University of Arkansas Department of Mathematical Sciences

46th Spring Lecture Series

David Keyes

Extreme Computing Research Center

King Abdullah University of Science and Technology

5-9 April 2021

Lecture 4

Spatial Statistics Applications of HRL, TRL, and Mixed Precision

Motivation & challenge of geospatial statistics

- **Geospatial statistics predicts desired quantities directly from spatially distributed data, presumed to be drawn from a random process**
  - **For data, it may draw upon observations or simulations**
- **Alternative statistical approaches, such as wrapping many simulations in a Monte Carlo loop, are expensive given the slow convergence of Monte Carlo** ($\sim 1/\sqrt{M}$, for M trials)
  - **One can instead sample from a parameterized distribution learned from a smaller number of simulations**
- **This trades many sparse PDE simulations running at a small % of peak for a large dense linear algebra problem**
  - **"Smack in the wheelhouse" of today's HPC systems**
  - **However, memory goes as N^2, for N spatially distributed points**

Sample data sets

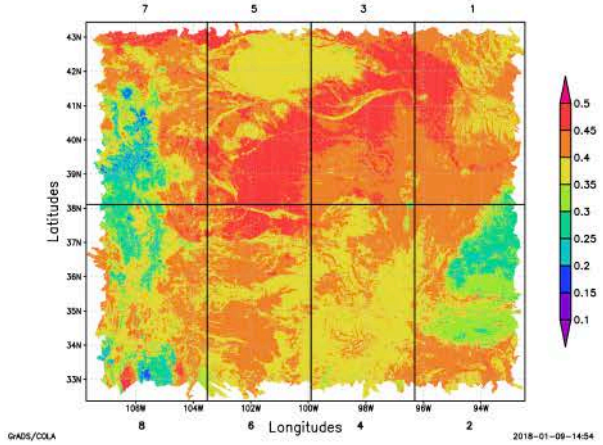
We use synthetic data of readily specified size and statistical parameters for testing and tuning algorithms

Goal is environmental modeling, particular of conditions required for engineering design of sustainable energy solutions

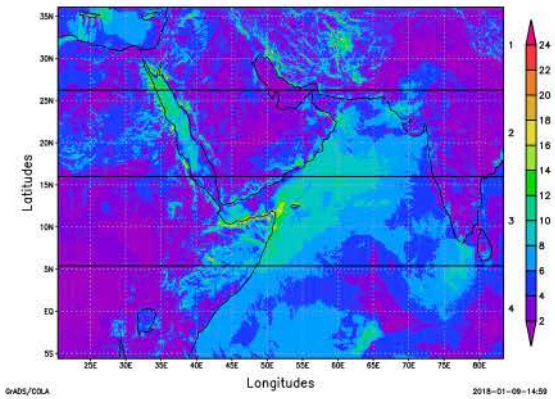
Where to plant which kind of crops, install wind farms, install solar photovoltaics?

How to estimate electricity loads for air conditioning, etc.?

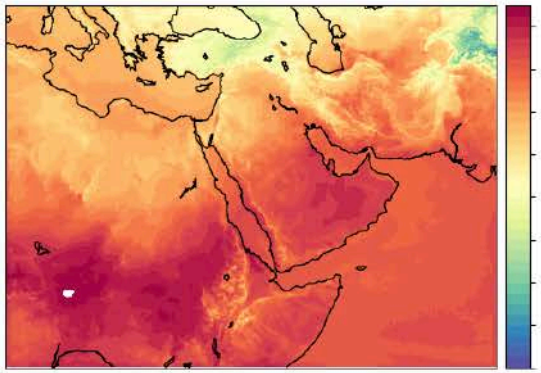
- Soil Moisture data at the top layer of the Mississippi River Basin, US, on January 1st, 2004.
- ~ 2M Locations.



- Wind Speed data at Middle East, on September 1st, 2017.
- ~ 1M Locations.



- Temperature data at Middle East, on September 1st, 2017.
- ~ 1M Locations.



Lecture 2

Geospatial statistics challenge

"Increasing amounts of data are being
by remote sensing instruments and nun
while techniques to handle millions of o
historically lagged behind…

Computational implementations that w
irregularly-spaced observations are still

- Dorit Hammerling, N

1M × 1M dense sym DP matrix requires 4 TB, $N^3 \sim 10^{18}$ Flops

Traditional approaches:

- **Global low rank**
- **Zero outer diagonals**

Better approaches:

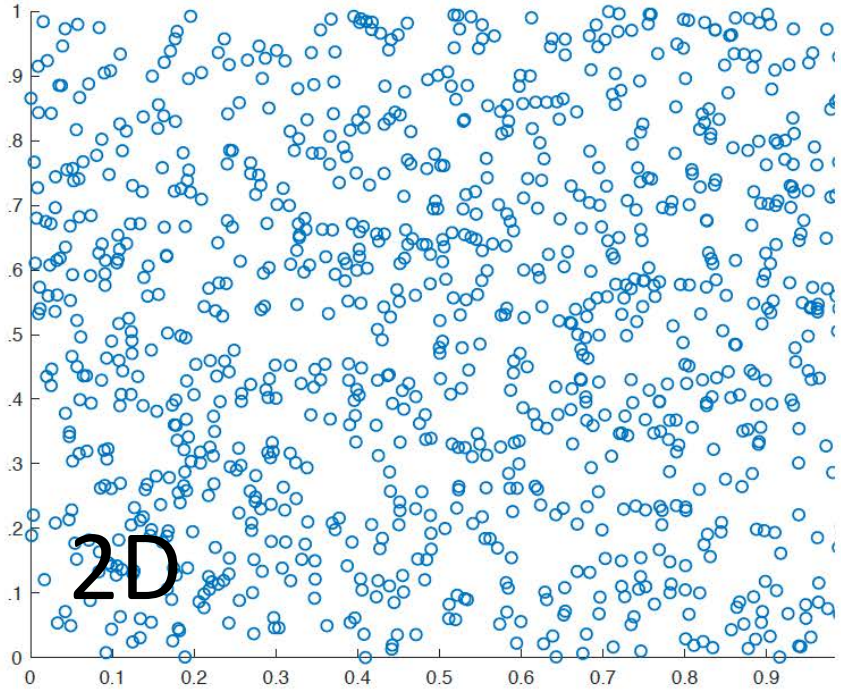
- **Hierarchical low rank**
- **Reduced precision outer diagonals**

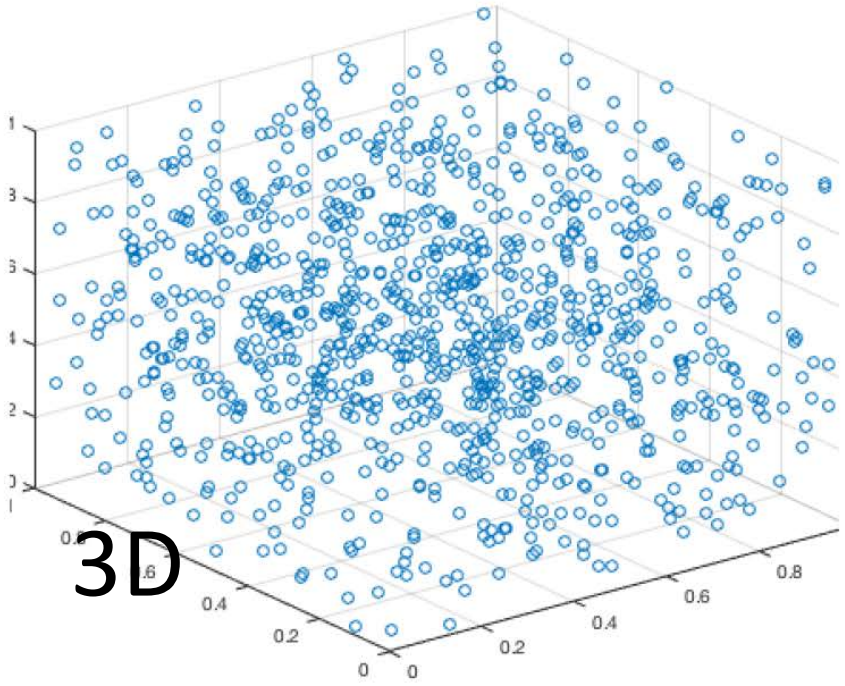
Lecture 2

Geospatial statistics test data

Synthetic test matrix: random coordinate generation within the unit square or unit cube with Matérn kernel decay, each pair of points connected by

- **linear exp to square exp decay,** $a_{ij} \sim \exp(-c|x_i - x_j|^p)$, $p = 1,2$





Maximum Likelihood Estimation

In statistics, **maximum likelihood estimation** (**MLE**) is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate.[1] The logic of maximum likelihood is both intuitive and flexible, and as such the method has become a dominant means of statistical inference.[2][3][4]

Associated with each probability distribution is a unique vector $\theta = [\theta_1, \theta_2, \ldots, \theta_k]^{\mathsf{T}}$ of parameters that index the probability distribution within a parametric family $\{f(\cdot\,;\theta) \mid \theta \in \Theta\}$ where Θ is called the parameter space, a finite-dimensional subset of Euclidean space. Evaluating the joint density at the observed data sample $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ gives a real-valued function,

$$L_n(\theta) = L_n(\theta; \mathbf{y}) = f_n(\mathbf{y}; \theta)$$

which is called the likelihood function.

Maximum likelihood estimation can be traced to Gauss when fit to a Gaussian function, with two parameters, the mean and the variance, $\theta = \{ \mu, \sigma^2 \}$.

c/o wikipedia

Maximum Likelihood Estimation

In practice, it is often convenient to work with the natural logarithm of the likelihood function, called the log-likelihood:

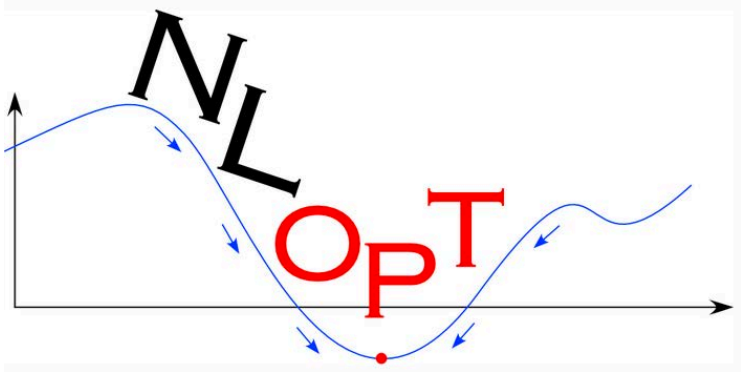
$$\ell(\theta\,;\mathbf{y}) = \ln L_n(\theta\,;\mathbf{y}).$$

Since the logarithm is a monotonic function, the maximum of $\ell(\theta\,;\mathbf{y})$ occurs at the same value of θ as does the maximum of L_n.[8] If $\ell(\theta\,;\mathbf{y})$ is differentiable in θ, the necessary conditions for the occurrence of a maximum (or a minimum) are

$$\frac{\partial \ell}{\partial \theta_1} = 0, \quad \frac{\partial \ell}{\partial \theta_2} = 0, \quad \ldots, \quad \frac{\partial \ell}{\partial \theta_k} = 0,$$

known as the likelihood equations. For some models, these equations can be explicitly solved for $\widehat{\theta}$, but in general no closed-form solution to the maximization problem is known or available, and an MLE can only be found via numerical optimization.

We use NLopt and treat it as a blackbox for the rest of this lecture. NLopt is a free/open-source library for nonlinear optimization, providing a common interface for various freely available optimization routines. Although n is typically very large, k is typically very small, like 3 or 4 in our examples, so the cost of the optimization logic is insignificant, but each optimization iteration is expensive – a large dense linear algebra problem.



c/o Wikipedia and NLopt documentation, Steven G. Johnson, http://github.com/stevengj/nlopt

Parameter identification problem

Given:

Let $s_1, \ldots, s_n$ be locations.

$Z = \{Z(s_1), \ldots, Z(s_n)\}^\top$, where $Z(s)$ is a Gaussian random field indexed by a spatial location $s \in \mathbb{R}^d$, $d \geq 1$.

Let $r_{ij} = |\, s_i - s_j \,|$.

Assumption: Z has mean zero and stationary parametric covariance function, e.g. Matérn:

$$C(\boldsymbol{\theta}) = \frac{2\sigma^2}{\Gamma(\nu)} \left(\frac{r}{2\ell}\right)^\nu K_\nu \left(\frac{r}{\ell}\right) + \tau^2 I, \quad \boldsymbol{\theta} = (\sigma^2, \nu, \ell, \tau^2).$$

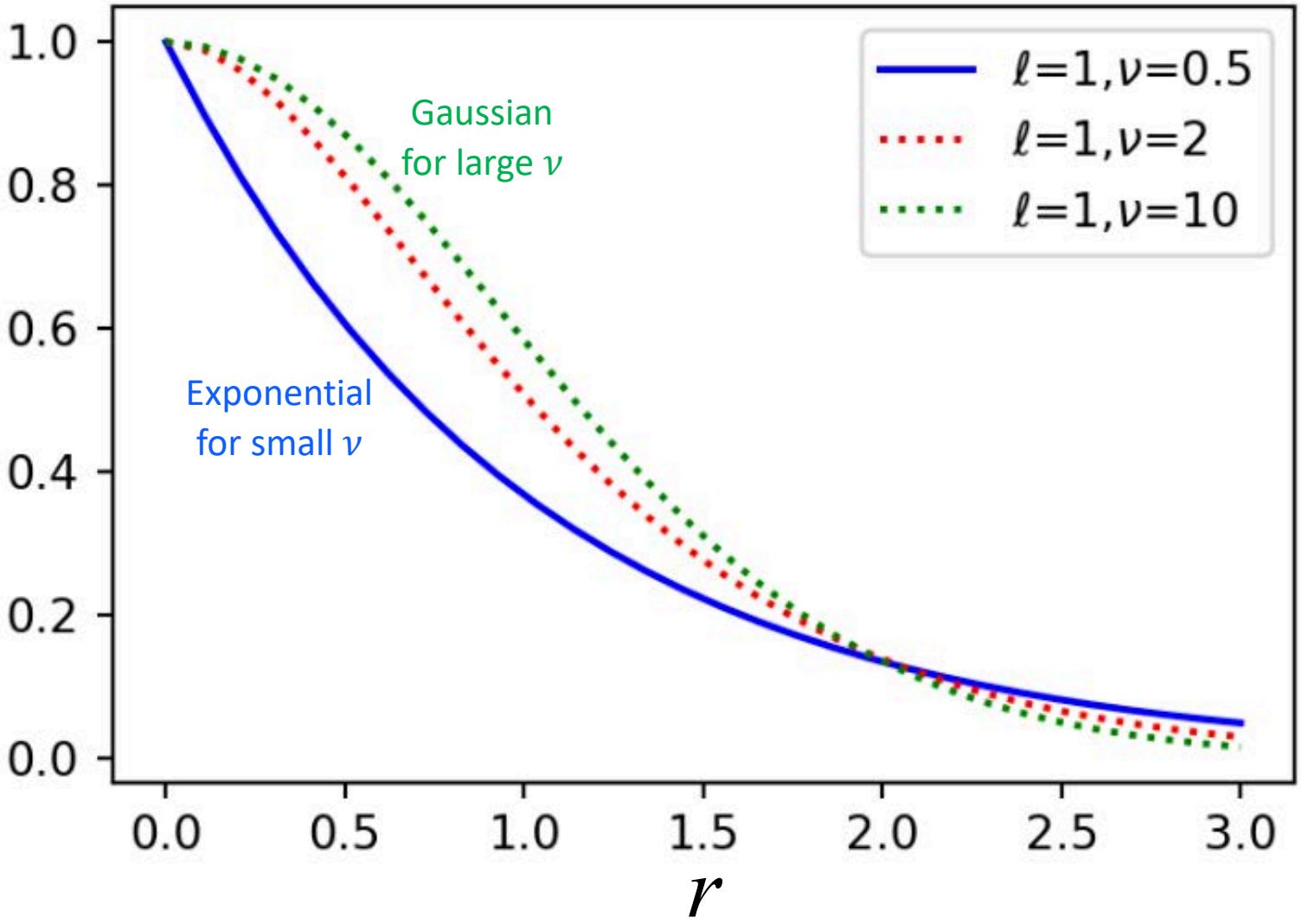
To identify: unknown parameters $\boldsymbol{\theta} := (\sigma^2, \nu, \ell, \tau^2)$.

The first three parameters are standard ("partial sill", "range", "smoothness").

The last parameter ("nugget") adds to the diagonal to preserve definiteness in upcoming approximations. (All four parameters are nonnegative, though only two are written that way.)

In the limit as $\nu \to \infty$, Matérn reduces to a Gaussian. The key shape parameters are ν and ℓ . Apart from the diagonal shift of τ^2, σ^2 is just a scaling factor.

Matérn distribution
as function of r for ℓ=1 and various ν



Mean and covariance of random fields

The mean function of $Z(\boldsymbol{s})$ is

$$\mu(\boldsymbol{s}) = \mathbb{E}\{Z(\boldsymbol{s})\}$$

The covariance function of $Z(\boldsymbol{s})$ is

$$C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \mathrm{cov}\{Z(\boldsymbol{s}_1), Z(\boldsymbol{s}_2)\} = \mathbb{E}[\{Z(\boldsymbol{s}_1) - \mu(\boldsymbol{s}_1)\}\{Z(\boldsymbol{s}_2) - \mu(\boldsymbol{s}_2)\}],$$

where $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are two spatial locations

The covariance matrix Σ is

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_1 - \mathrm{E}[X_1])] & \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_2 - \mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_n - \mathrm{E}[X_n])] \\ \mathrm{E}[(X_2 - \mathrm{E}[X_2])(X_1 - \mathrm{E}[X_1])] & \mathrm{E}[(X_2 - \mathrm{E}[X_2])(X_2 - \mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_2 - \mathrm{E}[X_2])(X_n - \mathrm{E}[X_n])] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mathrm{E}[X_n])(X_1 - \mathrm{E}[X_1])] & \mathrm{E}[(X_n - \mathrm{E}[X_n])(X_2 - \mathrm{E}[X_2])] & \cdots & \mathrm{E}[(X_n - \mathrm{E}[X_n])(X_n - \mathrm{E}[X_n])] \end{bmatrix}$$

Valid covariance functions/matrices

A covariance function must possess positive semi-definiteness

$$\sum_{j,k=1}^{n} c_j c_k C(\boldsymbol{s}_j, \boldsymbol{s}_k) \geq 0$$

for any finite n, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in D$, and real numbers $c_1, \ldots, c_n$

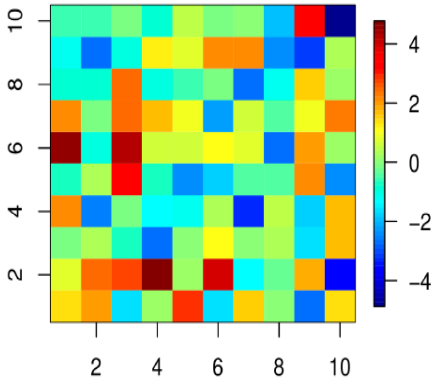
Or in matrix form,

$$x^{\mathsf{T}} \Sigma y \geq 0$$

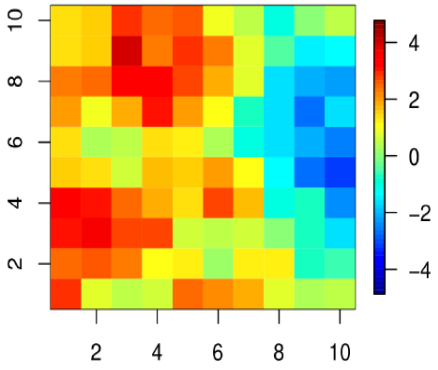
For any n-dimensional real vectors x, y .

