جامعة الملك عبدالله
للعلوم والتقنية

King Abdullah University of
Science and Technology

Accelerated Spatio-Temporal Bayesian Modeling for Multivariate Gaussian Processes

Lisa Gaedke-Merzhäuser
Bayesian Computational Statistics and Modeling Group
KAUST

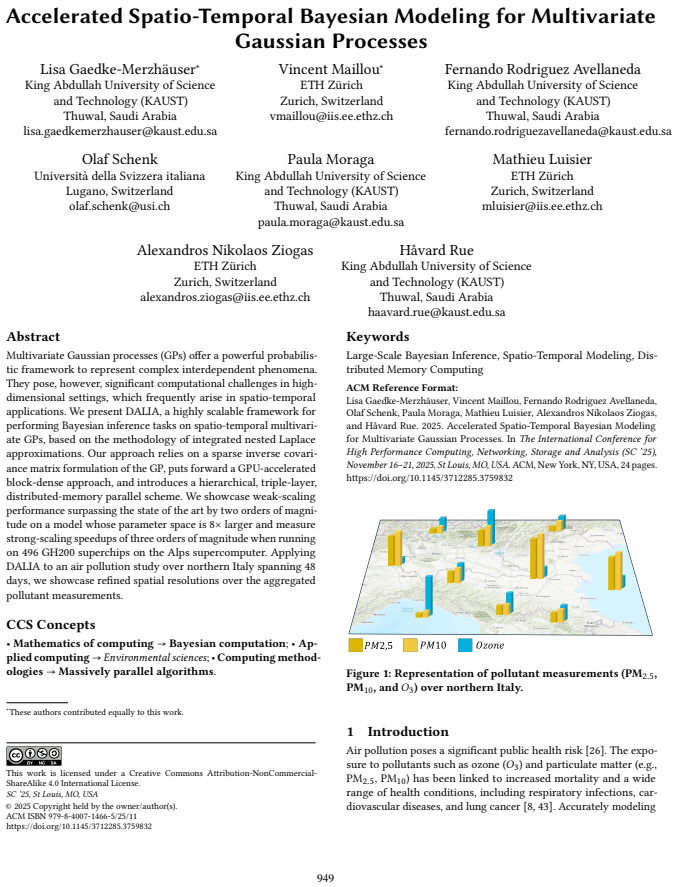
Feb 19, 2026

Accelerated Spatio-Temporal Bayesian Modeling for Multivariate Gaussian Processes

L. G-M*[1], Vincent Maillou*[2], Fernando Rodriguez Avellaneda[1], Olaf Schenk[3], Paula Moraga[1], Mathieu Luisier[2], Alexandros Nikolaos Ziogas[2], Haavard Rue[1]

[1]King Abdullah University of Science and Technology,
[2]ETH Zurich, [3]Università della Svizzera italiana

The International Conference for High Performance Computing, Networking, Storage, and Analysis 2025 (SC'25)



Accelerated Spatio-Temporal Bayesian Modeling for Multivariate Gaussian Processes

Lisa Gaedke-Merzhäuser*
King Abdullah University of Science and Technology (KAUST)
Thuwal, Saudi Arabia
lisa.gaedkemerzhauser@kaust.edu.sa

Vincent Maillou*
ETH Zürich
Zurich, Switzerland
vmaillou@iis.ee.ethz.ch

Fernando Rodriguez Avellaneda
King Abdullah University of Science and Technology (KAUST)
Thuwal, Saudi Arabia
fernando.rodriguezavellaneda@kaust.edu.sa

Olaf Schenk
Università della Svizzera italiana
Lugano, Switzerland
olaf.schenk@usi.ch

Paula Moraga
King Abdullah University of Science and Technology (KAUST)
Thuwal, Saudi Arabia
paula.moraga@kaust.edu.sa

Mathieu Luisier
ETH Zürich
Zurich, Switzerland
mluisier@iis.ee.ethz.ch

Alexandros Nikolaos Ziogas
ETH Zürich
Zurich, Switzerland
alexandros.ziogas@iis.ee.ethz.ch

Håvard Rue
King Abdullah University of Science and Technology (KAUST)
Thuwal, Saudi Arabia
haavard.rue@kaust.edu.sa

Abstract

Multivariate Gaussian processes (GPs) offer a powerful probabilistic framework to represent complex interdependent phenomena. They pose, however, significant computational challenges in high-dimensional settings, which frequently arise in spatio-temporal applications. We present DALIA, a highly scalable framework for performing Bayesian inference tasks on spatio-temporal multivariate GPs, based on the methodology of integrated nested Laplace approximations. Our approach relies on a sparse inverse covariance matrix formulation of the GP, puts forward a GPU-accelerated block-dense approach, and introduces a hierarchical, triple-layer, distributed-memory parallel scheme. We showcase weak-scaling performance surpassing the state of the art by two orders of magnitude on a model whose parameter space is 8× larger and measure strong-scaling speedups of three orders of magnitude when running on 496 GH200 superchips on the Alps supercomputer. Applying DALIA to an air pollution study over northern Italy spanning 48 days, we showcase refined spatial resolutions over the aggregated pollutant measurements.

CCS Concepts

• **Mathematics of computing** → **Bayesian computation**; • **Applied computing** → *Environmental sciences*; • **Computing methodologies** → **Massively parallel algorithms**.

*These authors contributed equally to this work.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
SC '25, St Louis, MO, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1466-5/25/11
https://doi.org/10.1145/3712285.3759832

Keywords

Large-Scale Bayesian Inference, Spatio-Temporal Modeling, Distributed Memory Computing

ACM Reference Format:

Lisa Gaedke-Merzhäuser, Vincent Maillou, Fernando Rodriguez Avellaneda, Olaf Schenk, Paula Moraga, Mathieu Luisier, Alexandros Nikolaos Ziogas, and Håvard Rue. 2025. Accelerated Spatio-Temporal Bayesian Modeling for Multivariate Gaussian Processes. In *The International Conference for High Performance Computing, Networking, Storage and Analysis (SC '25), November 16–21, 2025, St Louis, MO, USA.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3712285.3759832

Figure 1: Representation of pollutant measurements ($PM_{2.5}$, PM_{10}, and O_3) over northern Italy.

1 Introduction

Air pollution poses a significant public health risk [26]. The exposure to pollutants such as ozone (O_3) and particulate matter (e.g., $PM_{2.5}$, PM_{10}) has been linked to increased mortality and a wide range of health conditions, including respiratory infections, cardiovascular diseases, and lung cancer [8, 43]. Accurately modeling

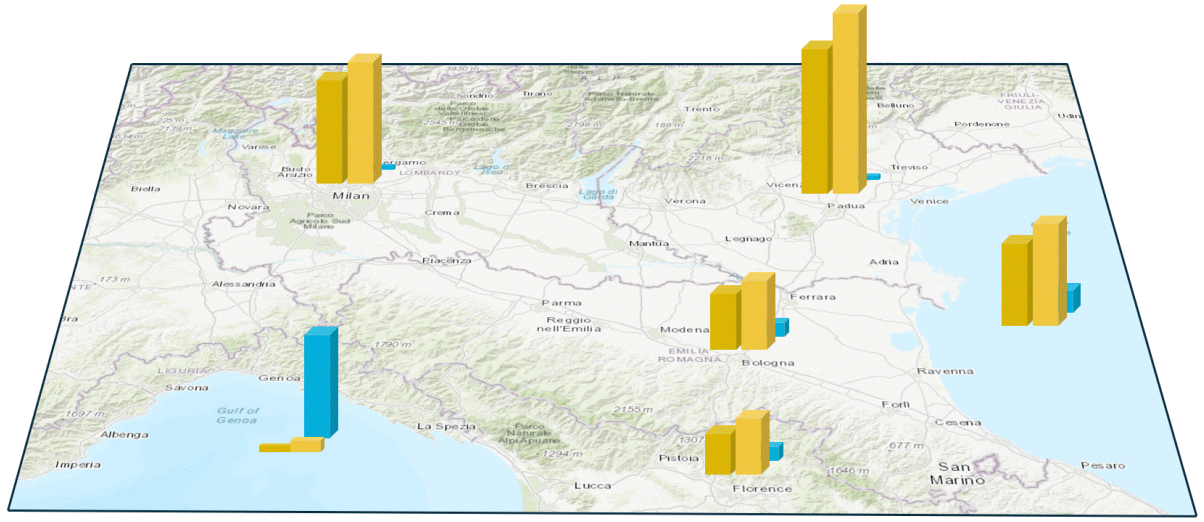
949

Motivation

Model the behavior of a process from data over space and time

Air Pollution over Italy

PM 2.5 PM 10 Ozone



- Jointly model air pollutants
- Refine spatial resolution
- Quantify uncertainty

