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

Extreme Computing Research Center

siam

Accelerating Space and Space-Time Statistical Modeling with Mixed-Precision Arithmetic

Sameh Abdulah
Extreme Computing Research Center (ECRC), King Abdullah, University of Science and Technology (KAUST), Saudi Arabia

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

OFFICE OF SPONSORED RESEARCH

ICL UT INNOVATIVE COMPUTING LABORATORY at UNIVERSITY of TENNESSEE

SIAM Conference on Parallel Processing for Scientific Computing (PP22), Seattle, 23-26 February, 2022.

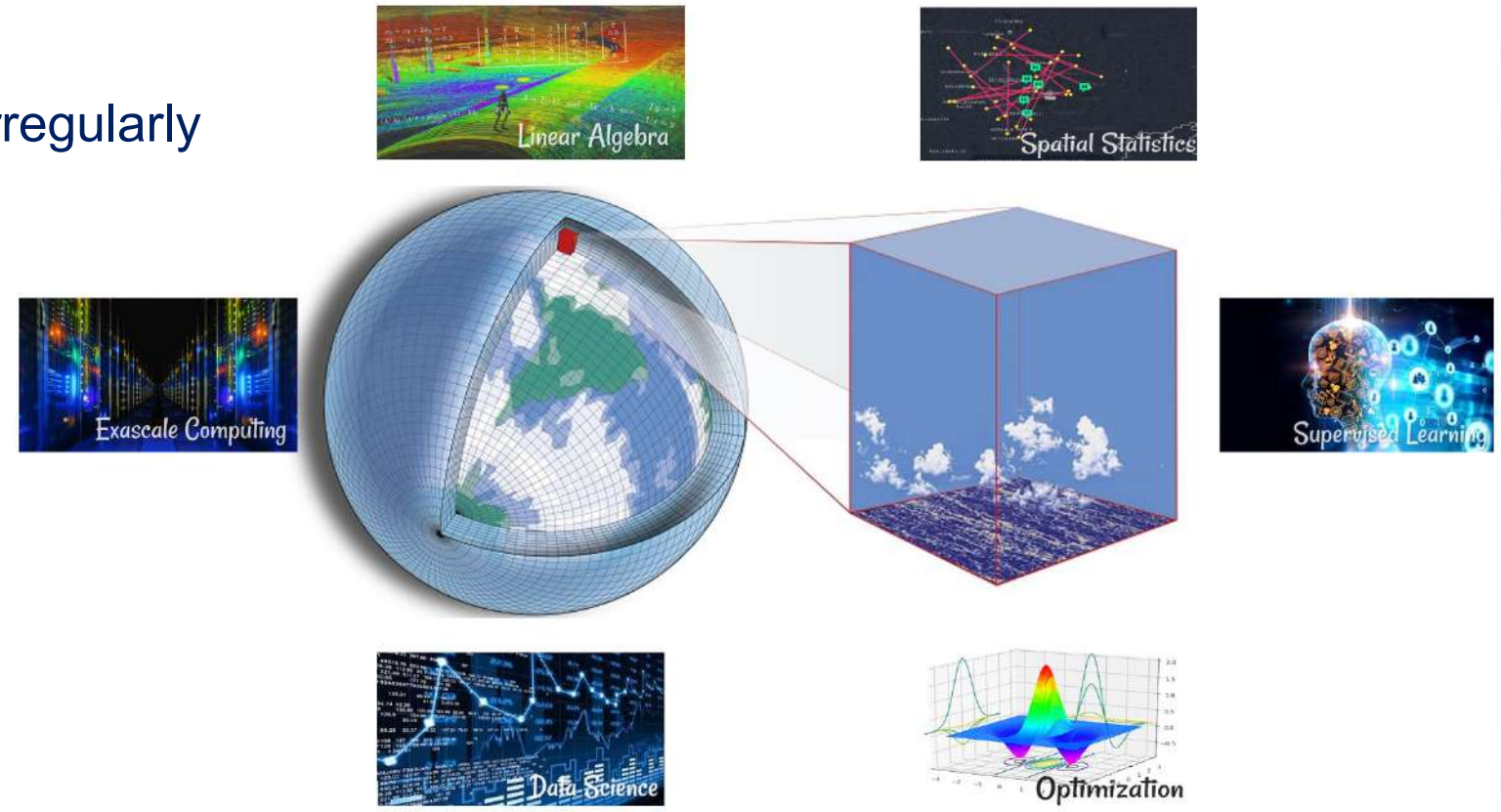
Acknowledgments

- Collaborators:
  - ECRC @ KAUST: Hatem Ltaief, Ying Sun, Marc Genton, and David Keyes
  - ICL @ UTK: Qinlei Cao, Yu Pei, George Bosilca, and Jack Dongarra

- Resource Allocations:
  - Shaheen-2 @ KSL, Saudi Arabia
  - HAWK @ HLRS, Germany
  - Summit @ ORNL, USA
  - Fugaku @ Riken, Japan

Pop Stats for Big Geodata

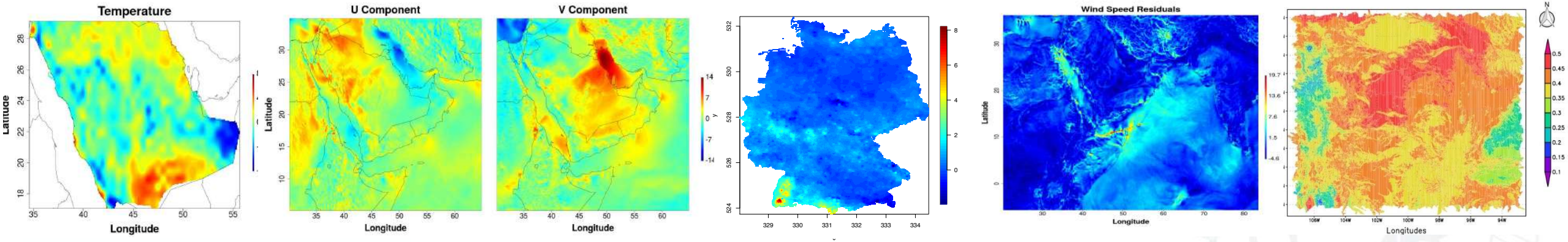
- Observing an increase of produced geodata.
- Techniques to process millions of observations have lagged behind.
- Implementations that work with irregularly spaced observations are rare.



Large-scale climate/weather modeling

Perform Climate/Weather Forecasting Simulations

- Applications for climate and environmental predictions are among the most time-consuming simulations workloads running on HPC facilities.
- Computational statistics: univariate/multivariate large spatial / space-time datasets in climate/weather modeling.



Examples of large climate/weather data

Problem Statement

- Today, weather and climate data are often huge!
  - Collect a large set of Z observations at a given n locations.
  - Process the Z observations, e.g., temperature and precipitation.
  - Predict missing observations in the remote locations which are related to the observed locations.
- Maximum Likelihood Function: an important machine learning technique for estimating statistical parameters required to perform prediction inference in climate and environmental applications.
- Complexity requirements:
  - Arithmetic cost: the linear solver and log-determinant involving n-by-n covariance matrix
  - $O(n^3)$ floating-point operations and $O(n^2)$ memory (assuming univariate case).
  - Memory footprint: 10^6 locations require 8 TB memory!

