A Machine Learning Framework for Large-Scale Weather and Climate Prediction using Exact and Approximate Linear Algebra Computation

Sameh Abdulah, **Hatem Ltaief, Ying Sun, Marc Genton and David Keyes**
Extreme Computing Research Center
King Abdullah University of Science and Technology

IXPUG Annual Fall Conference 2018, Oregon, US

September 25-28, 2018

KAUST

Agenda

Agenda

- Problem Statement.

Agenda

- Problem Statement.
- Materń Covariance Function.

Agenda

- Problem Statement.
- Mateŕn Covariance Function.
- Dynamic Runtime Systems.

Agenda

- Problem Statement.
- Mateŕn Covariance Function.
- Dynamic Runtime Systems.
- State-of-the-Art Linear Algebra Libraries.

Agenda

- Problem Statement.
- Materń Covariance Function.
- Dynamic Runtime Systems.
- State-of-the-Art Linear Algebra Libraries.
- ExaGeoStat Framework.

Agenda

- Problem Statement.
- Mateŕn Covariance Function.
- Dynamic Runtime Systems.
- State-of-the-Art Linear Algebra Libraries.
- ExaGeoStat Framework.
  - Synthetic Dataset Generator.

Agenda

- Problem Statement.
- Mateŕn Covariance Function.
- Dynamic Runtime Systems.
- State-of-the-Art Linear Algebra Libraries.
- ExaGeoStat Framework.
  - Synthetic Dataset Generator.
  - Maximum Likelihood Estimator (MLE).

Agenda

- Problem Statement.
- Mateŕn Covariance Function.
- Dynamic Runtime Systems.
- State-of-the-Art Linear Algebra Libraries.
- ExaGeoStat Framework.
    - Synthetic Dataset Generator.
    - Maximum Likelihood Estimator (MLE).
    - ExaGeoStat Predictor.

Agenda

- Problem Statement.
- Mateŕn Covariance Function.
- Dynamic Runtime Systems.
- State-of-the-Art Linear Algebra Libraries.
- ExaGeoStat Framework.
    - Synthetic Dataset Generator.
    - Maximum Likelihood Estimator (MLE).
    - ExaGeoStat Predictor.
- Tile Low-Rank (TLR) Approximation.

Agenda

- Problem Statement.
- Mateŕn Covariance Function.
- Dynamic Runtime Systems.
- State-of-the-Art Linear Algebra Libraries.
- ExaGeoStat Framework.
  - Synthetic Dataset Generator.
  - Maximum Likelihood Estimator (MLE).
  - ExaGeoStat Predictor.
- Tile Low-Rank (TLR) Approximation.
- Performance on Synthetic and Real Datasets.

Agenda

- Problem Statement.
- Mateŕn Covariance Function.
- Dynamic Runtime Systems.
- State-of-the-Art Linear Algebra Libraries.
- ExaGeoStat Framework.
    - Synthetic Dataset Generator.
    - Maximum Likelihood Estimator (MLE).
    - ExaGeoStat Predictor.
- Tile Low-Rank (TLR) Approximation.
- Performance on Synthetic and Real Datasets.
- Outcomes.

Agenda

- Problem Statement.
- Materń Covariance Function.
- Dynamic Runtime Systems.
- State-of-the-Art Linear Algebra Libraries.
- ExaGeoStat Framework.
  - Synthetic Dataset Generator.
  - Maximum Likelihood Estimator (MLE).
  - ExaGeoStat Predictor.
- Tile Low-Rank (TLR) Approximation.
- Performance on Synthetic and Real Datasets.
- Outcomes.
- Conclusion.

Climate/Weather Forecasting

Climate/Weather Forecasting

- Applications for climate and environmental predictions are among the most time-consuming simulations workloads running on today's supercomputer facilities.

Climate/Weather Forecasting

- Applications for climate and environmental predictions are among the most time-consuming simulations workloads running on today's supercomputer facilities.
- Today, weather and climate data are usually huge!

