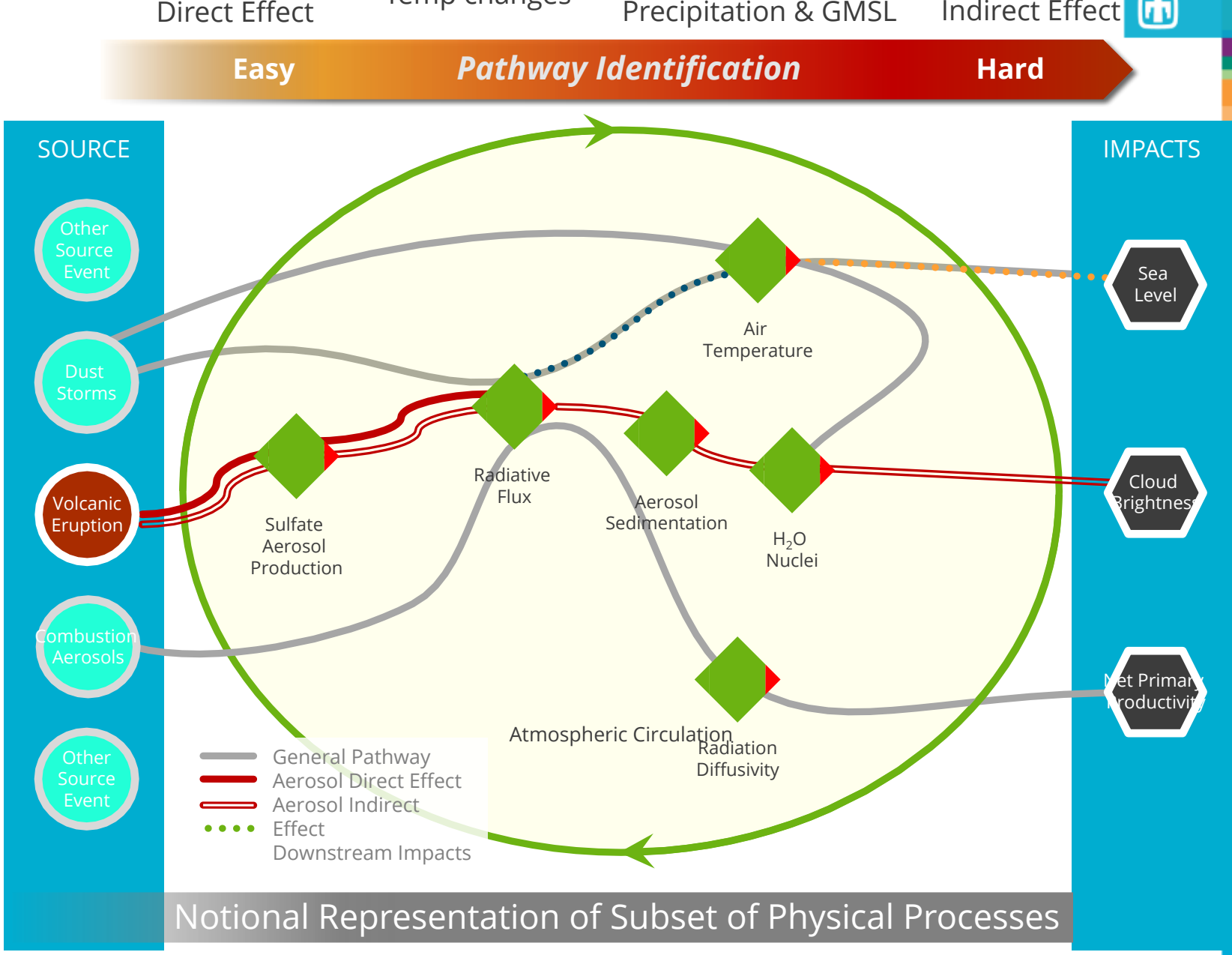
CLDERA Project

Can pathways be discovered that reveal relationships between source and impacts?

Aerosols → Radiative Flux → Temperature

Inverse formulation:

- Can CLDERA identify location and magnitude of Mt. Pinatubo eruption from the temperature perturbation?
- How does the attribution change as a function of eruption characteristics and lag from eruption time?



Project Thrusts

- Simulated Pathways
  - Random Forest Regressions
  - Sensitivity Analysis
  - Tracers
  - Profiling
  - Simulation-Based Analysis
  - Clustering
- Observed Pathways
  - Echo State Networks
  - Change-Point Detection
  - Space-Time Statistical Methods
  - Data Fusion
- Attribution
  - Inverse Optimization
  - Enhanced Fingerprinting
  - Causal Modeling
  - Dimension Reduction

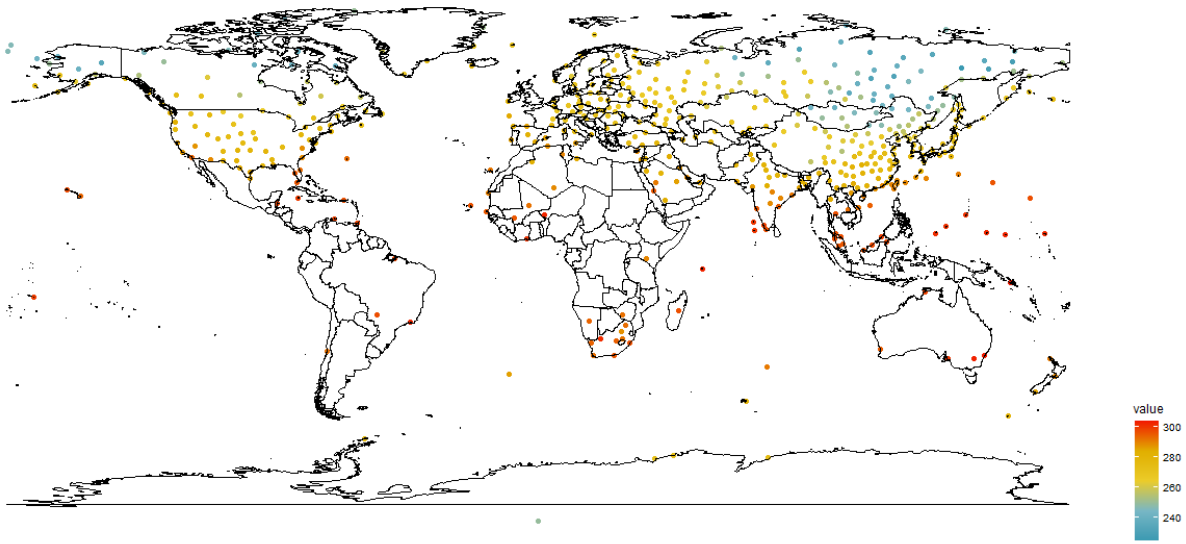
Discover simulated pathways in ESMs between a source and its impacts to uncover sound causal relationships, rather than piecemeal source-impact correlations

Develop statistical methods to simultaneously account for dynamic spatio-temporal evolution and key atmospheric processes

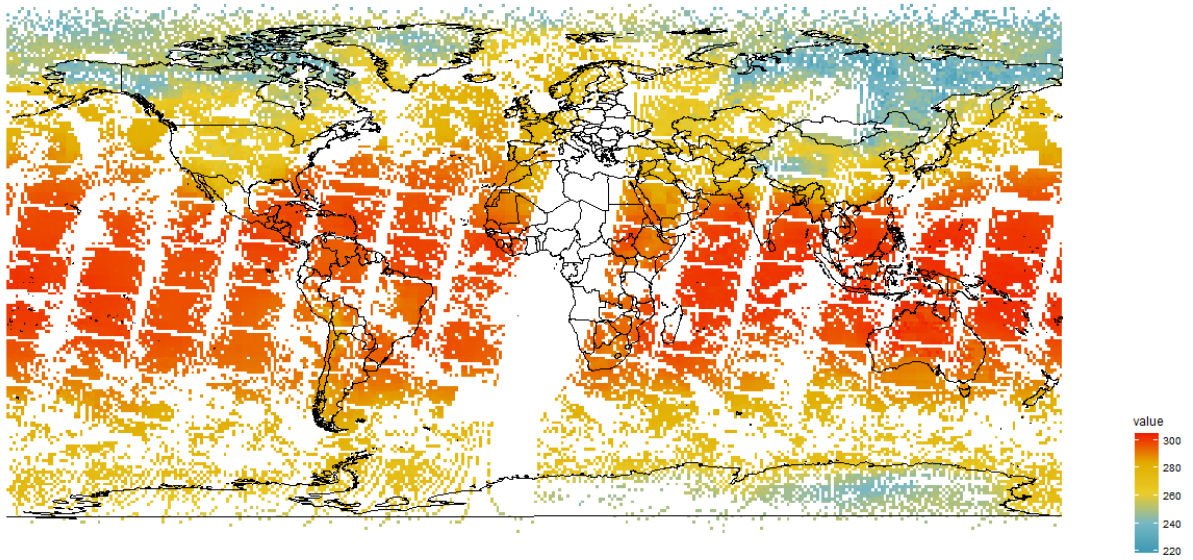
Develop new approaches that preserve important pathway features in the system, but cull the high-dimensional space to enable dominant source-to-impact attribution

Data Sources: Temperature

IGRA: January 1, 1990

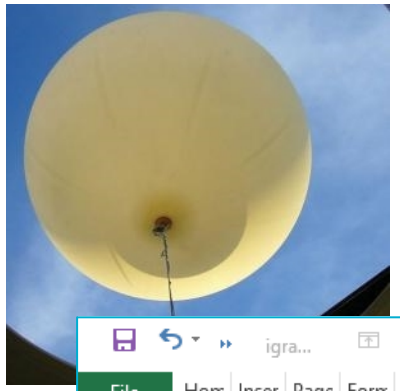


TOVS: January 1, 1990



Source	What is it?	Date Range	Spatial Resolution	Data Type
Integrated Global Radiosonde Archive (**IGRA**)	Radiosonde	1905-present varies by station	Point coverage Vertical: finely resolved	Point
NOAA-**TOVS**	Vertical sounder suite	1988-1998	1° x 1° Vertical: surface to 30mb	Areal

Change of Support Problem



IGRA

igra...