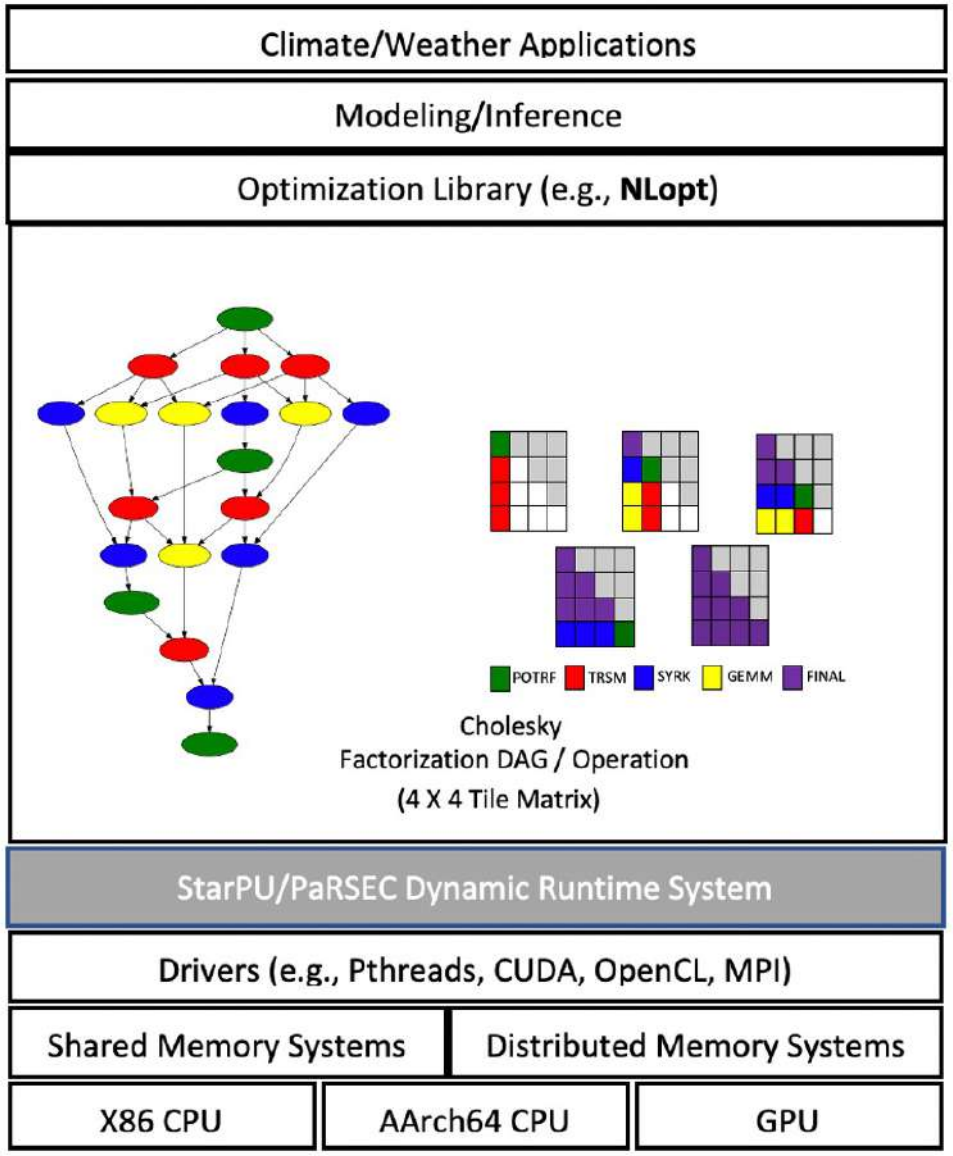
The ExaGeoStat Software Stack

X86 CPU

AArch64

Fujitsu A64FX

NVIDIA V100



#1 Fugaku

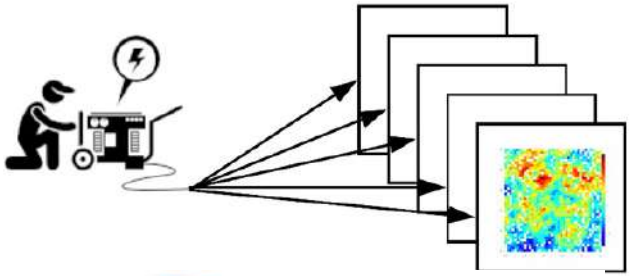
#2 Summit

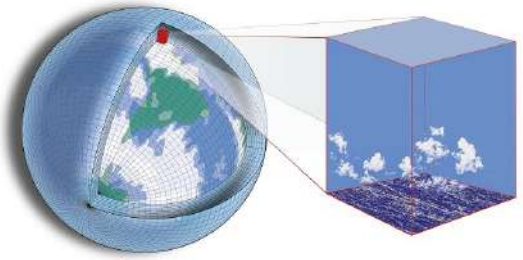
#24 HAWK

#89 Shaheen-II

The ExaGeoStat Framework

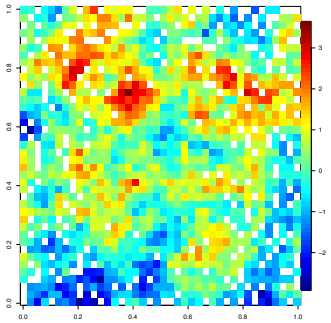
- Synthetic Dataset Generator
  - Generates large-scale geospatial datasets which can be separately used as benchmark datasets for other software packages.
- Maximum Likelihood Estimator (MLE)
  - Evaluates the maximum likelihood function on large-scale geospatial datasets.
  - Supports full machine precision (full-matrix), Tile Low-Rank (TLR) approximation, low-precision approximation accuracy.
- ExaGeoStat Predictor
  - Predicts unknown measurements at known geospatial locations by leveraging the MLE estimated parameters.



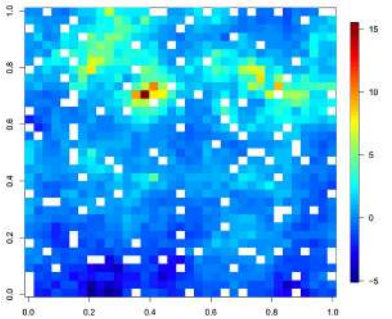


Synthetic Dataset Generator

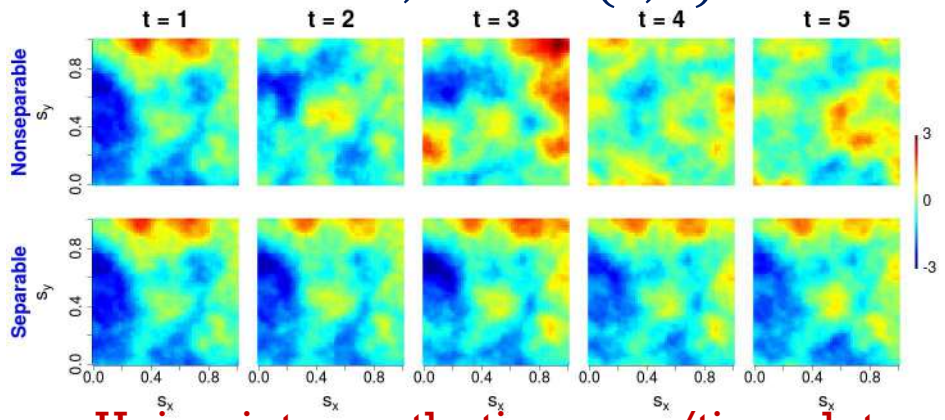
- Builds the covariance matrix $\sum(\boldsymbol{\theta_t})$ using a specific kernel and truth parameter vector $\boldsymbol{\theta_t}$.
- Computes Cholesky factorization of $\sum(\boldsymbol{\theta_t})$: $\sum(\boldsymbol{\theta_t})$ = **V** . **V**^{T}
- Generates **Z** vector: $\boldsymbol{Z} = \boldsymbol{V}.\boldsymbol{e}, \quad e \sim N(0,1)$



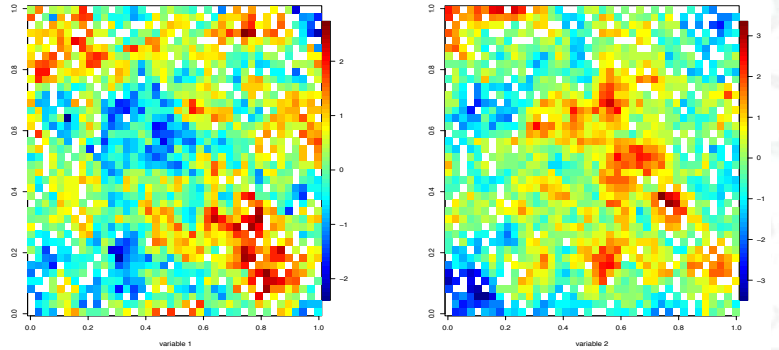
Univariate synthetic spatial data



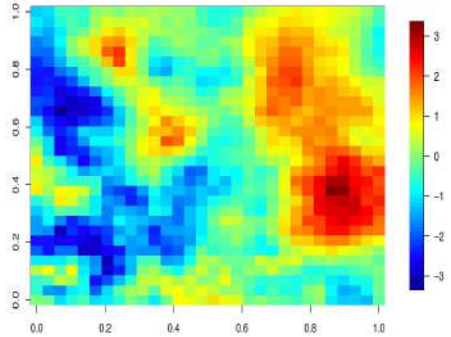
Univariate non-Gaussian synthetic spatial data



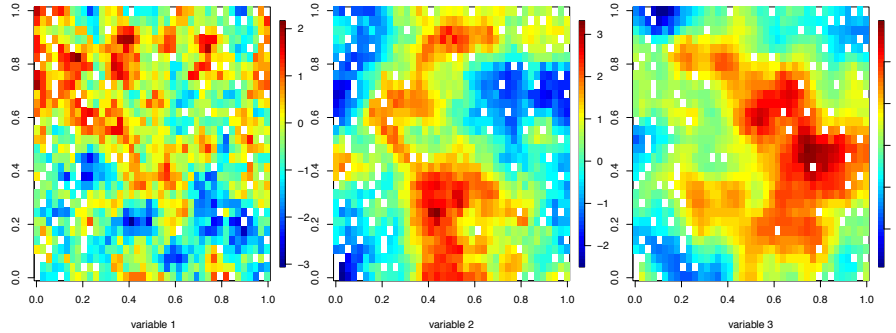
Univariate synthetic space/time data



Bivariate synthetic spatial data



Univariate non-stationary synthetic spatial data



Trivariate synthetic spatial data

Maximum Likelihood Estimator (MLE)

- The likelihood function: $\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}\mathbf{Z}^{\top}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{Z}.$
- Optimization loop with different $\boldsymbol{\theta}$ to maximize the likelihood function estimation until convergence
  - Generate the covariance matrix $\boldsymbol{\theta}$ using a specific kernel and the parameter vector $\boldsymbol{\theta}$($\boldsymbol{\theta}$ comes from the optimization function)
  - Solve the log determinant and the inverse operations requires a Cholesky factorization of the given covariance matrix $\sum(\boldsymbol{\theta_t})$ Cholesky factorization requires $O(n^3)$ floating-point operations $O(n^2)$ memory storage.
- NLOPT optimization library has been used to maximize the likelihood function until convergence in both cases. Recently, we improved the efficiency of the optimization process by using parallel PSwarm optimization algorithm to run several likelihood estimation step at the same time.