To understand the importance of covariance, consider making the “best” guess for a missing pixel in the two 10 x 10 arrays below. Both are zero-mean random fields, but (a) is completely uncorrelated and nearby cells in (b) have modest correlation, or positive covariance.

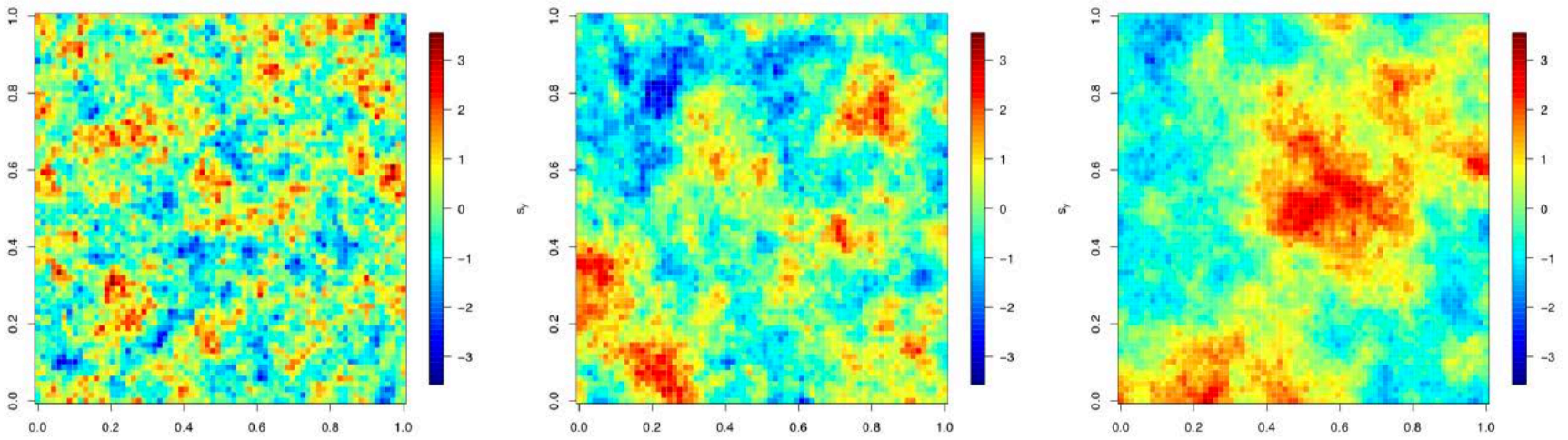
(a)



(b)

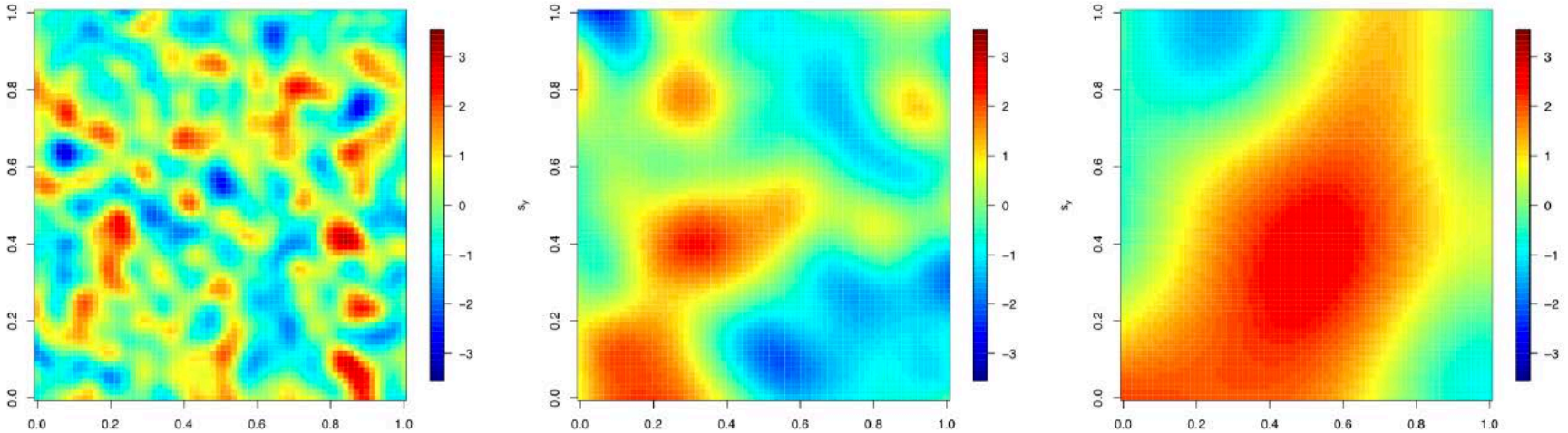


Range parameter ℓ sets correlation length



ℓ=0.033 ℓ=0.100 ℓ=0.234

WEAK Exponential covariance ($\nu = 0.5$) STRONG



ℓ=0.058 ℓ=0.173 ℓ=0.404

Gaussian covariance ($\nu = \infty$)

Covariance parameter estimation

For simplicity, we focus on zero-mean stationary Gaussian random fields. The log-likelihood for n locations:

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}\boldsymbol{Z}^\top\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{Z},$$

determinant inverse

where

$$\boldsymbol{Z} = \begin{pmatrix} Z(\boldsymbol{s}_1) \\ \vdots \\ Z(\boldsymbol{s}_n) \end{pmatrix}, \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} C(\boldsymbol{s}_1, \boldsymbol{s}_1; \boldsymbol{\theta}) & \dots & C(\boldsymbol{s}_1, \boldsymbol{s}_n; \boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ C(\boldsymbol{s}_n, \boldsymbol{s}_1; \boldsymbol{\theta}) & \dots & C(\boldsymbol{s}_n, \boldsymbol{s}_n; \boldsymbol{\theta}) \end{pmatrix}$$

- Log determinant and linear solver require a Cholesky factorization of the given covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$
- Cholesky factorization requires $O(n^3)$ floating point operations and $O(n^2)$ memory

Prediction

- Assuming $\Sigma_{11} \in \mathbb{R}^{m \times m}$, $\Sigma_{12} \in \mathbb{R}^{m \times n}$, $\Sigma_{21} \in \mathbb{R}^{n \times m}$, and $\Sigma_{22} \in \mathbb{R}^{n \times n}$.

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N_{m+n}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \tag{1}$$

- The associated conditional distribution can be represented as

$$Z_1 | Z_2 \sim N_m(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(Z_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}). \tag{2}$$

Schur complement

- Assuming that the known measurements vector Z_2 has a zero-mean function (i.e., $\boldsymbol{\mu}_1 = 0$ and $\boldsymbol{\mu}_2 = 0$), the unknown measurements vector Z_1 can be predicted using,

$$Z_1 = \Sigma_{12}\Sigma_{22}^{-1} Z_2. \tag{3}$$

- Solution of system of linear equation $(\Sigma_{22}^{-1} Z_2)$ requires also the Cholesky factorization of Σ_{22}.

Parameter identification complexity

Maximum Likelihood Estimator

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \{\log |\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \boldsymbol{Z}^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \boldsymbol{Z}\}\end{aligned}$$

$O(n^3)$ floating point operations and $O(n^2)$ memory requirements for exact computations

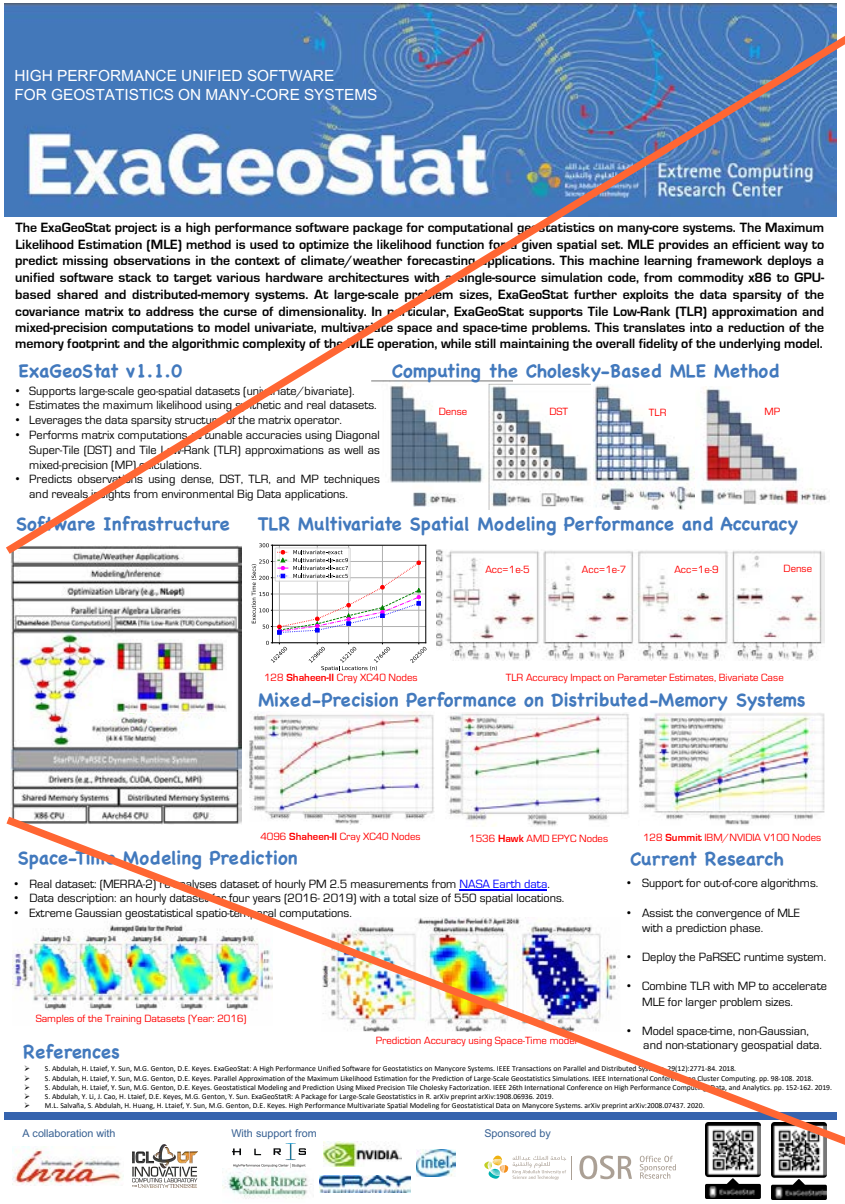
With cubic complexity, a 1-minute computation made 10 times larger at constant computing rate would require nearly 17 hours, and computing rate would likely degrade due to greater memory traffic, prolonging the run further.

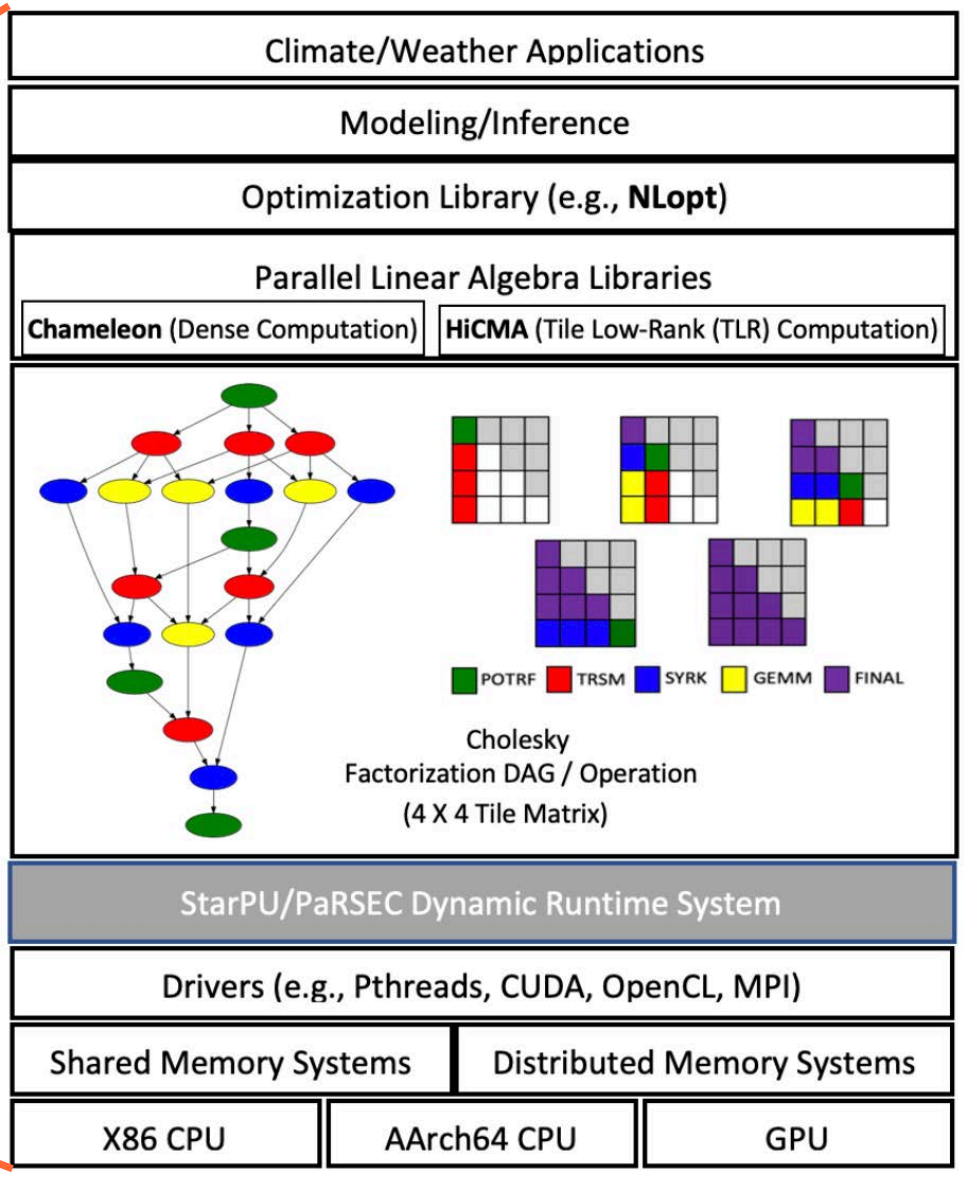
Enter, of course, data sparsity considerations…

In lecture 2, Tile Low Rank

In lecture 4, mixed precision

https://github.com/ecrc/exageostat



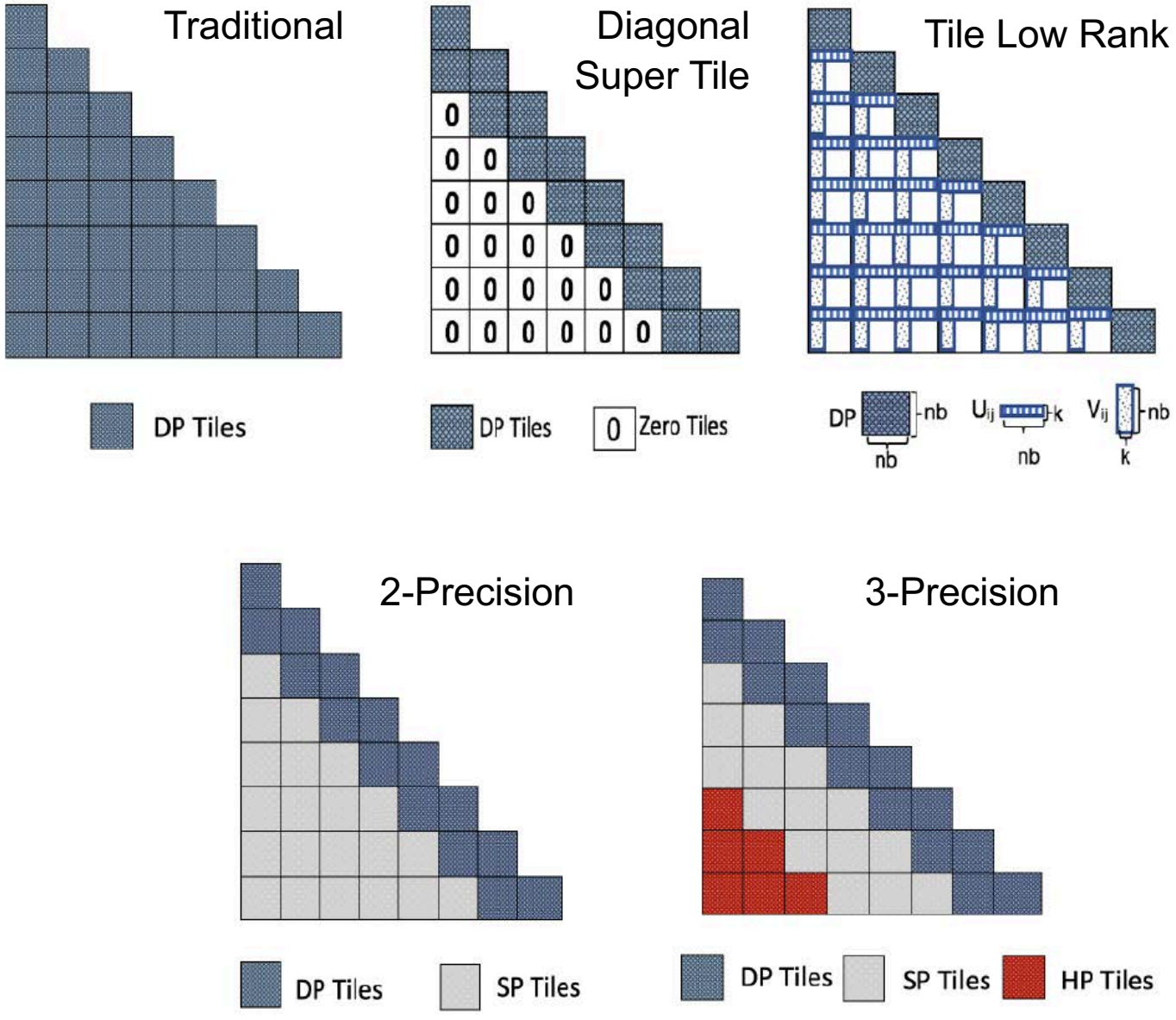


ExaGeoStat framework

Sameh Abdulah

- **Synthetic Dataset Generator**
  - generate large-scale geospatial datasets which can separately be used as benchmark datasets for other software packages
- **Maximum Likelihood Estimator (MLE)**
  - evaluate the maximum likelihood function on large-scale geospatial datasets within the family of tile algorithms
  - support full machine precision accuracy (full-matrix) and Tile Low-Rank (TLR) approximation
  - support mixed precision optimizations
- **ExaGeoStat Predictor**
  - predict unknown measurements on known geospatial locations by leveraging the MLE estimated parameters