File | Hom | Inser | Page | Form | Data | Revie | View | JMP

A1 | lat

	A	B	C
1	lat	lon	temp
2	-1.3036	36.7597	289.75
3	-1.3833	-48.4833	298.15
4	-12.4239	130.893	298.75
5	-14.3383	-170.719	299.35
6	-18.8	47.4833	296.95
7	-20.15	28.617	292.55
8	-20.4667	-54.6667	297.15
9	-22.5667	17.1	294.15
10	-22.8167	-43.25	302.55
11	-23.8667	29.45	292.95
12	-25.91	28.2111	28[illegible]8.15
13	-28.25	28.3333	29[illegible]1.95

This is an observation of temperature at this location

TOVS

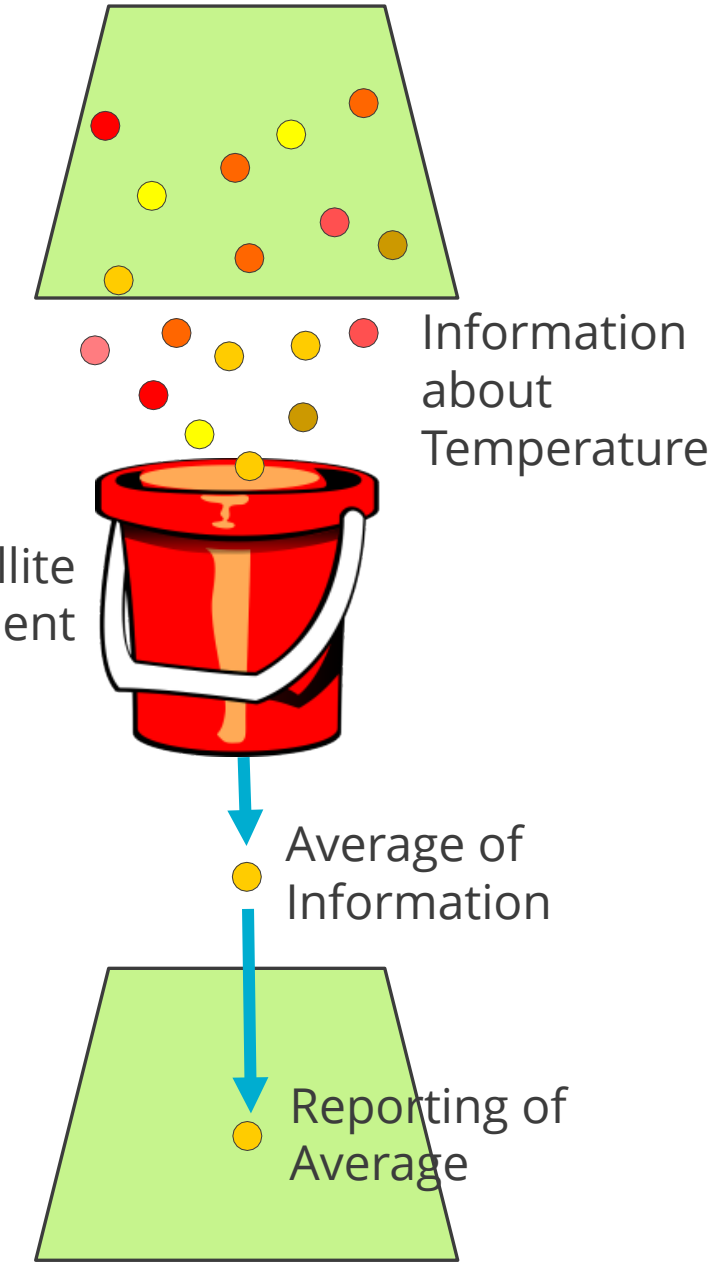
tovs_1...

File | Hom | Inser | Page | Form | Data | Revie | View | JMP

A2 | -75.5

	A	B	C	D
1	lat	lon	temp	
2	-75.5	-179.5	274.723	
3	-50.5	-179.5	283.883	
4	-49.5	-179.5	282.274	
5	-47.5	-179.5	284.264	
6	-46.5	-179.5	284.643	
7	-43.5	-179.5	286.071	
8	-42.5	-179.5	287.149	
9	-41.5	-179.5	287.943	
10	-40.5	-179.5	289.159	
11	-36.5	-179.5	291.959	
12	-35.5	-179.5	292[illegible]139	
13	-34.5	-179.5	29[illegible].46	

This is not



Change of Support Problem

9	7	4	9	5	4
1	8	7	9	1	5
9	1	3	1	6	9
6	5	2	7	7	8
9	5	4	4	8	9
5	7	6	1	4	1

Mean = 5.4
Std. Dev. = 2.8

6.3	7.3	3.8
5.3	3.3	7.5
6.5	3.8	5.5

Mean = 5.4
Std. Dev. = 1.6

Sources of Uncertainty
- Measurement uncertainty (IGRA and TOVS)
- Uncertainty due to averaging (TOVS only)
- Inaccurate uncertainty quantification (TOVS only)

Data Fusion Goal:

Fuse multiple observational datasets
- Spatially and temporally complete
- With uncertainty quantification
- In near-real-time

Accounting for
- Change of support
- Spatial and temporal auto-correlation
- Non-stationarity

For use in
- Causal pathway analysis
- Validation of synthetic and re-analysis data

Interpolation Methods

Heaton, M.J., et. al. 2019. A case study competition among methods for analyzing large spatial data.

General methodology	Examples (not comprehensive)	Spatio-temporal model?	Handles COS?	Handles large datasets?	R package?
Low-rank methods	1. Fixed-rank kriging (Cressie and Johannesson) 2. Lattice kriging (Nychka) 3. Predictive processes (Finley) 4. Nguyen, Cressie et al	3. **spBayes** can do space-time or COS but not both 4: Nguyen et. al 2013 extend to spatio-temporal Others spatial only	1, 2, 4: Yes 3. **spBayes** can do space-time or COS but not both	Yes	1. **FRK** 2. **LatticeKrig** 3. **spBayes** 4. No
Sparse covariance methods	1. Spatial partitioning (Heaton) 2. Covariance tapering (Furrer)	Spatial only	No	Yes	1. **spBayes** 2. No
Sparse precision methods	1. Multiresolution approximations (Katzfuss and Hammerling) 2. SPDE/INLA (Lindgren) 3. Nearest neighbor (Datta and Finley) 4. Periodic embedding (Guinness)	3. Yes Others spatial only	2, 3. Yes Others No	Yes	1. **Gpveccia** 2. **INLA** 3. **spNNGP** 4. **GpGp**
Algorithmic approaches	1. Metakriging (Guhaniyogi) 2. Gapfill (Gerber) 3. Local approximation Gaussian process (Gramacy and Sun)	Spatial only	No	Yes	1. No 2. No 3. **laGP, deepGP, quack**
Re-scaling approaches	1. Spatial downscaling (Mugglin and Carlin) 2. Spatial upscaling (Bradley, Wikle, Holan)	Spatial only	Yes	No	1. No 2. **stcop**

laGP

Gaussian Process

For an *N*-dimensional random vector $\boldsymbol{x}, \boldsymbol{x} \sim MVN(\boldsymbol{\mu}, \Sigma)$

We can partition the vector

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{pmatrix} \quad \text{with sizes} \quad \begin{pmatrix} q \text{ x } 1 \\ (N-q) \text{ x } 1 \end{pmatrix}$$

Similarly

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix} \quad \text{with sizes} \quad \begin{pmatrix} q \text{ x } 1 \\ (N-q) \text{ x } 1 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} \Sigma_{12} \\ \Sigma_{21} \Sigma_{22} \end{pmatrix} \quad \begin{pmatrix} q \text{ x } q & q \text{ x } (N-q) \\ (N-q) \text{ x } q & (N-q) \text{ x } (N-q) \end{pmatrix}$$

The conditional distribution is: $\boldsymbol{x_1} | \boldsymbol{x_2} \sim MVN(\bar{\boldsymbol{\mu}}, \bar{\Sigma})$

where $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu_1} + \Sigma_{12} \Sigma_{22}^{-1} (\boldsymbol{x_2} - \boldsymbol{\mu_2})$ and $\bar{\Sigma} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$

Matrix inversion: $O(n^3)$

In **laGP**, Σ is a standard squared-exponential distance function with possible tuning using a scale and/or nugget

The **laGP** package solves the matrix inversion problem by introducing sparsity into the covariance matrix, using a nearest-neighbor approach.