Supported Covariance Functions

Univariate Matern Kernel

$$C(r;\boldsymbol{\theta}) = \frac{\theta_1}{2^{\theta_3-1}\Gamma(\theta_3)}\left(\frac{r}{\theta_2}\right)^{\theta_3}\mathcal{K}_{\theta_3}\left(\frac{r}{\theta_2}\right)$$

Space/Time Non-Separable Kernel

$$C(\mathbf{h},u) = \frac{\sigma^2}{a_t|u|^{2\alpha}+1}\mathcal{M}_\nu\left\{\frac{\|\mathbf{h}\|/a_s}{(a_t|u|^{2\alpha}+1)^{\beta/2}}\right\},$$

Multivariate Parsimonious Kernel

$$C_{ij}(\|\mathbf{h}\|;\boldsymbol{\theta}) = \frac{\rho_{ij}\sigma_{ii}\sigma_{jj}}{2^{\nu_{ij}-1}\Gamma(\nu_{ij})}\left(\frac{\|\mathbf{h}\|}{a}\right)^{\nu_{ij}}\mathcal{K}_{\nu_{ij}}\left(\frac{\|\mathbf{h}\|}{a}\right)$$

Tukey g-and-h Non-Gaussian Field with Kernel

$$\rho_Z(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(4\sqrt{2\nu}\frac{h}{\phi}\right)^{\nu}\mathcal{K}_\nu\left(4\sqrt{2\nu}\frac{h}{\phi}\right)$$

Multivariate Flexible Kernel

$$C(\mathbf{h};u) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)\,(a|u|^{2\alpha}+1)^{\delta+\beta d/2}}\left(\frac{c\|\mathbf{h}\|}{(a|u|^{2\alpha}+1)^{\beta/2}}\right)^{\nu}$$
$$\times K_\nu\left(\frac{c\|\mathbf{h}\|}{(a|u|^{2\alpha}+1)^{\beta/2}}\right), \quad (\mathbf{h};u)\in\mathbb{R}^d\times\mathbb{R},$$

Powered Exponential Kernel

$$C(r;\boldsymbol{\theta}) = \theta_0\exp\left(\frac{-r^{\theta_2}}{\theta_1}\right)$$

ExaGeoStat Predictor

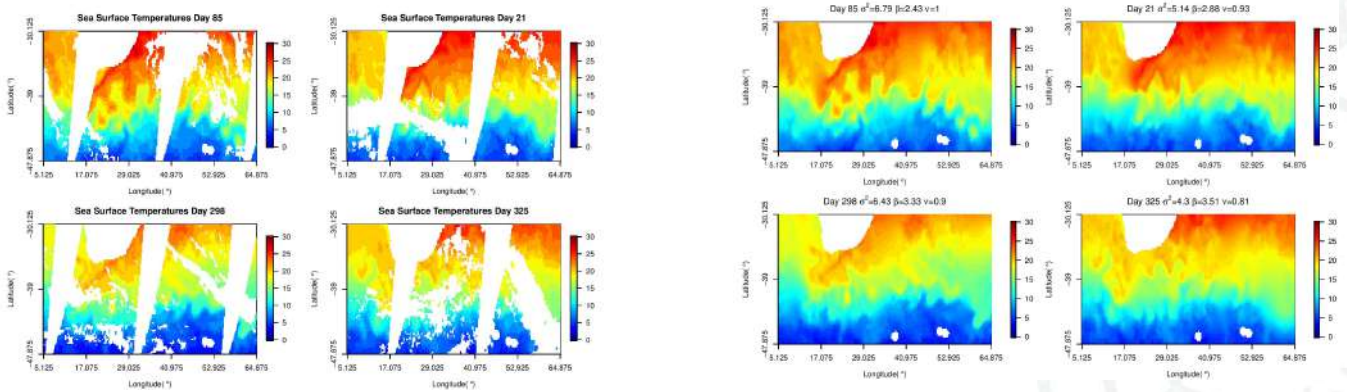
- Assuming $\sum_{11} \in \mathbb{R}^{m \times n}$, $\sum_{12} \in \mathbb{R}^{m \times n}$, $\sum_{21} \in \mathbb{R}^{n \times m}$, and $\sum_{22} \in \mathbb{R}^{n \times n}$

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N_{m+n}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{bmatrix}\right)$$

- The associated conditional distribution can be represented as

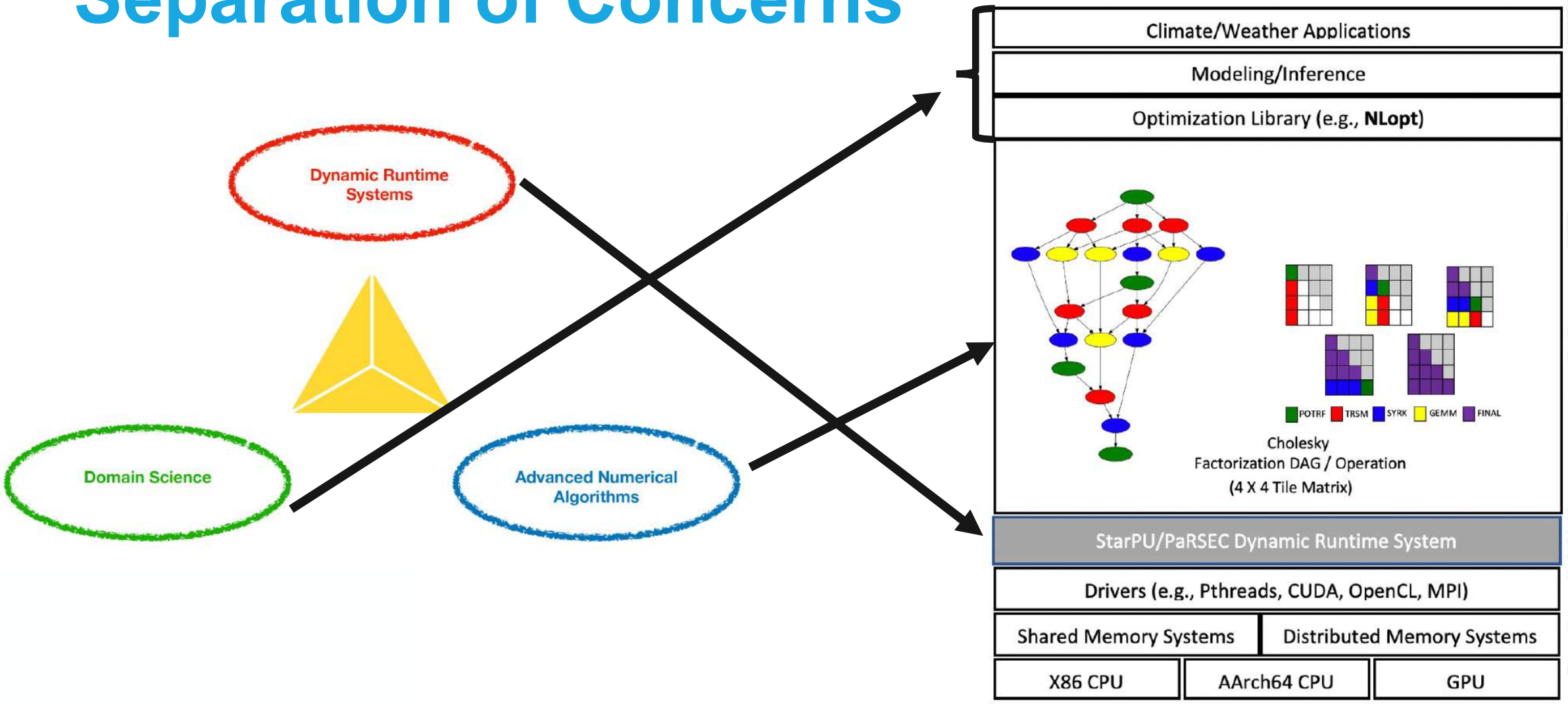
$$Z_1 | Z_2 \sim N_m(\mu_1 + \sum(\theta)_{12} \sum(\theta)_{22}^{-1} (Z_2 - \mu_2), \sum(\theta)_{11} - \sum(\theta)_{12} \sum(\theta)_{22} \sum(\theta)_{21}$$

- Assuming that the known measurements vector Z_2 has a zero-mean function (i.e., μ_1=0 , μ_2=0), the unknown measurements vector Z_1 can be predicted using, $Z_1 = \sum_{12} \sum_{22}^{-1} Z_2$, assuming Z_2 has a zero-mean function (μ_1=0 , μ_2=0)
- Solution of system of linear equation ($\sum_{22}^{-1} Z_2$) needs also a Cholesky factorization of $\sum_{22}$.