Block Cholesky approaches



Illustrative application

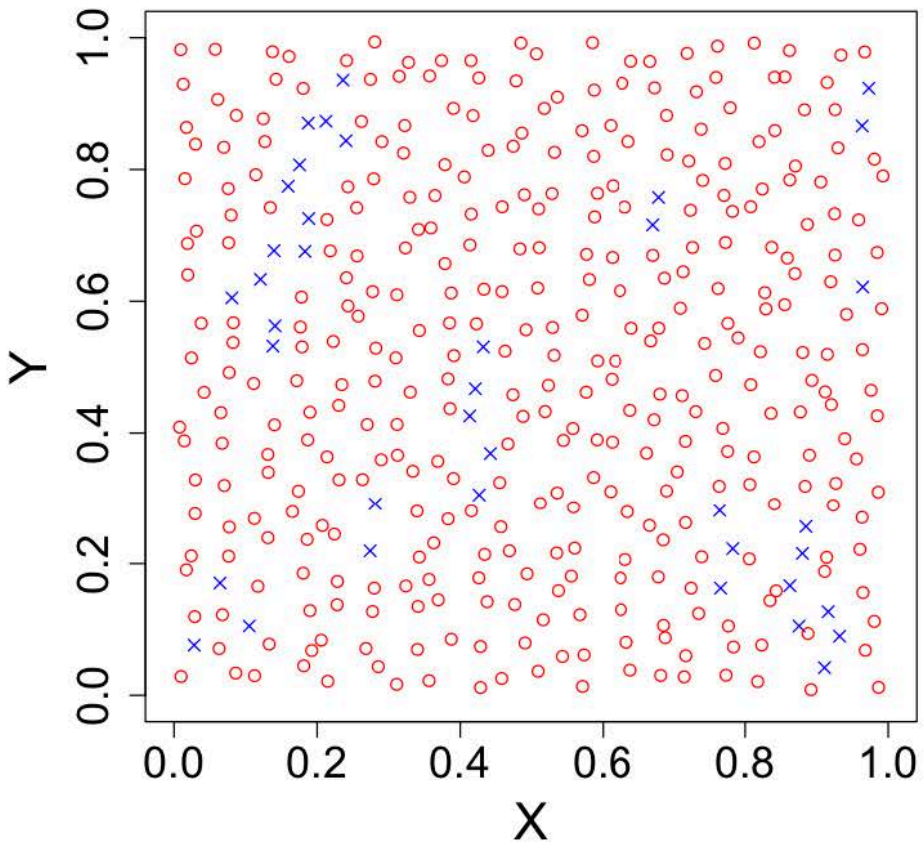
- Cholesky factorization of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{V} \cdot \boldsymbol{V}^{\top}$$

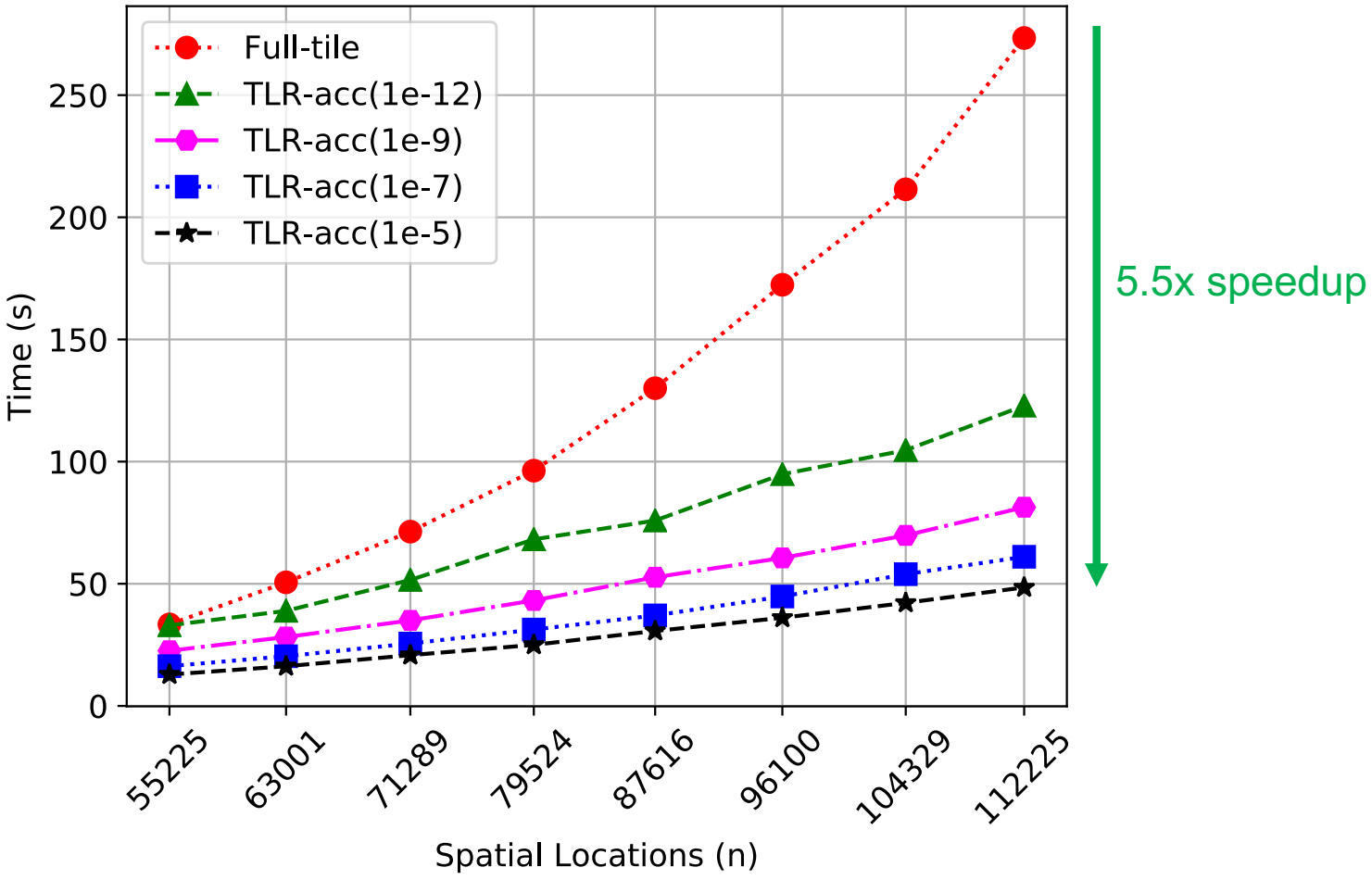
- $\boldsymbol{Z}$ vector generation

$$\boldsymbol{Z} = \boldsymbol{V} \cdot \boldsymbol{e},\ \boldsymbol{e} \sim N(0,1)$$

- An example of 400 points irregularly distributed in space with 362 points ($\circ$) for maximum likelihood estimation and 38 points ($\times$) for prediction validation.

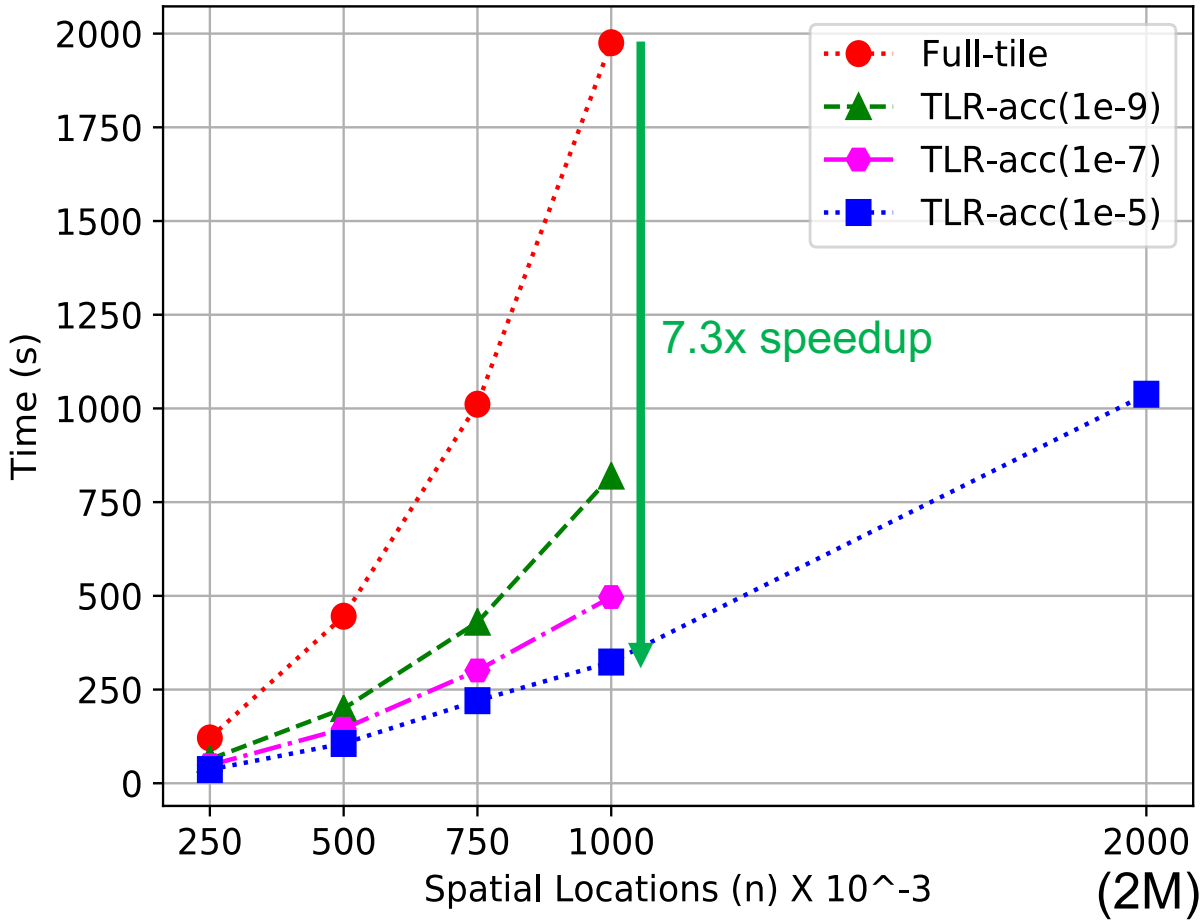


ExaGeoStat V1.0.0, shared memory (TLR, Nov. 2018)



Time for one MLE iteration with different problem sizes for different accuracies per block on 56-core Intel Skylake

ExaGeoStat V1.0.0, distributed memory (TLR approx., Nov. 2018)



Time for one MLE iteration with different problem sizes on Cray XC40 (Shaheen 1024 nodes)

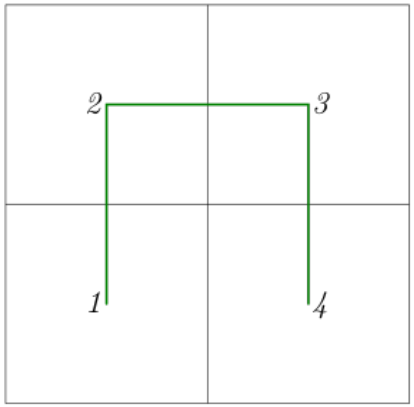
Motivations for mixed precision

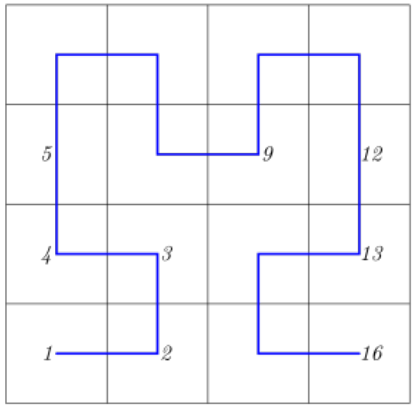
- **Mathematical: better than "zero precision"**
  - **if statisticians are used to treating remote diagonals as zero after performing a diagonally clustered space-filling curve ordering, then their coefficients must often be orders of magnitude down from the diagonals**
  - **not just "smooth" in the low-rank sense, but actually small**
- **Computational: faster time to solution**
  - **hence lower energy consumption and higher performance, especially by exploiting heterogeneity**

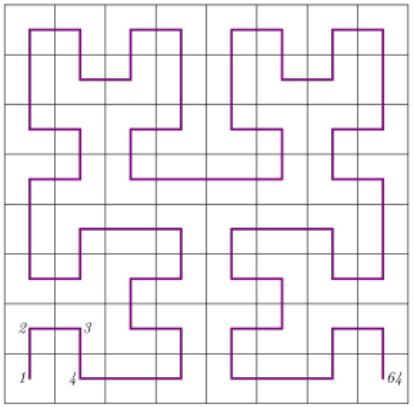
	V100 NVIDIA NVLink	A100 NVIDIA NVLink
Peak FP64 Performance	7.5 TF	9.7 TF
Peak FP64 Tensor Core	—	19.5 TF
Peak FP32 Performance	15 TF	19.5 TF
Peak Tensor Float 32 (TF32)	—	156 TF
Peak FP16 Tensor Performance	120 TF	312 TF
	rel. Dec 2017	rel. May 2020

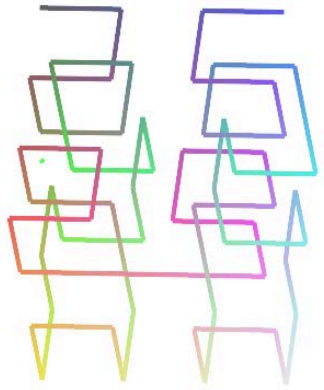
Diagonal-based reasoning depends upon good orderings

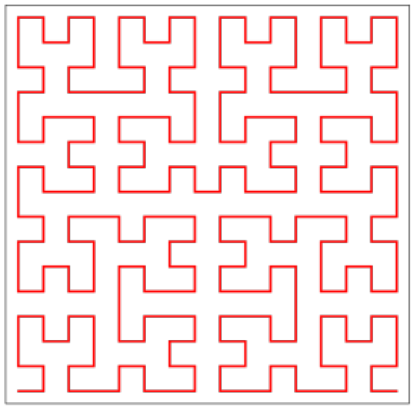
Points near each other in 1D memory must be near each other, on average, in N-dimensional space, using space-filling curves

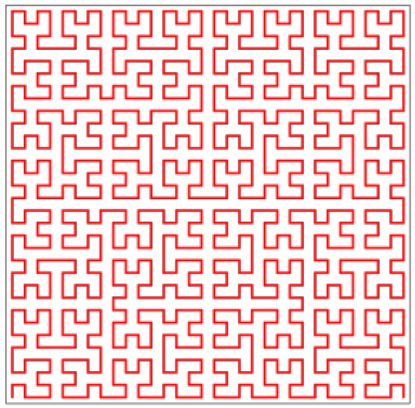


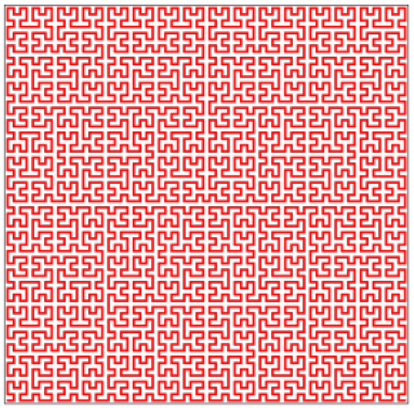




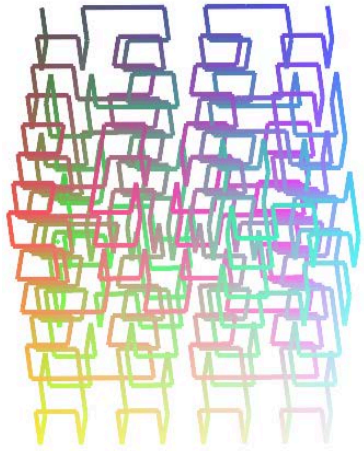








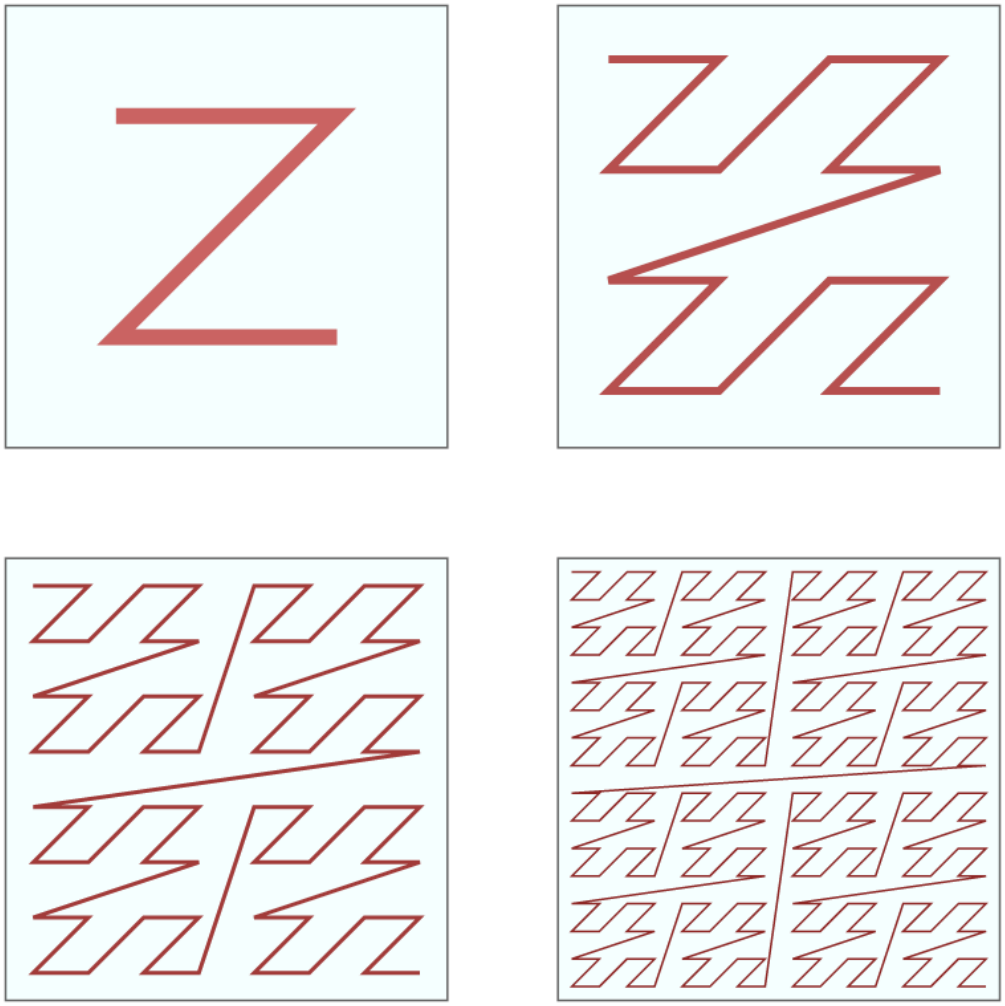
2D Hilbert curves



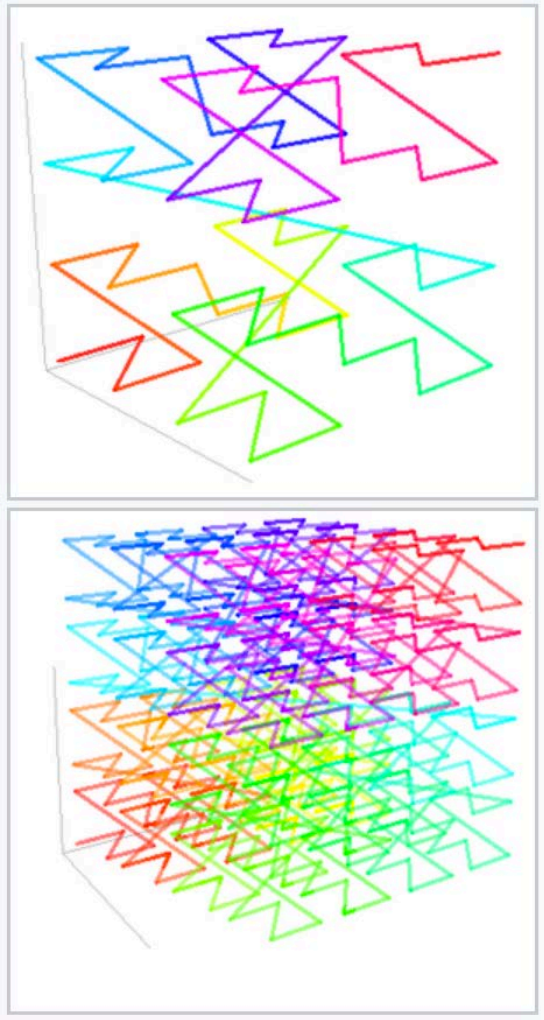
3D Hilbert curves equipartitioned by color