LatticeKrig

- Spatial process is the sum of radial basis functions
  - Constructed using a Wendland compactly-supported correlation function
  - Nodes arranged on a rectangular grid
- Coefficients on basis functions are assumed to be correlated
  - Distributed according to a Gaussian Markov random field (covariance also Wendland)
- Linear model fit using least squares, then basis functions fit to the residuals

$$\boldsymbol{y} = X\boldsymbol{d} + \Phi\mathbf{c} + \mathbf{e}$$

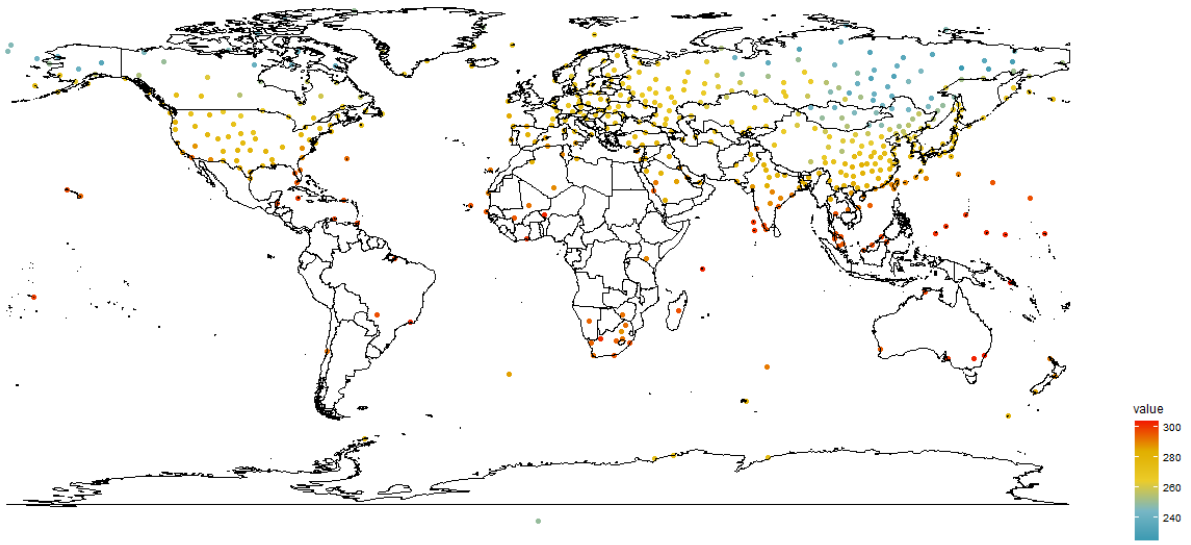
$$g(\boldsymbol{s}) = \sum_{k=1}^{n} \phi^{k}(\boldsymbol{s})\hat{\boldsymbol{d}}_k + \sum_{k=1}^{m} \psi_k(\boldsymbol{s})\hat{\boldsymbol{c}}_k$$

y: observations
X: matrix of locations, covariates
d: vector of coefficients for the linear model
Φ: matrix of radial basis functions
(evaluated at data points)
c: vector of coefficients on the basis functions
e: measurement error
g(**s**): true process

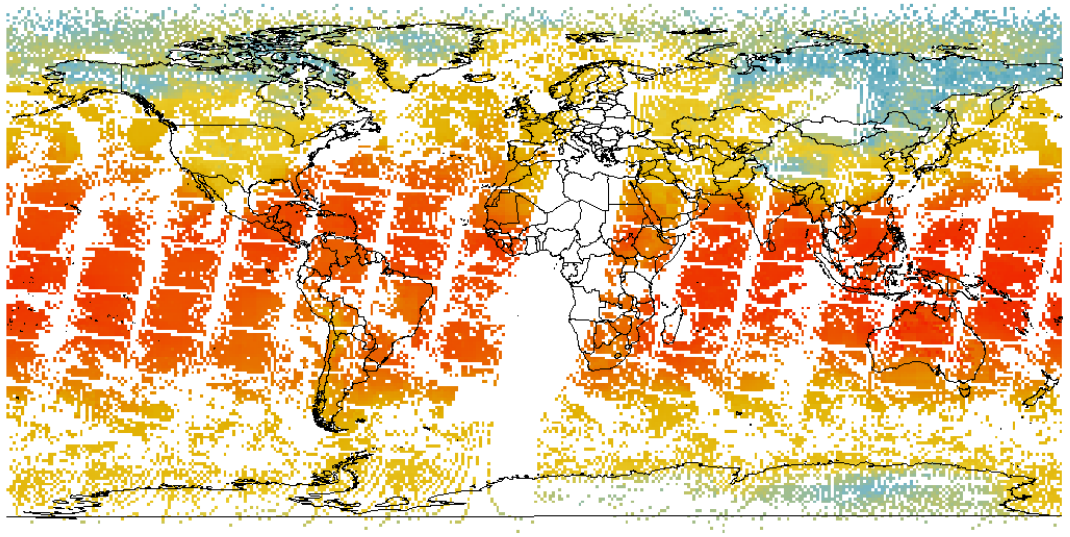
The **LatticeKrig** package solves the matrix inversion problem by introducing sparsity using basis functions and covariance functions that are nonzero only on a compact interval