Motivation

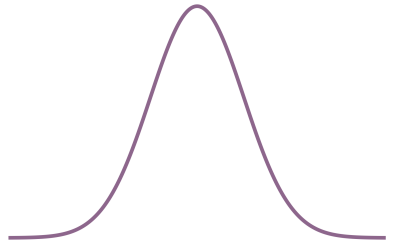
Model the behavior of a process from data over space and time

Air Pollution over Italy

PM 2.5 PM 10 Ozone

- Jointly model air pollutants
- Refine spatial resolution
- Quantify uncertainty

Spatio-temporal multivariate Gaussian processes (GPs)



Bayesian approach

- Include prior knowledge

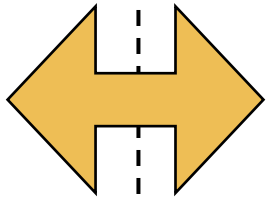
➢ **Problem: High computational cost !**

Spatial Correlation and the SPDE Approach

Matérn Covariance function

$$c(s_1, s_2) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{8\nu}\, \|s_1 - s_2\| / \rho\right) K_\nu \left(\sqrt{8\nu}\, \|s_1 - s_2\| / \rho\right)$$

- Used to construct covariance matrix
- Describes (in)dependence between variables



Linear fractional SPDE

$$\left(\kappa^2 - \Delta\right)^{\alpha/2} (\tau u(s)) = \mathscr{W}(s), \quad s \in \mathbb{R}^d$$
$$\alpha = \nu + d/2, \quad \kappa > 0, \quad \nu > 0,$$

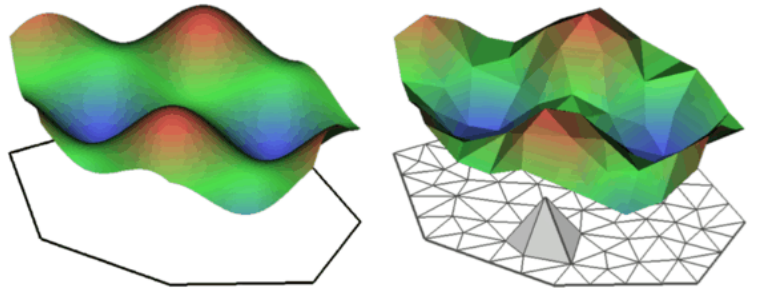
- The solution $u(s)$ has a Matérn covariance function[1]
- Used to construct precision (inverse covariance) matrix
- Describes conditional (in)dependence between variables

[1]Whittle, P. "Stochastic-processes in several dimensions." *Bulletin of the International Statistical Institute* 40.2 (1963): 974-994.

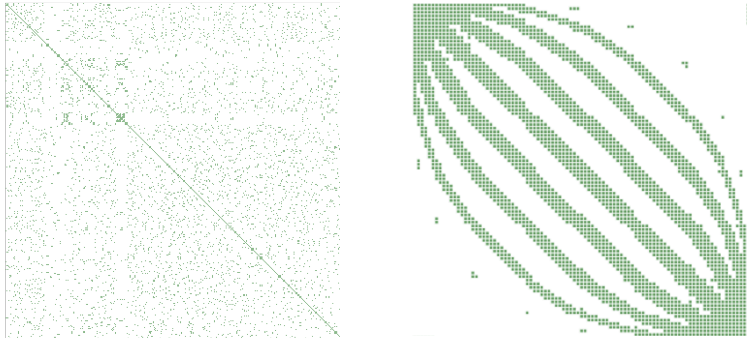
Spatial Correlation and the SPDE Approach

$$\left(\kappa^2 - \Delta\right)^{\alpha/2}(\tau u(s)) = \mathcal{W}(s), \quad s \in \mathbb{R}^d$$

- Discretize SPDE using first-order finite elements
- FE approximation of $u(s)$ forms GMRF[1]
- GMRF has sparse precision matrix whose inverse (covariance) matrix is dense
- Mesh is independent from observation locations
- In the limit both approaches converge to the same solution



Cameletti, M., Lindgren, F., Simpson, D. et al. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. AStA Adv Stat Anal 97, 109–131 (2013).

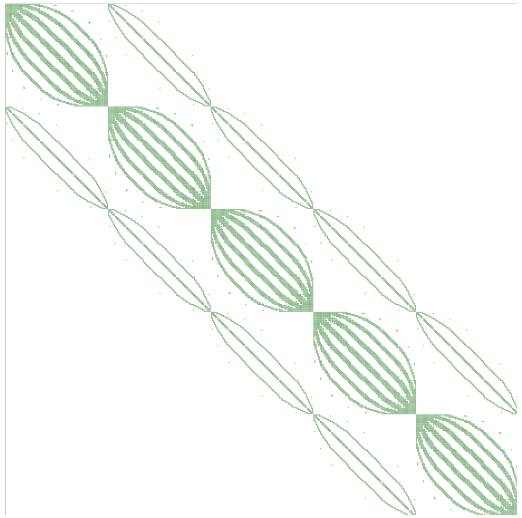


[1]Lindgren F, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2011 Sep;73(4):423-98.

Spatial Correlation and the SPDE Approach

$$\left(\kappa^2 - \Delta\right)^{\alpha/2}(\tau u(s)) = \mathscr{W}(s), \quad s \in \mathbb{R}^d$$

- Discretize SPDE using first-order finite elements[1]
- FE approximation of $u(s)$ forms GMRF
- GMRF has sparse precision matrix whose inverse (covariance) matrix is dense
- Mesh is independent from observation locations
- In the limit both approaches converge to the same solution
- Use SPDE-derived spatio-temporal extension[2]



[1]Lindgren F, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2011 Sep;73(4):423-98.

[2]Lindgren F, Bakka H, Bolin D, Krainski E, Rue H. A diffusion-based spatio-temporal extension of Gaussian Matérn fields: Sort: Statistics and Operations Research Transactions. 2024;48(1):3-66.

Model Formulation

Data

$$y = Ax + \epsilon$$

Likelihood
$p(y \mid x, \theta)$

Latent Parameters

$$x \mid \theta \sim N(0, Q^{-1}(\theta))$$

Prior
$p(x \mid \theta)$

Hyperparameters

$$\theta$$

Prior
$p(\theta)$

Linear Models of Coregionalization[1]

- Multiple observed quantities

PM 2.5 $\longrightarrow$

$$\underbrace{\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \mathbf{y}_3\end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix}\sigma_1\mathbf{I} & 0 & 0\\ \lambda_1\sigma_1\mathbf{I} & \sigma_2\mathbf{I} & 0\\ (\lambda_3+\lambda_1\lambda_2)\sigma_1\mathbf{I} & \lambda_2\sigma_2\mathbf{I} & \sigma_3\mathbf{I}\end{bmatrix}}_{\mathbf{\Lambda}} \cdot \underbrace{\begin{bmatrix}\mathbf{A}_1 & 0 & 0\\ 0 & \mathbf{A}_2 & 0\\ 0 & 0 & \mathbf{A}_3\end{bmatrix}}_{\mathbf{A}} \cdot \underbrace{\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \mathbf{x}_3\end{bmatrix}}_{\mathbf{x}} + \boldsymbol{\epsilon}$$