mons.wikimedia.org/w/index.php?curid=47570255

Diagonal based reasoning depends upon good orderings



2D Morton curves

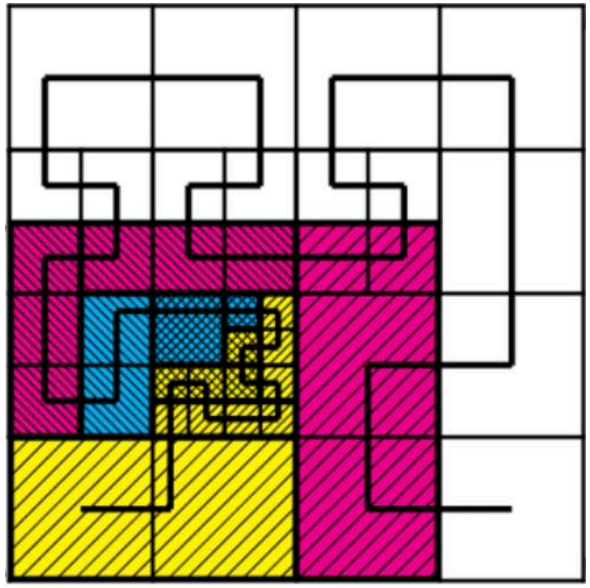


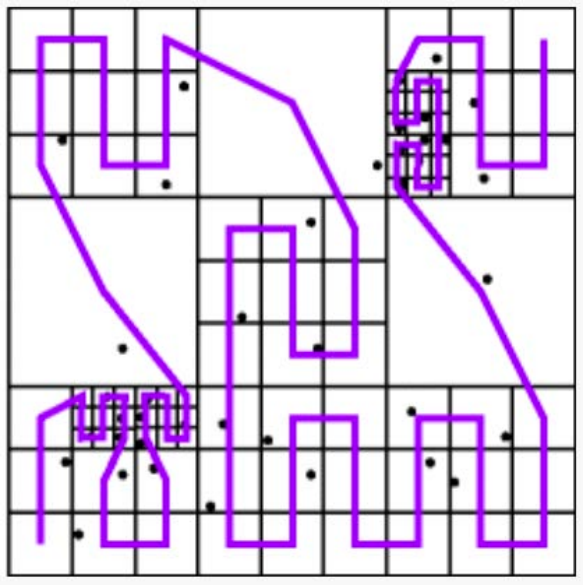
3D Morton curves equipartitioned by color

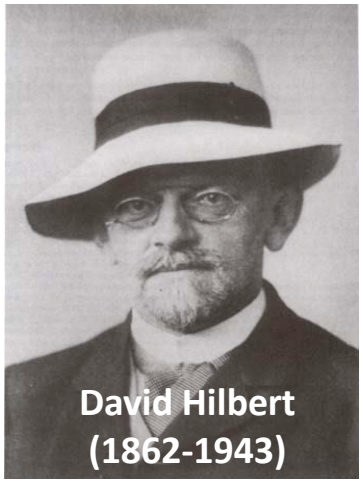
https://commons.wikimedia.org/w/index.php?curid=3879675

Extension to nonuniform distributions

Recursively subdivide within Cartesian blocks by density





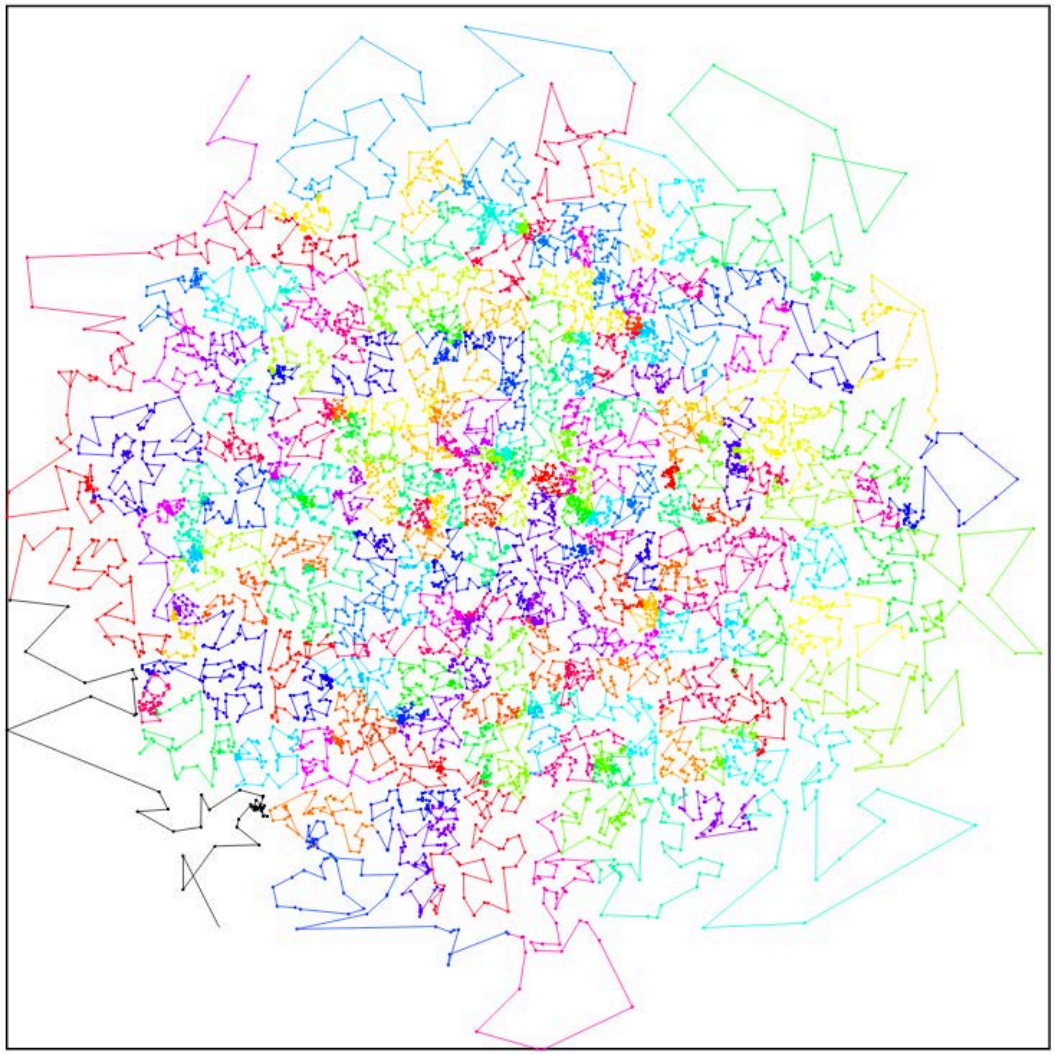


David Hilbert
(1862-1943)

- We live in his spaces ("Hilbert Space" – a function space equipped with an inner product)
- His statement of 23 problems, first presented at the Paris International Mathematical Congress in 1900, gave definition to much of 20th century mathematical research
- Invented what we now call the Hilbert (or Peano-Hilbert) curve in 1891
- Brought into computer science in 1997

Extension to nonuniform distributions

Ordered by points instead of cells



10,000 points in 200 groups of 50 each

- **John Salmon (now of D E Shaw Research, then of LLNL) "connected the dots" back to Hilbert for his N-body computations with Mike Warren at SC'97**
- **Won "Test of Time" award at SC'18**

John Salmon,
with early universe simulation

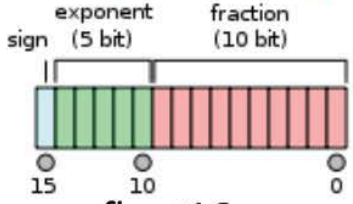
Adapt precision to accuracy requirements

Precision	Type	Bits Signif (t)	Exp	Range	$u = 2^{-t}$
half	bfloat16	8	8	$10^{\pm 38}$	3.9×10^{-3}
half	fp16	11	5	$10^{\pm 5}$	4.9×10^{-4}
single	fp32	24	8	$10^{\pm 38}$	6.0×10^{-8}
double	fp64	53	11	$10^{\pm 308}$	1.1×10^{-16}

fp64, fp32, fp16 defined by IEEE standard

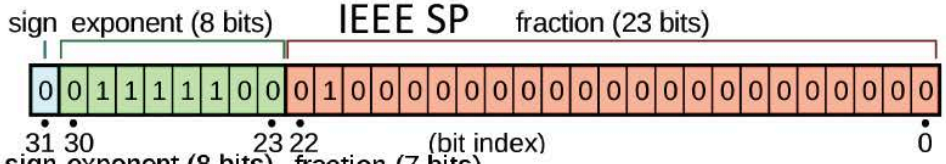
Bfloat16: Google, Intel, ARM, NVIDIA

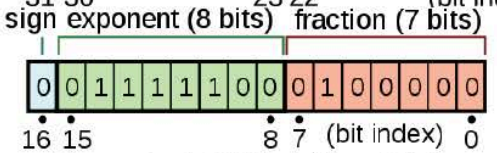
- **Note the number range with half precision (16 bit fl.pt.)**



float16

largest fl pt number 65,504



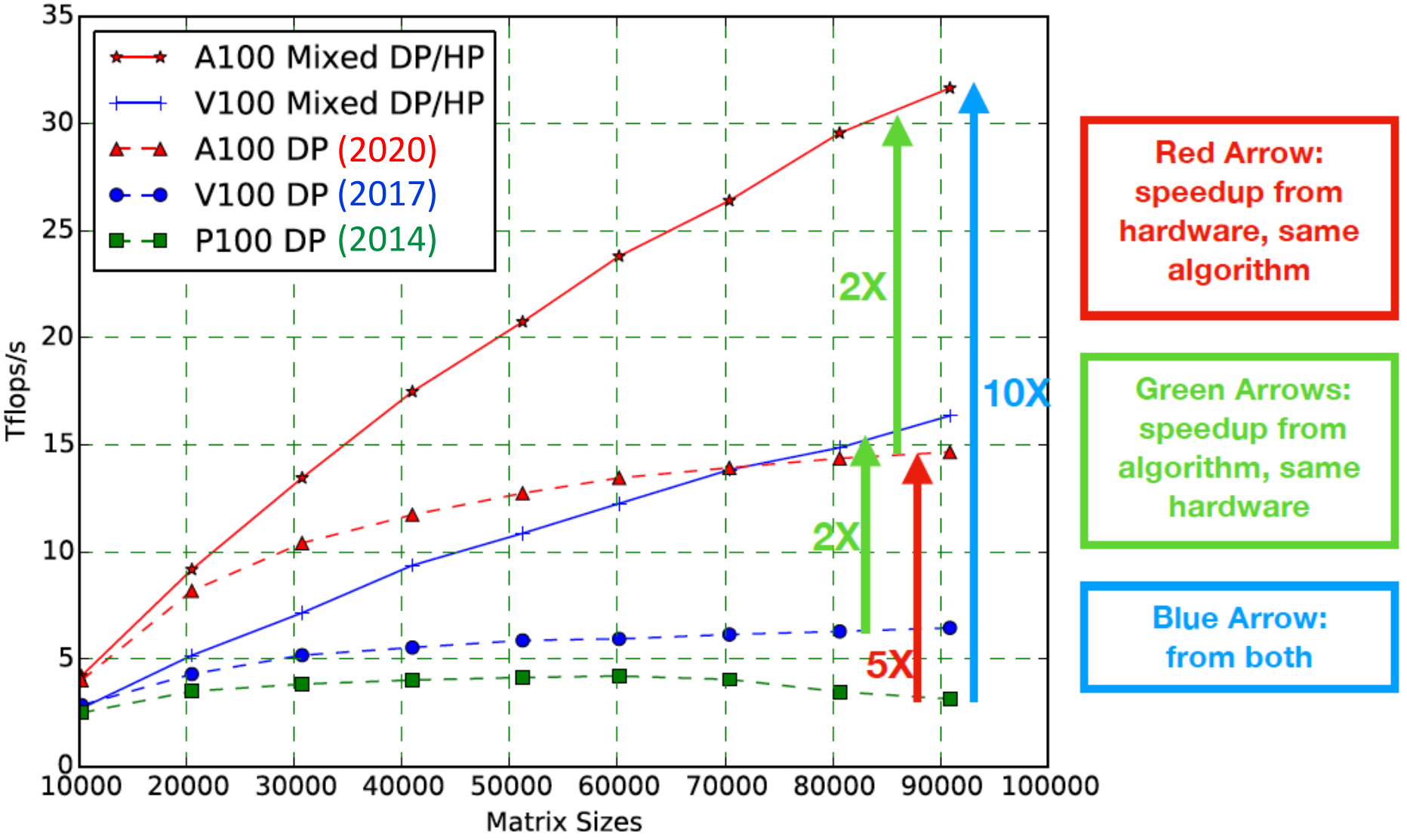


Google TPU: bfloat16

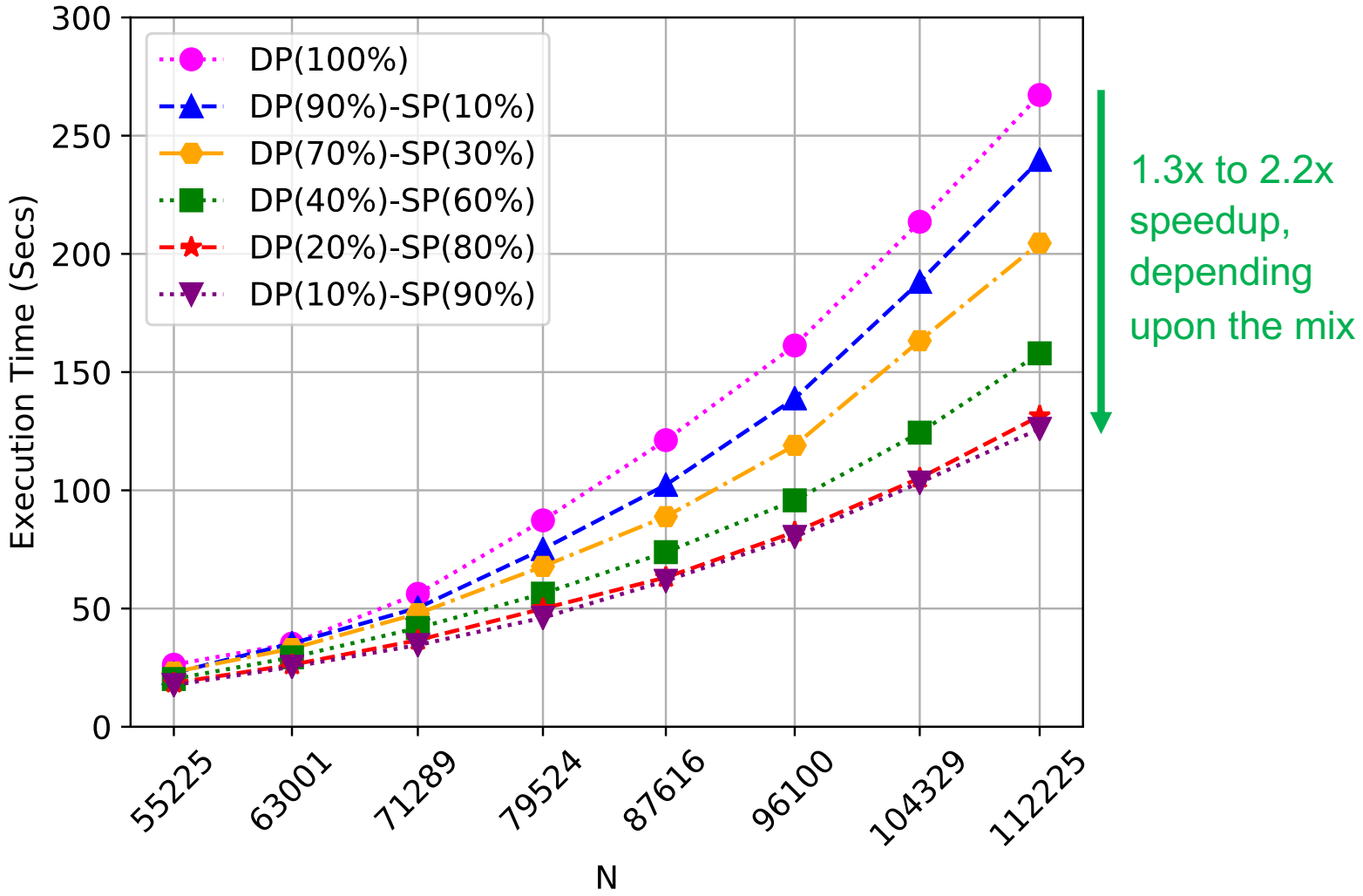
largest fl pt number $O(10^{38})$

c/o Nick Higham (2021), Jack Dongarra (2021)

Mixed precision dense Cholesky



Mixed-precision ExaGeoStat (MP approx., 2019)