laGP

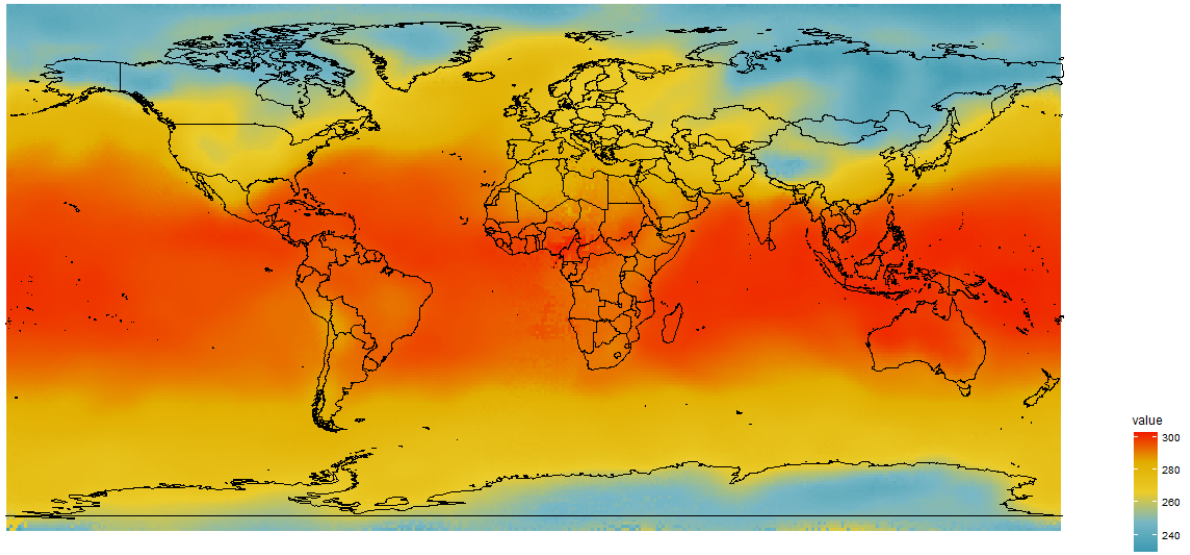
IGRA: January 1, 1990



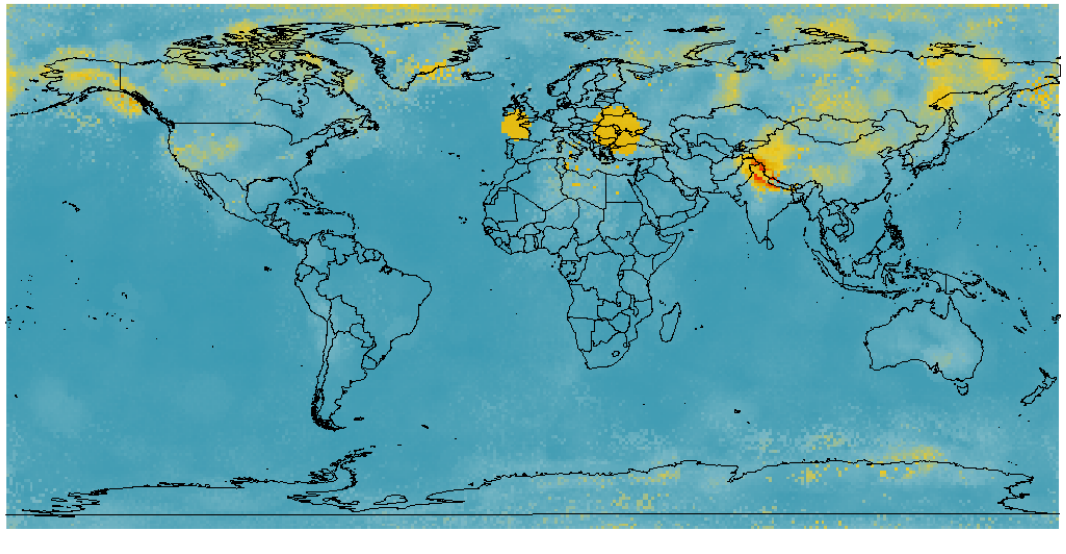
TOVS: January 1, 1990



Predicted Mean

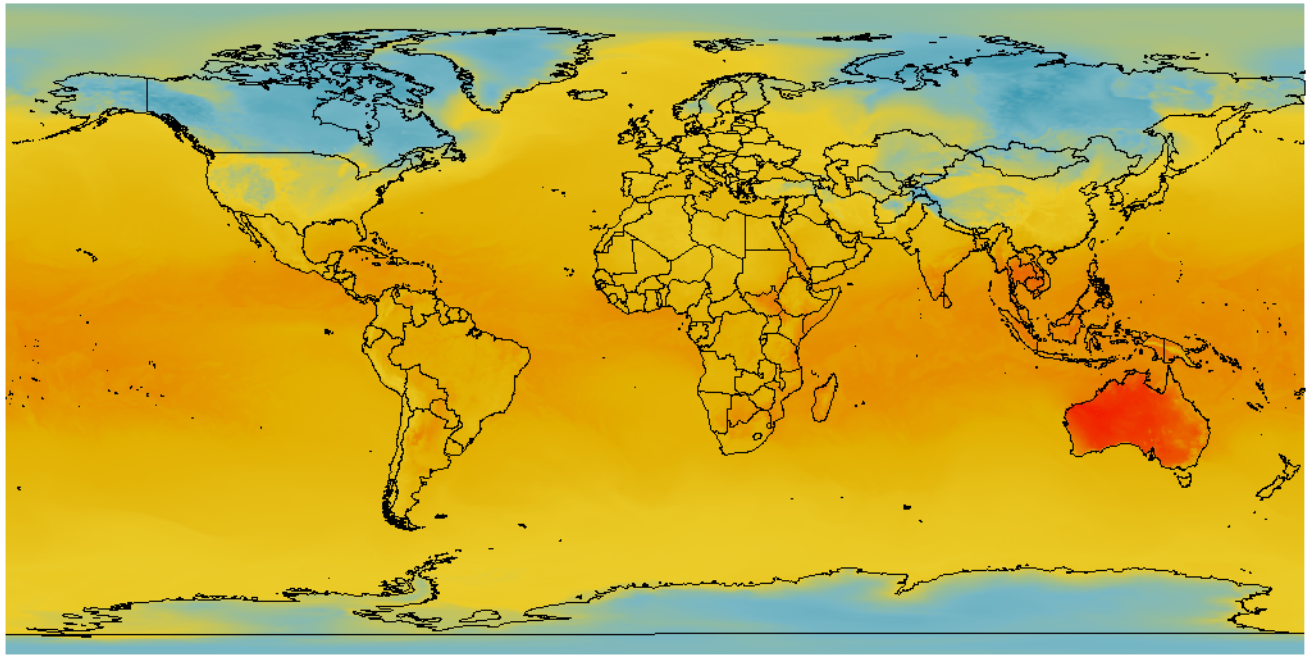


Prediction Error

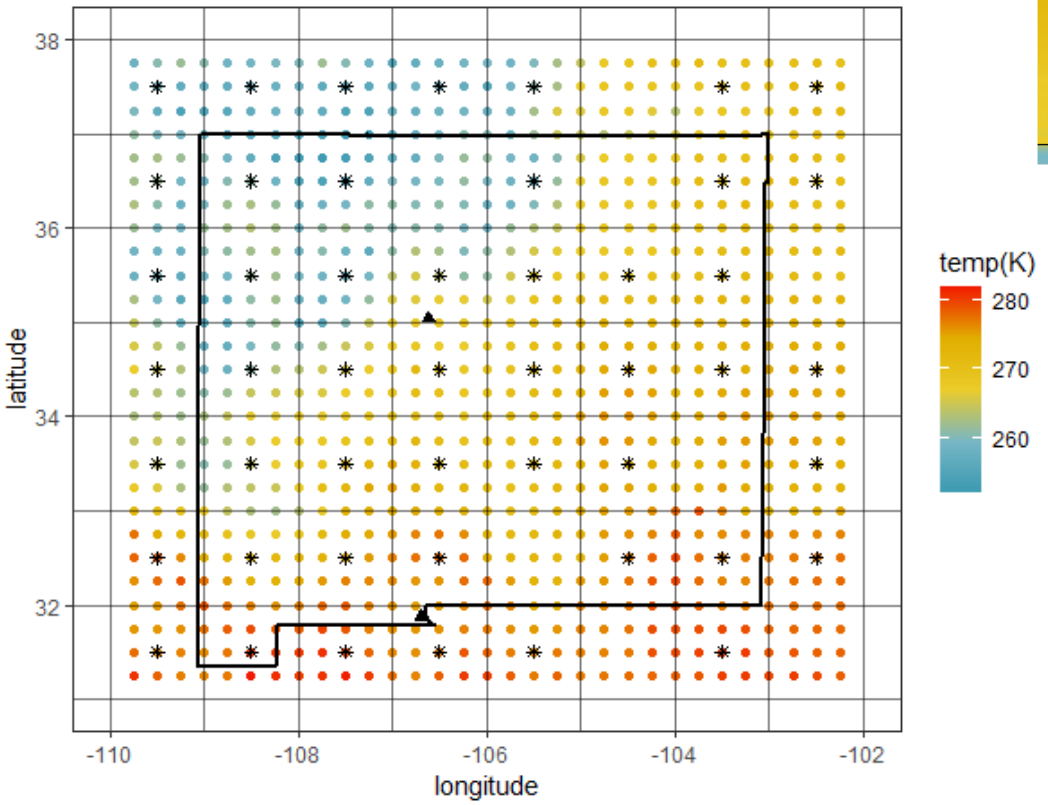


laGP Boostrap Experiment

ERA5 Reanalysis Data: Jan 1, 1990

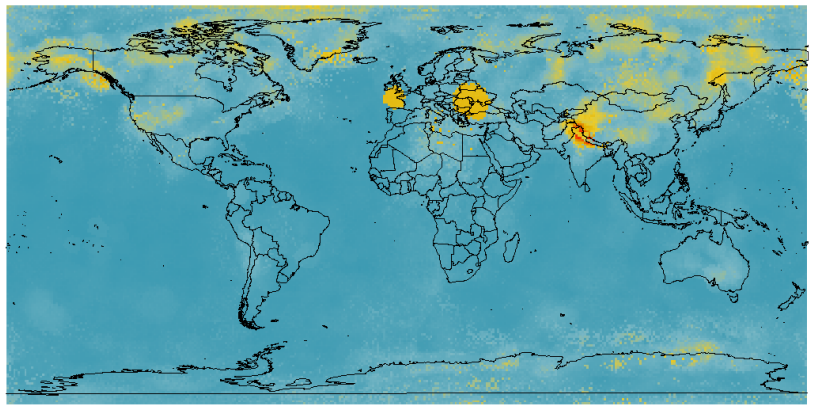


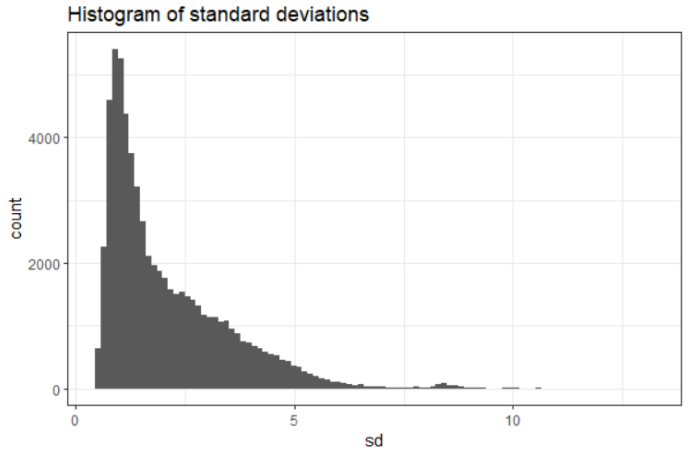
ERA5, TOVS, IGRA
New Mexico Jan 1, 1990



Bootstrap procedure:

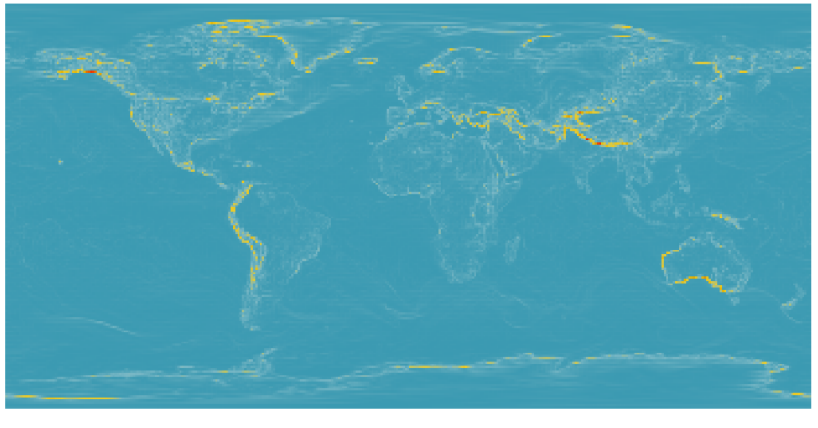
1) Calculate variance of ERA5 reanalysis data in 1-deg grid square
2) Sample from a Normal
   - mean = TOVS/IGRA obs
   - var = ERA5 var
3) Create 199 bootstrap samples, plus one with obs values (200 total dfs)
4) Each bootstrap dataset modeled using **laGP**
   - Each model takes about 11 min using 46 cores: ~ 36 hrs total
5) Calculate variance across 200 predicted means for each grid cell

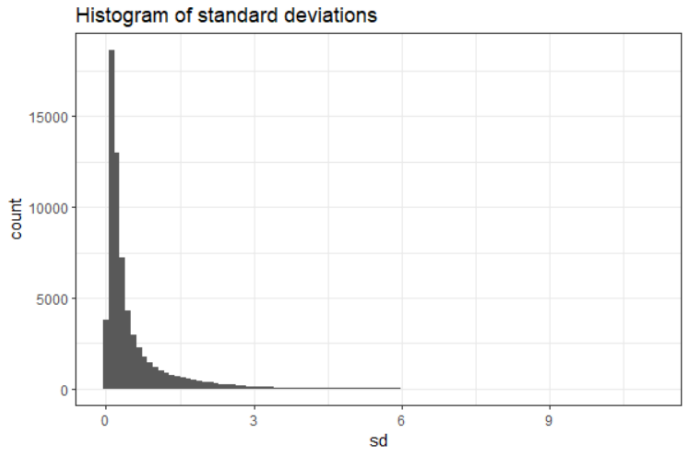




laGP Pred errors

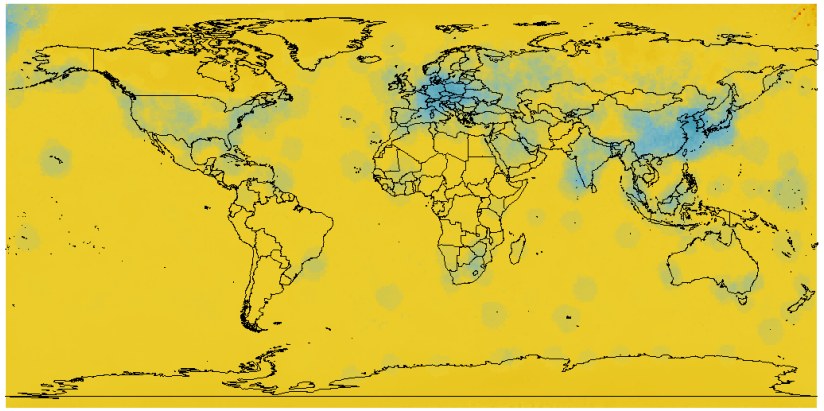
Min: 0.461
1st Q: 1.032
Median: 1.595
4 * median = 6.38
Mean: 2.097
3rd Q: 2.830
Max: 13.102

