2023 Barnett Lecture: Exascale Geostatistics for Environmental Data Science

Marc G. Genton

marc.genton@kaust.edu.sa

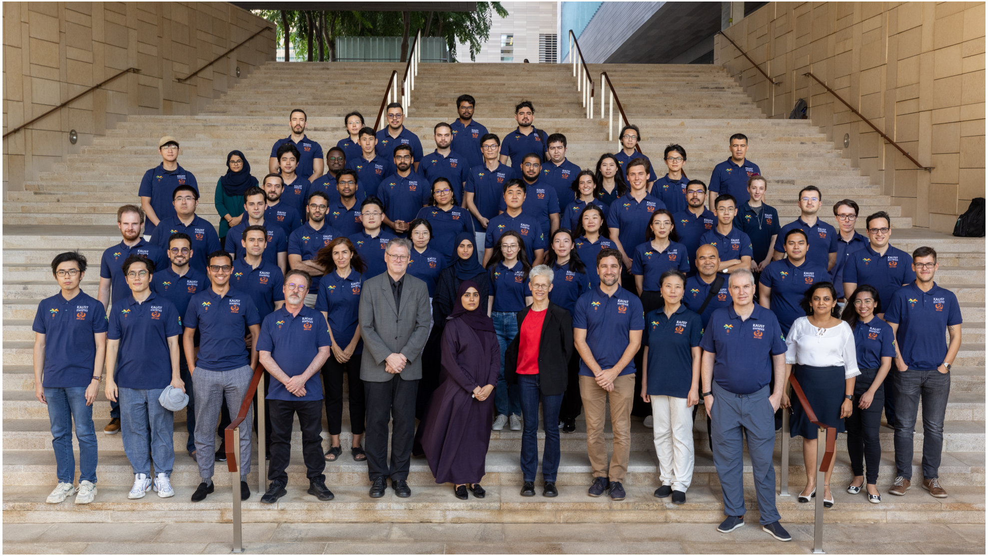
stsds.kaust.edu.sa
stat.kaust.edu.sa

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

STATISTICS
@KAUST

September 7, 2023

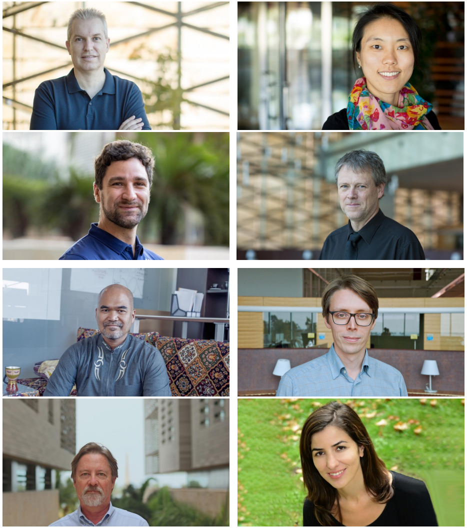
Statistics groups at KAUST in October 2022: stat.kaust.edu.sa



My research group in 2012

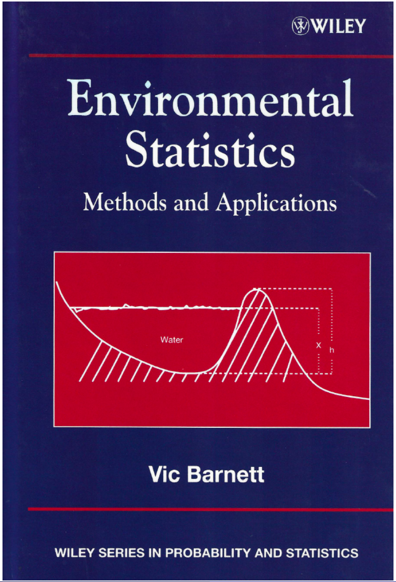
STATISTICS
@KAUST

Current Statistics core faculty



Dr Sameh Abdulah (ECRC)

Vic Barnett (1938-2014)



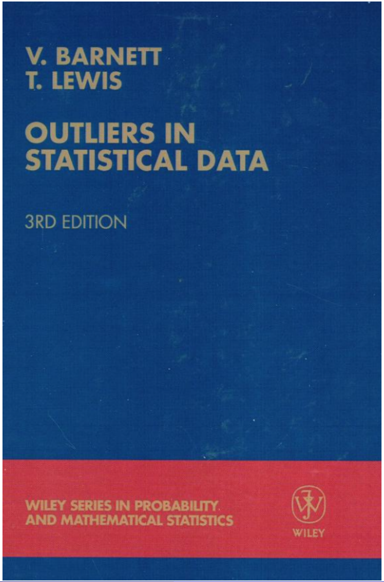
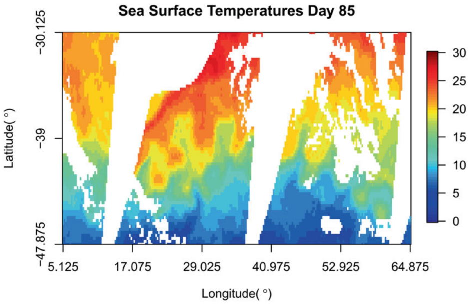
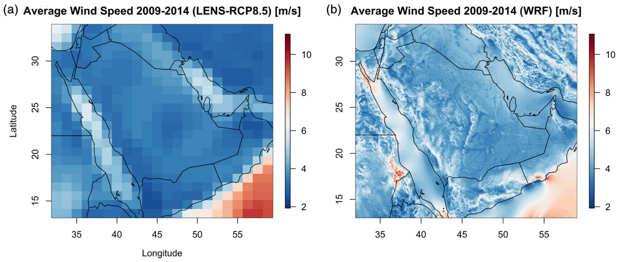


Table of contents

1. Some Fundamental Problems in Environmental Data Science
   - 1.1 Spatial Gaussian likelihood inference
   - 1.2 Spatial kriging
   - 1.3 Gaussian random field simulations
   - 1.4 Multivariate Gaussian probabilities
   - 1.5 Robust inference for spatial data
2. Large-Scale Environmental Data Science with *ExaGeoStat*
   - 2.1 What is HPC?
   - 2.2 Task-based parallelism and dynamic runtime systems
   - 2.3 Tile-based linear algebra
   - 2.4 Tile low-rank (TLR) linear algebra
   - 2.5 Multi- and mixed-precision computational statistics
   - 2.6 *ExaGeoStat* software: Exascale geostatistics
3. Competitions on Spatial Statistics for Large Datasets
   - 3.1 In 2021: Gaussian and non-Gaussian
   - 3.2 In 2022: Nonstationary, space-time, multivariate
   - 3.3 In 2023: Irregular locations, confidence/prediction intervals