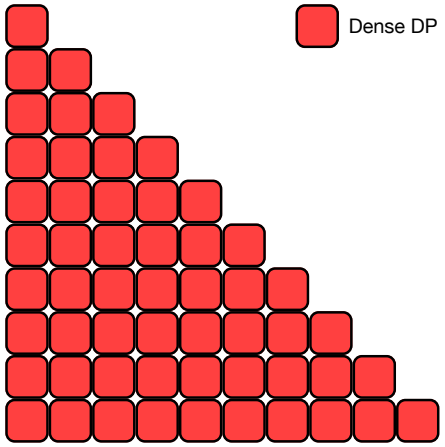
Sea Surface temperature Agulhas, South Africa.

An Effective Approach Based on a Separation of Concerns

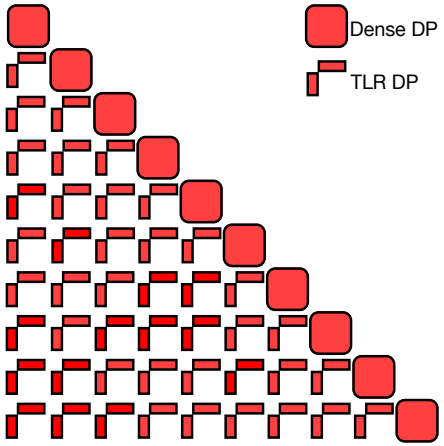


Matrix Data Structure in ExaGeoStat

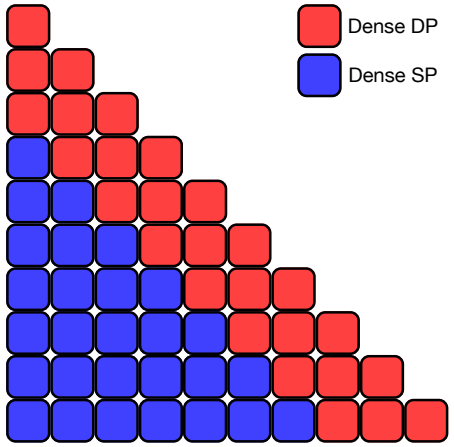
$$\sum(\boldsymbol{\theta})$$



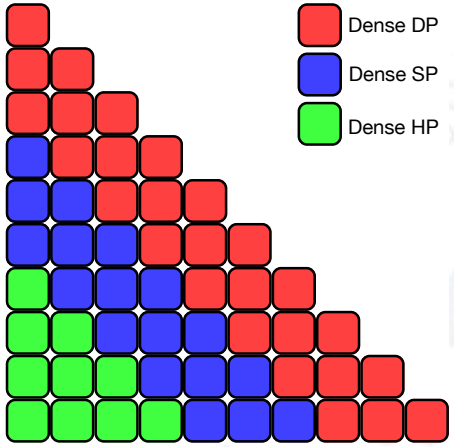
Exact Computation



TLR Computation



Double/Single
Precision Approximation



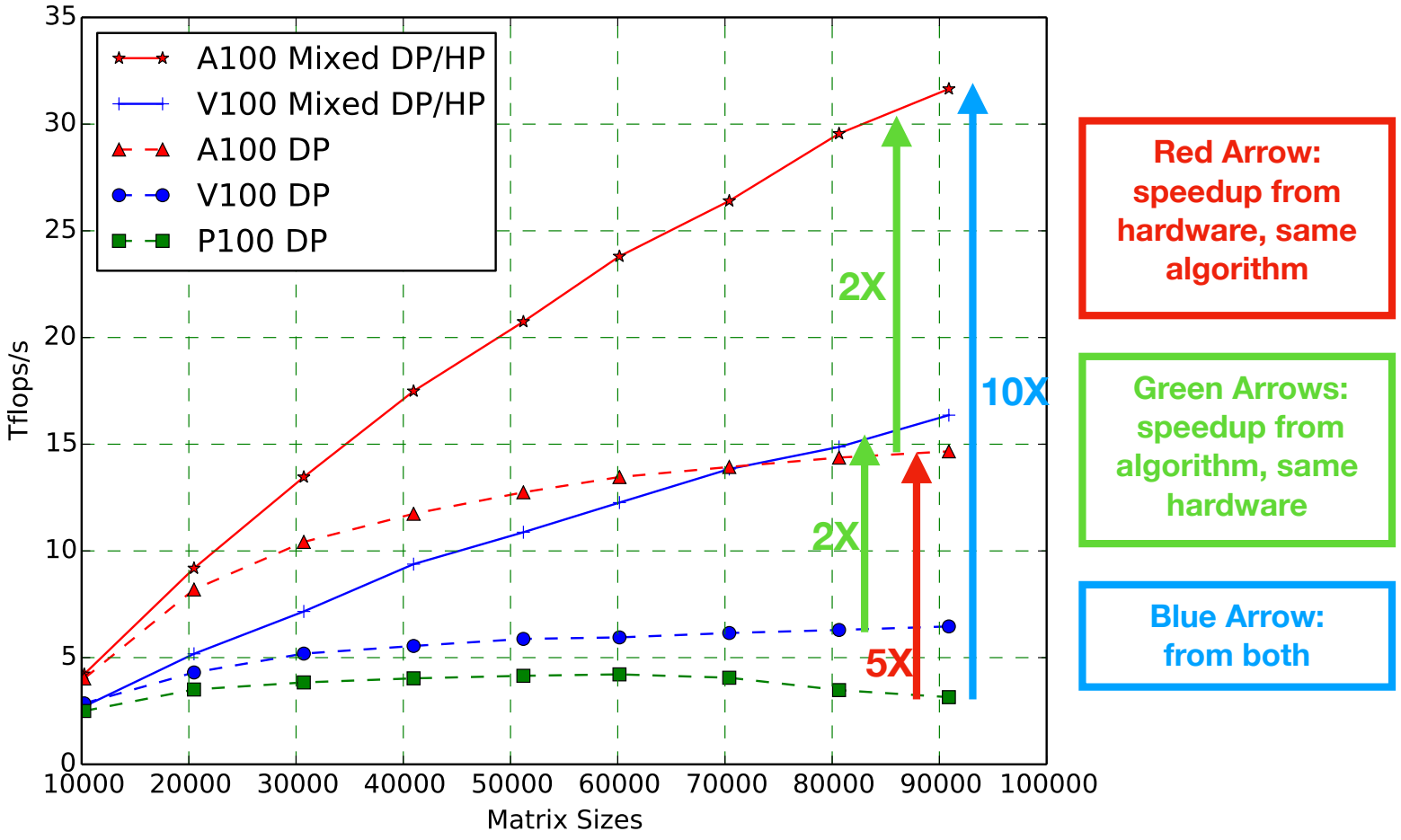
Double/Single/Half
Precision Approximation

Mixed-Precision Algorithms

- Achieve higher performance, faster time to solution (benefits from reduction of operations and data movement).
- Satisfy the trade-off between precision arithmetic combinations and ultimate application accuracy.
- Reduce power consumption by decreasing the execution time (energy saving).
- Extract performance from NVIDIA GPUs:

	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	---	156 TF
Peak FP16 Tensor Performance	120 TF	312 TF