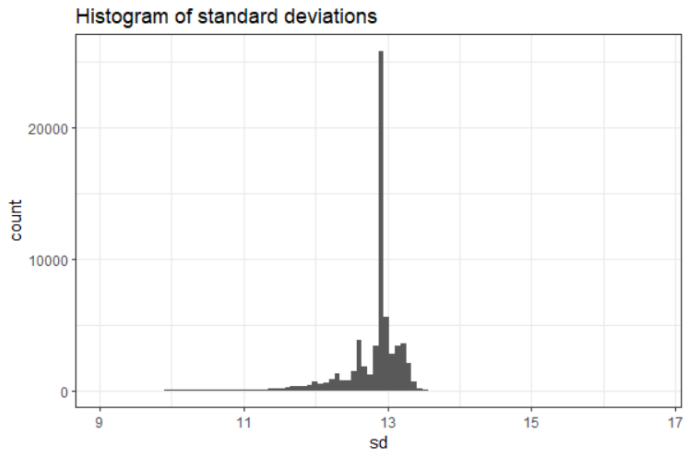




ERA5 SDs

Min: 0.005
1st Q: 0.130
Median: 0.250
4 * median = 1.00
Mean: 0.529
3rd Q: 0.576
Max: 11.074



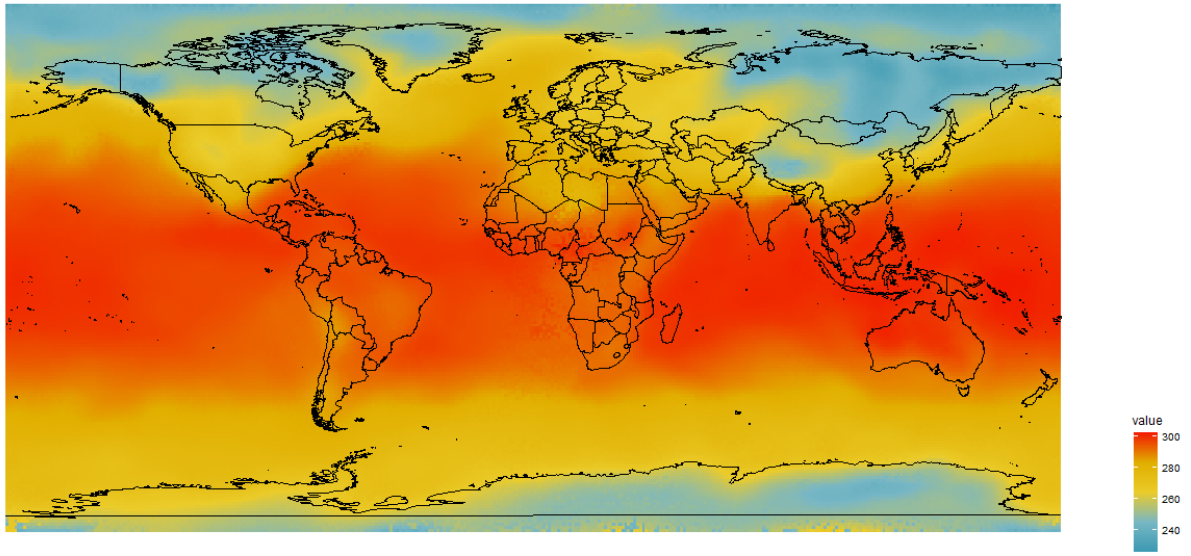


Bootstrap Pred errors

Min: 9.113
1st Q: 12.752
Median: 12.910
4 * median = 51.64
Mean: 12.805
3rd Q: 12.966
Max: 16.667

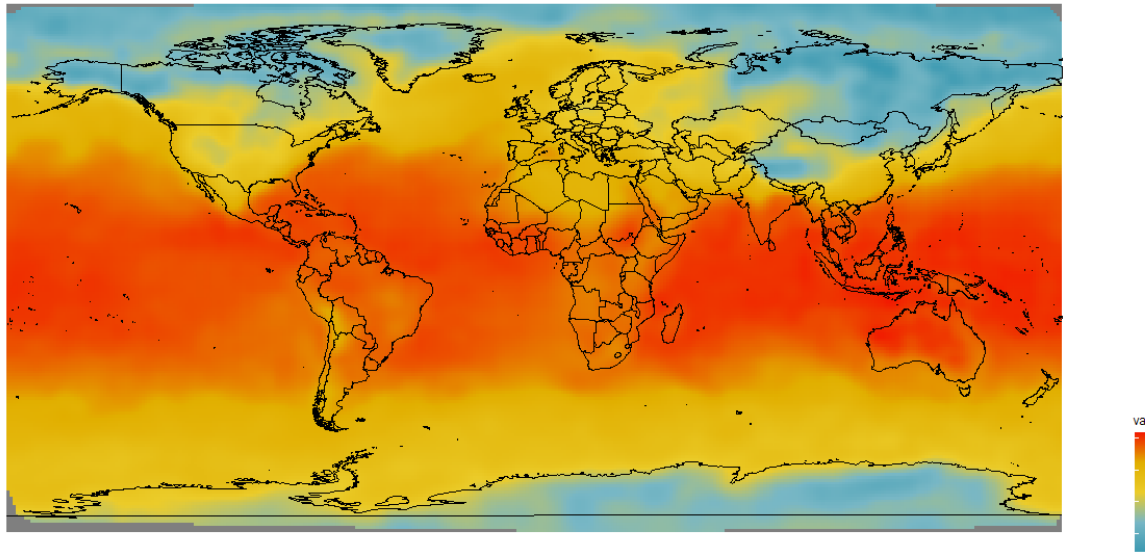
Comparison of Interpolators

Predictions from **laGP** using TOVS data




```
system.time()
user       system  elapsed
8142.075    0.972  230.416
```

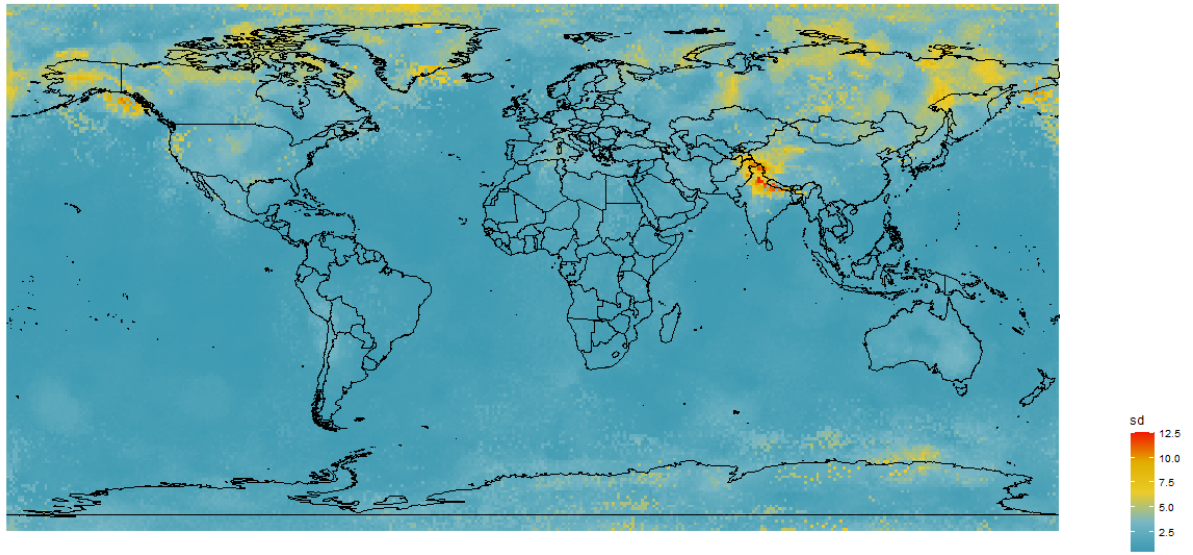
Predictions from **LatticeKrig** using TOVS data




```
system.time()
user    system  elapsed
96.272   2.957  99.448
```