Spatial data follow law of geography:
"nearby things tend to be more alike than those far apart"



1.1 Spatial Gaussian likelihood inference

- n irregularly-spaced observations from zero-mean Gaussian random field: $\boldsymbol{Z} = \{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}^\top$
- Matérn spatial covariance function:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})_{ij} = \mathrm{cov}\{Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)\} = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|}{\beta} \right)^{\nu} \mathcal{K}_\nu \left(\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|}{\beta} \right) + \tau^2 I\{i = j\}$$

  where $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν, $\Gamma(\cdot)$ is the Gamma function, and I is the indicator function
- The four components of the parameter vector $\boldsymbol{\theta}$: partial sill σ^2, range $\beta > 0$, smoothness $\nu > 0$, and nugget τ^2
- Spatial Gaussian log-likelihood:

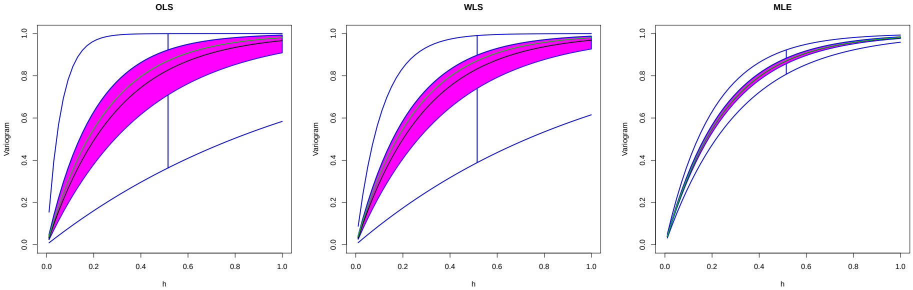
$$\boxed{\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}\boldsymbol{Z}^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{Z}}$$

- Log determinant and linear solver require a Cholesky factorization of the symmetric positive definite covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$
- Cholesky factorization requires $O(n^3)$ floating point operations and $O(n^2)$ memory
- Computations become challenging for large n

Likelihood vs least squares in spatial covariance estimation

- **Functional boxplots for:** functional data, functional simulations
- **Other functions**: variogram; covariogram; extremal coefficient; return level curve; log-periodogram; forecasting skill curve; etc.

- **Example**: exponential variogram $1 - \exp(-h/\theta)$ with $\theta = 0.25$
- Mean-zero GP generated at 400 random locations in unit square
- Estimate θ by OLS, WLS, MLE; 1000 replicates



- Kriging is spatial interpolation (Best Linear Unbiased Predictor, BLUP)
- Let

$$\begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} \sim \mathcal{N}_{n+m}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right)$$

then:

$$(\mathbf{Z}_2 | \mathbf{Z}_1 = \mathbf{z}_1) \sim \mathcal{N}_m\left(\boxed{\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1)}, \boxed{\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}} \right)$$

- Kriging with conditional mean
- Uncertainty quantification with conditional variance
- Computations become challenging for large n and/or m

Unconditional simulations:

- $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ where Y_i are iid from $\mathcal{N}(0,1)$
- $\boldsymbol{\Sigma}$ is an $n \times n$ covariance matrix with $(\boldsymbol{\Sigma})_{ij} = \text{cov}\{Z(\mathbf{s}_i), Z(\mathbf{s}_j)\}$
- $\boldsymbol{\Sigma}^{1/2}$ from spectral decomposition or Cholesky decomposition of $\boldsymbol{\Sigma}$
- Then: $\boxed{\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{Y} \text{ is } \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})}$

Conditional simulations: If

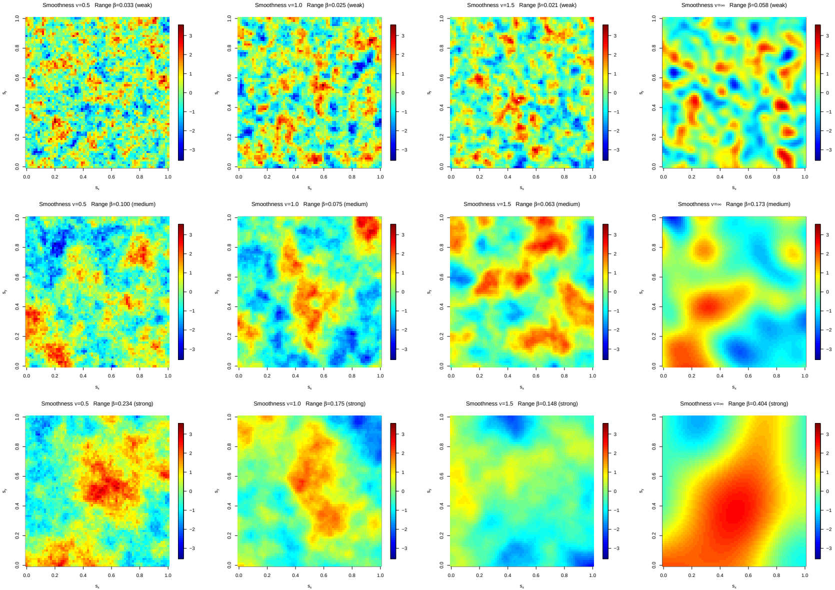
$$\begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} \sim \mathcal{N}_{n+m}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right)$$

then:

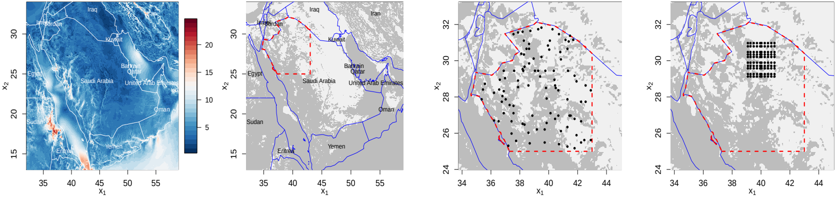
$$\boxed{(\mathbf{Z}_2 | \mathbf{Z}_1 = \mathbf{z}_1) \sim \mathcal{N}_m\left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \right)}$$