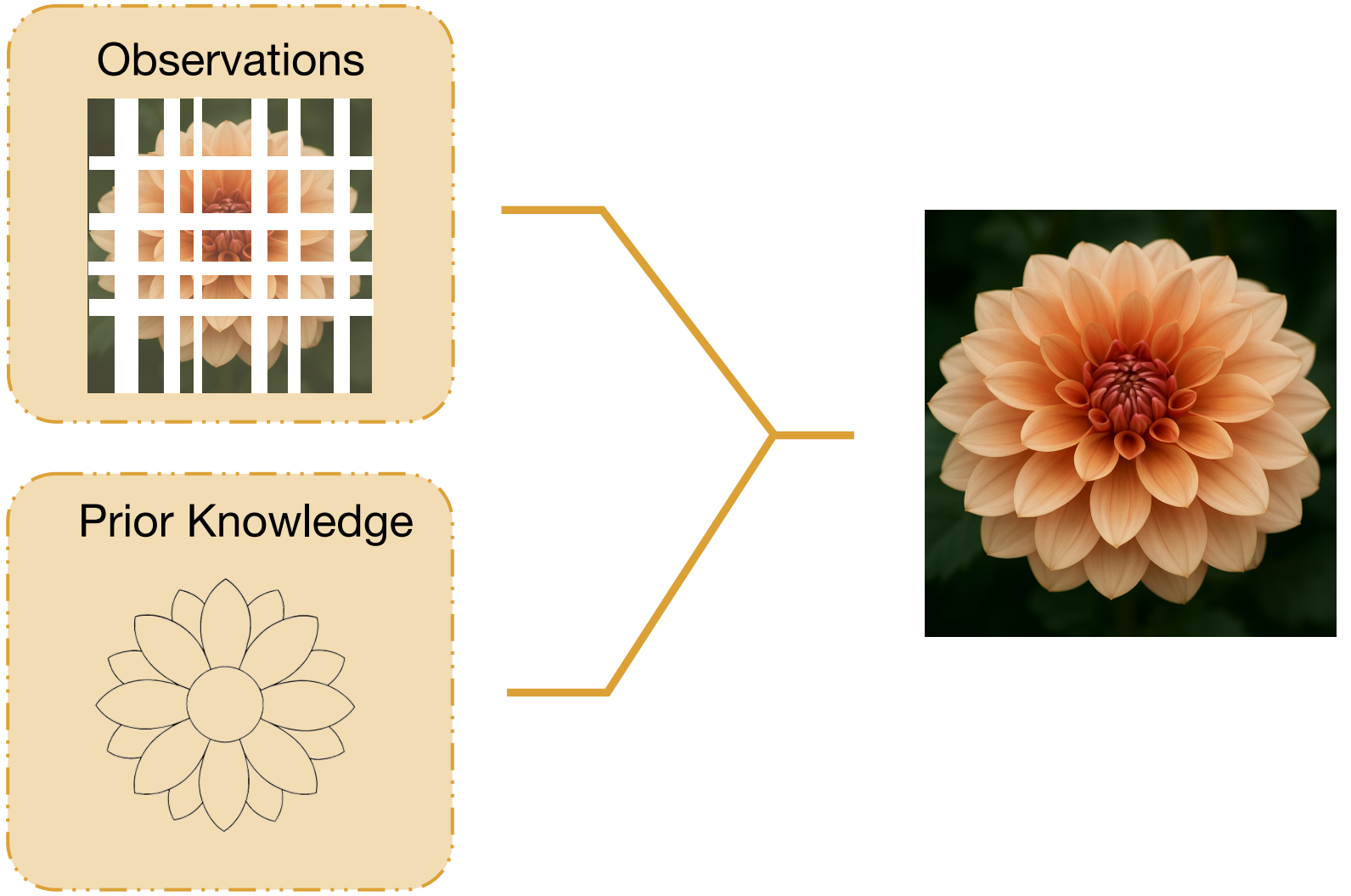
$$\mathbf{Q} = \mathbf{\Lambda}^{-\mathbf{T}} \begin{bmatrix}\mathbf{Q}_1 & 0 & 0\\ 0 & \mathbf{Q}_2 & 0\\ 0 & 0 & \mathbf{Q}_3\end{bmatrix} \mathbf{\Lambda}^{-1}$$

univariate spatio-temporal precision matrix

➢ Q remains sparse

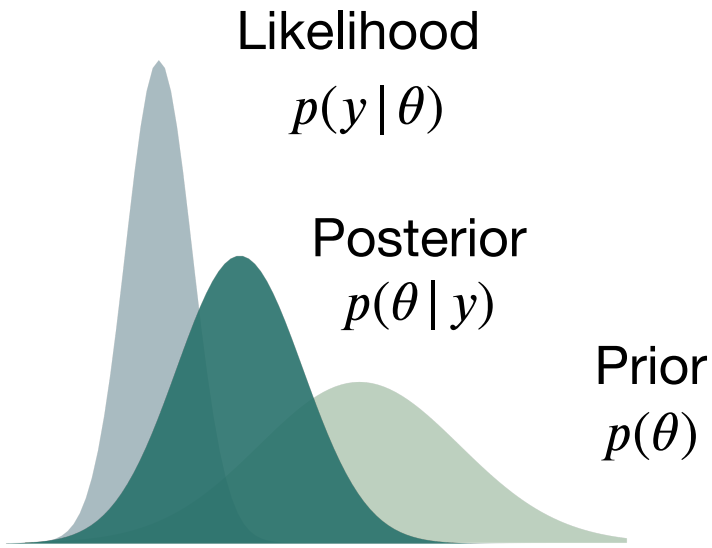
[1]Schmidt, AM., and AE. Gelfand. "A Bayesian coregionalization approach for multivariate pollutant data." *Journal of Geophysical Research: Atmospheres* 108.D24 (2003).

Bayesian inference



Bayesian inference

Aim: *Update our beliefs upon observing data using Bayes' rule.*



Bayes Rule

$$p(\theta \mid y) \propto \frac{p(\theta)\, p(x \mid \theta)\, p(y \mid x, \theta)}{p(x \mid \theta, y)}$$

⇨ Problem: No closed-form of the Posterior!

Bayesian inference

	Sampling-based	Deterministic	
	Markov Chain Monte Carlo *(MCMC)*	Variational Inference *(VI)*	Integrated Nested Laplace Approximation *(INLA)*
Accuracy	Asympt. Exact	⊖	⊕
Parallelization potential	⊖	⊕	⊕
Applicability	⊕	⊕	Latent Gaussian Models *(LGM)*

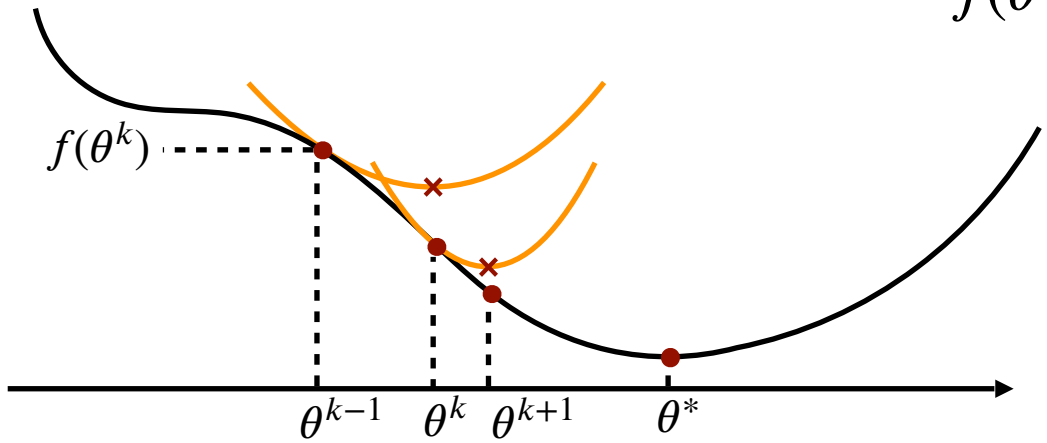
Integrated Nested Laplace Approximations (INLA)

Construct point-wise approximation to the posterior

$$f(\theta) := -\log \tilde{p}(\theta \mid y)$$

Phase I: Optimization

- BFGS-algorithm, requires $f(\theta^k), \nabla f(\theta^k)$ for k-th iterate



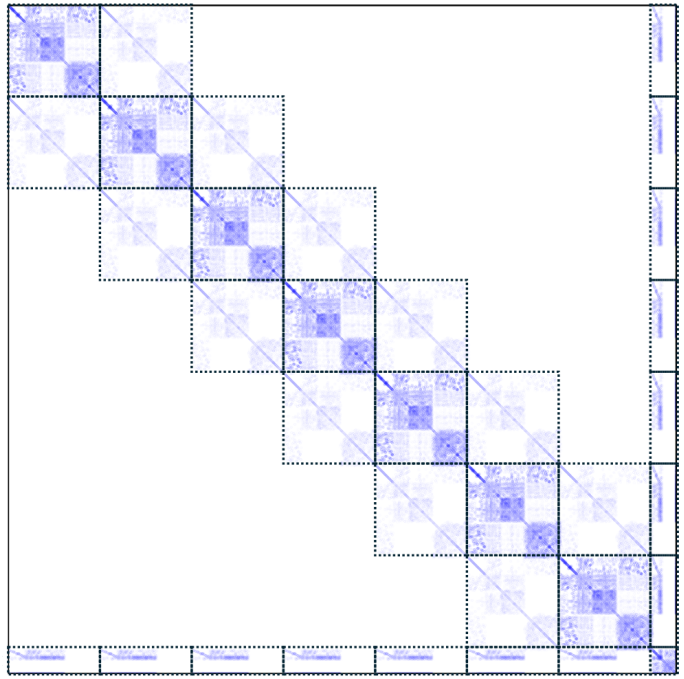
Integrated Nested Laplace Approximations (INLA)