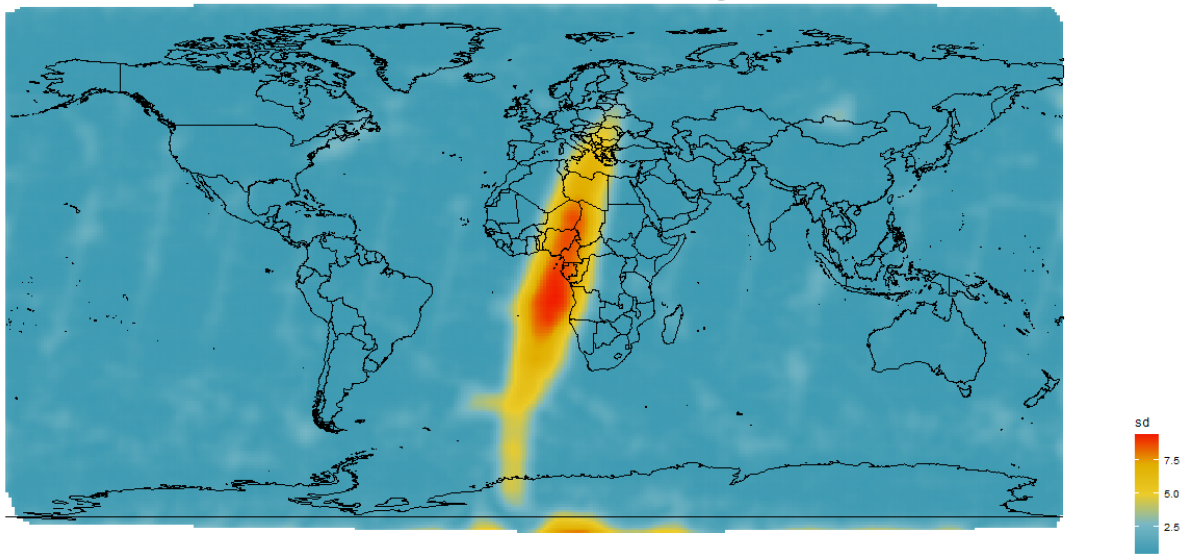
Model Uncertainty

Prediction error from **laGP**



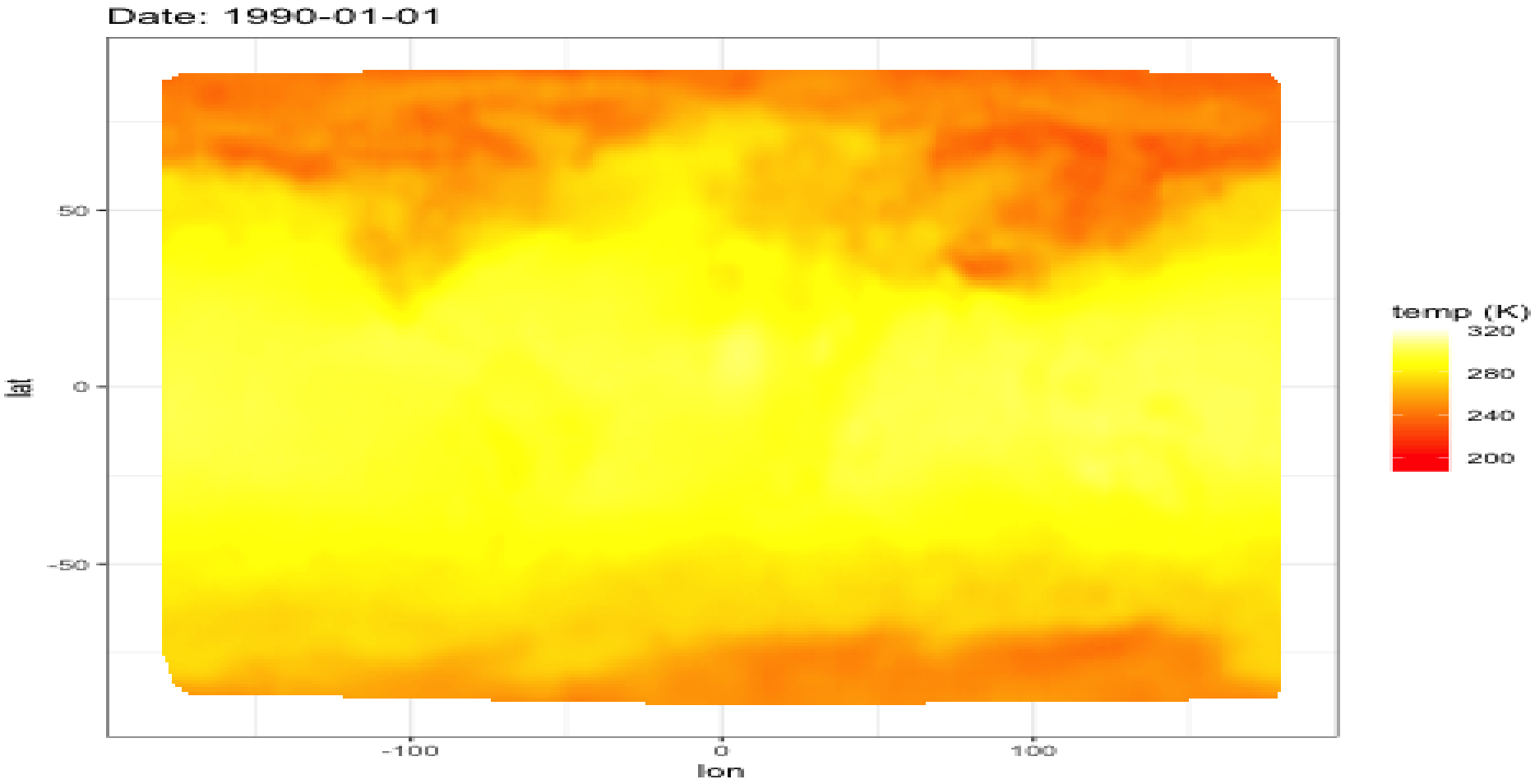
Min: 0.4779
1st Quart: 1.017
Median: 1.539
Mean: 2.002
3rd Quart: 2.685
Max: 12.57

Prediction error from **LatticeKrig**



Min: 0.393
1st Quart: 0.570
Median: 0.677
Mean: 0.995
3rd Quart: 0.930
Max: 9.38

LatticeKrig predictions, 1990



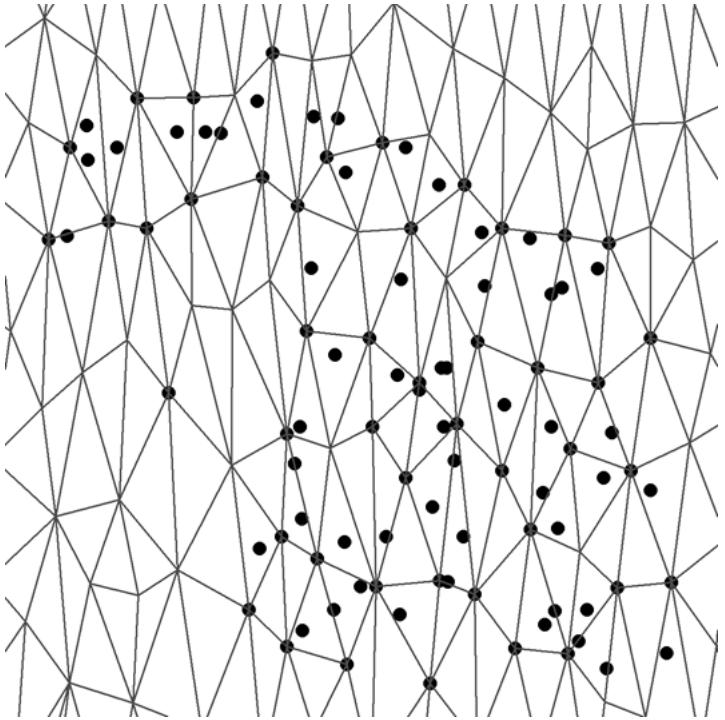
Daily model runs in ~6-7 min
Jan 1, 1990 – June 30, 1993: 6.5 days

INLA

Moraga, P, et al. 2017.

- INLA: Approximate Bayesian inference in a latent Gaussian model
  - LaPlace approximations for posterior marginals
  - Numerical algorithms for sparse matrices
- SPDE: the continuously indexed Gaussian field $S(\boldsymbol{x})$ is represented as a discretely indexed Gaussian Markov random field (GMRF) using basis functions
  - Basis functions defined on a triangulation of the region

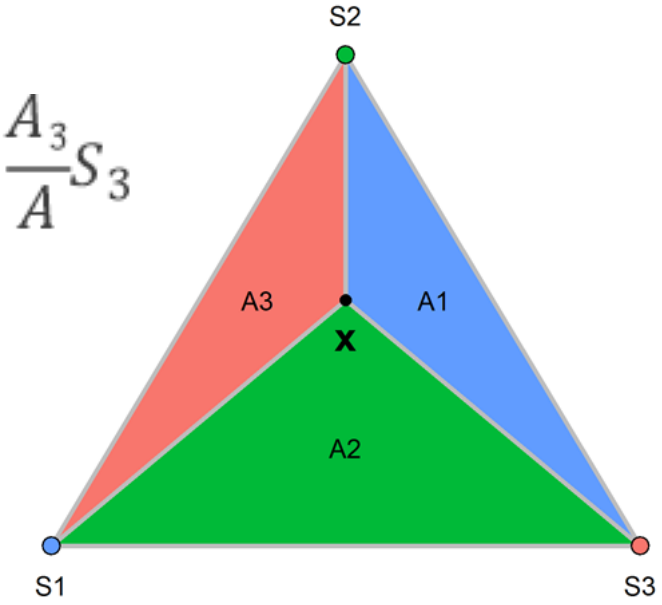
$$S(\boldsymbol{x}) = \sum_{g=1}^{G} \psi_g(\boldsymbol{x}) S_g$$



- Matérn covariance for Gaussian field S

INLA

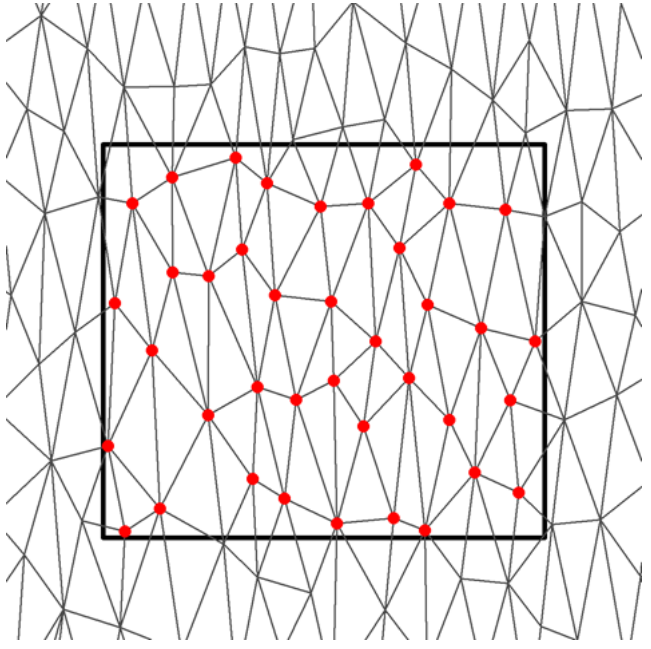
$$S(x) = \frac{A_1}{A}S_1 + \frac{A_2}{A}S_2 + \frac{A_3}{A}S_3$$