Computations become challenging for large n and/or m

Gaussian random field simulations with Matérn correlation function



- Probit Gaussian process models
- Application: windspeed exceeds a threshold for energy production
- Windspeed at 140 m on January 21, 2014; threshold of 4 m/s
- Region includes NEOM and Dumat Al Jandal



$$\Phi_n(\mathbf{a}, \mathbf{b}; \boldsymbol{\Sigma}) = \int_{\mathbf{a}}^{\mathbf{b}} \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}\right) d\mathbf{x}$$

- Also: Bayesian probit regression; unified skew-normal (SUN) distributions ($\propto \phi_n \Phi_n$); excursion/contour regions
- Computations become challenging for large n

- Vic Barnett and Toby Lewis book: **Outliers in Statistical Data**

- More **challenging** for spatial data (position of outliers matters)
- Spatial breakdown point
- Highly robust variogram estimator

- Maximum L_q-likelihood estimator (MLqE) for Gaussian random fields:

$$\boxed{\ell_q(\boldsymbol{\theta}) = \sum_{j=1}^{R} L_q \left[\frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}(\boldsymbol{\theta})|}} \exp\left(-\frac{1}{2} \boldsymbol{Z}_j^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \boldsymbol{Z}_j \right) \right]}$$

where

$$L_q(u) = \begin{cases} \log u, & \text{if} \quad q = 1 \\ \left(u^{1-q} - 1\right)/(1-q), & \text{if} \quad 0 < q < 1 \end{cases}$$

- Computations become challenging for large n and/or R, and many q's

2. Large-Scale Environmental Data Science with *ExaGeoStat*

When the size n of datasets becomes large:

- $O(n^3)$ floating point operations and $O(n^2)$ memory for **exact computations** of Cholesky factorization
  - High-Performance Computing (HPC) can help when n is large
  - *ExaGeoStat* software:
    https://github.com/ecrc/exageostat
    https://github.com/ecrc/exageostatr

  Note: $n = 1'000'000$ then $n^3 = 10^{18}$ = 1 billion billions

- *ExaGeoStat* for:
  1. Likelihood inference/learning for Matérn covariance function (among others)
  2. Spatial kriging (interpolation)
  3. Random field simulations
  4. Multivariate Gaussian probabilities
  5. Robust spatial inference

- Various **approximation methods** have been proposed in literature to ease computation & memory burden

- **2021/2022/2023 KAUST Competitions on Spatial Statistics for Large Datasets** investigate the performance of different approximation methods with large synthetic data generated by *ExaGeoStat*

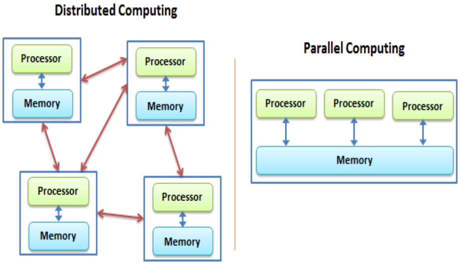
- **High-Performance Computing (HPC)** is the use of advanced computing techniques and technologies to solve complex problems that require significant computational power
- **HPC systems** are designed to deliver high processing speeds, large-scale storage capacities, and high-speed data transfer capabilities
- They often have **multiple processors** (each with multi-cores) and may have **accelerators** (such as Graphics Processing Units (GPUs))
- HPC term applies to systems that **function above a TFLOPS** or $O(10^{12})$ floating-point operations per second (Flops/s)

Name	Unit	Value
kiloFLOPS	kFLOPS	10^3
megaFLOPS	MFLOPS	10^6
gigaFLOPS	GFLOPS	10^9
teraFLOPS	TFLOPS	10^{12}
petaFLOPS	PFLOPS	10^{15}
exaFLOPS	EFLOPS	10^{18}
zettaFLOPS	ZFLOPS	10^{21}
yottaFLOPS	YFLOPS	10^{24}

Modern hardware architectures

- **Shared-memory systems:** a type of computer architecture where multiple processors or cores access a common physical memory, e.g., x86-64 (Intel and AMD processors), IBM POWER, Graphics Processing Units (GPUs)
- **Distributed-memory systems:** a type of computer architecture in which multiple processors or nodes have their own local memory and communicate with each other through message passing, e.g., clusters and supercomputers

- With sufficiently **fast network** we can in principle extend this approach to millions of CPU-cores and beyond
- **Benefits:** Scalability, reliability, and performance
- **Challenges:** Complex architectural, construction, and debugging processes

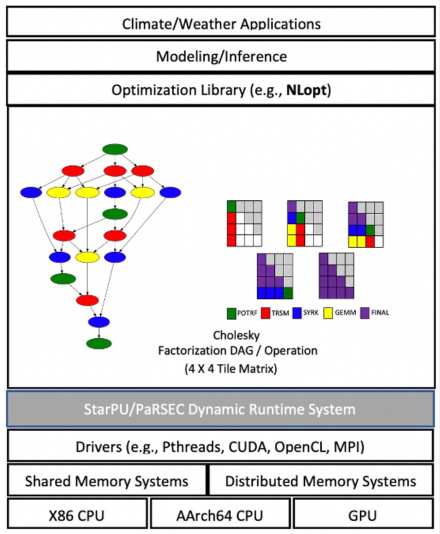


TOP 500 Supercomputers June 2023 (https://www.top500.org)

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016
4	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos EuroHPC/CINECA Italy	1,824,768	238.70	304.47	7,404
5	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096

You can think of Fugaku (cost $1 billion!) as putting 20 million smartphones in a single room, or equivalently 300,000 standard servers in a single room

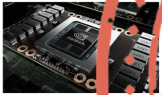
Example: *ExaGeoStat* software for exascale geostatistics



X86 CPU

AArch64

Fujitsu A64FX



NVIDIA V100

AMD MI250X

#1 Frontier

#2 Fugaku

#5 Summit

#32 HAWK

#113 Shaheen-II

Science for the Future

KAUST RESEARCH OPEN WEEK

SUSTAINABILITY

- **Task-based parallelization** is a parallel computing technique in which a large task or problem is divided into smaller subtasks that can be executed concurrently on multiple processors or cores