Mixed-Precision for Environmental Applications

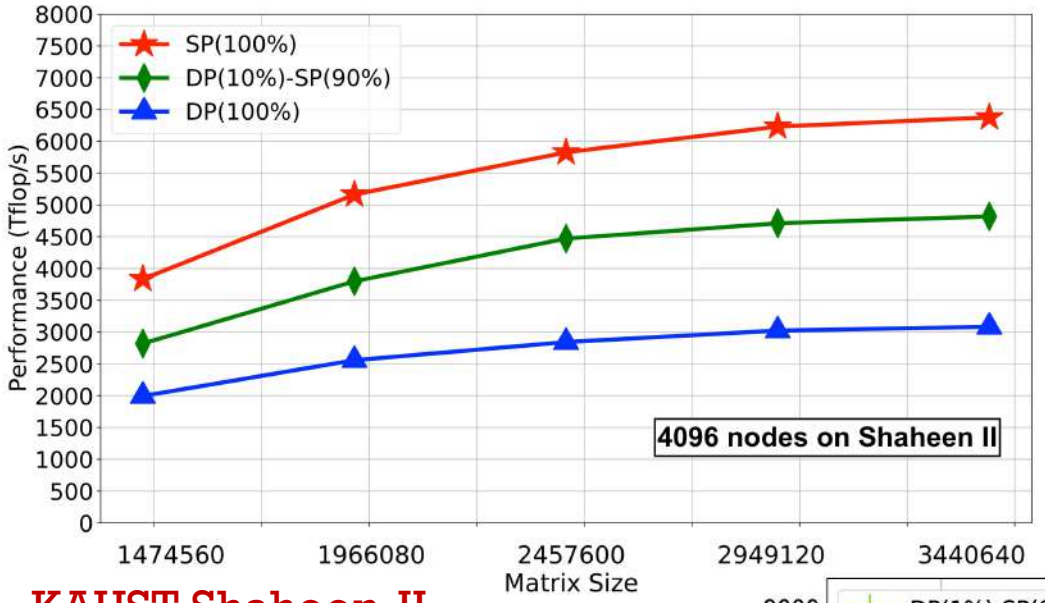


Red Arrow: speedup from hardware, same algorithm

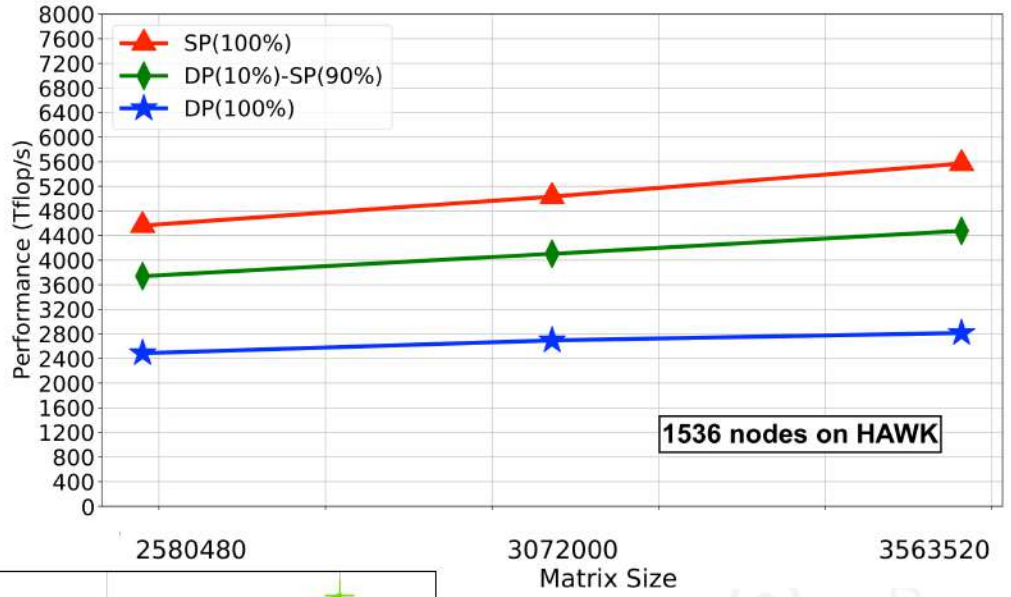
Green Arrows: speedup from algorithm, same hardware

Blue Arrow: from both

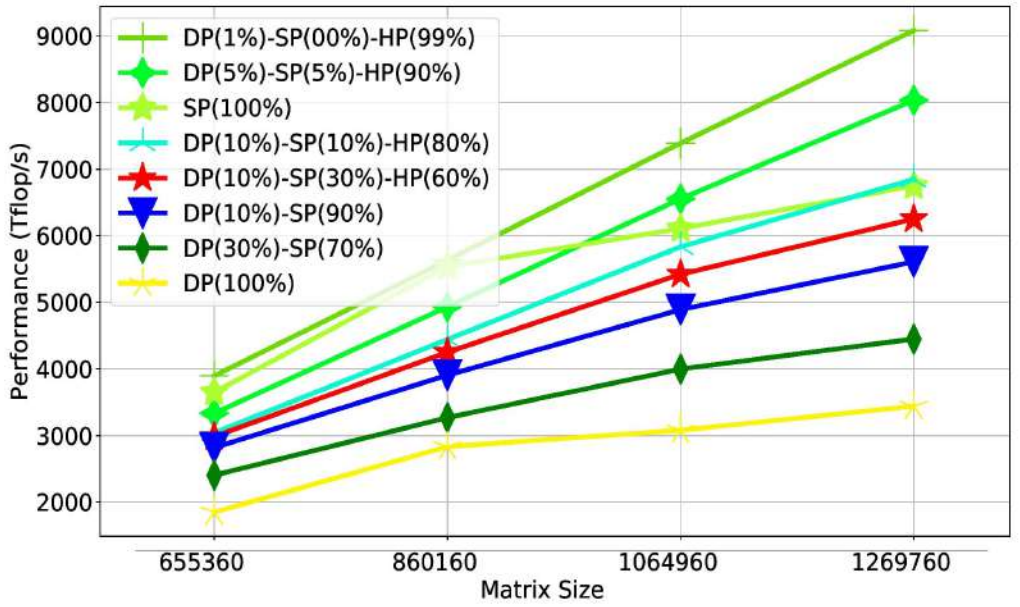
Mixed-Precision for Environmental Applications