Phase I: Optimization

- Bottleneck $f(\theta^k)$: Evaluate 2 high-dim. multivariate normal distributions

$$\frac{1}{2}\left(\log\underbrace{|Q(\theta)|}_{\text{Determinant}} - (x-\mu(\theta))^T Q(\theta)(x-\mu(\theta))\right)$$

Sparse s.p.d precision matrix

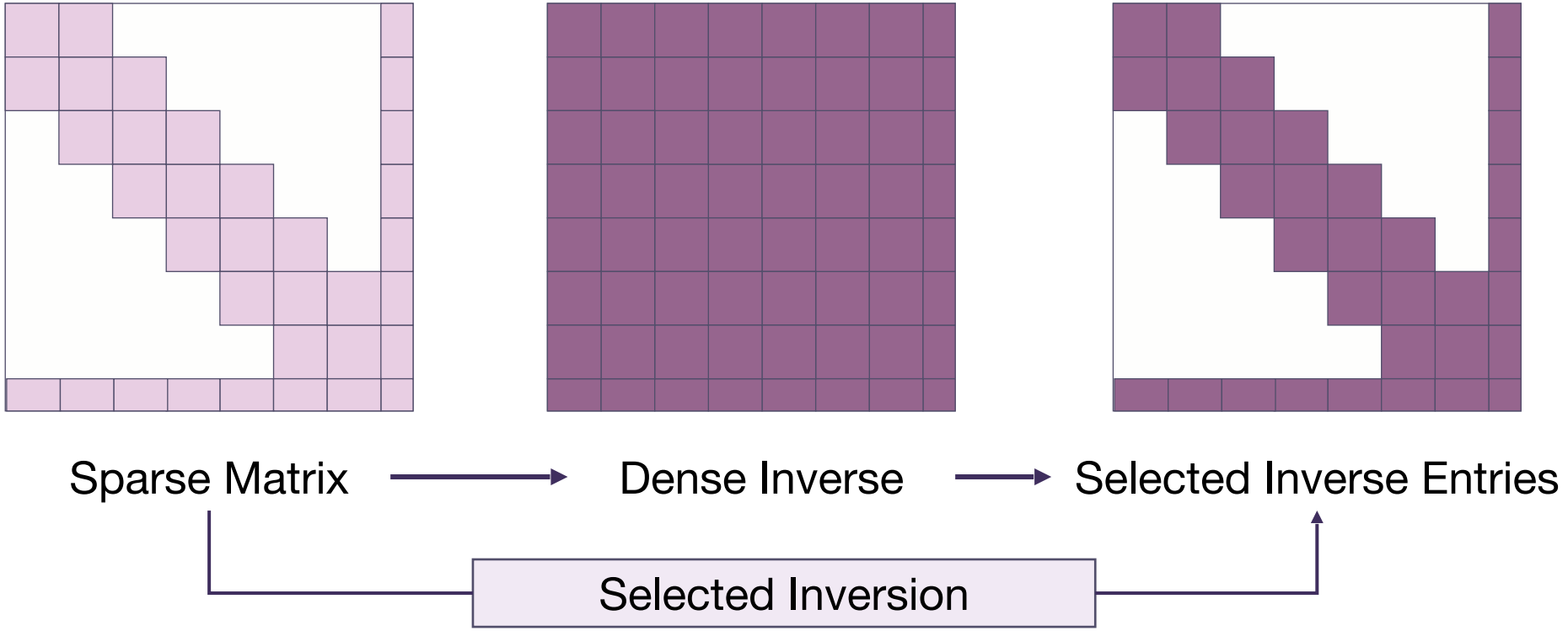


- Cholesky decomposition
- Triangular solve

Integrated Nested Laplace Approximations (INLA)

Phase II: Approximation of the marginals $p(\theta_i \mid y), p(x_j \mid y)$

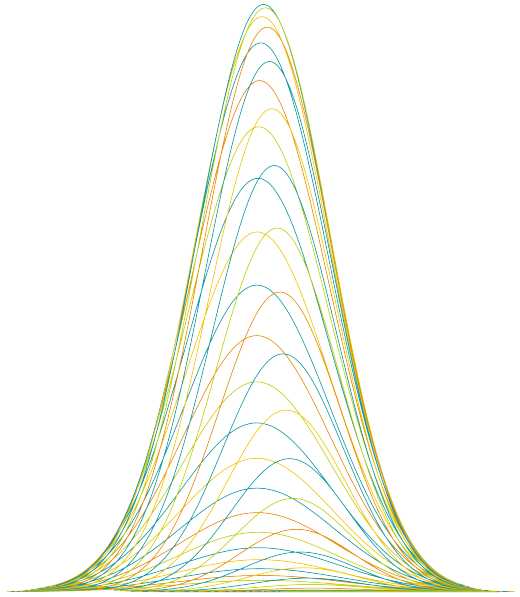
- Evaluate f around θ^*: $f(\theta^{*1}), \theta^* f(\theta^{*2}), \ldots$
- Compute inverse : $Q^{-1}(\theta^*), Q^{-1}(\theta^{*1}), \ldots, Q^{-1}(\theta^{*K})$



Current Software

Reference implementation : R-INLA[1]

- Widely used in Statistics community
- Written in C with R interface
- Shared memory parallelism (nested OpenMP)
- Relies on sparse direct solver PARDISO



Limitations

- ⊗ Distributed memory parallelism
- ⊗ GPU support
- ⊗ Modularity

[1]H. Rue, S. Martino, and N. Chopin, INLA: Approximate Bayesian Inference using Integrated Nested Laplace Approximations, 2011, www.r-inla.org

DALIA

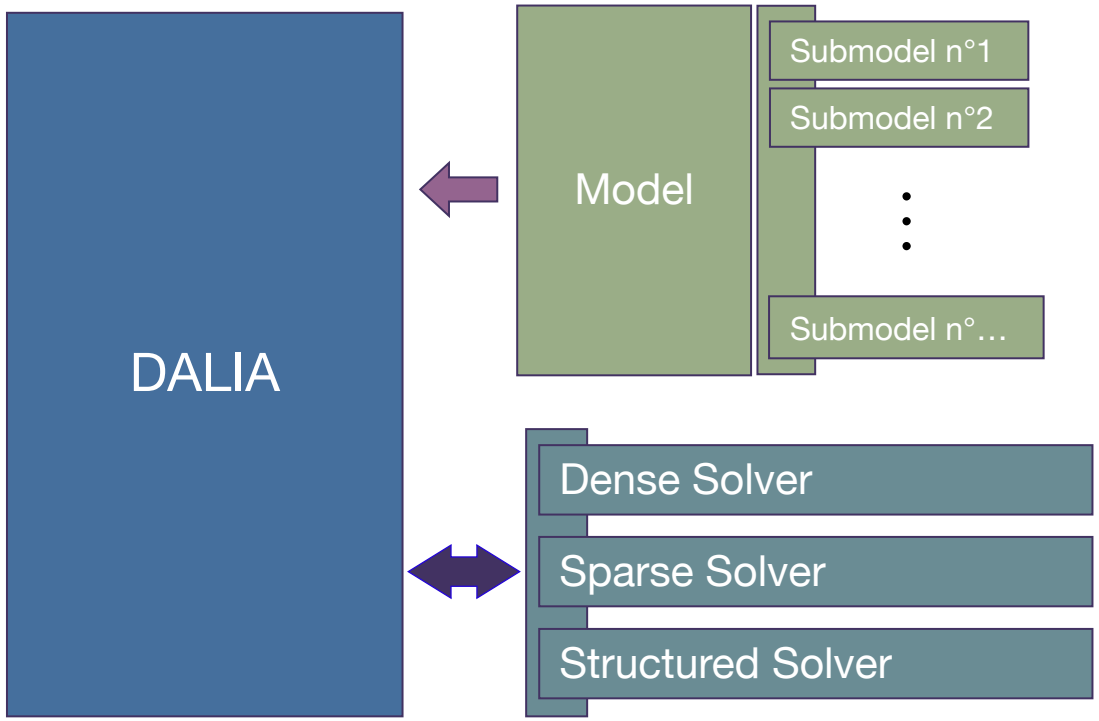
Distributed Accelerated Laplace Inference Approximations