Point Observations:

- Weighted average at the vertices of the triangle containing the point.
- Weights = barycentric coordinates

$$S(B) = \frac{1}{m}\sum_{g \in B} S_g$$



Areal Observations:

- Weighted average at the vertices of the triangles within the area.
- Weights = (number of vertices in area)$^{-1}$

INLA

point

$$S(\boldsymbol{x}_i) \approx \sum_{g=1}^{G} A_{ig} S_g$$

areal

$$S(B_j) \approx \sum_{g=1}^{G} A_{js} S_g$$

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1G} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2G} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \cdots & A_{nG} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & \cdots & 0 \\ .2 & .2 & 0 & \cdots & .6 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{m} & \frac{1}{m} & \frac{1}{m} & \cdots & 0 \end{bmatrix}$$

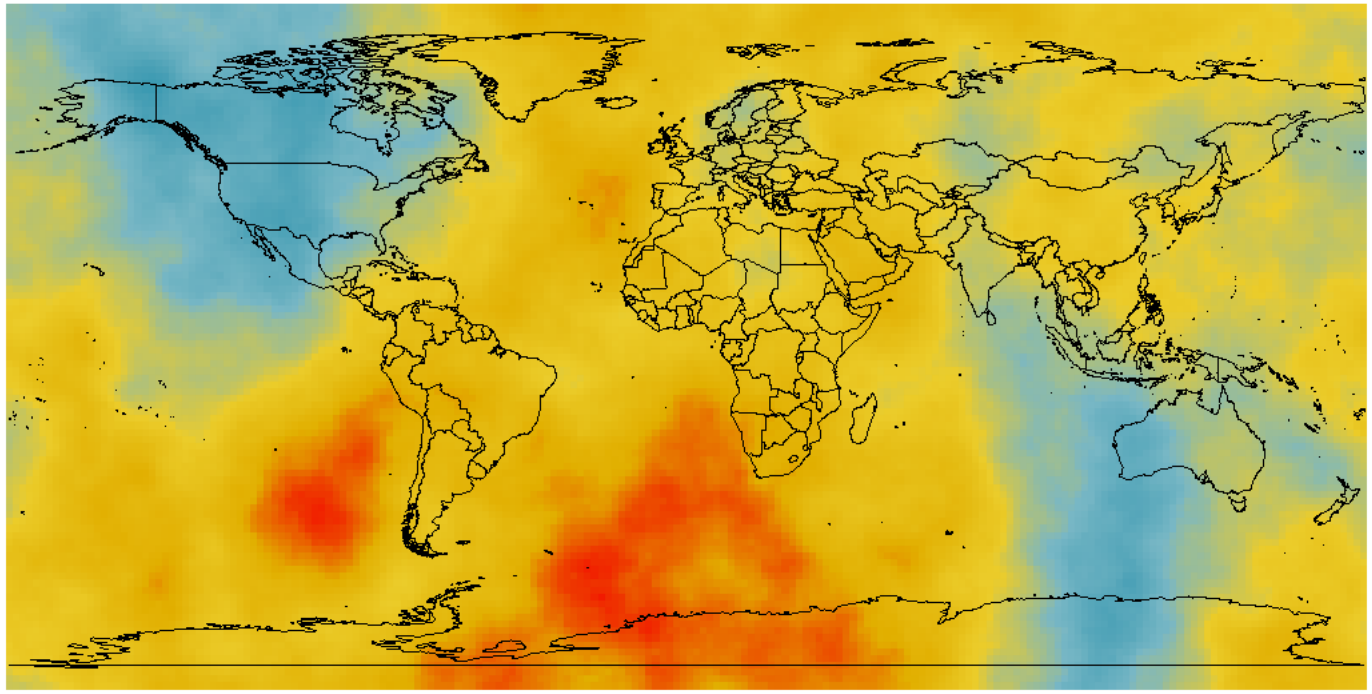
point

areal

A is a projection matrix that maps the GMRF from the observations (rows) to the triangulation nodes (columns).

- Point observation
  - Up to 3 non-zero values at the columns that represent the vertices of the triangle containing the point
  - Value is barycentric coordinates
- Areal observation:
  - Non-zero values at the *m* vertices inside the area
  - Value is 1/*m*

INLA Simulations: Areal data



True Simulated Field:

- 100 x 200 = 20,000 points

Areal data blocks:

- 10 x 20 (each block 100 values)
- 32 x 64 (each block 9-16 values)
- 90 x 180 (each block 1-4 values)

Prediction Error: Standard deviation of posterior predictive distribution

RMSE: $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{S}(x_i) - S(x_i)\right)^2}$

- COSP: we want prediction error to be larger when treating areal data properly than when treating it as point data
- Difference between the two should disappear as the resolution becomes finer

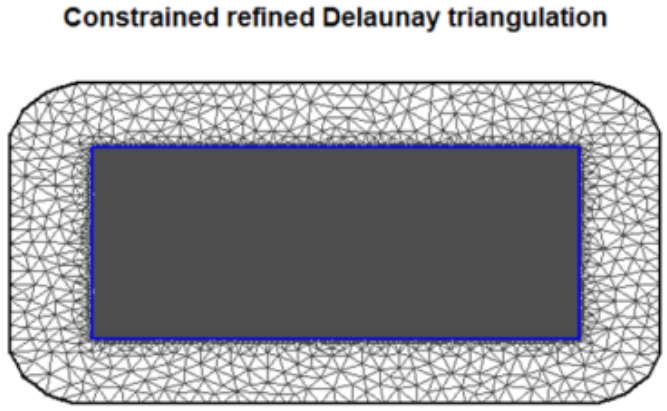
INLA Simulation Results: Areal data

Median Prediction Error

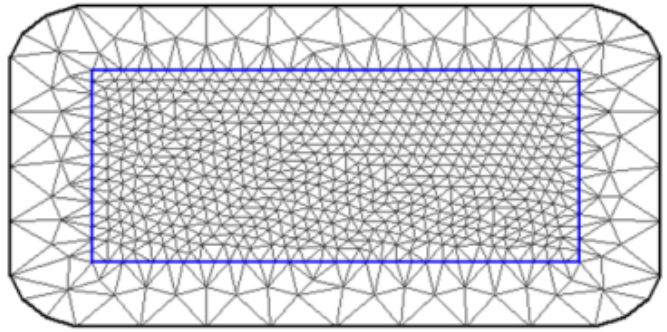
Areal blocks	Areal (fine mesh)	Point (fine mesh)	Point (coarse mesh)
10 x 20 (coarsest)	0.115	0.099	0.006
32 x 64	0.031	0.028	0.034
90 x 180 (finest)	0.0018	0.0016	0.0138

RMSE

Areal blocks	Areal (fine mesh)	Point (fine mesh)	Point (coarse mesh)
10 x 20 (coarsest)	0.128	0.142	0.164
32 x 64	0.078	0.051	0.094
90 x 180 (finest)	0.0261	0.0260	0.0759