KAUST Shaheen-II

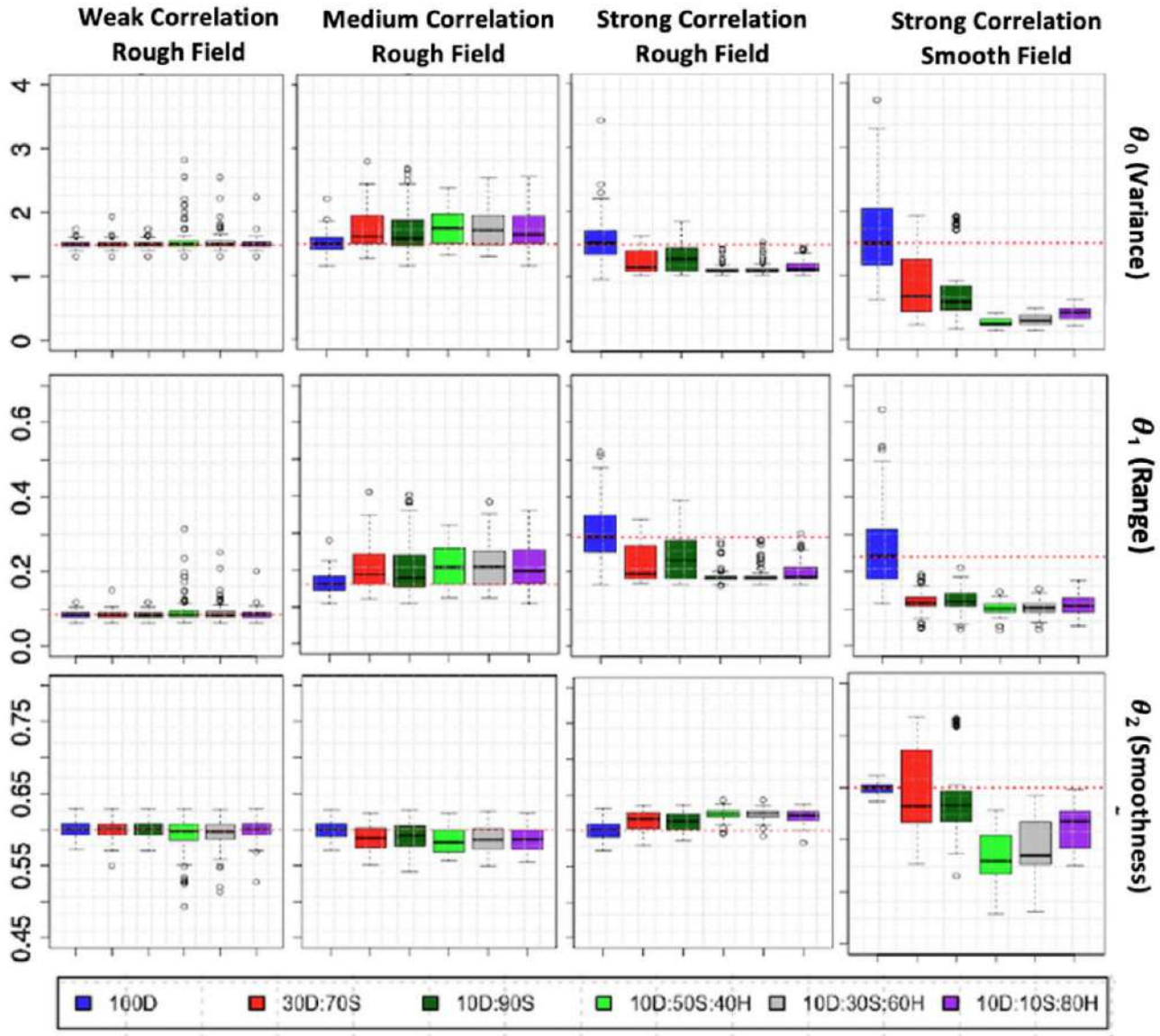


HLRS HAWK

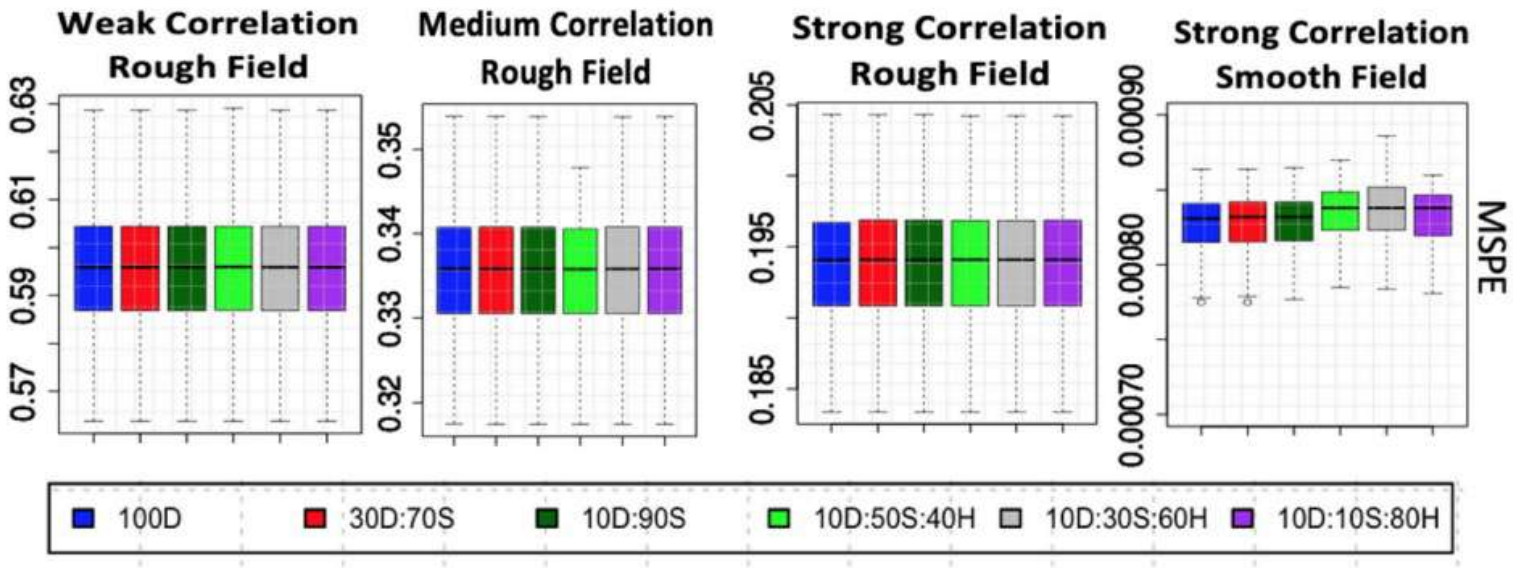


ORNL Summit

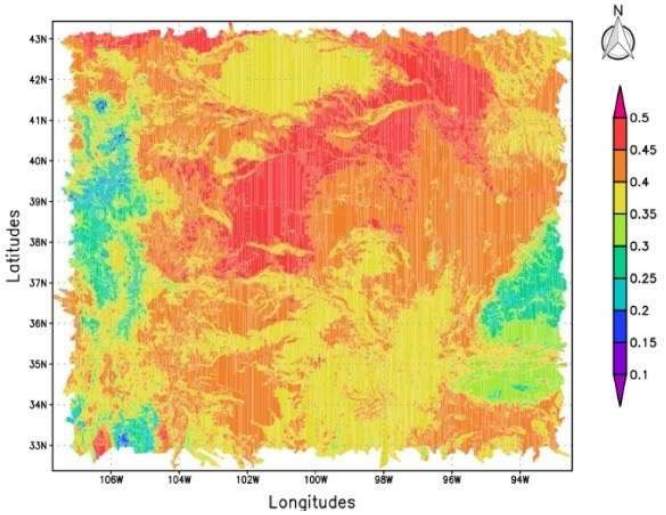
Accuracy Assessment using Synthetic Datasets



Prediction Assessment using Synthetic Datasets



Assessment on Real Datasets

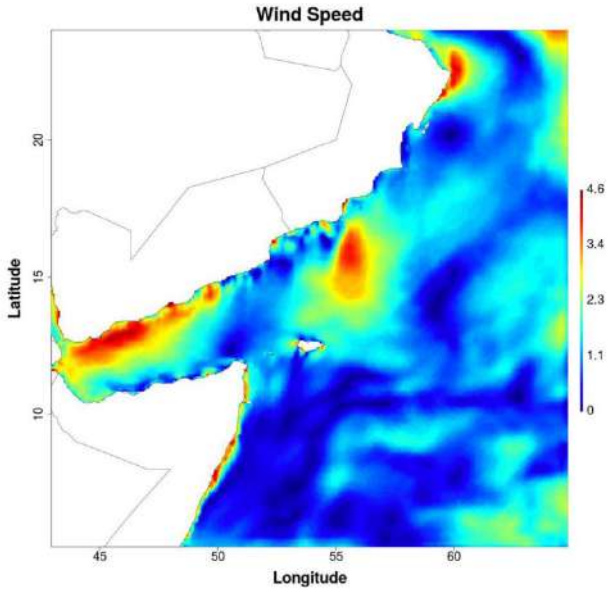


· 2M soil moisture data - the US

Qualitative Assessment of the MLE Based on the Mixed-Precision Approach Using 2D Soil Moisture Dataset

Variants	Variance (θ_0)	Range (θ_1)	Smoothness (θ_2)	Log-Likelihood (llh)	MSPE	Prediction Uncertainty	Iterations
100D	0.7223	0.0933	0.9983	−59740.65974	0.044926	4.734439e+03	180
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
5D:5S:90H	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

Assessment on Real Datasets



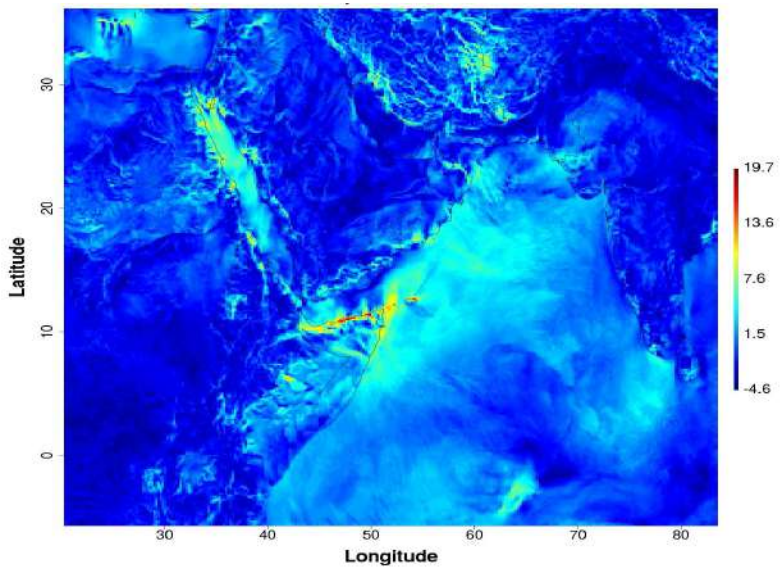
· 2M wind speed- Arabian sea

Qualitative Assessment of the MLE Based on the Mixed-Precision Approach Using 2D Wind Speed Dataset

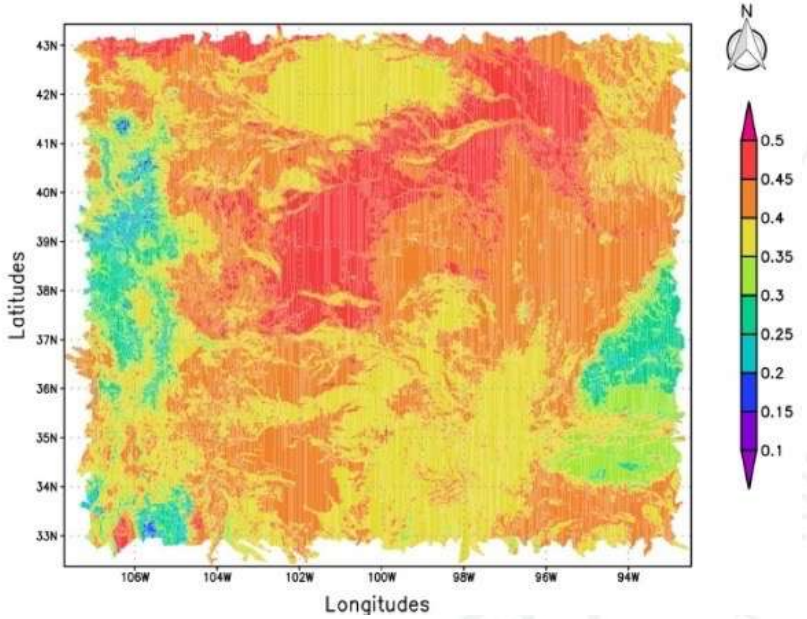
Variants	Variance (θ_0)	Range (θ_1)	Smoothness (θ_2)	Log-Likelihood (llh)	MSPE	Prediction Uncertainty	Iterations
100D	0.8407	0.0751	1.9905	241480.9994	1.752914E-02	2.2855E+00	666
10D:90S	0.9924	0.1794	1.9757	239908.1004	1.766194E-02	2.9170E+00	91
10D:30S:60H	0.9761	0.1804	1.9576	232783.9932	1.765651E-02	5.2836E+00	94

Qualitative Results with ExaGeoStat

Univariate Modeling (Exact/TLR approximation):



1M wind speed data – Middle-East
(MSPE:0.043201)