Climate/Weather Forecasting

- Applications for climate and environmental predictions are among the most time-consuming simulations workloads running on today's supercomputer facilities.
- Today, weather and climate data are usually huge!
  - A set of Z observations at given n locations.

Climate/Weather Forecasting

- Applications for climate and environmental predictions are among the most time-consuming simulations workloads running on today's supercomputer facilities.
- Today, weather and climate data are usually huge!
  - A set of Z observations at given n locations.
  - Z observations could be temperature, precipitation, ... etc.

Climate/Weather Forecasting

- Applications for climate and environmental predictions are among the most time-consuming simulations workloads running on today's supercomputer facilities.
- Today, weather and climate data are usually huge!
  - A set of Z observations at given n locations.
  - Z observations could be temperature, precipitation, ... etc.
- Maximum Likelihood Function

Climate/Weather Forecasting

- Applications for climate and environmental predictions are among the most time-consuming simulations workloads running on today's supercomputer facilities.
- Today, weather and climate data are usually huge!
  - A set of Z observations at given n locations.
  - Z observations could be temperature, precipitation, ... etc.
- Maximum Likelihood Function
  - An important statistical technique for predicting unknown measurements in climate and environmental applications.

Climate/Weather Forecasting

- Applications for climate and environmental predictions are among the most time-consuming simulations workloads running on today's supercomputer facilities.
- Today, weather and climate data are usually huge!
  - A set of Z observations at given n locations.
  - Z observations could be temperature, precipitation, ... etc.
- Maximum Likelihood Function
  - An important statistical technique for predicting unknown measurements in climate and environmental applications.
    - Prohibitive computational Cost and memory requirements:

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{Z}^T\Sigma^{-1}(\boldsymbol{\theta})\mathbf{Z} - \frac{1}{2}\log|\Sigma(\boldsymbol{\theta})|$$

Climate/Weather Forecasting

- Applications for climate and environmental predictions are among the most time-consuming simulations workloads running on today's supercomputer facilities.
- Today, weather and climate data are usually huge!
  - A set of Z observations at given n locations.
  - Z observations could be temperature, precipitation, ... etc.
- Maximum Likelihood Function
  - An important statistical technique for predicting unknown measurements in climate and environmental applications.
    - Prohibitive computational Cost and memory requirements:

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{Z}^T\Sigma^{-1}(\boldsymbol{\theta})\mathbf{Z} - \frac{1}{2}\log|\Sigma(\boldsymbol{\theta})|$$

  - The linear solver and log-determinant involving floating point operations on n-by-n covariance matrix with $O(n^3)$ complexity and $O(n^2)$ memory footprint.

Climate/Weather Forecasting

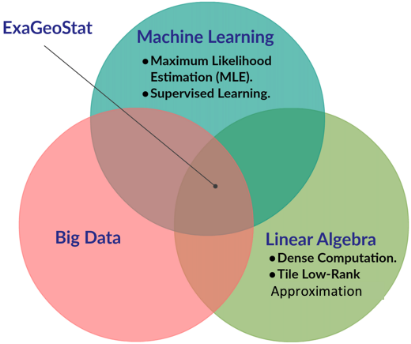
- Applications for climate and environmental predictions are among the most time-consuming simulations workloads running on today's supercomputer facilities.
- Today, weather and climate data are usually huge!
  - A set of Z observations at given n locations.
  - Z observations could be temperature, precipitation, ... etc.
- Maximum Likelihood Function
  - An important statistical technique for predicting unknown measurements in climate and environmental applications.
    - Prohibitive computational Cost and memory requirements:

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{Z}^T\Sigma^{-1}(\boldsymbol{\theta})\mathbf{Z} - \frac{1}{2}\log|\Sigma(\boldsymbol{\theta})|$$

    - The linear solver and log-determinant involving floating point operations on n-by-n covariance matrix with $O(n^3)$ complexity and $O(n^2)$ memory footprint.
    - For instance: 10^6 locations require $8TB$ Memory!