- **Parallel coding** on different hardware architectures requires different skills and coding tools:
  - Shared-memory systems (e.g., OpenMP)
  - GPUs (e.g., OpenCL, CUDA)
  - Distributed systems (e.g., Message Passing Interface (MPI))

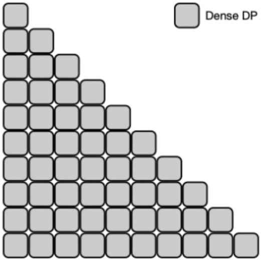
- **Dynamic runtime systems** are software frameworks or environments that **provide a layer of abstraction above the hardware and operating system**, aiming to simplify the management and coordination of parallel and concurrent computations, e.g., StarPU (INRIA Bordeaux, France) and PaRSEC (UTK, USA)

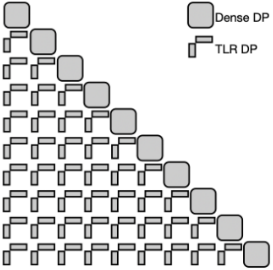
- Dynamic runtime systems **facilitate** the creation, scheduling, and execution of tasks on available processing units (such as CPU cores or GPUs).

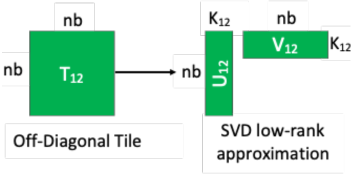
2.3 Tile-based linear algebra

- **Tile-based linear algebra** refers to a technique used to optimize the execution of linear algebra operations on parallel architectures
- It involves breaking down large matrices into smaller submatrices, called **tiles**, to exploit the memory hierarchy and parallelism of modern processors
- It aims to **enhance cache utilization and minimize data movement** between different levels of memory, such as cache and main memory
- The **size of the tiles** is chosen based on factors such as cache size, memory bandwidth, and computational requirements
- Tile-based linear algebra algorithms can be parallelized to take advantage of **multi-core processors, GPUs, or distributed computing environments**
- Existing **tile-based algorithms rely on task-based parallelism and runtime systems** (e.g., StarPU and PaRSEC) to optimize the performance of existing linear algebra solvers over the modern HPC hardware

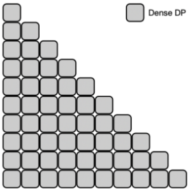
- Tile-based low-rank approximation refers to an approach for approximating matrices by decomposing them into **low-rank structures** using a tile-based framework
- This technique aims to **reduce the computational complexity and storage requirements** associated with working with large-scale data by representing the original matrix as a combination of low-rank factor







- **Double-precision computation** (64-bit) has been widely used as the primary representation for floating-point numbers in computations
- There has been a recent surge in studies driven by the demand from applications to use reduced representations, such as **single** (32-bit) or **half** (16-bit), in order to **accelerate computations while maintaining an acceptable level of accuracy**
- The concepts of multi- and mixed-precision computation have emerged:
  - **Multi-precision** computation uses a combination of different precisions in different parts of an algorithm
  - **Mixed-precision** computation uses varying precisions within the same algorithm's operation
- We introduced the new concept of **mixed-precision tile-based linear solvers for spatial statistics:**

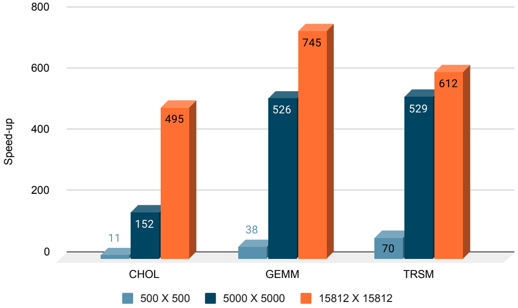


- **Benefits:** Faster computations; memory savings, energy efficiency; scalability

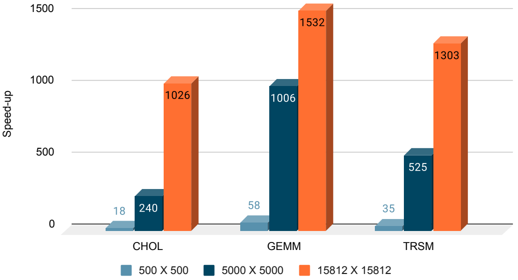
MMPR: R package for Multi- and Mixed-Precision computational statistics

- MMPR is an advanced package designed to provide R users with a customized data structure
- MMPR is tailored for researchers and data scientists working with multi- or mixed-precision arithmetic
- The package provides support for **three distinct precisions**: half, single, and double. It also offers a **mixed-precision** data structure organized in a **tile-based format**
- MMPR achieves fast execution for lower precisions by **leveraging highly optimized libraries** such as MKL and OpenBLAS, whereas R uses Rblas
- Download: https://github.com/stsds/MMPR Soon on CRAN!