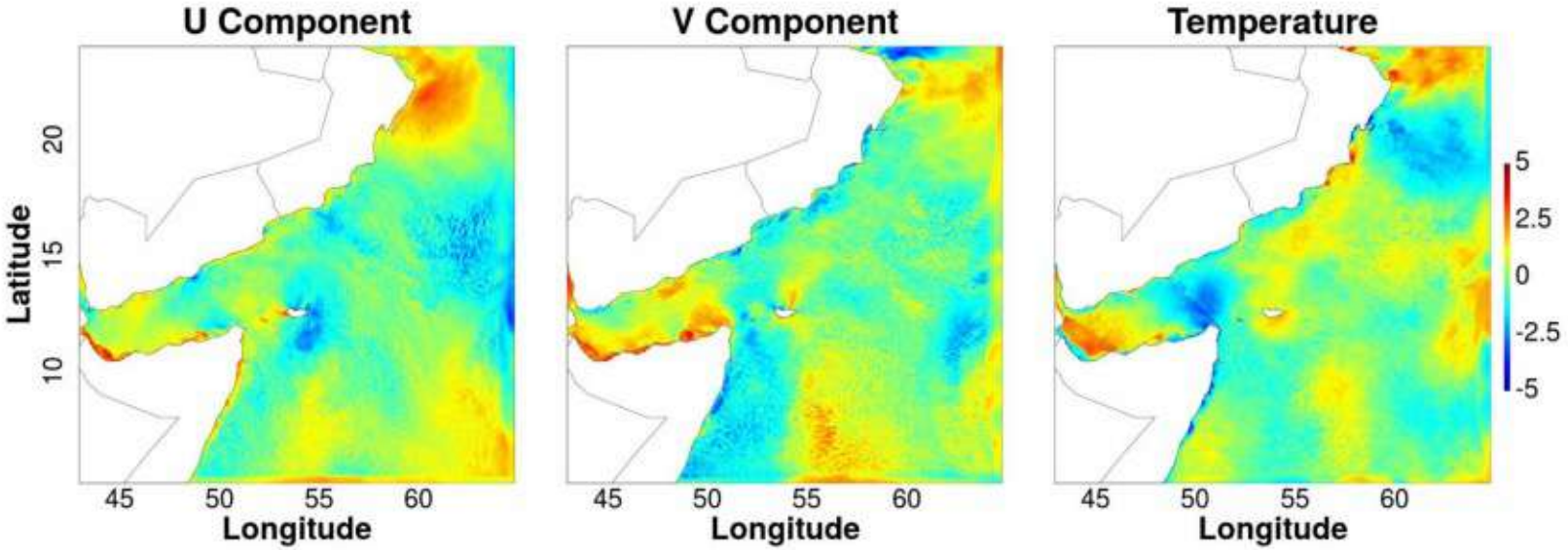


2M soil moisture data - the US
(MSPE:0.03507)

- Sameh Abdulah , Hatem Ltaief, Ying Sun, Marc G. Genton, and David E. Keyes. "ExaGeoStat: A high performance unified software for geostatistics on manycore systems." IEEE Transactions on Parallel and Distributed Systems 29, no. 12 (2018): 2771-2784.
- Sameh Abdulah, Hatem Ltaief, Ying Sun, Marc G. Genton, and David E. Keyes. "Parallel approximation of the maximum likelihood estimation for the prediction of large-scale geostatistics simulations." In 2018 IEEE International Conference on Cluster Computing (CLUSTER), pp. 98-108. IEEE, 2018.

Qualitative Results with ExaGeoStat

- Multivariate Modeling (Exact /TLR approximation):

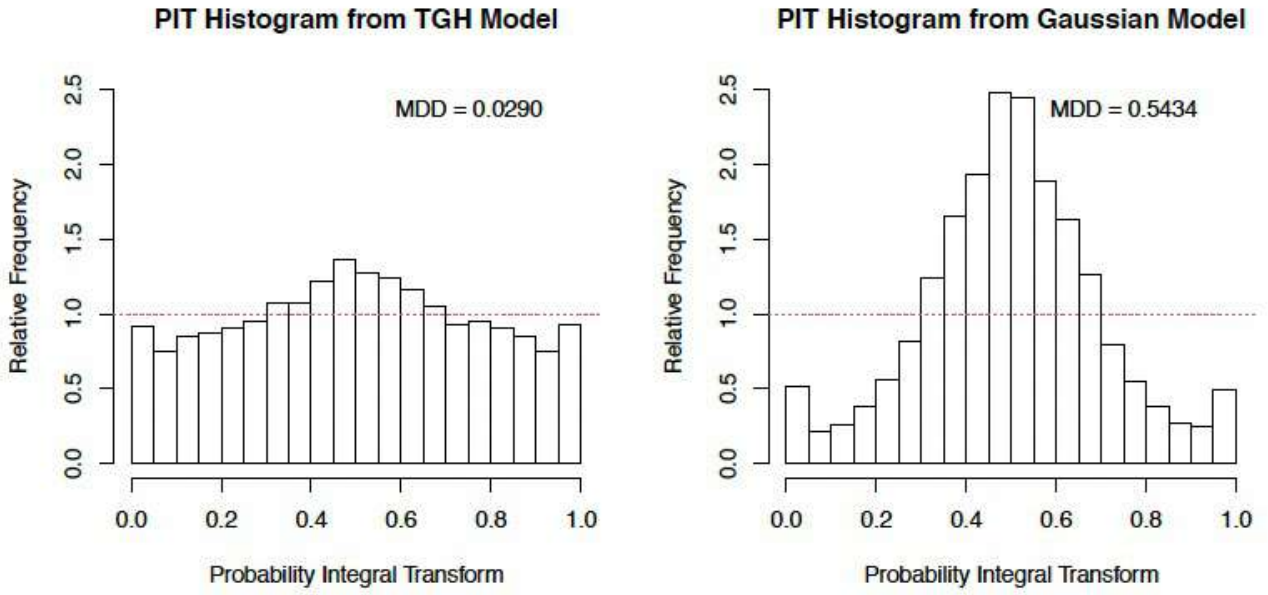


- 116K Trivariate dataset over the Arabian Sea (MSPE:0.00990)

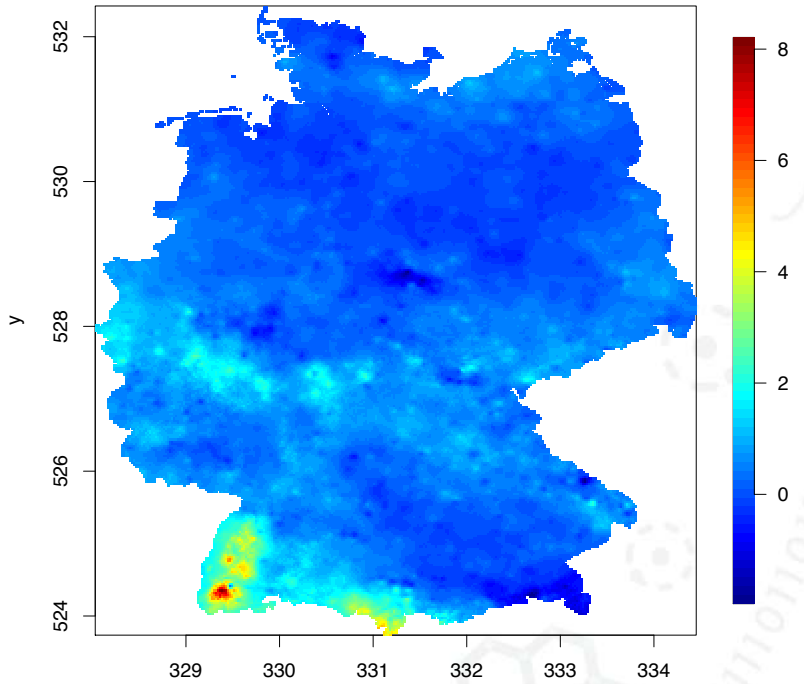
- Mary L. O. Salvaña, Sameh Abdulah, Huang Huang, Hatem Ltaief, Ying Sun, Marc M. Genton, and David Keyes. "High Performance Multivariate Geospatial Statistics on Manycore Systems," in IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 11, pp. 2719-2733, 1 Nov. 2021.

Qualitative Results with ExaGeoStat

Non-Gaussian Modeling



PIT histograms for both TGH and Gaussian model

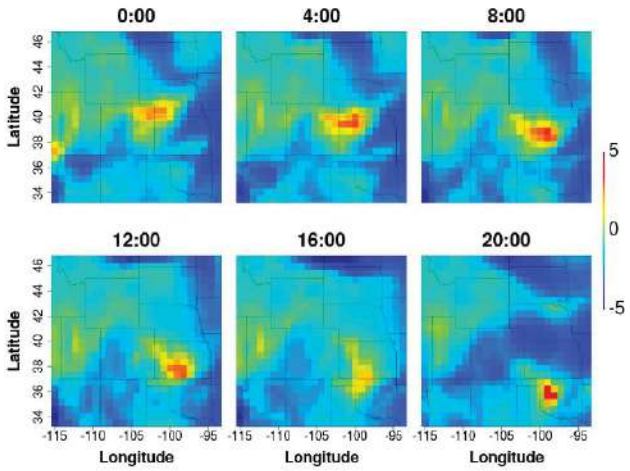


Visualization of the average daily precipitation over Germany (358 K locations)

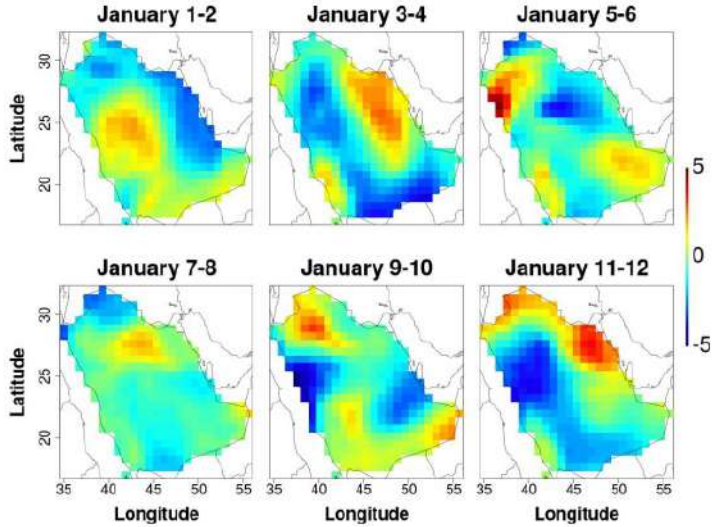
- Sagnik Mondal, Sameh Abdulah, Hatem Ltaief, Ying Sun, Marc M. Genton, and David Keyes. " Parallel Approximations of the Tukey g-and-h Likelihoods and Predictions for Non-Gaussian Geostatistics". IPDPS, 2022, accepted.