Overview of our Framework: DALIA

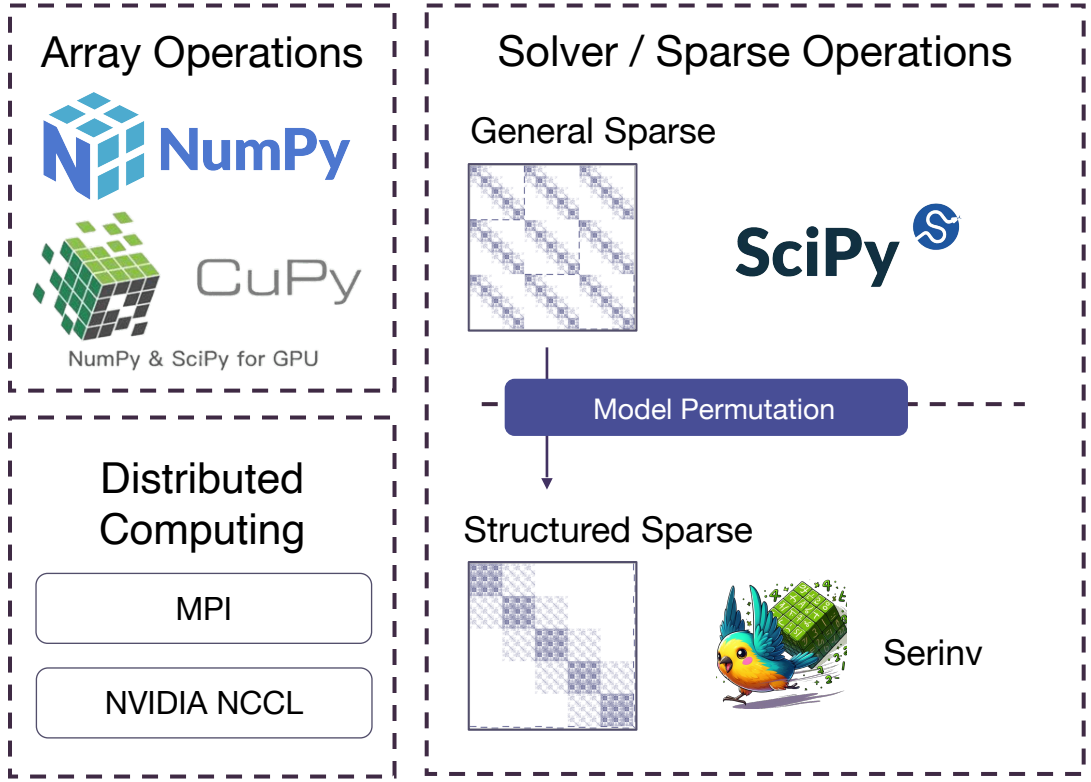
A Modular Approach

Precision matrices optimizer-model interface

Unified solver interface



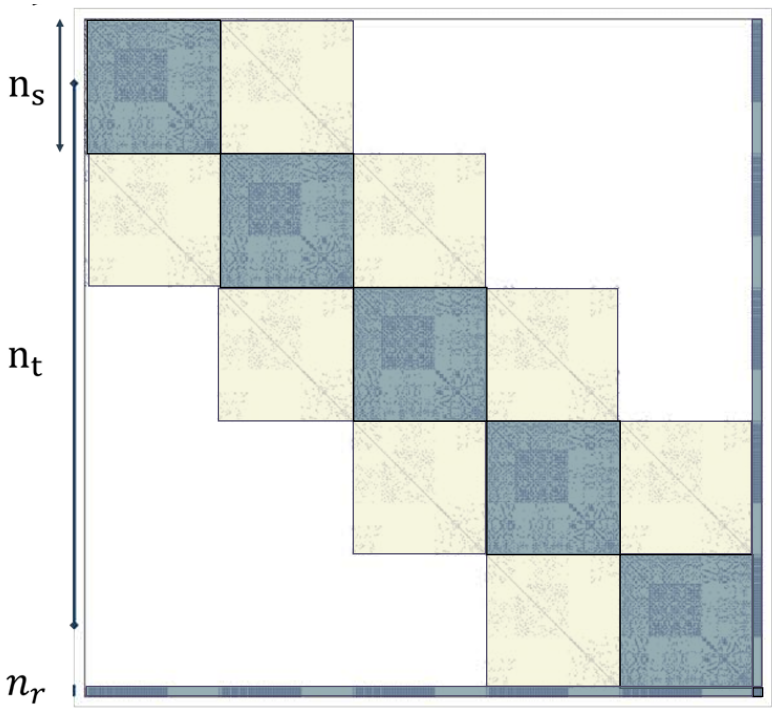
Implementation



Precision Matrices (Q) Sparsity

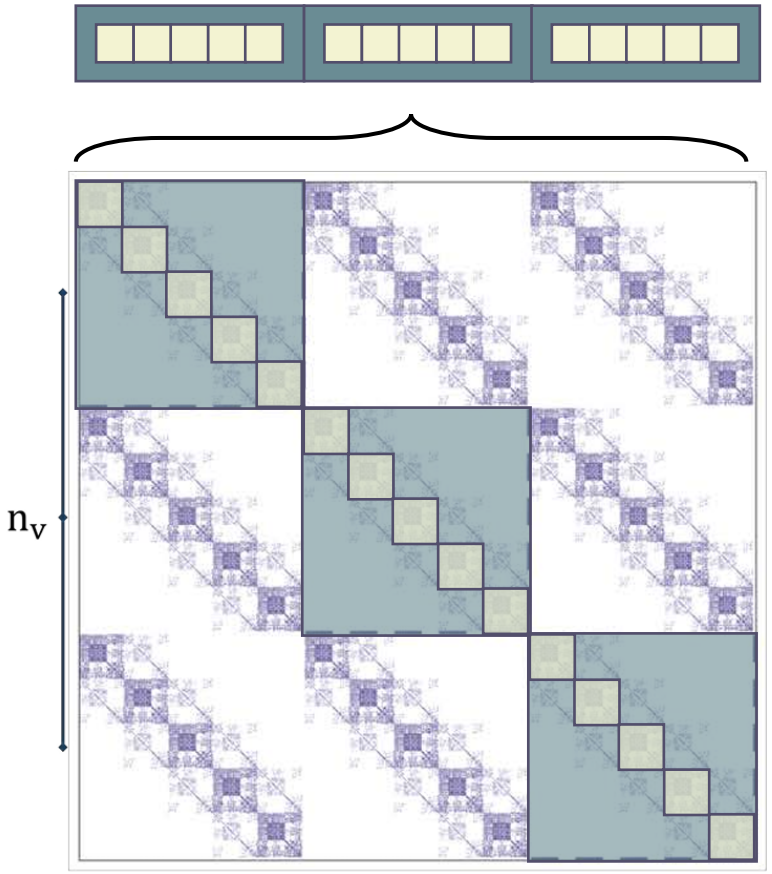
Univariate Spatio-Temporal Model

$$n_v = 1, n_t = 5$$

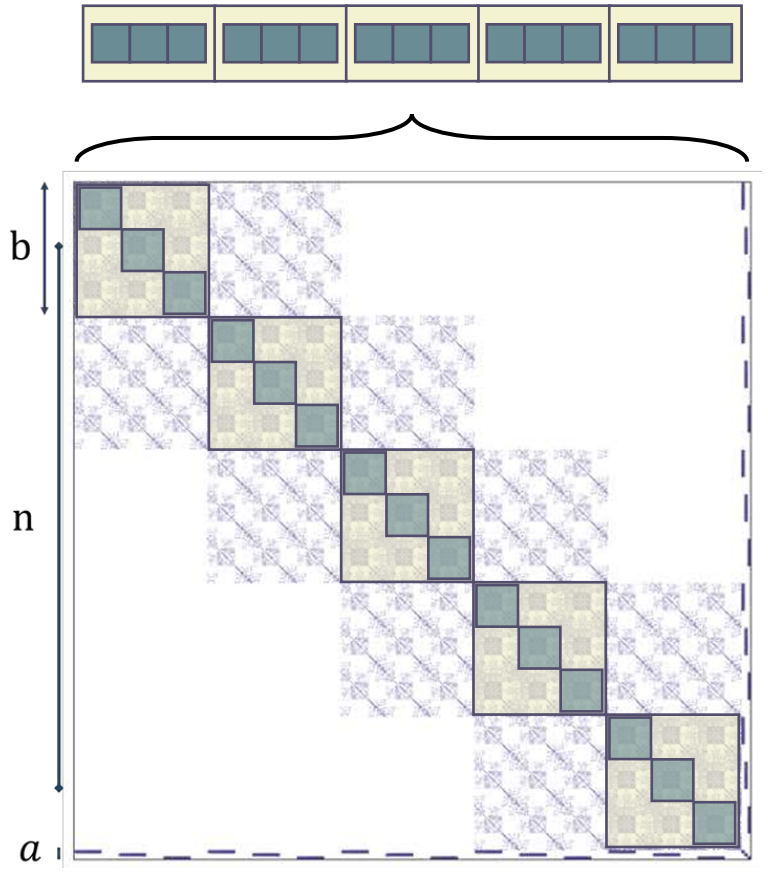


Trivariate Spatio-Temporal Model $(n_v = 3, n_t = 5)$

Ordering: $n_v \cdot n_t$

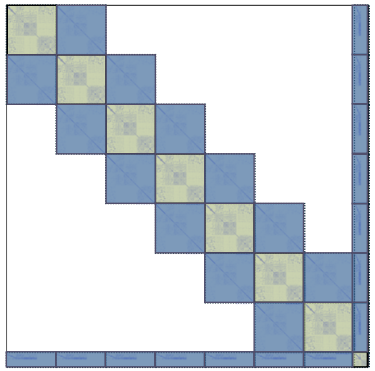


Ordering: $n_t \cdot n_v$

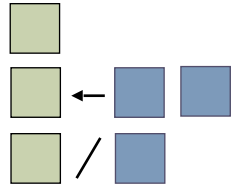


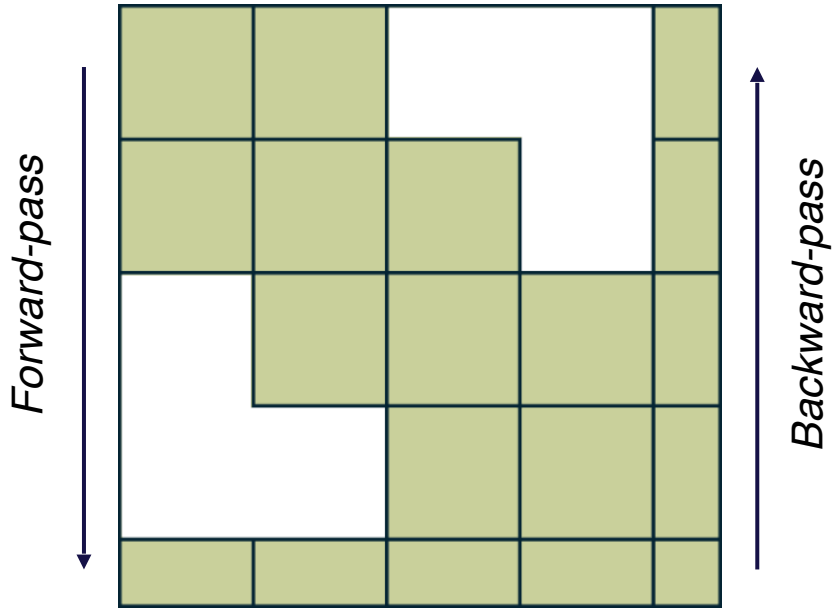