Exascale Geostatistics (ExaGeoStat)

- A framework which exploits machine learning, statistical modeling and forecasting, and the state-of-the-art linear algebra techniques to handle large-scale Geostatistics data.



Mateŕn Covariance Function

- In climate/weather applications different covariance functions can be used to generate $\Sigma(\theta)$.

Mateŕn Covariance Function

- In climate/weather applications different covariance functions can be used to generate $\Sigma(\theta)$.
- A Generic covariance function can be directly used (Matérn function):

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

Mateŕn Covariance Function

- In climate/weather applications different covariance functions can be used to generate $\Sigma(\theta)$.
- A Generic covariance function can be directly used (Matérn function):

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

  - $\theta_1 > 0$ (Variance).

Mateŕn Covariance Function

- In climate/weather applications different covariance functions can be used to generate $\Sigma(\theta)$.
- A Generic covariance function can be directly used (Matérn function):

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

  - $\theta_1 > 0$ (Variance).
  - $\theta_2 > 0$ (Spatial Range, larger values $\rightarrow$ strong correlation).

Mateŕn Covariance Function

- In climate/weather applications different covariance functions can be used to generate $\Sigma(\theta)$.
- A Generic covariance function can be directly used (Matérn function):

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

  - $\theta_1 > 0$ (Variance).
  - $\theta_2 > 0$ (Spatial Range, larger values $\rightarrow$ strong correlation).
  - $\theta_3 > 0$ (Smoothness, larger values $\rightarrow$ smoother field).

Mateŕn Covariance Function

- In climate/weather applications different covariance functions can be used to generate $\Sigma(\theta)$.
- A Generic covariance function can be directly used (Matérn function):

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

  - $\theta_1 > 0$ (Variance).
  - $\theta_2 > 0$ (Spatial Range, larger values $\rightarrow$ strong correlation).
  - $\theta_3 > 0$ (Smoothness, larger values $\rightarrow$ smoother field).
- Apparently dense matrices arising in climate/weather applications,

Mateŕn Covariance Function

- In climate/weather applications different covariance functions can be used to generate $\Sigma(\theta)$.
- A Generic covariance function can be directly used (Matérn function):

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

  - $\theta_1 > 0$ (Variance).
  - $\theta_2 > 0$ (Spatial Range, larger values $\rightarrow$ strong correlation).
  - $\theta_3 > 0$ (Smoothness, larger values $\rightarrow$ smoother field).
- Apparently dense matrices arising in climate/weather applications,
  - Rely on leading edge parallel architectures.

Mateŕn Covariance Function

- In climate/weather applications different covariance functions can be used to generate $\Sigma(\theta)$.
- A Generic covariance function can be directly used (Matérn function):

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

  - $\theta_1 > 0$ (Variance).
  - $\theta_2 > 0$ (Spatial Range, larger values $\rightarrow$ strong correlation).
  - $\theta_3 > 0$ (Smoothness, larger values $\rightarrow$ smoother field).
- Apparently dense matrices arising in climate/weather applications,
  - Rely on leading edge parallel architectures.
  - Compress dense matrices with tile low-rank approximation

Mateŕn Covariance Function

- In climate/weather applications different covariance functions can be used to generate $\Sigma(\theta)$.
- A Generic covariance function can be directly used (Matérn function):

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

  - $\theta_1 > 0$ (Variance).
  - $\theta_2 > 0$ (Spatial Range, larger values $\rightarrow$ strong correlation).
  - $\theta_3 > 0$ (Smoothness, larger values $\rightarrow$ smoother field).
- Apparently dense matrices arising in climate/weather applications,
  - Rely on leading edge parallel architectures.
  - Compress dense matrices with tile low-rank approximation
    - Huge performance improvement via cutting down flops.