Qualitative Results

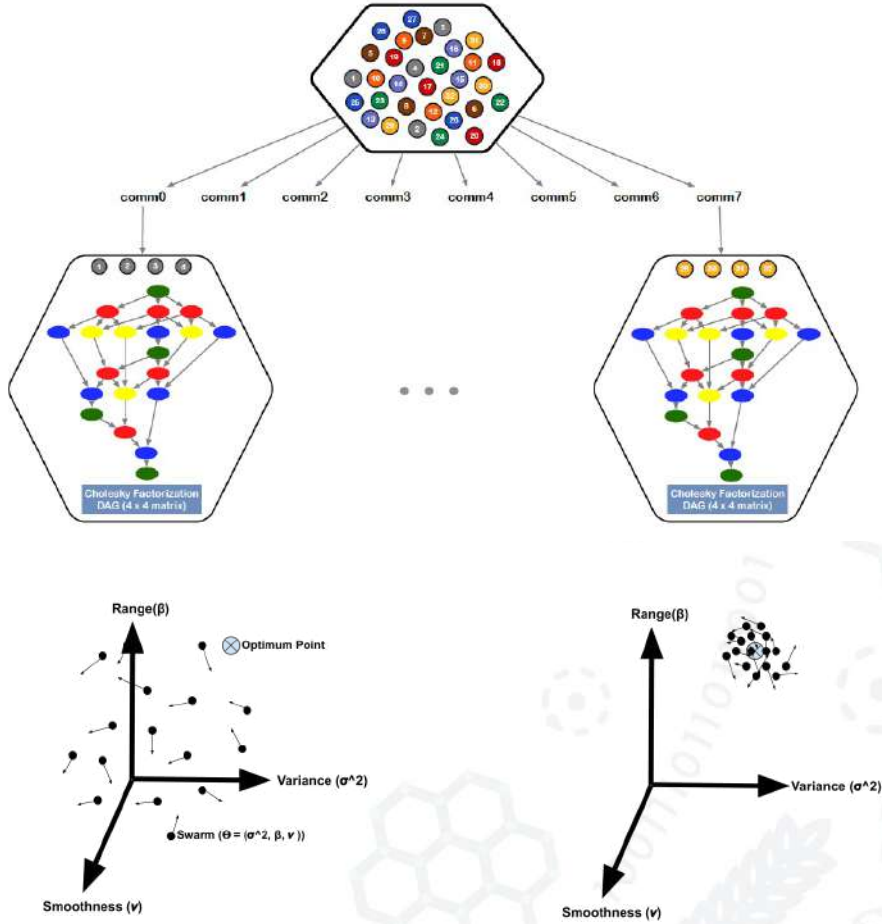
Spatio-Temporal Modeling



· Visualization of the log PM2.5 dataset (air pollution) over Midwest US (MSPE: 0.00280)



· Visualization of the log PM2.5 dataset (air pollution) over Saudi Arabia (MSPE: 0.001798)



Parallel Optimization Using PSwarm Algorithm

- Mary L. O. Salvaña, Sameh Abdulah, Hatem Ltaief, Ying Sun, Marc M. Genton, and David Keyes. "Massively Parallel Likelihood Function Optimization to Accelerate Air Pollution Prediction on Large-Scale Systems." Submitted to PASC22.

2021 KAUST Competition on Spatial Statistics for Large Datasets

- 29 research teams worldwide registered and 21 teams successfully submitted their results.

	Task	Data Model	Data size
1a	Parameters Estimation	Univariate Stationary Matern	90,000
1b	Prediction	Univariate Stationary Matern	Predict 10,000 conditional on 90,000
2a	Prediction	Tukey g-and-h	Predict 10,000 conditional on 90,000
2b	Prediction	Univariate Stationary Matern & Tukey g-and-h	Predict 100,000 conditional on 900,000

Country	US	Saudi Arabia	Germany	France	Switzerland	China	Australia	Japan	UAE	Russia	Taiwan	Ecuador	Russia	Chile
# Participants	30	10	5	4	8	3	6	3	1	1	7	1	1	2

- Huang Huang, Sameh Abdulah, Ying Sun, Hatem Ltaief, David Keyes, Marc Genton, "Competition on Spatial Statistics for Large Datasets." JABES 26, 580–595 (2021).

- More details: https://cemse.kaust.edu.sa/stsds/2021-kaust-competition-spatial-statistics-large-datasets

- 2nd KAUST SS competition: https://cemse.kaust.edu.sa/stsds/2022-kaust-competition-spatial-statistics-large-datasets

2022 KAUST Competition on Spatial Statistics for Large Datasets

- In total, the competition has twenty-eight (28) datasets generated using the ExaGeoStat software based on different models and settings as follows:
  - Sub-competition 1a includes two datasets which have been generated using a univariate spatial model in 2D with 100K (10^5) geospatial data points.
  - Sub-competition 1b includes two datasets which have been generated using a univariate spatial model in 2D with 1M (10^6) geospatial data points.
  - Sub-competition 2a includes nine datasets which have been generated using a space-time model in 2D×time with 1K geospatial data points and 100 time points.
  - Sub-competition 2b includes nine datasets which have been generated using a space-time model in 2D×time with 10K geospatial data points and 100 time points.
  - Sub-competition 3a includes three datasets which have been generated using a bivariate spatial model in 2D with 50K geospatial data points.
  - Sub-competition 3b includes three datasets which have been generated using a bivariate spatial model in 2D with 500K geospatial data points.

List of Publications

- 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(12), 2771-2784.
- Abdulah, S., Ltaief, H., Sun, Y., Genton, M. G., & Keyes, D. E. (2019, December). Geostatistical Modeling and Prediction using Mixed Precision Tile Cholesky Factorization. In 2019 IEEE 26th International Conference on High Performance Computing, Data, and Analytics (HiPC) (pp. 152-162). IEEE.
- Abdulah, S., Ltaief, H., Sun, Y., Genton, M. G., & Keyes, D. E. (2018, September). Parallel Approximation of the Maximum Likelihood Estimation for the Prediction of Large-scale Geostatistics Simulations. In 2018 IEEE International Conference on Cluster Computing (CLUSTER) (pp. 98-108).
- Abdulah, S., Li, Y., Cao, J., Ltaief, H., Keyes, D. E., Genton, M. G., & Sun, Y. (2019). ExaGeoStatR: A Package for Large-scale Geostatistics in R. (arXiv:1908.06936).
- Hong, Y., Abdulah, S., Genton, M. G., & Sun, Y. (2021). Efficiency assessment of approximated spatial predictions for large datasets. Spatial Statistics, 43, 100517.
- Salvana ML, Abdulah S, Huang H, Ltaief H, Sun Y, Genton MG, Keyes DE. High performance multivariate geospatial statistics on manycore systems. IEEE Transactions on Parallel and Distributed Systems. 2021 Apr 6;32(11):2719-33.
- Abdulah S, Cao Q, Pei Y, Bosilca G, Dongarra J, Genton MG, Keyes DE, Ltaief H, Sun Y. Accelerating Geostatistical Modeling and Prediction With Mixed-Precision Computations: A High-Productivity Approach With PaRSEC. IEEE Transactions on Parallel Distributed Systems. 2022 Apr 1;33(04):964-76
- Mondal S, Abdulah S, Ltaief H, Sun Y, Genton MG, Keyes DE. Parallel Approximations of the Tukey g-and-h Likelihoods and Predictions for Non-Gaussian Geostatistics, IPDPS22, to be published.
- Salvana ML, Abdulah S, Huang H, Ltaief H, Sun Y, Genton MG, Keyes DE. Parallel Space-Time Likelihood Optimization to Improve Air Pollution Prediction Large-Scale System. IEEE Transactions on Parallel and Distributed Systems. PASC22, under review.

ExaGeoStat is an open-source software which is available at https://github.com/ecrc/exageostat.

ExaGeoStatR is available at: https://github.com/ecrc/exageostatR.

Thanks,

QUESTIONS?