Structured Sparse (Selected) Solver

1. Block Algorithms



- Dense Tiling of Sparsity Pattern
- Level-3 BLAS
  - Factorization
  - Matrix Mult.
  - Triangl. Solve





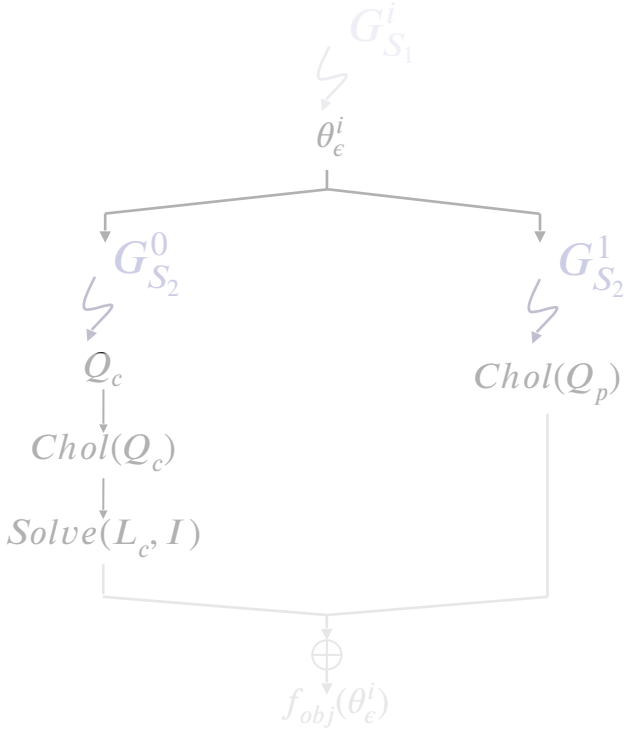
2. Basic Idea

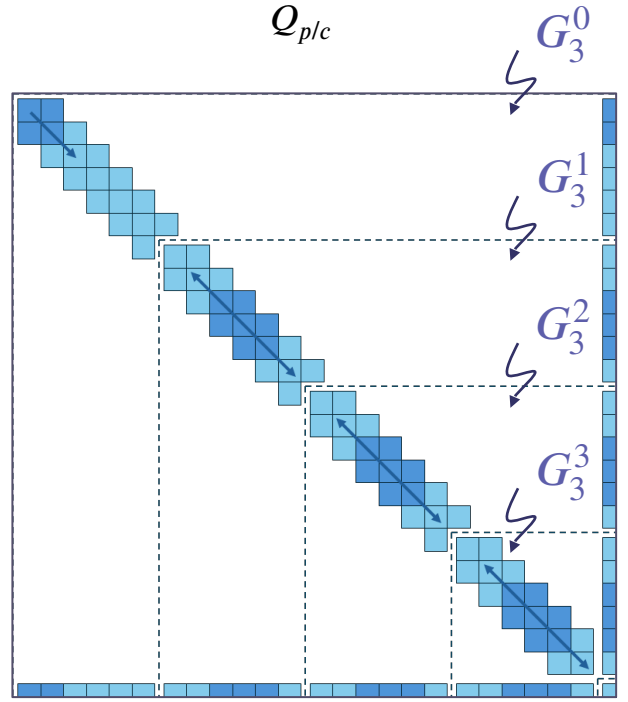
Cholesky

- Forward Decomposition
- Backward Selected Inversion

V. Maillou, L. Gaedke-Merzhäuser, A. N. Ziogas, O. Schenk and M. Luisier, "Parallel Selected Inversion of Block-Tridiagonal with Arrowhead Matrices," 2025 IEEE International Conference on Cluster Computing (CLUSTER), United Kingdom, 2025, pp. 1-12, doi: 10.1109/CLUSTER59342.2025.11186484.