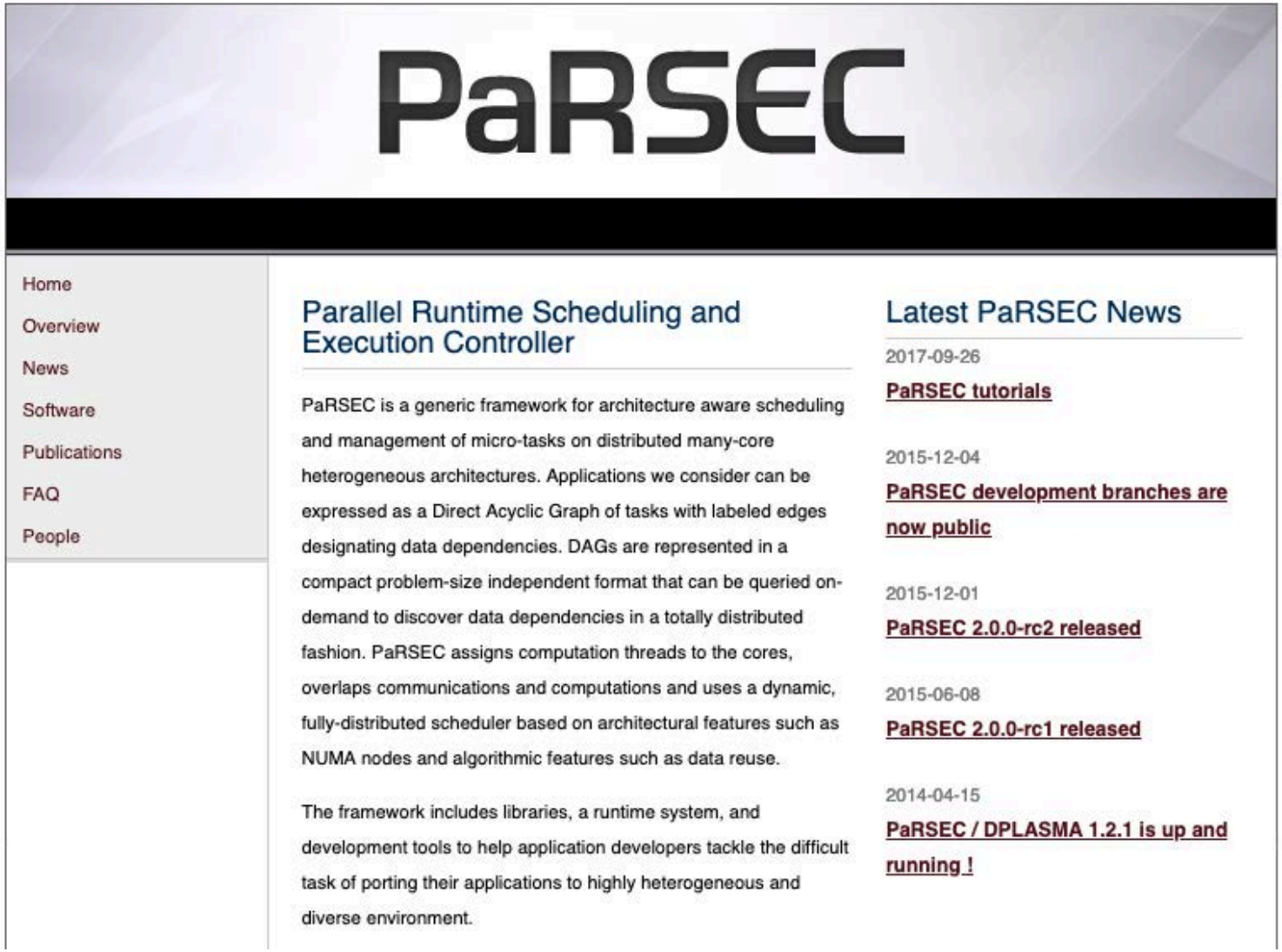
Time for one MLE iteration with different problem sizes on Skylake/V100

ExaGeoStat with PaRSEC runtime system

- StarPU runtime has scalability issues as more nodes are used
- Implemented ExaGeoStat's MLE and prediction steps with PaRSEC's parameterized task graph (PTG) interface with better scalability
- Novel mixed-precision implementation extends to three precisions, exploiting the Nvidia GPU Tensor Core
- A specialized lookahead scheme and workload balancer among the GPUs on the same node ensured performance
- All kernels run on GPU
- PaRSEC coordinates data transfer between host and device, inter-node communication
- Banded distribution for different precisions reduce memory footprint while managing workload balancing
- Features operator overloading with type conversion

PaRSEC runtime system

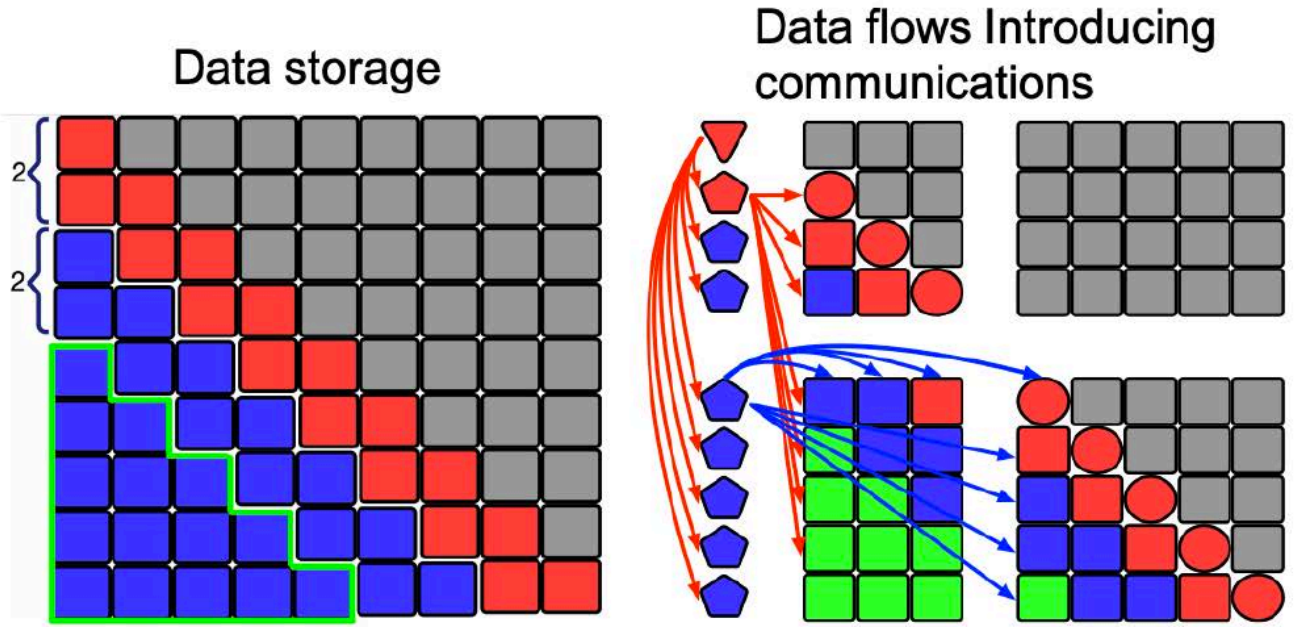
http://icl.cs.utk.edu/PaRSEC/index.html



George Bosilca
ICL, University of Tennessee

Three-precision dense Cholesky

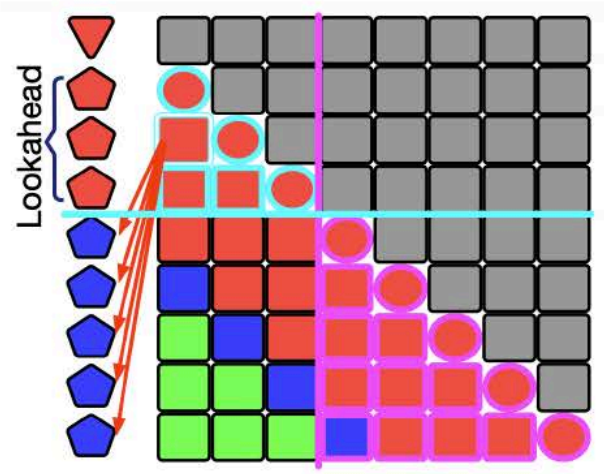
- Store a single copy of the matrix
- Support collective communications
- Encapsulate the datatype into the data-flow
- Mitigate load imbalance overheads in computation / communication using lookahead and hybrid (and nested) data distribution



Abdulah, Cao, Pei, Bosilca, Dongarra, Genton, K., Ltaief & Sun, *Accelerating Geostatistical Modeling and Prediction with Mixed-Precision Computations*. IEEE TDPS, 2021.

Three-precision dense Cholesky

- Possibility of tasks far away from the critical path may be scheduled first, e.g., tiles with magenta boundary.
- The concept of control dependency between tasks in PaRSEC guides the task execution order and priorities by adding an empty dependency (without data encapsulated).
- Expediting the discovery of tasks on/near the critical path, and enough workloads could be guaranteed.



PaRSEC Lookahead Technique.

Abdulah, Cao, Pei, Bosilca, Dongarra, Genton, K., Ltaief & Sun, *Accelerating Geostatistical Modeling and Prediction with Mixed-Precision Computations.* IEEE TDPS, 2021.

Three-precision dense Cholesky

- Possibility of tasks far away from the critical path may be scheduled first, e.g., tiles with magenta boundary.
- The concept of control dependency between tasks in PaRSEC guides the task execution order and priorities by adding an empty dependency (without data encapsulated).
- Expediting the discovery of tasks on/near the critical path, and enough workloads could be guaranteed.

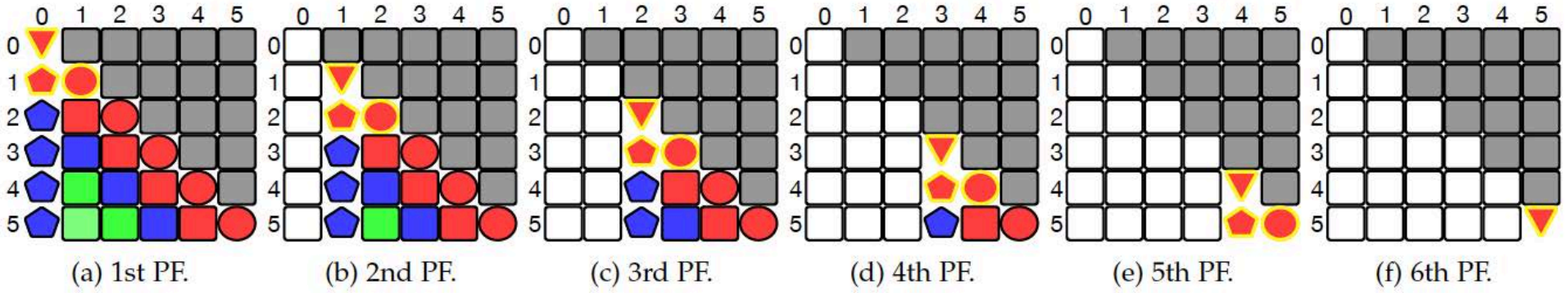
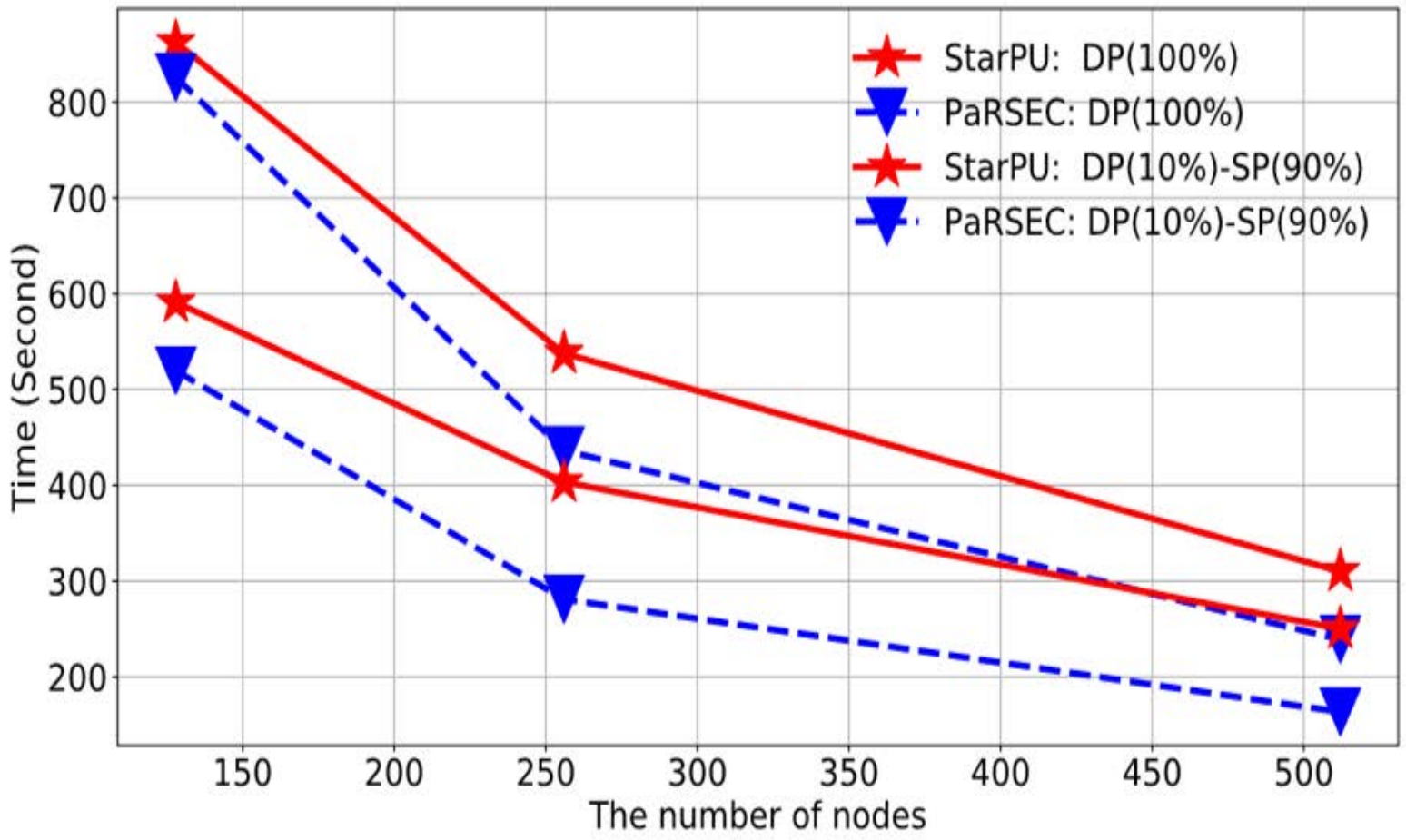


Fig. 2: Mixed-precision Cholesky factorization with 6×6 tiles, `band_size_dp` $= 2$, and `band_size_sp` $= 1$. White tiles represent the completed task. Other colors represent different precisions for each tile: `DP` in red, `SP` in blue, and `HP` in green. Different shapes indicate different kernels: triangle `POTRF`, square `GEMM`, pentagon `TRSM`, and circle `SYRK`.

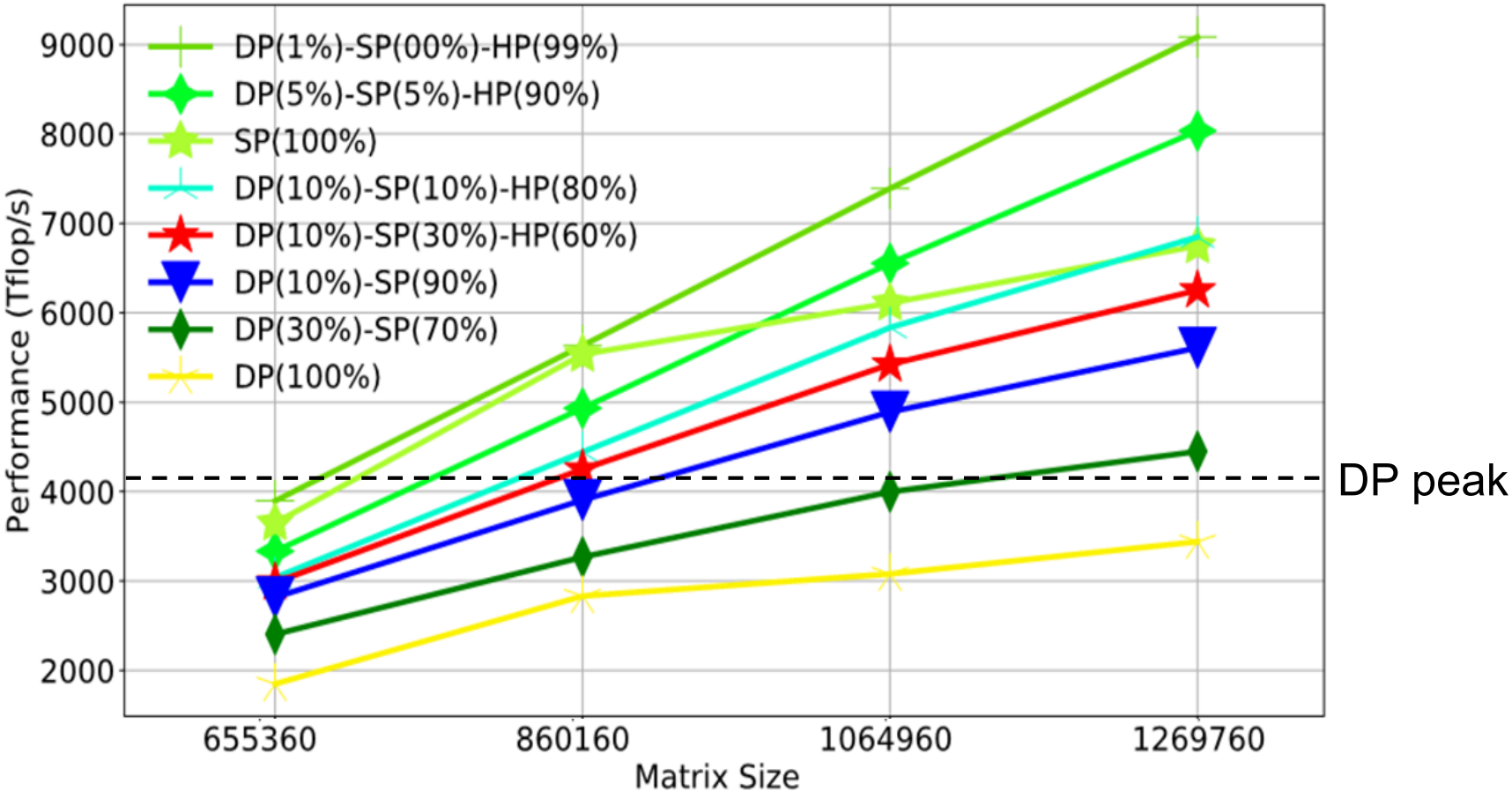
Abdulah, Cao, Pei, Bosilca, Dongarra, Genton, K., Ltaief & Sun, *Accelerating Geostatistical Modeling and Prediction with Mixed-Precision Computations*. IEEE TDPS, 2021.