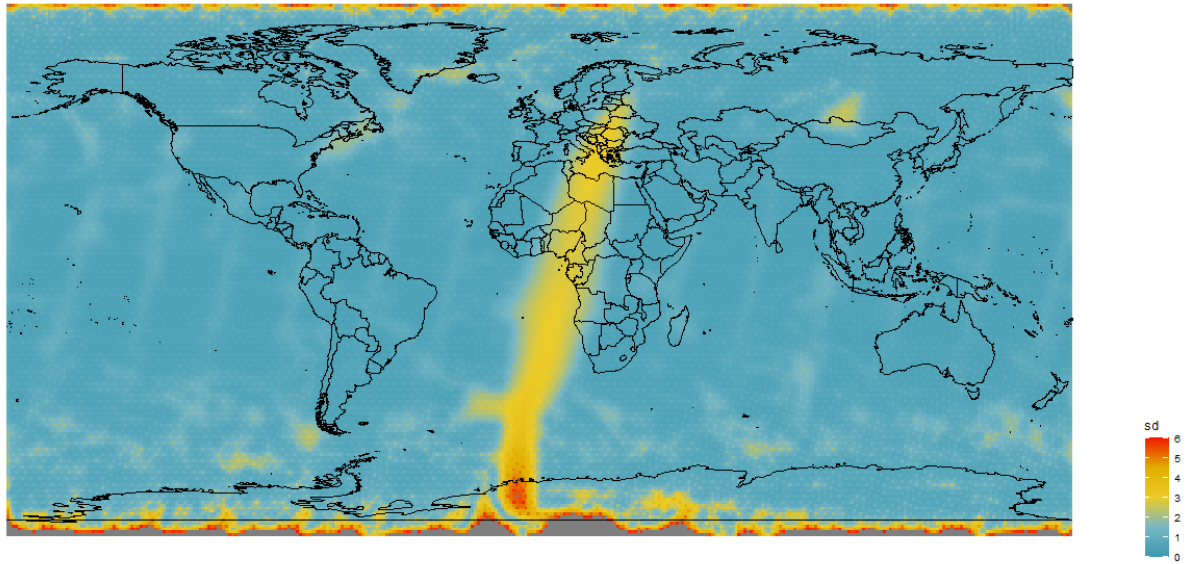
Fine Mesh



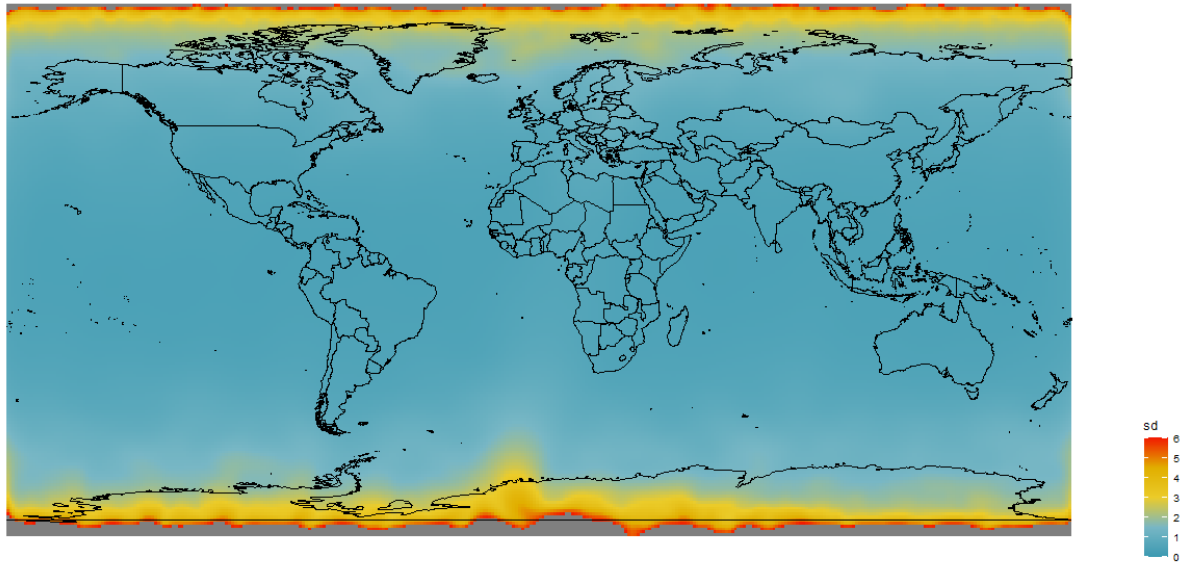
Coarse Mesh

INLA Model Results: Observational Data

Prediction Error: Fine Mesh



Prediction Error: As Aerial Data



Point data

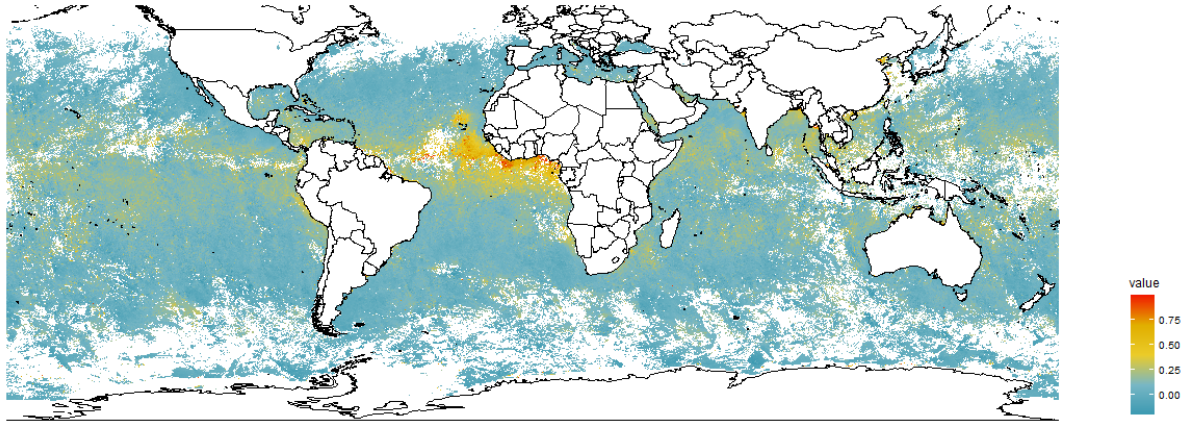
Model	Min	1st Q	Median	4*Median	Mean	3rd Q	Max	Time
Coarse mesh	0.218	0.328	0.399	1.596	0.529	0.529	9.590	41 sec
Medium mesh	0.321	0.502	0.629	2.516	0.878	0.886	11.452	50 sec
Fine mesh	0.395	0.589	0.745	2.982	1.069	1.063	14.744	156 sec
Mesh size 1	0.398	0.446	0.842	3.367	2.091	2.320	38.744	57 min
As aerial	**0.407**	**0.511**	**0.782**	**3.127**	**1.345**	**1.469**	**20.593**	**44 min**

Conclusions

- Interpolators:
  - Prediction error is too small
  - Results from bootstrap using reanalysis aren't useful
  - **LatticeKrig** vs. **laGP**
    - Better quantitative results
    - Better qualitative results
    - Faster
  - LatticeKrig: 3.5 years of data in ~6.5 days
- INLA
  - Handles change-of-support
  - Running areal data appropriately probably takes too long
    - 3.5 years of data in ~1 month
  - Run as point data with fine mesh is a good approximation
    - 3.5 years of data in <3 days

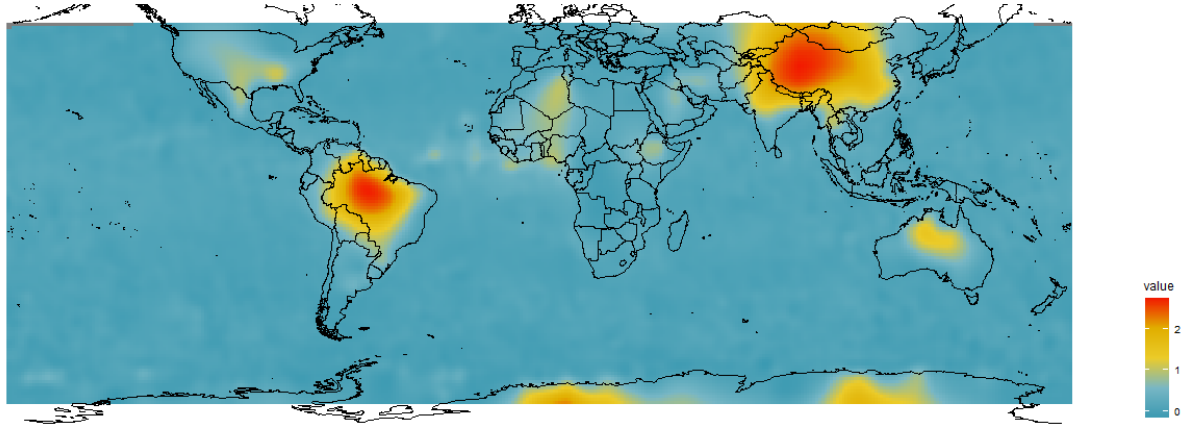
Challenges and Next Steps

AOD data: Jan 1990

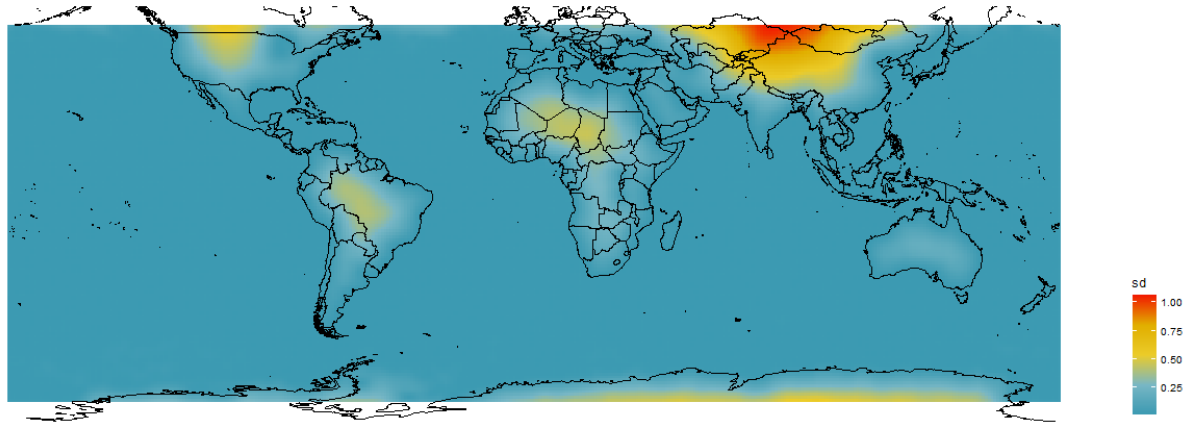


Data: 0.1-degree resolution
averaged into 1-degree blocks
Model: 1-degree resolution
time: ~ 1 min
Predictions: 1-degree resolution
time: ~ 4 seconds
Prediction error: 200 bootstrap samples
time: ~ 25 min

LatticeKrig predictions



Prediction error



Acknowledgements

This presentation describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories; a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

References

Gramacy, R.B., 2020. *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. CRC press.

Heaton, M.J., et. al. 2019. A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24, pp.398-425.

Lindgren, F., H. Rue, and J. Lindstrom. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach (with discussion). *Journal of the Royal Statistical Society, Series B*, 73 (4):423–498.

Moraga, P., Cramb, S.M., Mengersen, K.L. and Pagano, M., 2017. A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. *Spatial Statistics*, 21, pp.27-41.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F. and Sain, S., 2015. A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2), pp.579-599.

Rue, H. , S. Martino, and N. Chopin. 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71(2):319-392.