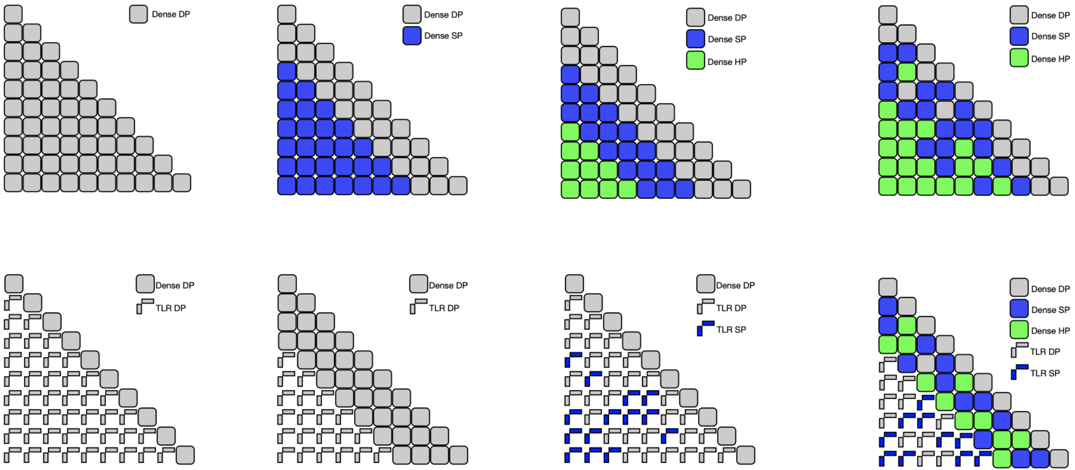
MMPR-DP vs R-DP



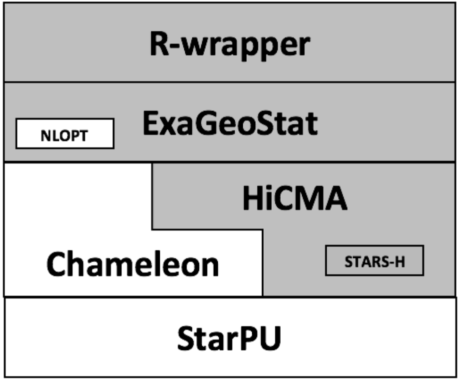
MMPR-SP vs R-DP



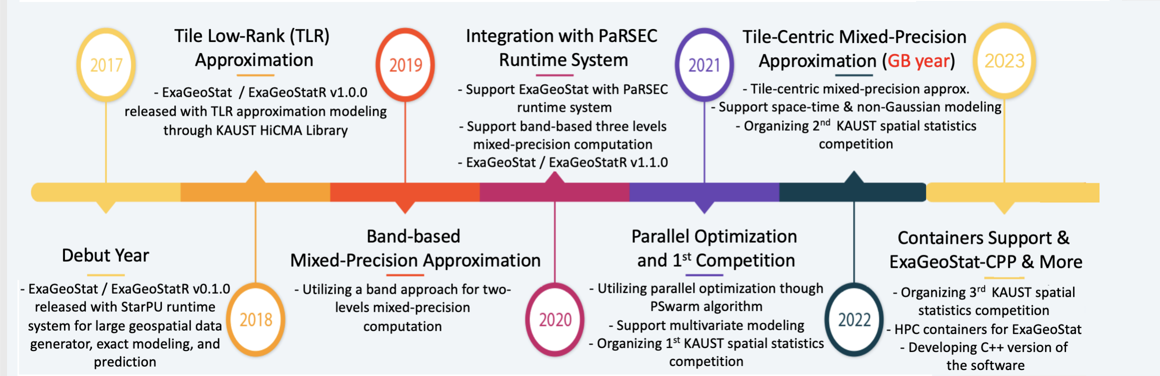
ExaGeoStat covariance matrix representation



ExaGeoStatR is a package for large-scale Geostatistics in R that supports parallel computation of the Gaussian maximum likelihood function, kriging and simulations on shared memory, GPU, and distributed memory systems

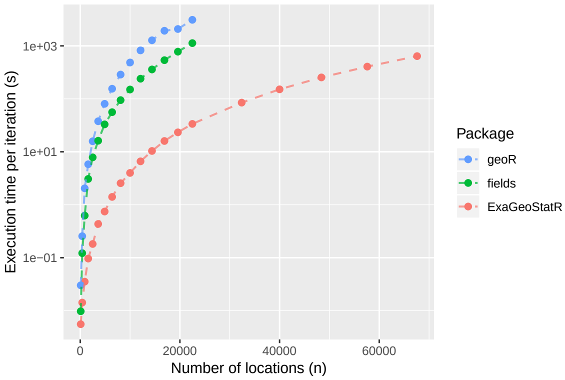


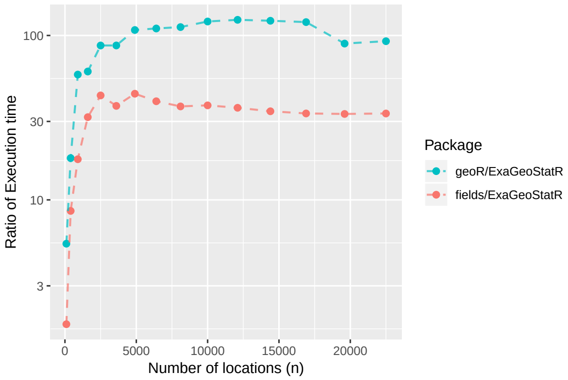
ExaGeoStat development timeline



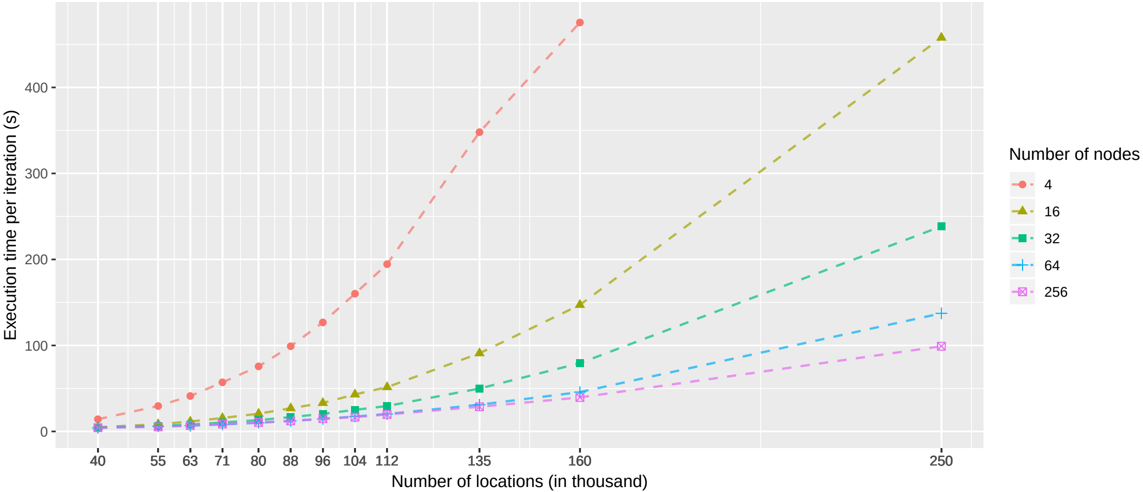
geoR - fields - *ExaGeoStatR* comparison (speedup)