ExaGeoStat with PaRSEC runtime system



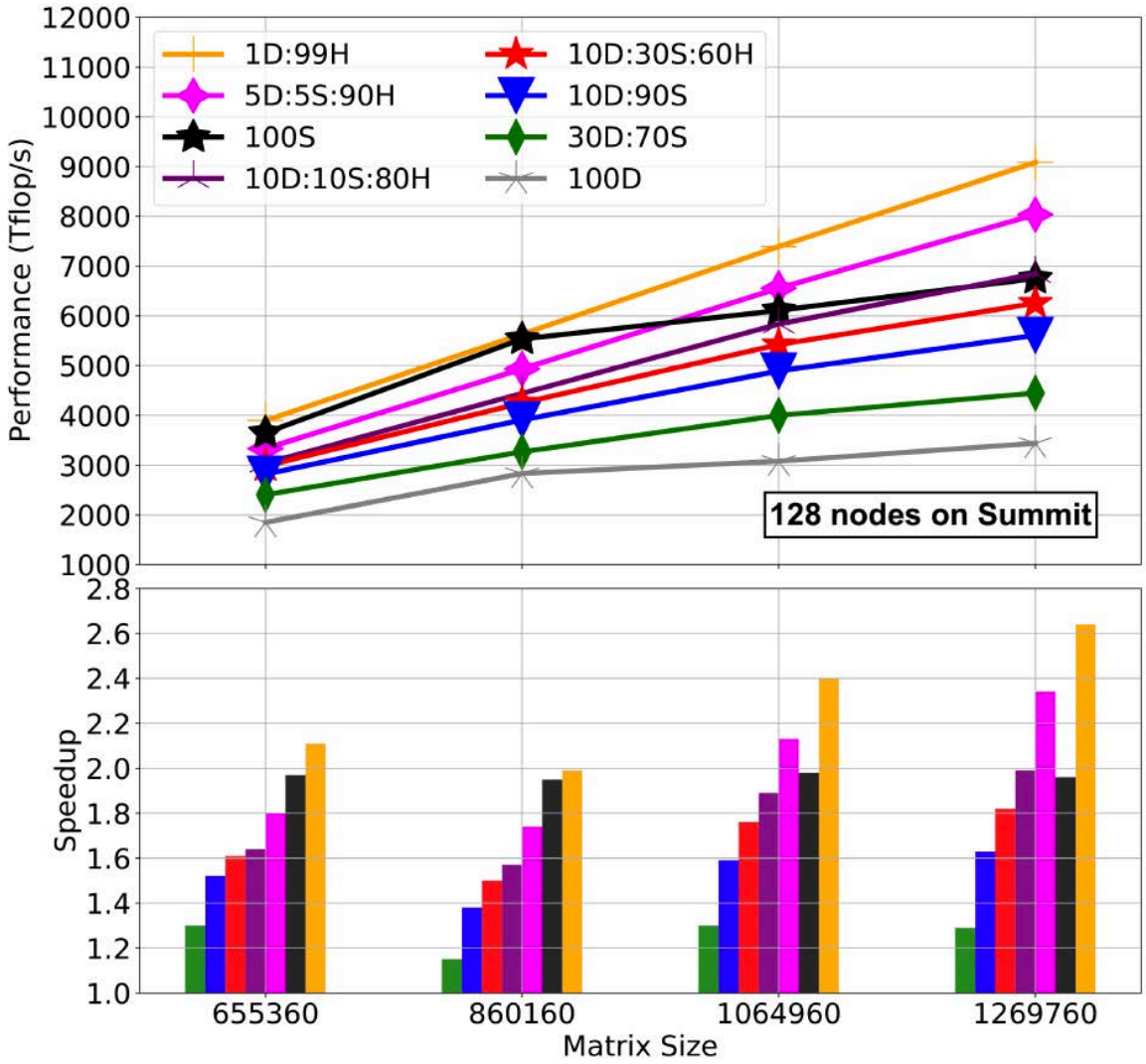
Cholesky factorization Scalability on Cray XC40 – Shaheen-II with different runtime systems

ExaGeoStat with 3 precisions and PaRSEC runtime system on Summit



128 ORNL Summit nodes (V100 GPUs) (DP on single node can achieve 32+ Tflop/s)

Three-precision dense Cholesky

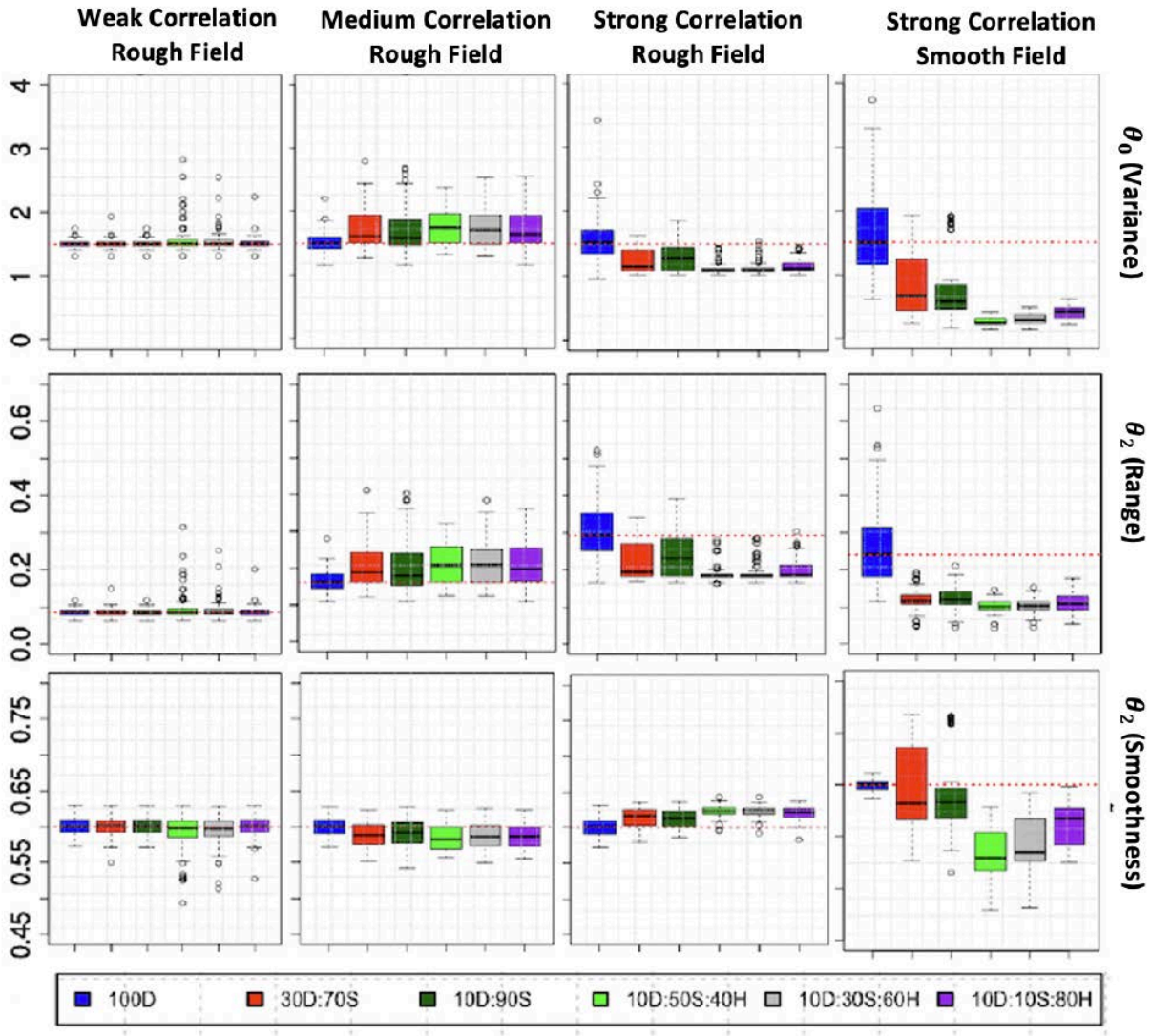


128 ORNL Summit nodes (V100 GPUs) (DP on single node can achieve 32+ Tflop/s)

Abdulah, Cao, Pei, Bosilca, Dongarra, Genton, K., Ltaief & Sun, *Accelerating Geostatistical Modeling and Prediction with Mixed-Precision Computations.* IEEE TDPS, 2021.

Three-precision dense Cholesky

- 5 blends of mixed precision approximation
- 4 synthetic field types
- 40K points each
- Some of the combinations are unacceptable



Abdulah, Cao, Pei, Bosilca, Dongarra, Genton, K., Ltaief & Sun, *Accelerating Geostatistical Modeling and Prediction with Mixed-Precision Computations.* IEEE TDPS, 2021.

Three-precision dense Cholesky

1M 2D Soil Moisture Dataset: The estimation of the model parameters are close to the pure DP MLE, except for the $1D$:$99H$ variant. We observed that this dataset has a medium correlated data with an average smooth field. This corroborates the analysis made with synthetic datasets that concludes on the effectiveness of the mixed-precision MLE for such data characteristics

Variants	Variance	Range	Smoothness	LLH	MSPE	Prediction Uncertainty	Iterations
100D	0.7223	0.0933	0.9983	-59740.65974	0.044926	4.734439e+03	180
30D:70S	0.7230	0.0935	0.9982	-59740.66579	0.044919	4.741754e+02	206
10D:90S	0.7314	0.0953	0.9969	-59741.37532	0.044933	4.736149e+03	207
10D:30S:60H	0.7239	0.0936	0.9982	-59740.65200	0.044927	4.734435e+03	244
10D:10S:80H	0.7328	0.0947	0.9983	-59741.03423	0.044926	4.733337e+03	207
5D:5S:90S	0.7106	0.0927	0.9967	-59741.35348	0.044935	4.736572e+03	204
1D:99H	0.9330	0.1286	0.9863	-59867.53239	0.044980	4.750953e+03	159

Abdulah, Cao, Pei, Bosilca, Dongarra, Genton, K., Ltaief & Sun, *Accelerating Geostatistical Modeling and Prediction with Mixed-Precision Computations.* IEEE TDPS, 2021.

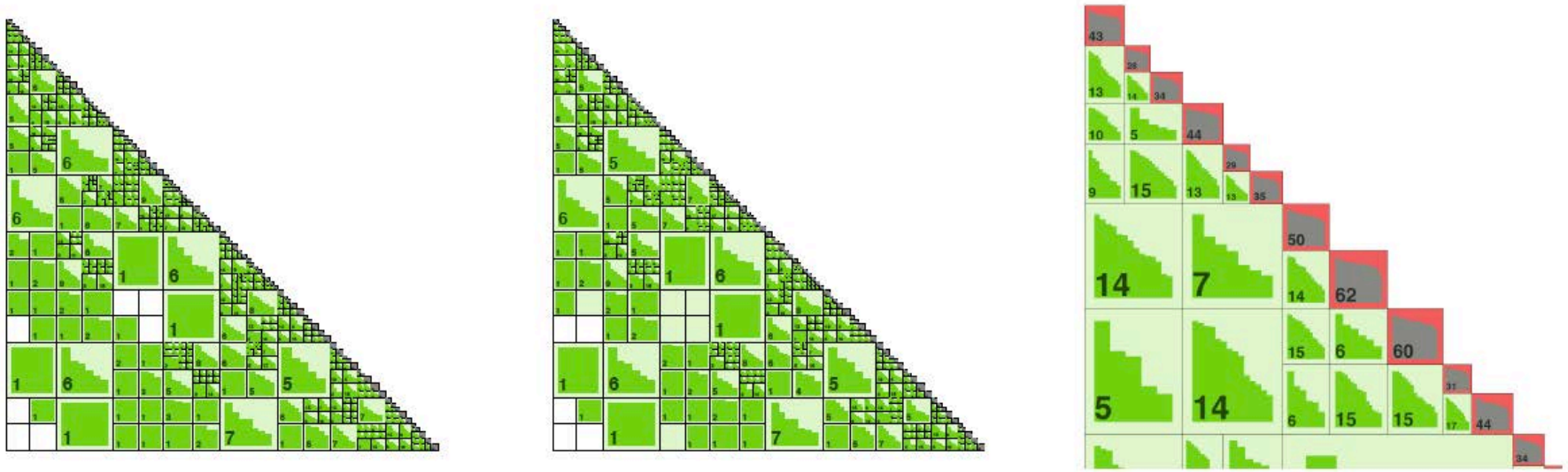
Three-precision dense Cholesky

116K 2D Wind Speed Dataset: This dataset comes from a highly smooth field. Thus, the estimation of the model parameters is impacted starting from the first mixed-precision 30D:70S variant and further deteriorates with lower precision configurations

Variants	Variance	Range	Smoothness	LLH	MSPE	Prediction Uncertainty	Iterations
100D	0.8407	0.0751	1.9905	241480.9994	1.752914E-02	2.2855E+00	666
30D:70S	0.9925	0.1794	1.9757	239908.5983	1.766191E-02	2.9221E+00	108
10D:90S	0.9924	0.1794	1.9757	239908.1004	1.766194E-02	2.9170E+00	91
10D:50S:40H	0.9911	0.1810	1.9754	239884.4173	1.766318E-02	2.9313E+00	89
10D:30S:60H	0.9761	0.1804	1.9576	232783.9932	1.765651E-02	5.2836E+00	94
10D:10S:80H	0.9774	0.1802	1.9588	233438.8691	1.765624E-02	5.1232E+00	107

Abdulah, Cao, Pei, Bosilca, Dongarra, Genton, K., Ltaief & Sun, *Accelerating Geostatistical Modeling and Prediction with Mixed-Precision Computations.* IEEE TDPS, 2021.

Hierarchical matrix approximations



$\mathcal{H}$-matrix approximations of the exponential covariance matrix (left), its hierarchical Cholesky factor $\tilde{L}$ (middle), and the zoomed upper-left corner of the matrix (right), $n = 4000$, $\ell = 0.09$, $\nu = 0.5$, $\sigma^2 = 1$.

c/o Alex Litvinenko (RTHW-Aachen, SIAM CSE 2021)

Recall covariance parameter estimation

For simplicity, we focus on zero-mean stationary Gaussian random fields. The log-likelihood for n locations:

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}\boldsymbol{Z}^\top\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{Z},$$

determinant inverse

where

$$\boldsymbol{Z} = \begin{pmatrix} Z(\boldsymbol{s}_1) \\ \vdots \\ Z(\boldsymbol{s}_n) \end{pmatrix}, \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} C(\boldsymbol{s}_1, \boldsymbol{s}_1; \boldsymbol{\theta}) & \dots & C(\boldsymbol{s}_1, \boldsymbol{s}_n; \boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ C(\boldsymbol{s}_n, \boldsymbol{s}_1; \boldsymbol{\theta}) & \dots & C(\boldsymbol{s}_n, \boldsymbol{s}_n; \boldsymbol{\theta}) \end{pmatrix}$$

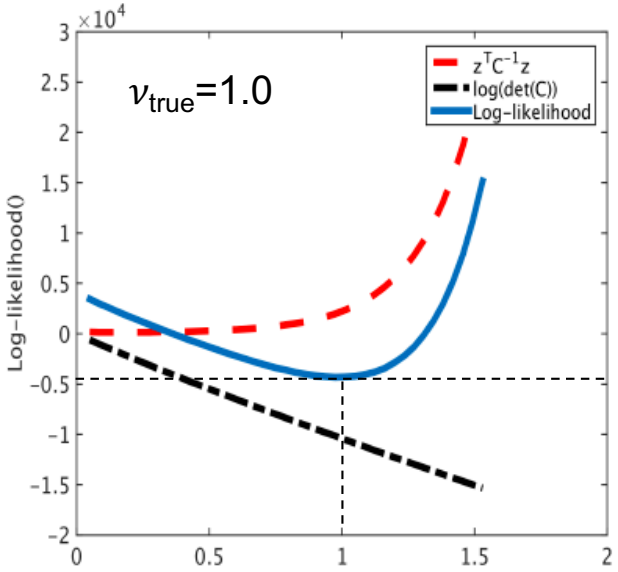
- Log determinant and linear solver require a Cholesky factorization of the given covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$
- Cholesky factorization requires $O(n^3)$ floating point operations and $O(n^2)$ memory

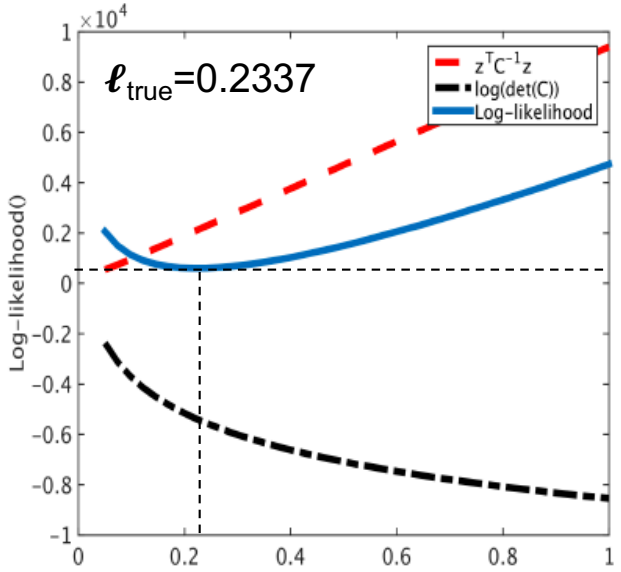
Balance of log-likelihood terms in learning

Missippi soil moisture example

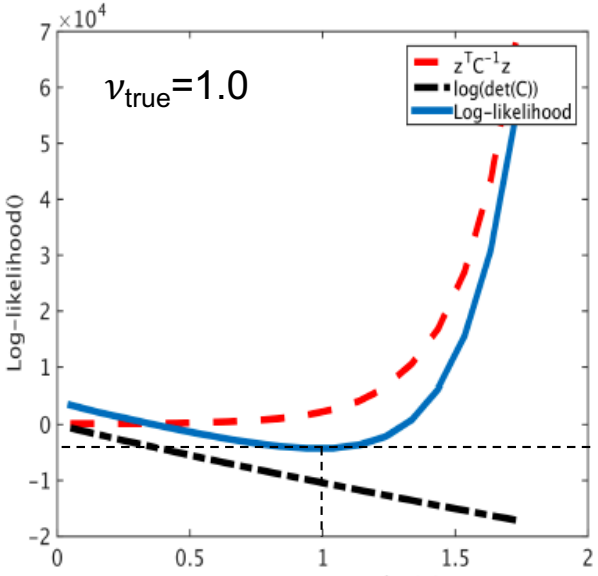
N=4225

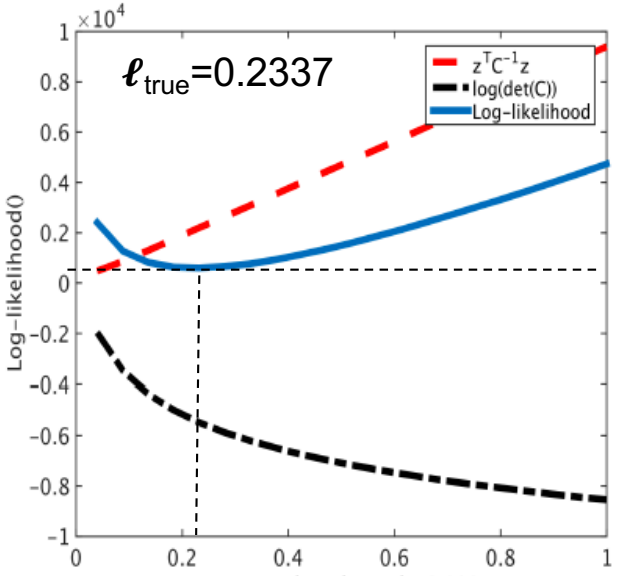
Max rank *k*=8





Max rank *k*=16





c/o Alex Litvinenko (RTHW-Aachen, SIAM CSE 2021)

Quantifying the $\mathscr{H}$-approximation

Theorem (1)

Let $\widetilde{\mathsf{C}}$ *be an* $\mathcal{H}$*-matrix approximation of matrix* $\mathsf{C} \in \mathbb{R}^{n \times n}$ *such that*

$$\rho(\widetilde{\mathsf{C}}^{-1}\mathsf{C} - \mathsf{I}) \leq \varepsilon < 1.$$

Then

$$|log \det \mathsf{C} - log \det \widetilde{\mathsf{C}}| \leq -nlog(1-\varepsilon), \qquad (3)$$

Proof: See [Ballani, Kressner 14] and [Ipsen 05].
Remark: factor n is pessimistic and is not really observed numerically.

c/o Alex Litvinenko (RTHW-Aachen, SIAM CSE 2021)

Quantifying the $\mathscr{H}$-approximation

Theorem (2)

Let $\widetilde{\mathsf{C}} \approx \mathsf{C} \in \mathbb{R}^{n\times n}$ *and* Z *be a vector,* $\|\mathsf{Z}\| \leq c_0$ *and* $\|\mathsf{C}^{-1}\| \leq c_1$.
Let $\rho(\widetilde{\mathsf{C}}^{-1}\mathsf{C} - \mathsf{I}) \leq \varepsilon < 1$. *Then it holds*

$$
\begin{aligned}
|\widetilde{\mathcal{L}}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| &= \frac{1}{2}(log|\mathsf{C}| - log|\widetilde{\mathsf{C}}|) + \frac{1}{2}|\mathsf{Z}^{\mathrm{T}}\left(\mathsf{C}^{-1} - \widetilde{\mathsf{C}}^{-1}\right)\mathsf{Z}| \\
&\leq -\frac{1}{2}\cdot nlog(1-\varepsilon) + \frac{1}{2}|\mathsf{Z}^{\mathrm{T}}\left(\mathsf{C}^{-1}\mathsf{C} - \widetilde{\mathsf{C}}^{-1}\mathsf{C}\right)\mathsf{C}^{-1}\mathsf{Z}| \\
&\leq -\frac{1}{2}\cdot nlog(1-\varepsilon) + \frac{1}{2}c_0^2\cdot c_1\cdot\varepsilon.
\end{aligned}
$$

c/o Alex Litvinenko (RTHW-Aachen, SIAM CSE 2021)

Accuracy results

ε accuracy in each sub-block, $n = 16641$, $\nu = 0.5$,
c.r.=compression ratio.