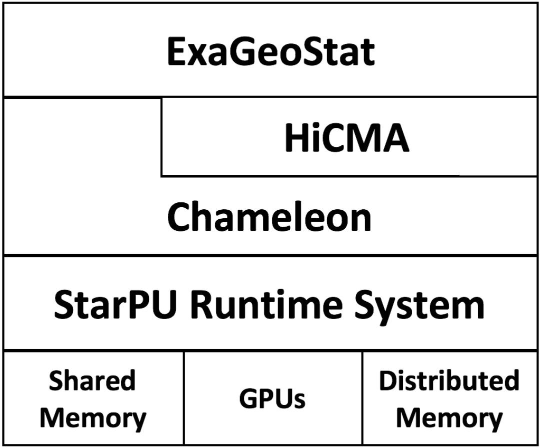
Mateŕn Covariance Function

- In climate/weather applications different covariance functions can be used to generate $\Sigma(\theta)$.
- A Generic covariance function can be directly used (Matérn function):

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

  - $\theta_1 > 0$ (Variance).
  - $\theta_2 > 0$ (Spatial Range, larger values $\rightarrow$ strong correlation).
  - $\theta_3 > 0$ (Smoothness, larger values $\rightarrow$ smoother field).
- Apparently dense matrices arising in climate/weather applications,
  - Rely on leading edge parallel architectures.
  - Compress dense matrices with tile low-rank approximation
    - Huge performance improvement via cutting down flops.
    - Preserving the accuracy requirements of the scientific application.

ExaGeoStat Framework



Dynamic Runtime Systems

- Operate directly on the sequential code and schedule the various tasks across the underlying hardware resources.

Dynamic Runtime Systems

- Operate directly on the sequential code and schedule the various tasks across the underlying hardware resources.
- Its role is to ensure that the data dependencies are not violated.

Dynamic Runtime Systems

- Operate directly on the sequential code and schedule the various tasks across the underlying hardware resources.
- Its role is to ensure that the data dependencies are not violated.
- Enhance the software productivity by abstracting the hardware complexity from the end users.

Dynamic Runtime Systems

- Operate directly on the sequential code and schedule the various tasks across the underlying hardware resources.
- Its role is to ensure that the data dependencies are not violated.
- Enhance the software productivity by abstracting the hardware complexity from the end users.
- StarPU

Dynamic Runtime Systems

- Operate directly on the sequential code and schedule the various tasks across the underlying hardware resources.
- Its role is to ensure that the data dependencies are not violated.
- Enhance the software productivity by abstracting the hardware complexity from the end users.
- StarPU
  - INRIA Bordeaux, France.

Dynamic Runtime Systems

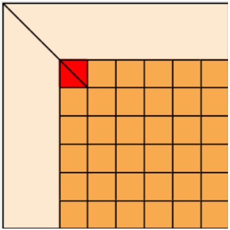
- Operate directly on the sequential code and schedule the various tasks across the underlying hardware resources.
- Its role is to ensure that the data dependencies are not violated.
- Enhance the software productivity by abstracting the hardware complexity from the end users.
- StarPU
  - INRIA Bordeaux, France.
  - A unified Runtime System for Heterogeneous Multicore Architectures.

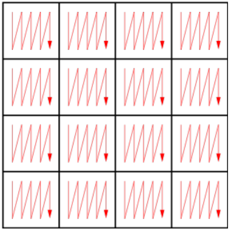
Dynamic Runtime Systems

- Operate directly on the sequential code and schedule the various tasks across the underlying hardware resources.
- Its role is to ensure that the data dependencies are not violated.
- Enhance the software productivity by abstracting the hardware complexity from the end users.
- StarPU
  - INRIA Bordeaux, France.
  - A unified Runtime System for Heterogeneous Multicore Architectures.
  - Shared Memory, GPUs, Distributed Systems.