Nested Parallelization Strategies





S_1: Function Evaluation

S_2: Precision Matrices Factorization

S_3: Distributed Structured Solvers

Experimental Setup

R-INLA + PARDISO

OpenMP

Fritz @ FAU
2 x Intel Xeon Platinum 8470
2 x 52 cores @ 2.0 GHz
2 TB DDR5

NHR FAU

DALIA + Serinv

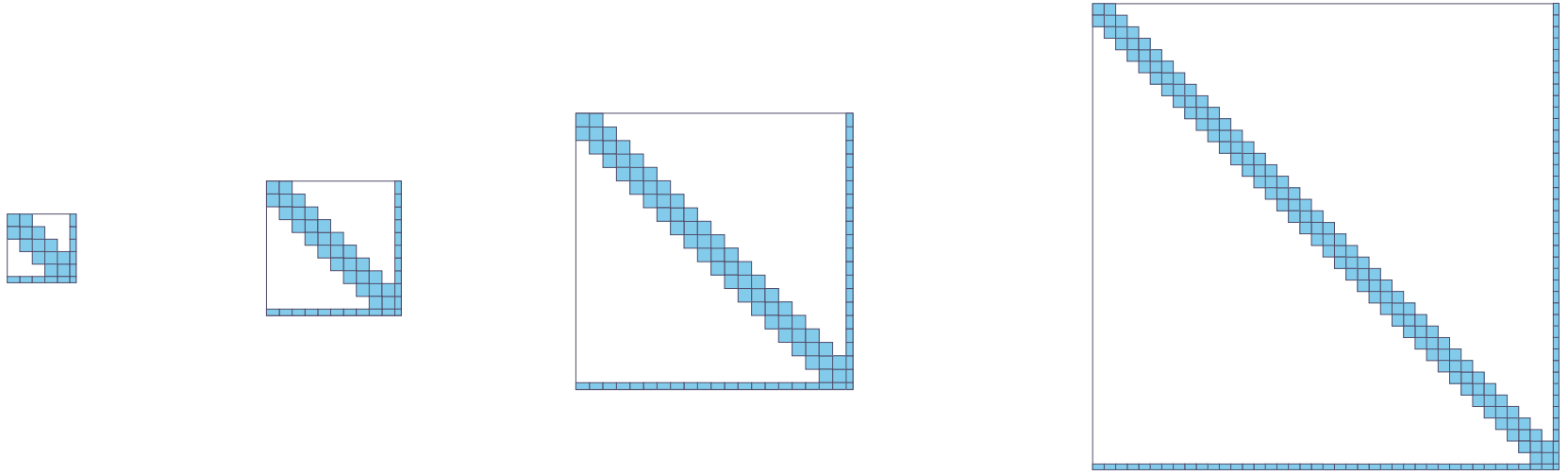
MPI/NCCL + GPU Acc

Alps @ CSCS
GH200 Superchip: 72 ARM cores, 128 GB RAM, 96 GB HBM3

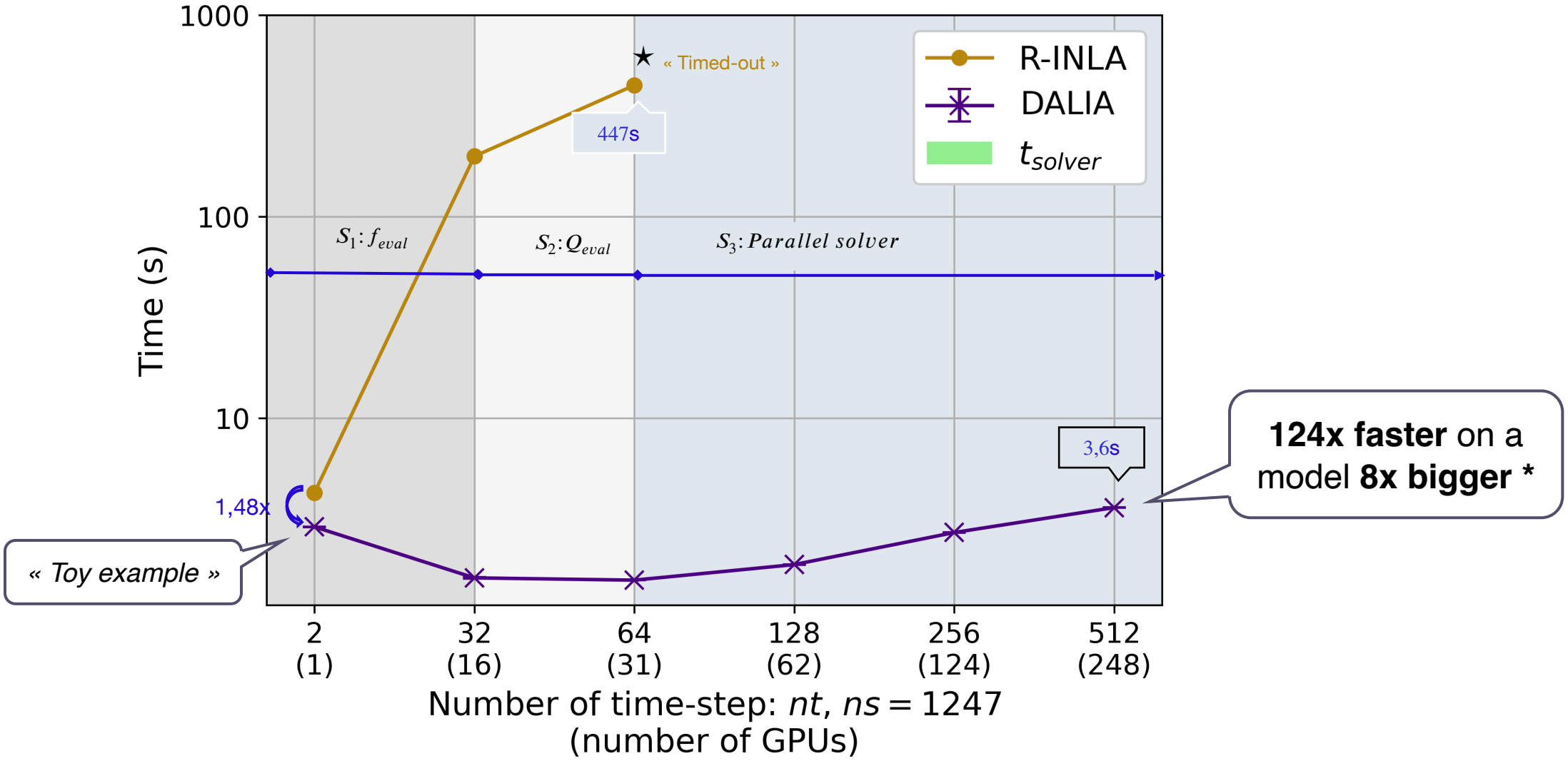
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Weak Scaling in Time

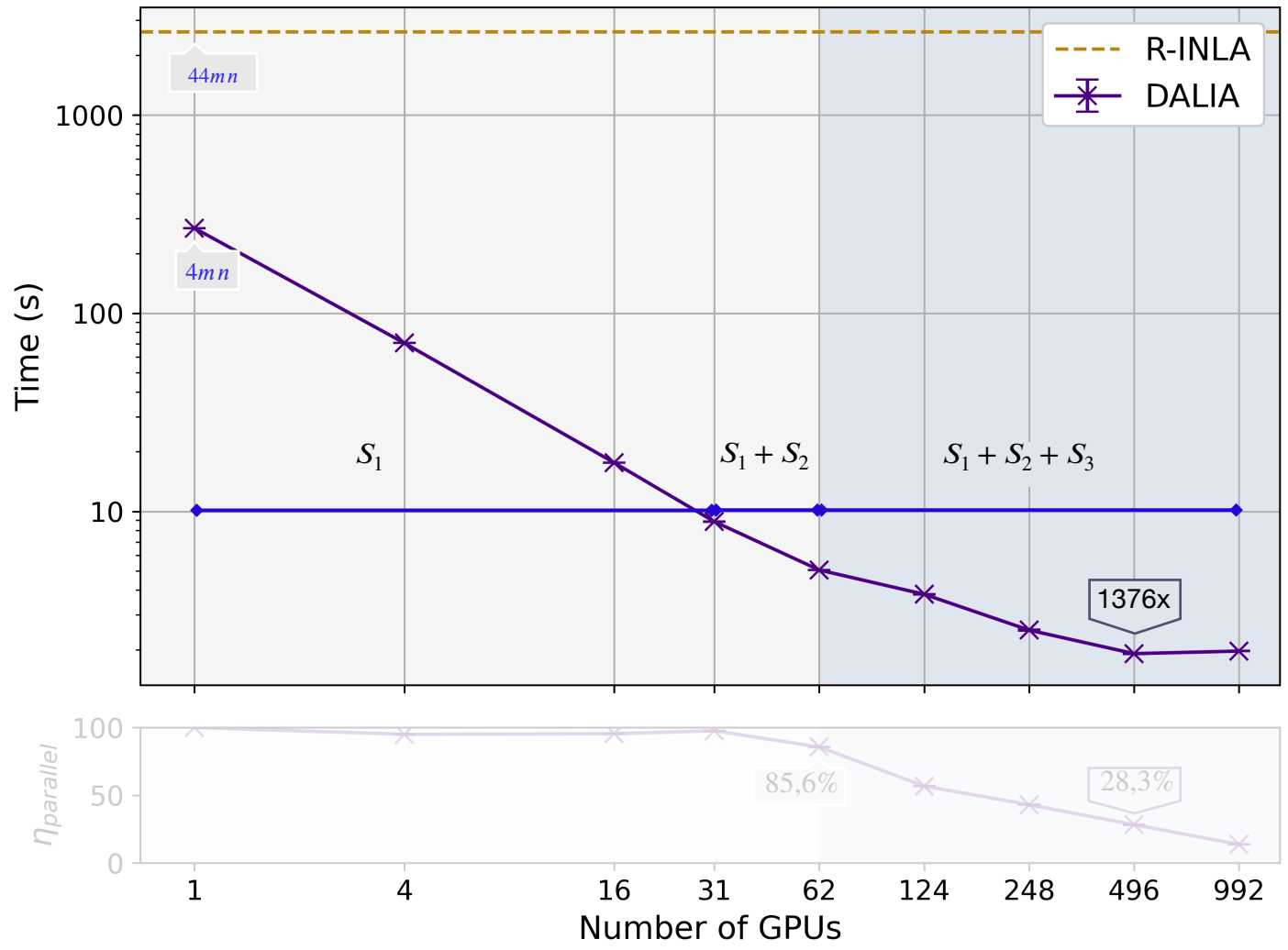
	$\dim(\theta)/n_v$	n_s/n_r	n_t	N	Mem
MB_1	4/1	4002/6	250	1M	64GB
MB_2	-/1	1675/6	128 - 2048	214k - 3.3M	5.7 – 92GB
$\mathbf{WA_1}$	**15/3**	**1247/1**	**2 - 512**	**7.5k - 1.9M**	**0.3 – 115GB**
WA_2	15/3	[72, 282, 1119, 4485]/1	48	10k - 646k	36MB-138GB
SA_1	15/3	1675/1	192	965k	77.4GB
AP_1	15/3	4210/2	48	606k	121GB