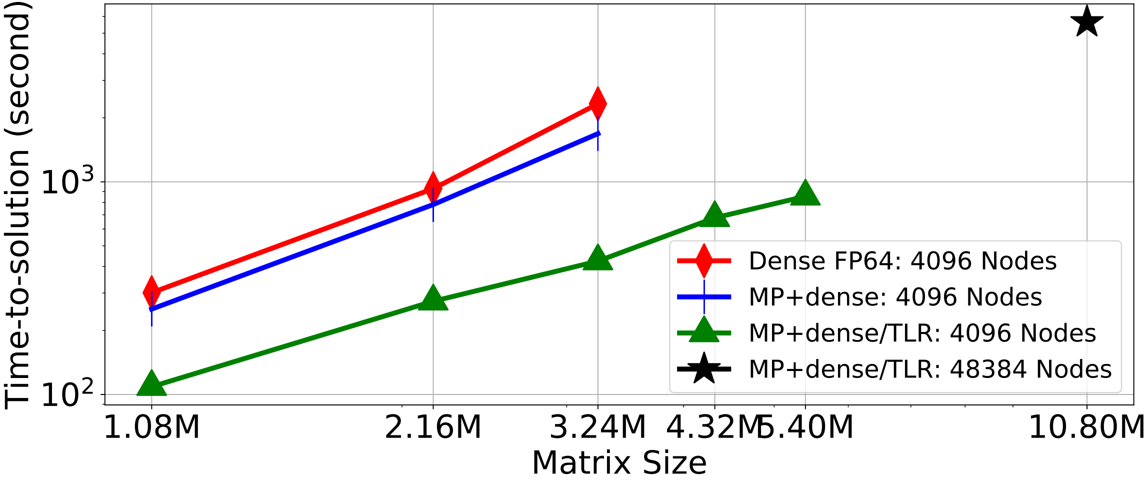
Average over 100 samples





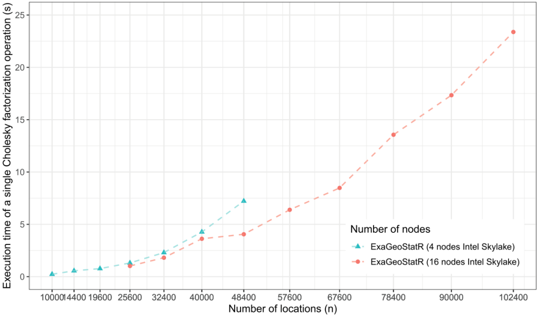
ExaGeoStatR performance on distributed-memory system (Shaheen-II)



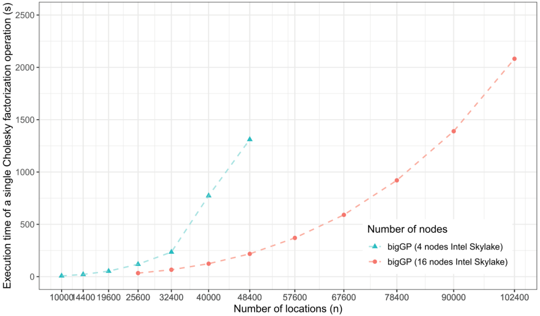


Performance of space-time Matérn of strong correlation on 4096 & 48384 Fugaku nodes

ExaGeoStatR comparison with bigGP



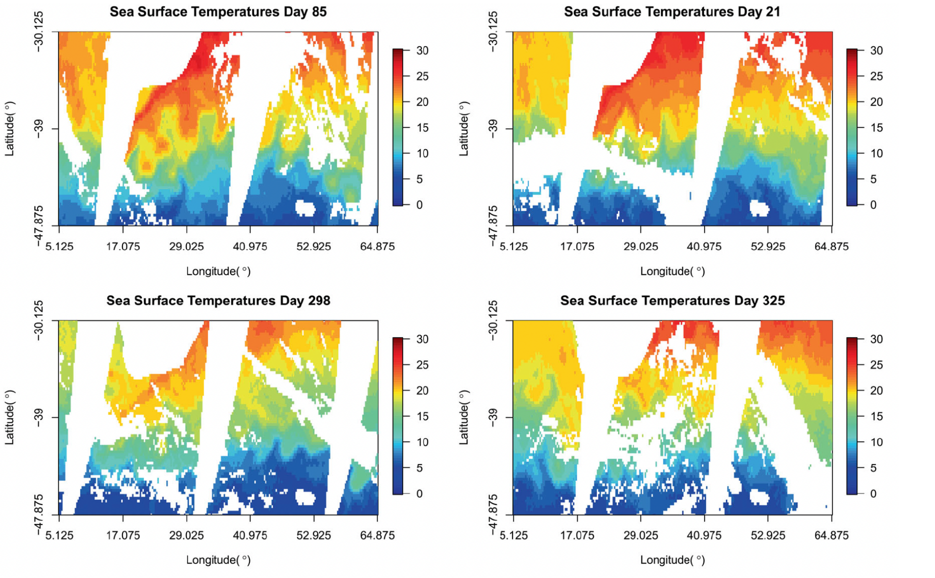
(a) *ExaGeoStatR* on Intel Skylake



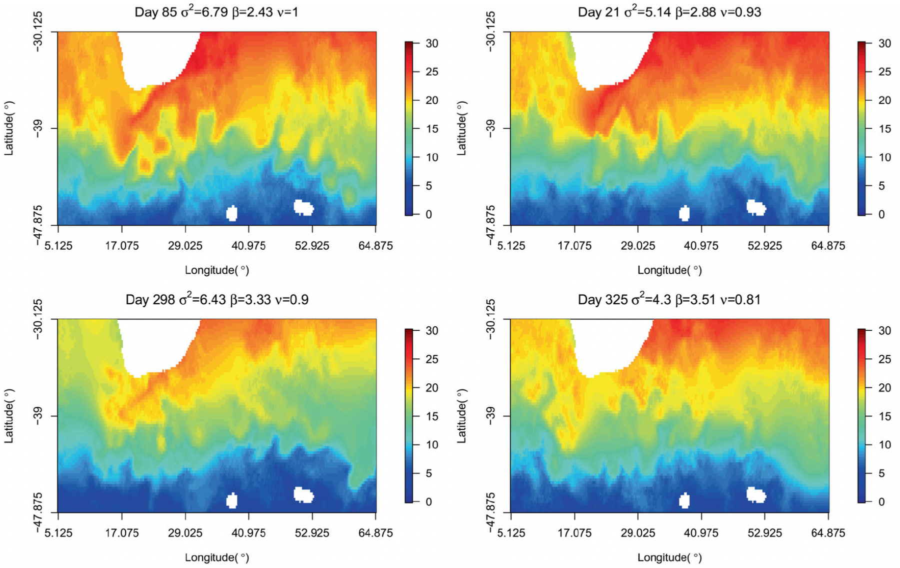
(b) *bigGP* on Intel Skylake

Comparison of *ExaGeoStatR* with *bigGP* from a performance perspective for Cholesky factorization on a distributed system (Ibex HPC cluster from KAUST using up to 16 40-core Intel Skylake nodes)