ε $\ell = 0.0334$	$\lvert\log\lvert C\rvert - \log\lvert\widetilde{C}\rvert\rvert$	$\lvert\frac{\log\lvert C\rvert - \log\lvert\widetilde{C}\rvert}{\log\lvert\widetilde{C}\rvert}\rvert$	$\lVert C - \widetilde{C}\rVert_F$	$\frac{\lVert C-\widetilde{C}\rVert_2}{\lVert C\rVert_2}$	$\lVert I - (\widetilde{L}\widetilde{L}^\top)^{-1} C\rVert_2$	c.r. in %
1e-1	3.2e-4	1.2e-4	7.0e-3	7.6e-3	2.9	9.16
1e-2	1.6e-6	6.0e-7	1.0e-3	6.7e-4	9.9e-2	9.4
1e-4	1.8e-9	7.0e-10	1.0e-5	7.3e-6	2.0e-3	10.2
1e-8	4.7e-13	1.8e-13	1.3e-9	6e-10	2.1e-7	12.7
$\ell = 0.2337$						
1e-4	9.8e-5	1.5e-5	8.1e-5	1.4e-5	2.5e-1	9.5
1e-8	1.45e-9	2.3e-10	1.1e-8	1.5e-9	4e-5	11.3

$\log|C| = 2.63$ for $\ell = 0.0334$ and $\log|C| = 6.36$ for $\ell = 0.2337$.

c/o Alex Litvinenko (RTHW-Aachen, SIAM CSE 2021)

Memory and execution time

$$\nu = 0.325,\ \ell = 0.64,\ \sigma^2 = 0.98$$

n	$\tilde{C}$			$\tilde{L}\tilde{L}^\top$		
	time sec.	size MB	kB/dof	time sec.	size MB	$\|I - (\tilde{L}\tilde{L}^\top)^{-1}\tilde{C}\|_2$
32.000	3.3	162	5.1	2.4	172.7	$2.4 \cdot 10^{-3}$
128.000	13.3	776	6.1	13.9	881.2	$1.1 \cdot 10^{-2}$
512.000	52.8	3420	6.7	77.6	4150	$3.5 \cdot 10^{-2}$
2.000.000	229	14790	7.4	473	18970	$1.4 \cdot 10^{-1}$

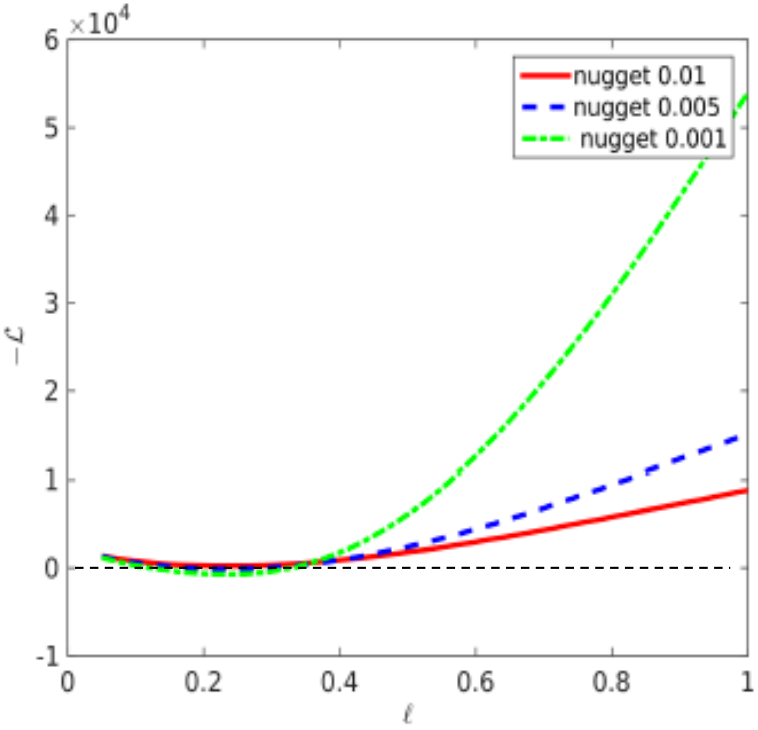
Dell Station, 20×2 cores, 128 GB RAM

c/o Alex Litvinenko (RTHW-Aachen, SIAM CSE 2021)

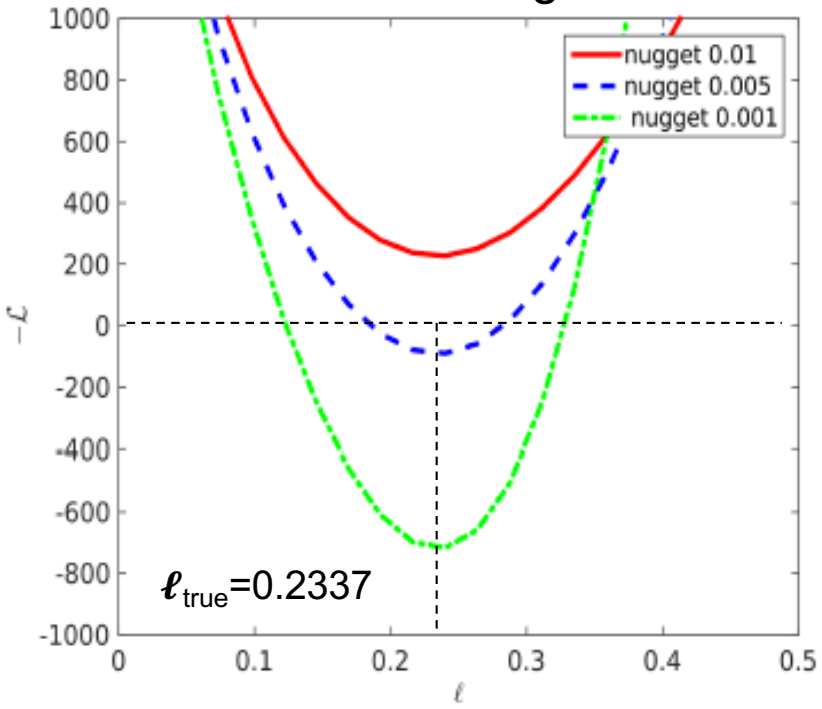
Importance of the nugget parameter

Missippi soil moisture example

N=2000, ν=0.5



zoom of left figure

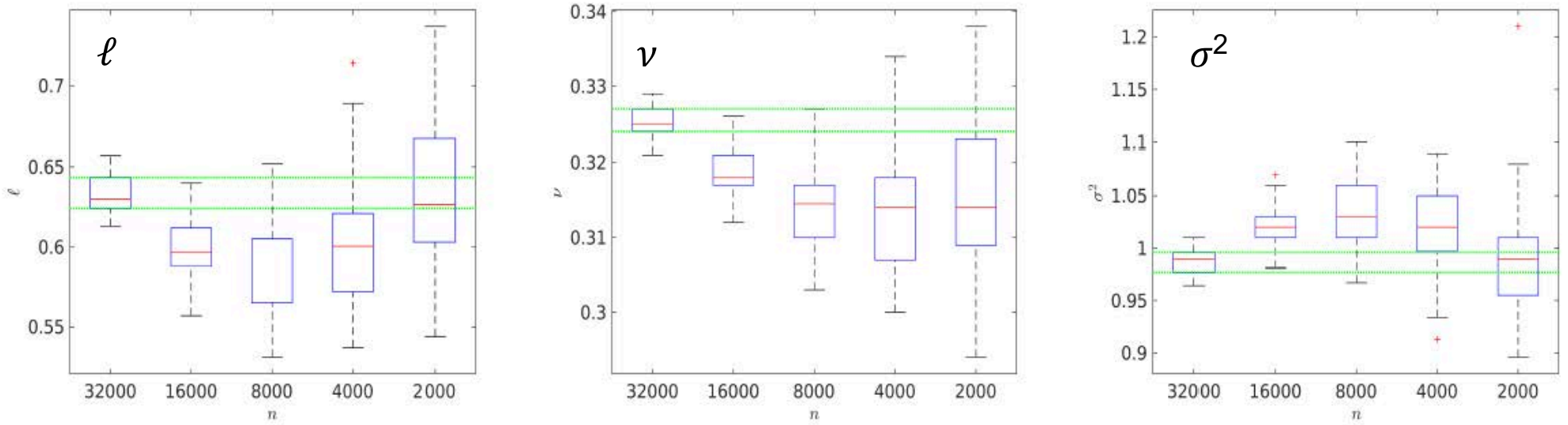


c/o Alex Litvinenko (RTHW-Aachen, SIAM CSE 2021)

Convergence with sample size

Missippi soil moisture example

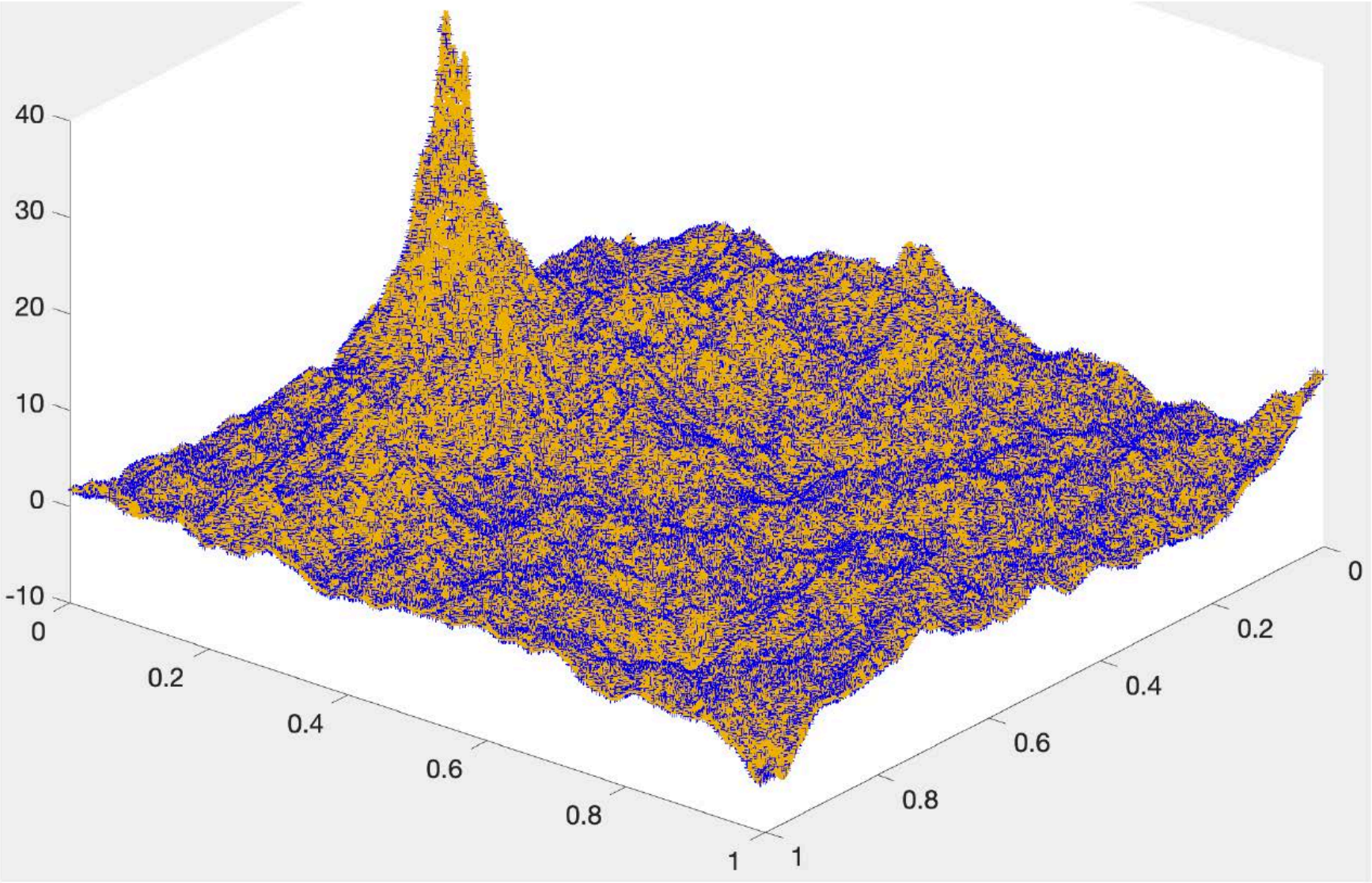
fixed k=11



N = 2K, 4K, 8K, 16K, 32K

c/o Alex Litvinenko (RTHW-Aachen, SIAM CSE 2021)

Prediction



The yellow points at 900,000 locations were used for training and the blue points were predicted at 100,000 locations.

Generalizations to ExaGeoStat underway

- **Multivariate**
- **Nonstationary**
- **Time-varying**

Multivariate Modeling (Exact & TLR)

- **Multivariate random variables, where a vector of variables is measured at each location**
- **Mathematically, this means that at location $s \in R^d$, $d \geq 1$, each variable is considered as one component of the p-dimensional vector $Z(s)$**
- **Parsimonious multivariate Matérn**

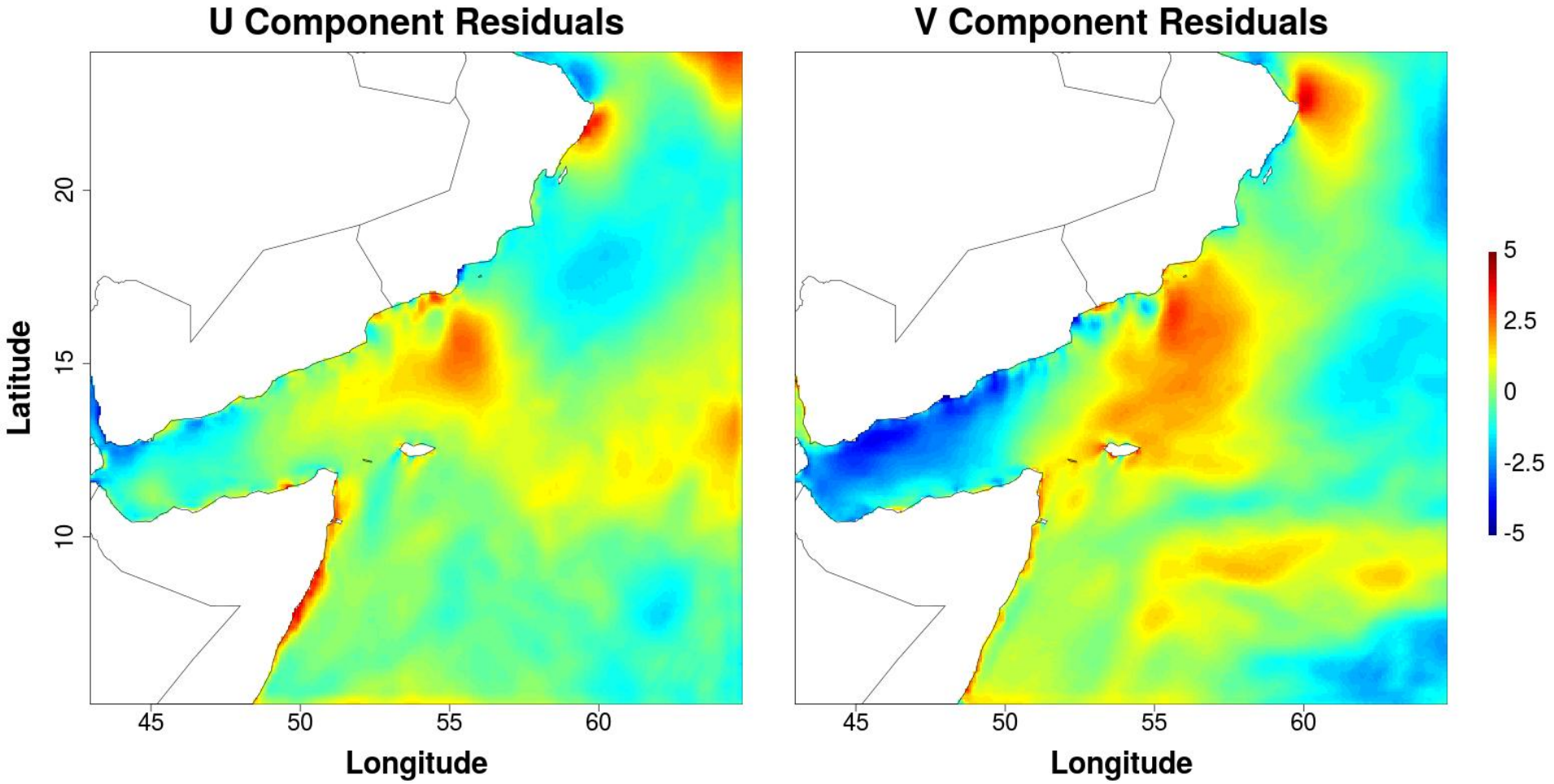
$$a_{ij} = a, \nu_{ij} = \frac{1}{2}\left(\upsilon_{ii} + \nu_{jj}\right), \sigma_{ij}^2 = \rho_{ij}\sigma_{ii}\sigma_{jj}, where$$

$$\rho_{ij} = \beta_{ij} \frac{\Gamma(\upsilon_{ii}+\frac{d}{2})^{1/2}}{\Gamma(\upsilon_{ii})^{1/2}} \frac{\Gamma(\upsilon_{jj}+\frac{d}{2})^{1/2}}{\Gamma(\upsilon_{jj})^{1/2}} \frac{\Gamma(1/2(\upsilon_{ii}+\upsilon_{jj}))^{1/2}}{\Gamma\left(1/2(\upsilon_{ii}+\upsilon_{jj})+\frac{d}{2}\right)}$$

Salvaña, Abdulah, Huang, Ltaief, Sun, Genton & K., *High Performance Multivariate Spatial Modeling for Geostatistical Data on Manycore Systems*. IEEE TDPS, 2021.