State-of-the-art Linear Algebra Libraries

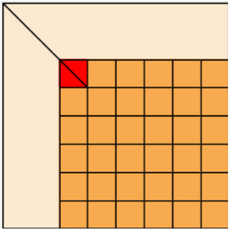
- Tile Algorithms
  - PLASMA, Chameleon, and FLAME,

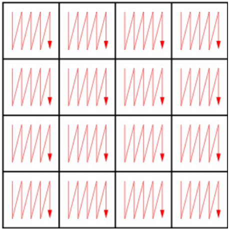




State-of-the-art Linear Algebra Libraries

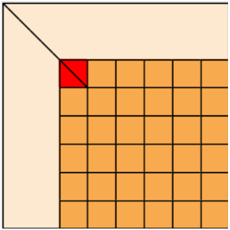
- Tile Algorithms
  - PLASMA, Chameleon, and FLAME,
  - The dense matrix is broken into tiles.

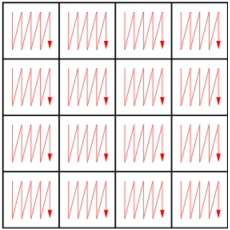




State-of-the-art Linear Algebra Libraries

- Tile Algorithms
  - PLASMA, Chameleon, and FLAME,
  - The dense matrix is broken into tiles.
  - Weaken the synchronization points by bringing the parallelism in multithreaded BLAS.





ExaGeoStat

- Synthetic Dataset Generator

ExaGeoStat

- Synthetic Dataset Generator
  - Generate large-scale geospatial datasets which can be separately used as benchmark datasets for other software packages.

ExaGeoStat

- Synthetic Dataset Generator
    - Generate large-scale geospatial datasets which can be separately used as benchmark datasets for other software packages.
- Maximum Likelihood Estimator (MLE)

ExaGeoStat

- Synthetic Dataset Generator
  - Generate large-scale geospatial datasets which can be separately used as benchmark datasets for other software packages.
- Maximum Likelihood Estimator (MLE)
  - Evaluate the maximum likelihood function on large-scale geospatial datasets.

ExaGeoStat

- Synthetic Dataset Generator
  - Generate large-scale geospatial datasets which can be separately used as benchmark datasets for other software packages.
- Maximum Likelihood Estimator (MLE)
  - Evaluate the maximum likelihood function on large-scale geospatial datasets.
  - Support full machine precision accuracy (full-matrix) and Tile Low-Rank (TLR) approximation.

ExaGeoStat

- Synthetic Dataset Generator
  - Generate large-scale geospatial datasets which can be separately used as benchmark datasets for other software packages.
- Maximum Likelihood Estimator (MLE)
  - Evaluate the maximum likelihood function on large-scale geospatial datasets.
  - Support full machine precision accuracy (full-matrix) and Tile Low-Rank (TLR) approximation.
- ExaGeoStat Predictor

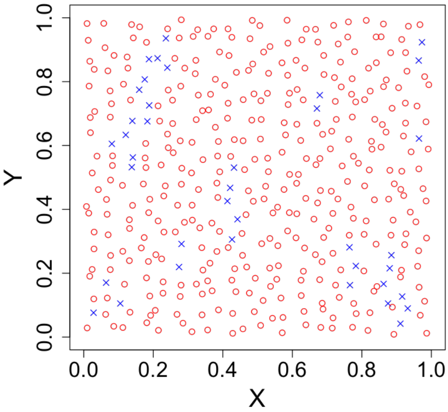
ExaGeoStat

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

Synthetic Dataset Generator

- Cholesky factorization of $\Sigma(\theta)$:

$$\boldsymbol{\Sigma(\theta) = V.V^T}$$



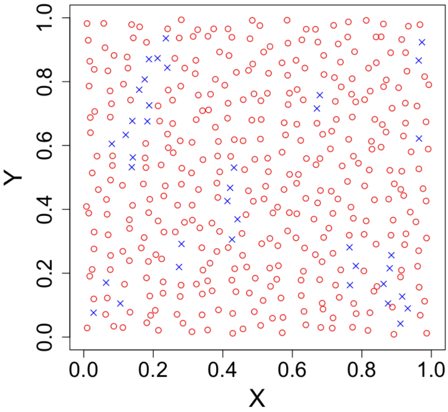
Synthetic Dataset Generator

- Cholesky factorization of $\Sigma(\theta)$:

$$\boldsymbol{\Sigma(\theta) = V.V^T}$$

- $\boldsymbol{Z}$ Vector Generation

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



Synthetic Dataset Generator

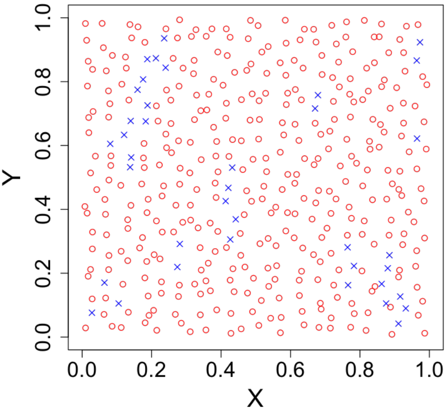
- Cholesky factorization of $\Sigma(\theta)$:

$$\boldsymbol{\Sigma(\theta) = V.V^T}$$

- $\boldsymbol{Z}$ Vector Generation

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

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



Maximum Likelihood Estimator (MLE)

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

- Log determinant and linear solver requires a Cholesky factorization of the given covariance matrix $\Sigma(\theta)$.

Maximum Likelihood Estimator (MLE)

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

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

Maximum Likelihood Estimator (MLE)

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

- Log determinant and linear solver requires a Cholesky factorization of the given covariance matrix $\Sigma(\theta)$.
- Cholesky factorization requires $O(n^3)$ floating point operations and $O(n^2)$.
- Dense ExaGeoStat uses Chameleon linear algebra libraries to provide a large-scale exact solution for the MLE operation.

Maximum Likelihood Estimator (MLE)

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

- Log determinant and linear solver requires a Cholesky factorization of the given covariance matrix $\Sigma(\theta)$.
- Cholesky factorization requires $O(n^3)$ floating point operations and $O(n^2)$.
- Dense ExaGeoStat uses Chameleon linear algebra libraries to provide a large-scale exact solution for the MLE operation.
- NLOPT optimization library has been used to maximize the likelihood function till convergence in both cases.

ExaGeoStat Predictor

$$\begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix} \sim N_{m+n}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right) \tag{2}$$

$$\mathbf{Z}_1 | \mathbf{Z}_2 \sim N_m(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{Z}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}). \tag{3}$$

- Assuming $\mathbf{Z_2}$ has a zero-mean function $(\mu_1 = 0, \mu_2 = 0)$

$$\mathbf{Z}_1 = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Z}_2. \tag{4}$$

ExaGeoStat Predictor

$$\begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix} \sim N_{m+n}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right) \tag{2}$$

$$\mathbf{Z}_1 | \mathbf{Z}_2 \sim N_m(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{Z}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}). \tag{3}$$

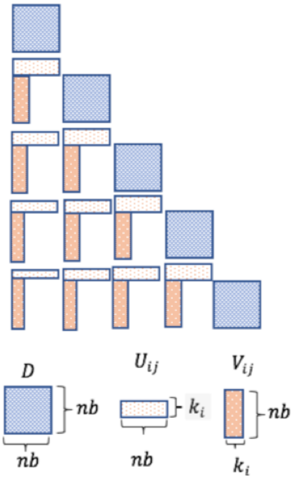
- Assuming $\mathbf{Z_2}$ has a zero-mean function $(\mu_1 = 0, \mu_2 = 0)$

$$\mathbf{Z}_1 = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Z}_2. \tag{4}$$

- Solution of system of linear equation $\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Z}_2$ needs also a Cholesky factorization of Σ_{22}

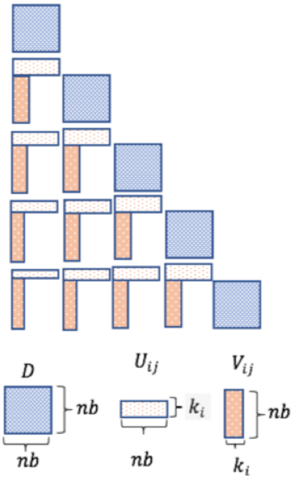
Tile-Low Rank Approximation

- Tile Low-Rank (TLR) Algorithms
  - HiCMA Library (KAUST, 2017).