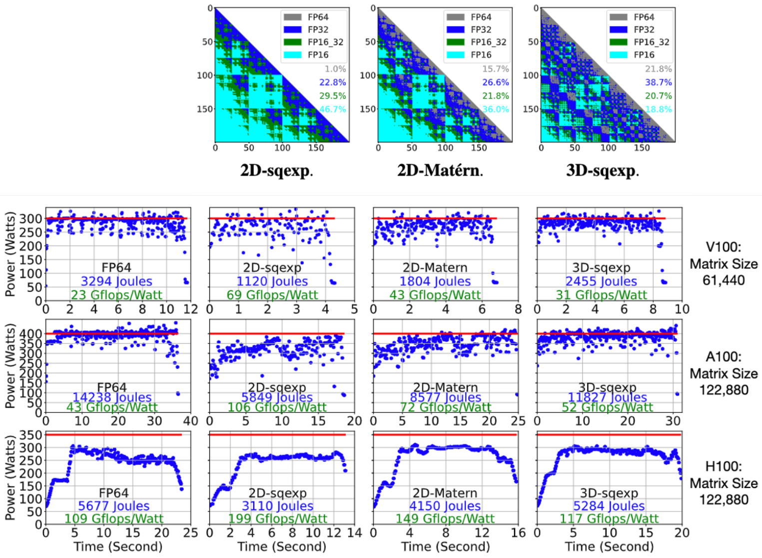
ExaGeoStatR for SST data kriging



ExaGeoStatR for SST data kriging



Adaptive mixed-precision maps & reduced power consumption on GPUs for Cholesky



3. Competitions on Spatial Statistics for Large Datasets

- **Goal:** investigate the **performance of different approximation methods** with large synthetic datasets generated by *ExaGeoStat*
- Through the competition, we can better understand when each approximation method is adequate
- The full datasets with one million locations are publicly available at:
  2021: https://doi.org/10.25781/KAUST-8VP2V
  2022: http://dx.doi.org/10.25781/KAUST-4ADYZ
  2023: ...
  which act as **benchmarking data** for future research
- The exact MLEs and lowest RMSEs achieved by researchers worldwide are released so that other/new methods can be easily compared

- Launched November 23, 2020; Ended February 1, 2021
- 29 research teams worldwide registered and 21 teams successfully submitted results
- Competition consists of four parts:

	Task	Data model	Data size
1a	GP estimation	GP	90,000
1b	prediction	GP	predict 10,000 conditional on 90,000
2a	prediction	Tukey g-and-h	predict 10,000 conditional on 90,000
2b	prediction	GP & Tukey g-and-h	predict 100,000 conditional on 900,000

- **Metric for GP estimation:**
  Mean Loss of Efficiency (MLOE) and
  Mean Misspecification of the Mean Square Error (MMOM)
- **Metric for prediction:** RMSE

In 2021: Gaussian estimation/prediction results

Sub-competition	Submission	Score	Rank
1a	ExaGeoStat(estimated-model)	154	0
	SpatStat-Fans	156	1
	GpGp	186	2
	RESSTE(CL/krig)	229	3
1b	ExaGeoStat(true-model)	72	0
	RESSTE(CL/krig)	78	1
	ExaGeoStat(estimated-model)	79	1.5
	HCHISS	93	2
	Chile-Team	113	3

3.2 In 2022: Nonstationary, space-time, multivariate

- Launched March 1, 2022; Ended May 1, 2022
- 20 research teams worldwide registered
- Hosted the competition on the **Kaggle** machine learning and data science platform

Sub-comp	Setting	True Data Model	# of Datasets	Training Data Size	Testing Data Size
1a	Univariate **Nonstationary** Spatial	GP with Nonstationary Mean or Cov	2	90K	10K
1b	Univariate **Nonstationary** Spatial	GP with Nonstationary Mean or Cov	2	900K	100K
2a	Univariate Stat. **ST**	GP with Non-Separable Cov	9	90K	10K
2b	Univariate Stat. **ST**	GP with Non-Separable Cov	9	900K	100K
3a	**Bivariate** Stationary Spatial	GP with Parsimonious/Flexible Matérn Cross-Cov	3	45K	5K
3b	**Bivariate** Stationary Spatial	GP with Parsimonious/Flexible Matérn Cross-Cov	3	450K	50K

3.3 In 2023: Irregular locations, confidence/prediction intervals

- Launched February 1, 2023; Ended May 1, 2023
- 11 research teams worldwide registered
- Five different **designs** considered for the locations of the observations:
  1. Chessboard; 2. Left-bottom; 3. Satellite; 4. Clusters; 5. Regular

Sub-comp	Model	Target	# designs	Training	Testing
1a	Gaussian Matérn	Estimation (95% conf interval)	5	90K	–
1b	Gaussian Matérn	Estimation (95% conf interval)	5	900K	–
2a	Gaussian Matérn	Prediction (95% pred interval)	5	90K	10K
2b	Gaussian Matérn	Prediction (95% pred interval)	5	900K	100K

List of publications related to *ExaGeoStat*

- Abdulah, S., Ltaief, H., Sun, Y., Genton, M. G., & Keyes., D. E. (2018)
  ExaGeoStat: A high performance unified software for geostatistics on manycore systems
  IEEE Transactions on Parallel and Distributed Systems, **29**, 2771-2784.

- Abdulah, S., Ltaief, H., Sun, Y., Genton, M. G., & Keyes., D. E. (2018)
  Parallel approximation of the maximum likelihood estimation for the prediction of large-scale geostatistics simulations
  IEEE International Conference on Cluster Computing, 98-108.