Weak Scaling in Time



Strong Scaling



Model Parameters

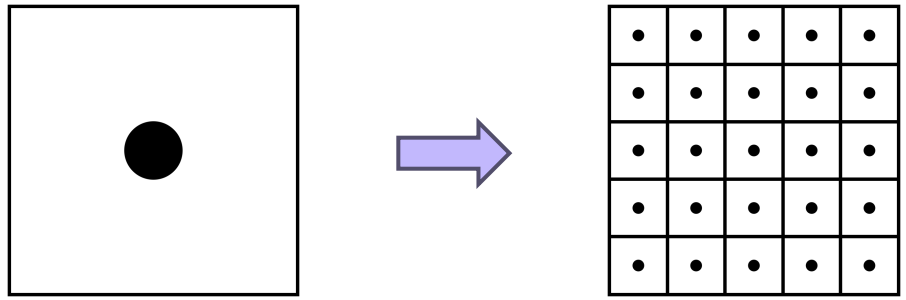
$n_s = 1675,\ n_r = 1$

$n_t = 192$

$N = 965k,\ Mem = 77GB$

Back to Case Study on Air Pollution

- Jointly model PM10, PM2.5, Ozone
- Spatial mesh: 4210 mesh nodes
- Temporal mesh: 48 days
- Covariates: Elevation, intercept
- Spatial downscaling: Increase resolution by 25x

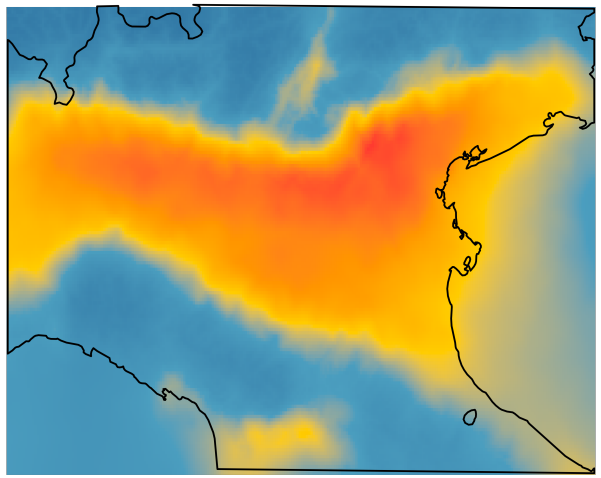


Back to Case Study on Air Pollution

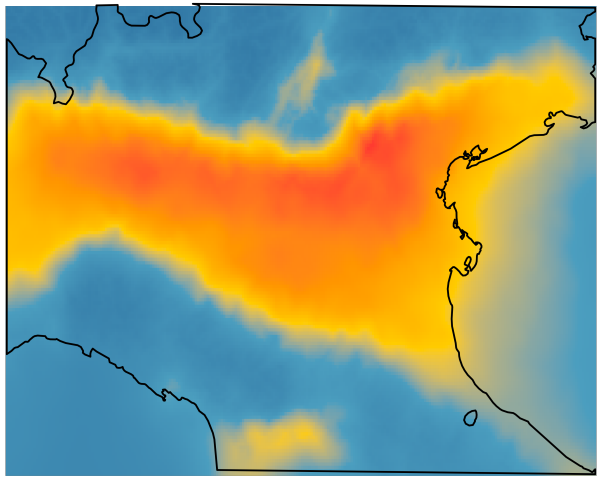
Blue: low concentration
Red: high concentration

Spatial Downscaling (Avg)

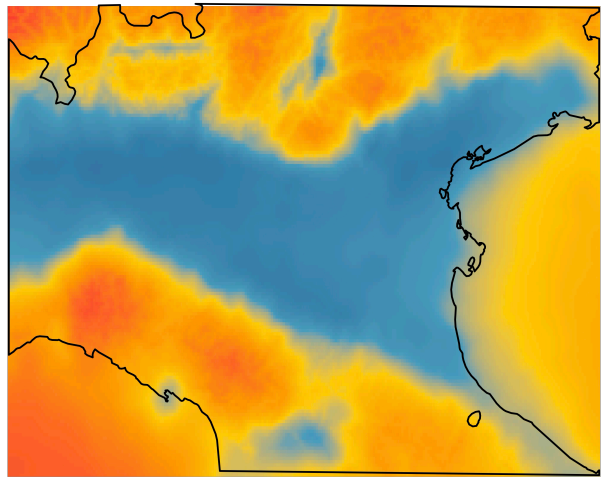
PM10



PM2.5

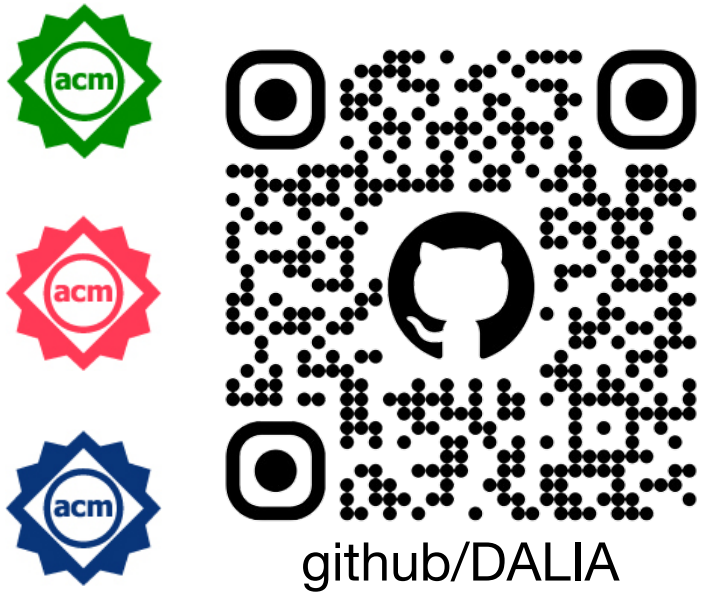


Ozone



Conclusion

- **DALIA**: Novel Bayesian framework for multivariate spatio-temporal Gaussian processes
- Allows us to scale beyond: high-resolution Bayesian Air Pollution Study
- We want to further extend DALIA!



github/DALIA

Swiss National Science Foundation
209358 (QuaTrEx)

DFG Deutsche Forschungsgemeinschaft
German Research Foundation
200021 (NumESC)

NHR FAU
n° 80227

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre
n° sm96, lp16