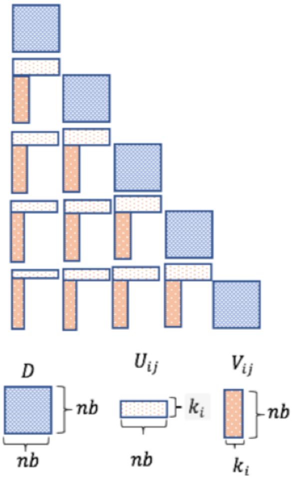
Tile-Low Rank Approximation

- Tile Low-Rank (TLR) Algorithms
  - HiCMA Library (KAUST, 2017).
  - Use *SVD*, approximate each off-diagonal tile, keep the most significant k (matrix rank) singular values and their left and right singular vectors, U and V.



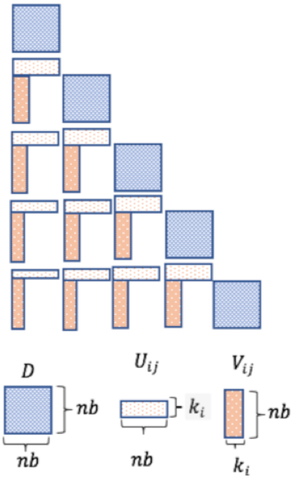
Tile-Low Rank Approximation

- Tile Low-Rank (TLR) Algorithms
  - HiCMA Library (KAUST, 2017).
  - Use *SVD*, approximate each off-diagonal tile, keep the most significant k (matrix rank) singular values and their left and right singular vectors, U and V.
  - Depend on Selected accuracy (application specific).



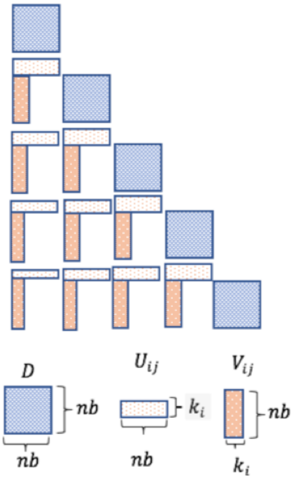
Tile-Low Rank Approximation

- Tile Low-Rank (TLR) Algorithms
  - HiCMA Library (KAUST, 2017).
  - Use *SVD*, approximate each off-diagonal tile, keep the most significant k (matrix rank) singular values and their left and right singular vectors, U and V.
  - Depend on Selected accuracy (application specific).
  - Two variations can be provided,



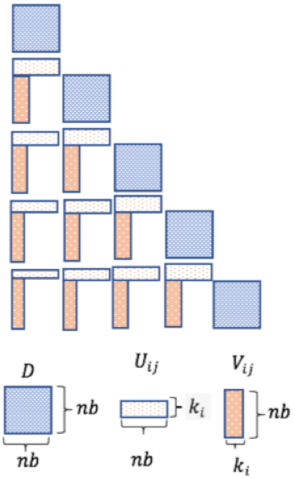
Tile-Low Rank Approximation

- Tile Low-Rank (TLR) Algorithms
  - HiCMA Library (KAUST, 2017).
  - Use *SVD*, approximate each off-diagonal tile, keep the most significant k (matrix rank) singular values and their left and right singular vectors, U and V.
  - Depend on Selected accuracy (application specific).
  - Two variations can be provided,
    - Fixed Rank.