- Abdulah, S., Ltaief, H., Sun, Y., Genton, M. G., & Keyes, D. E. (2019)
  Geostatistical modeling and prediction using mixed-precision tile Cholesky factorization
  IEEE 26th International Conference on High-Performance Computing, Data, Analytics, and Data Science, 152-162.

- Salvana, M. L., Abdulah, S., Huang, H., Ltaief, H., Sun, Y., Genton, M. G., & Keyes, D. E. (2021)
  High performance multivariate geospatial statistics on manycore systems
  IEEE Transactions on Parallel and Distributed Systems, **32**, 2719-2733.

- Huang, H., Abdulah, S., Sun, Y., Ltaief, H., Keyes, D. E., & Genton, M. G. (2021)
  Competition on spatial statistics for large datasets (with discussion)
  Journal of Agricultural, Biological, and Environmental Statistics, **26**, 580-595.

List of Publications related to *ExaGeoStat*

- Hong, Y., Abdulah, S., Genton, M. G., and Sun, Y. (2021)
  Efficiency assessment of approximated spatial predictions for large datasets
  Spatial Statistics, **43**:100517.

- Salvana, M. L., Abdulah, S., Ltaief, H., Sun, Y., Genton, M. G., and Keyes, D. E. (2022)
  Parallel space-time likelihood optimization for air pollution prediction on large-scale systems
  Platform for Advanced Scientific Computing Conference (PASC '22), Basel, Switzerland, Article No. 17, 1-11.

- Mondal, S., Abdulah, S., Ltaief, H., Sun, Y., Genton, M. G., and Keyes, D. E. (2022)
  Parallel approximations of the Tukey g-and-h likelihoods and predictions for non-Gaussian geostatistics
  International Parallel and Distributed Processing Symposium, 379-389.

- Ltaief, H., Genton, M. G., Gratadour, D., Keyes, D. E., and Ravasi, M. (2022)
  Responsibly reckless matrix algorithms for HPC scientific applications
  Computing in Science & Engineering, **24**, 12-22.

- Cao, Q., Abdulah, S., Alomairy, R., Pei, Y., Nag, P., Bosilca, G., Dongarra, J., Genton, M. G., Keyes, D. E., Ltaief, H., and Sun, Y. (2022)
  Reshaping geostatistical modeling and prediction for extreme-scale environmental applications *International Conference for High Performance Computing, Networking, Storage and Analysis (SC22)*, Dallas, TX, US, 13-24.
  Finalist for Gordon Bell prize.

List of Publications related to *ExaGeoStat*

- Abdulah, S., Cao, Q., Pei, Y., Bosilca, G., Dongarra, J., Genton, M. G., Keyes, D. E., Ltaief, H., and Sun, Y., (2022)
  Accelerating geostatistical modeling and prediction with mixed-precision computations: A high-productivity approach with PaRSEC
  IEEE Transactions on Parallel and Distributed Systems, **33**, 964-976.

- Abdulah, S., Alamri, F., Nag, P., Sun, Y., Ltaief, H., Keyes, D. E., and Genton, M. G. (2022)
  The second competition on spatial statistics for large datasets *Journal of Data Science*, **20**, 439-460.

- Mondal, S., Abdulah, S., Ltaief, H., Sun, Y., Genton, M. G., and Keyes, D. E. (2023)
  Tile low-rank approximations of non-Gaussian spatial and space-time Tukey g-and-h random field likelihoods and predictions on large-scale systems
  Journal of Parallel and Distributed Computing, **180**, 104715.

- Cao, Q., Abdulah, S., Ltaief, H., Genton, M. G., Keyes, D. E., and Bosilca, G. (2023)
  Reducing data motion and energy consumption of geospatial modeling applications using automated precision conversion
  IEEE International Conference on Cluster Computing, to appear.

- Abdulah, S., Li, Y., Cao, J., Ltaief, H., Keyes, D. E., Genton, M. G., and Sun, Y. (2023)
  Large-scale environmental data science with ExaGeoStatR
  Environmetrics, **34**:e2770.

Summary

- 1.1 Spatial Gaussian likelihood inference

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

- 1.2 Spatial kriging

$$\boxed{\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1)}, \boxed{\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}}$$

- 1.3 Gaussian random field simulations

$$\boxed{\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{Y} \text{ is } \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})}, \boxed{(\mathbf{Z}_2|\mathbf{Z}_1 = \mathbf{z}_1) \sim \mathcal{N}_m\left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right)}$$

- 1.4 Multivariate Gaussian probabilities

$$\boxed{\Phi_n(\mathbf{a}, \mathbf{b}; \boldsymbol{\Sigma}) = \int_{\mathbf{a}}^{\mathbf{b}} \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}\right)d\mathbf{x}}$$

- 1.5 Robust inference for spatial data

$$\boxed{\ell_q(\boldsymbol{\theta}) = \sum_{j=1}^{R} L_q\left[\frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}(\boldsymbol{\theta})|}} \exp\left(-\frac{1}{2}\boldsymbol{Z}_j^\top\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{Z}_j\right)\right]}$$

Summary

1. Some Fundamental Problems in Environmental Data Science
   - 1.1 Spatial Gaussian likelihood inference
   - 1.2 Spatial kriging
   - 1.3 Gaussian random field simulations
   - 1.4 Multivariate Gaussian probabilities
   - 1.5 Robust inference for spatial data
2. Large-Scale Environmental Data Science with *ExaGeoStat*
   - 2.1 What is HPC?
   - 2.2 Task-based parallelism and dynamic runtime systems
   - 2.3 Tile-based linear algebra
   - 2.4 Tile low-rank (TLR) linear algebra
   - 2.5 Multi- and mixed-precision computational statistics
   - 2.6 *ExaGeoStat* software: Exascale geostatistics
3. Competitions on Spatial Statistics for Large Datasets
   - 3.1 In 2021: Gaussian and non-Gaussian
   - 3.2 In 2022: Nonstationary, space-time, multivariate
   - 3.3 In 2023: Irregular locations, confidence/prediction intervals

QUESTIONS?

Supplement: *ExaGeoStatR* on shared-memory system

Tile size effect:

16-core Intel Sandy Bridge Xeon E5-2650 Chip