Multivariate Modeling

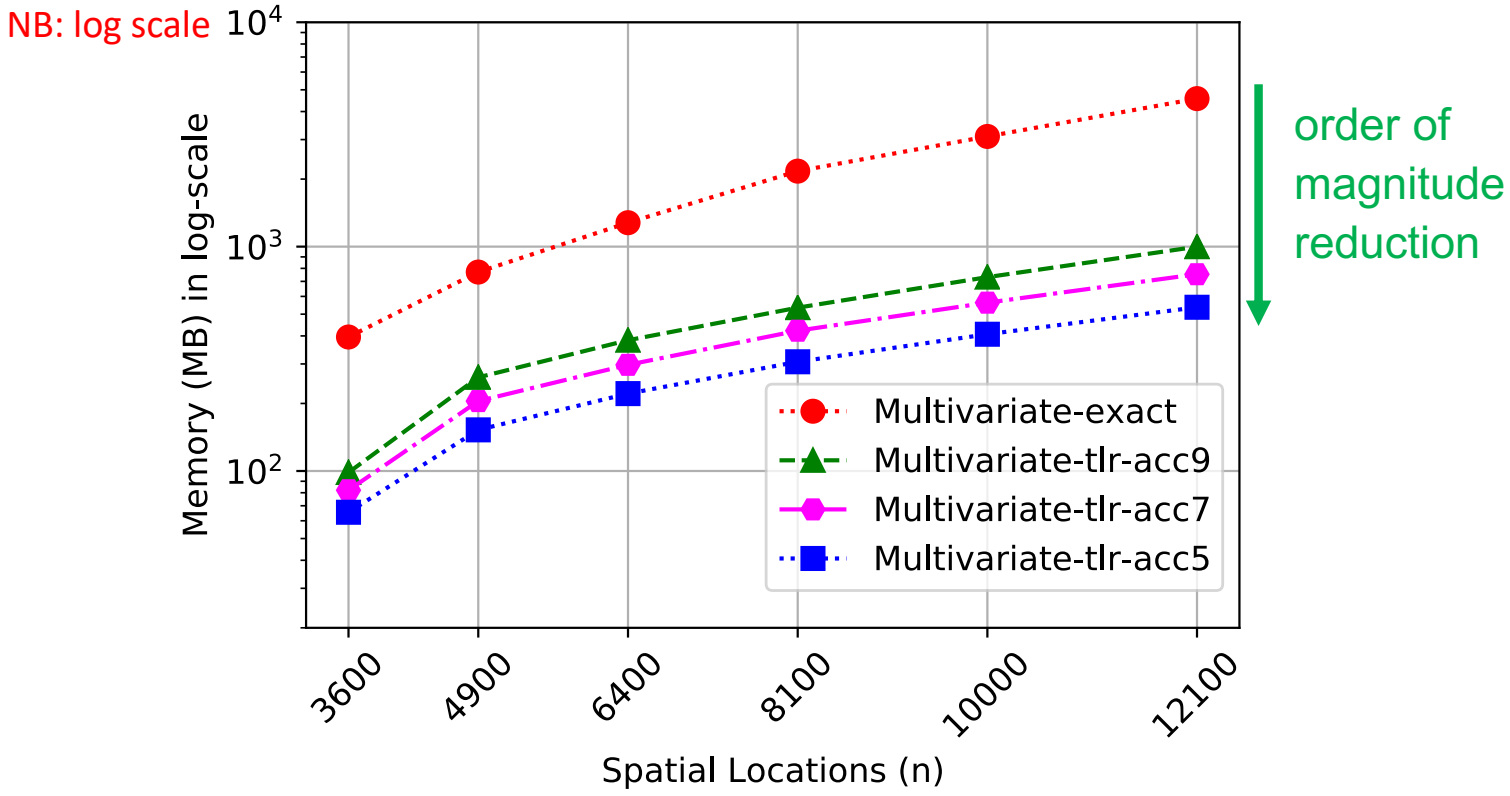
Spatial images of the U and V components (after mean removal) on 116, 100 locations over the Arabian Sea



Salvaña, Abdulah, Huang, Ltaief, Sun, Genton & K., *High Performance Multivariate Spatial Modeling for Geostatistical Data on Manycore Systems*. IEEE TDPS, 2021.

Multivariate Modeling (Exact & TLR)

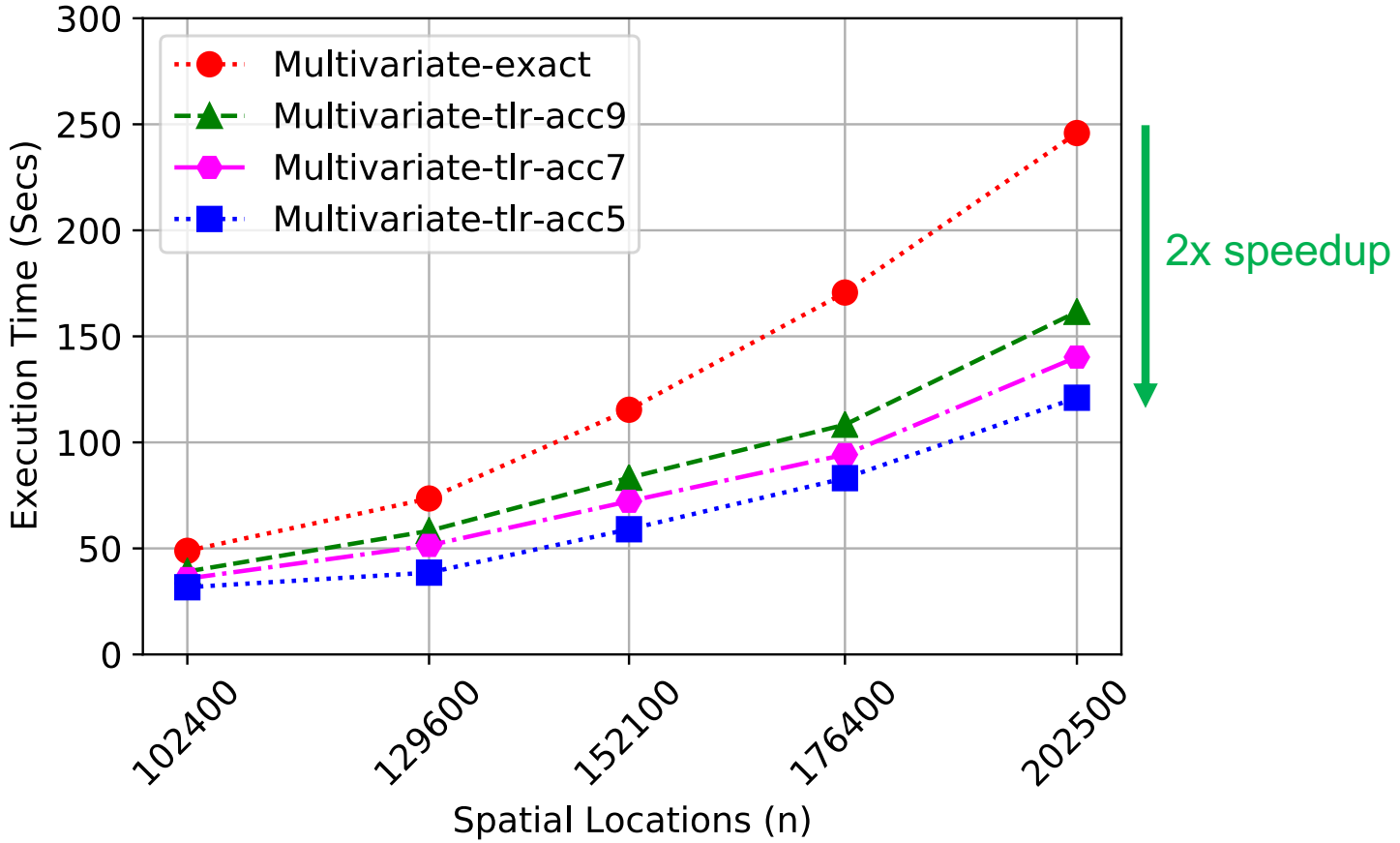
- **Memory footprint of exact and TLR-based MLE with varying number of spatial locations**
- **Bivariate Modeling two measurement vectors**



Salvaña, Abdulah, Huang, Ltaief, Sun, Genton & K., *High Performance Multivariate Spatial Modeling for Geostatistical Data on Manycore Systems.* IEEE TDPS, 2021.

Multivariate Modeling (Exact & TLR)

- **Execution time of exact and TLR-based MLE with varying number of spatial locations on 128 nodes of a Cray XC40 (2-socket 16-core Haswell)**
- **Bivariate Modeling two measurement vectors**



Salvaña, Abdulah, Huang, Ltaief, Sun, Genton & K., *High Performance Multivariate Spatial Modeling for Geostatistical Data on Manycore Systems.* IEEE TDPS, 2021.

Reference

Abdulah, Cao, Pei, Bosilca, Dongarra, Genton, K., Ltaief & Sun

IEEE TPDS (2021)

Accelerating Geostatistical Modeling and Prediction With Mixed-Precision Computations: A High-Productivity Approach with PaRSEC

Sameh Abdulah, Qinglei Cao, Yu Pei, George Bosilca, Jack Dongarra, Marc G. Genton, David E. Keyes, Hatem Ltaief, and Ying Sun

Abstract—Geostatistical modeling is an efficient technique for climate and environmental analysis and predicting desired quantities from geographically distributed data, based on statistical models and optimization of parameters. Spatial data presenting some known property of stationarity (or non-stationarity) requires a specific geostatistics kernel to handle. A primary computational kernel of stationary spatial statistics is the evaluation of the Gaussian maximum log-likelihood estimation (MLE) function, whose central data structure is a dense, symmetric, and positive definite covariance matrix of the dimension of the number of correlated observations. In the MLE approach considered herein, two essential operations are the application of its inverse and evaluation of its determinant. These can be rendered through the Cholesky decomposition and triangular solution. In this paper, we propose to reduce the precision of low correlated locations to single- or half- precision based on the distance. We migrate geostatistics to a three precision approximation by exploiting the mathematical structure of the dense covariance matrix. We illustrate application-expected accuracy worthy of double-precision from a majority half-precision computation, in a context where all single precision is by itself insufficient. We deploy PaRSEC runtime system with high productivity in mind to tackle the complexity and imbalance caused by the mixed three precisions. The PaRSEC delivers within a solo Cholesky factorization on-demand casting of precisions, while orchestrating tasks and data movement in a multi-GPU distributed-memory environment. We maintain the application-expected accuracy, while achieving against FP64-only operations up to $1.59X$ by mixing FP64/FP32 operations and $2.64X$ by mixing FP64/FP32/FP16 operations on 1536/4096 nodes of HAWK / Shaheen II and 128 nodes of Summit, respectively. This translates into up to 4.5 / 4.7 and 9.1 (mixed) PFlop/s sustained performance, respectively, demonstrating the effective synergism between PaRSEC dynamic runtime systems and challenging environmental HPC applications.

Index Terms—Climate/Weather Prediction, Dynamic Runtime Systems, Geospatial Statistics, High Performance Computing, Multiple Precisions, User-Productivity.

• *S. Abdulah, M. G. Genton, D. E. Keyes, H. Ltaief, and Y. Sun with Computer, Electrical, and Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. E-mail: {sameh.abdulah, marc.genton, david.keyes, hatem.ltaief, ying.sun}@kaust.edu.sa*
Q. Cao, Y. Pei, G. Bosilca, and J. Dongarra with the Innovative Computing Laboratory, University of Tennessee, Knoxville, TN 37996, US. E-mail: {qcao3, ypei2}@vols.utk.edu, {bosilca, dongarra}@icl.utk.edu

Manuscript received mm dd, yyyy; revised mm dd, yyyy.

1 Introduction

Geostatistics is a means of modeling and predicting desired quantities directly from data. It is based on statistical assumptions and optimization of parameters. It is complementary to first-principles modeling approaches rooted in conservation laws and typically expressed in PDEs. Alternative statistical approaches to predictions from first-principles methods, such as Monte Carlo sampling wrapped around simulations with a distribution of inputs, may be vastly more computationally expensive than sampling from an assumed parameterized distribution based on a much smaller number of simulations. Geostatistics is relied upon for economic and policy decisions for which billions of dollars or even lives are at stake, such as engineering safety margins into developments, mitigating hazardous air quality, locating fixed renewable energy resources, and planning agricultural yields or weather-dependent tourist revenues. Climate and weather predictions are among the principal workloads occupying supercomputers around the world, and even minor improvements for regular production applications pay large dividends. A wide variety of such predictive codes have opportunistically migrated or are migrating to mixed-precision environments; we describe a novel migration of one important class of such codes.

A main computational kernel of stationary spatial statistics considered herein is the evaluation of the Gaussian log-likelihood function, whose central data structure is a dense covariance matrix of the dimension of the number of (presumed) correlated observations, which is generally the product of the number of observation locations and the number of variables observed at each location. In the maximum log-likelihood estimation (MLE) technique considered herein, two essential operations on the covariance matrix are the application of its inverse and evaluation of its determinant. These operations can all be rendered through the classical Cholesky decomposition and triangular solution, occurring inside the optimization loop that fits statistical model parameters to the input data. The covariance matrix is dense, symmetric, and positive definite, and possesses a mathematical structure arising from its physical origin that motivates approximations of various kinds for high-

Reference

Salvaña, Abdulah, Huang, Ltaief, Sun, Genton & K.

IEEE TDPS (2021)

High Performance Multivariate Geospatial Statistics on Manycore Systems

Mary Lai O. Salvaña, Sameh Abdulah, Huang Huang, Hatem Ltaief, Ying Sun, Marc G. Genton, and David E. Keyes

Abstract—Modeling and inferring spatial relationships and predicting missing values of environmental data are some of the main tasks of geospatial statisticians. These routine tasks are accomplished using multivariate geospatial models and the cokriging technique. The latter requires the evaluation of the expensive Gaussian log-likelihood function, which has impeded the adoption of multivariate geospatial models for large multivariate spatial datasets. However, this large-scale cokriging challenge provides a fertile ground for supercomputing implementations for the geospatial statistics community as it is paramount to scale computational capability to match the growth in environmental data coming from the widespread use of different data collection technologies. In this paper, we develop and deploy large-scale multivariate spatial modeling and inference on parallel hardware architectures. To tackle the increasing complexity in matrix operations and the massive concurrency in parallel systems, we leverage low-rank matrix approximation techniques with task-based programming models and schedule the asynchronous computational tasks using a dynamic runtime system. The proposed framework provides both the dense and the approximated computations of the Gaussian log-likelihood function. It demonstrates accuracy robustness and performance scalability on a variety of computer systems. Using both synthetic and real datasets, the low-rank matrix approximation shows better performance compared to exact computation, while preserving the application requirements in both parameter estimation and prediction accuracy. We also propose a novel algorithm to assess the prediction accuracy after the online parameter estimation. The algorithm quantifies prediction performance and provides a benchmark for measuring the efficiency and accuracy of several approximation techniques in multivariate spatial modeling.

Index Terms—Gaussian log-likelihood, geospatial statistics, high-performance computing, large multivariate spatial data, low-rank approximation, multivariate modeling/prediction.

1 Introduction

The convergence of high-performance computing (HPC) and big data brings great promise in accelerating and improving large-scale applications [1], [2] on climate and weather modeling [3], astronomy [4], transportation [5], and bioinformatics [6]. Climate and weather modeling, in particular, is one of the first applications of HPC for big data [7]. The need to improve climate and weather models has pushed for advances in environmental data collection technologies such as spaceborne, airborne, and ground sensors [8]. The volume of data coming from these sources is huge and increasing. For instance, NASA's Earth Observing System Data and Information System (EOSDIS) is expected to archive more than 37 petabytes of data in 2020 [9]. By 2022, the yearly increase is projected at 47.7 petabytes.

Environmental data, such as climate and weather variables, are often recorded from different spatial locations, and thus indexed by $s \in \mathbb{R}^d, d \geq 1$, where s is the location of the measurement. Usually, there are multiple variables measured at each location, such as temperature, humidity, wind speed, and atmospheric pressure. These colocated variables may or may not depend on each other and on the variables at other locations.

• *The authors are with the Extreme Computing Research Center, Computer, Electrical, and Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. E-mail: {Marylai.Salvana, Sameh.Abdulah, Huang.Huang, Hatem.Ltaief, Ying.Sun, Marc.Genton, David.Keyes}@kaust.edu.sa.*

A major concern when dealing with environmental datasets is missing data on one or a few variables. For instance, when using environmental variables as inputs to climate and weather models, the gaps in areas with no measurements caused by poor atmospheric conditions or defective sensors, to name a few, need to be filled [10], [11]. Recently, with the existence of HPC capabilities, methods for big geospatial data analysis are sought in order to leverage these big geospatial data obtained from different sources such as satellite images, model simulations, sensors, and the Internet of Things for the purpose of missing data interpolation. These methods include the use of numerical models that solve a complex set of partial differential equations and generate large volumes of predictions on the quantities of interest, such as the concentrations of pollutants in the atmosphere [12], [13]. Other novel contemporary methods include applying machine learning and deep learning for analysis of big geospatial datasets [14], [15], [16], [17]. For instance, in [14], several machine learning methods to predict solar irradiation were reported. Although these methods show high predictions capabilities, they suffer the drawback of being unable to describe explicitly the spatial relationship among environmental variables [18].

In this work, we adopt the statistical approach using the Gaussian log-likelihood function to model the underlying multivariate geospatial data with the aid of a generated covariance matrix on large-scale systems. Multivariate geospatial statistics can interpolate environmental variables at unsampled locations by modeling the multivariate spatial dependencies. While every variable of interest can be

Very special thanks to…

Sameh Abdulah

Research Scientist
Extreme Computing Research Center

Alexander Litivinenko

formerly Research Scientist
Extreme Computing Research Center,
now at RTWH-Aachen


Covariances
In the billions require
ExaGeoStat


Thank you!
جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology
شكرا
david.keyes@kaust.edu.sa