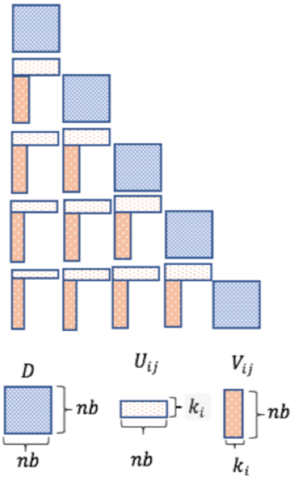
Tile-Low Rank Approximation

- Tile Low-Rank (TLR) Algorithms
  - HiCMA Library (KAUST, 2017).
  - Use *SVD*, approximate each off-diagonal tile, keep the most significant k (matrix rank) singular values and their left and right singular vectors, U and V.
  - Depend on Selected accuracy (application specific).
  - Two variations can be provided,
    - Fixed Rank.
    - Fixed Accuracy.



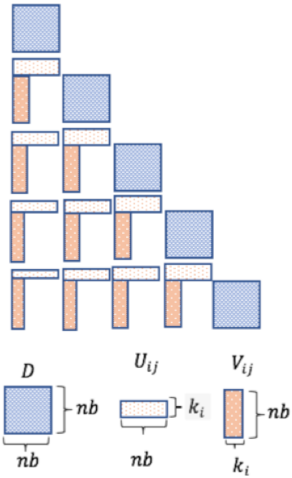
Tile-Low Rank Approximation

- Tile Low-Rank (TLR) Algorithms
  - HiCMA Library (KAUST, 2017).
  - Use *SVD*, approximate each off-diagonal tile, keep the most significant k (matrix rank) singular values and their left and right singular vectors, U and V.
  - Depend on Selected accuracy (application specific).
  - Two variations can be provided,
    - Fixed Rank.
    - Fixed Accuracy.
  - In the case of Fixed accuracy, k varies from one tile to another. Therefore, load imbalance issues appear.



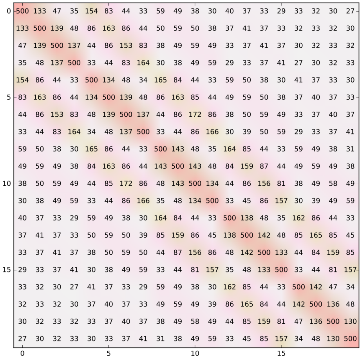
Tile-Low Rank Approximation

- Tile Low-Rank (TLR) Algorithms
  - HiCMA Library (KAUST, 2017).
  - Use *SVD*, approximate each off-diagonal tile, keep the most significant k (matrix rank) singular values and their left and right singular vectors, U and V.
  - Depend on Selected accuracy (application specific).
  - Two variations can be provided,
    - Fixed Rank.
    - Fixed Accuracy.
  - In the case of Fixed accuracy, k varies from one tile to another. Therefore, load imbalance issues appear.
  - Solution: rely on dynamic runtime systems.



Tile-Low Rank Approximation

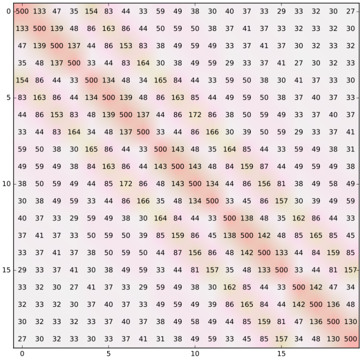
- Climate/ weather modeling applications requires 10^{-9} accuracy threshold.



Tile-Low Rank Approximation

- Climate/ weather modeling applications requires 10^{-9} accuracy threshold.
- Using Exponential covariance function,

$$C(r;\theta) = \theta_1 exp(\frac{r}{\theta_2})$$

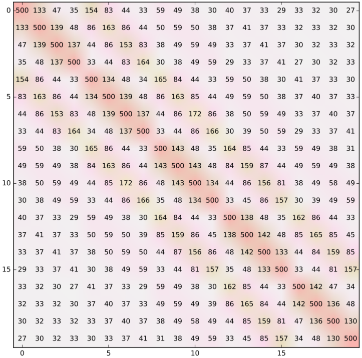


Tile-Low Rank Approximation

- Climate/ weather modeling applications requires 10^{-9} accuracy threshold.
- Using Exponential covariance function,

$$C(r;\theta) = \theta_1 exp(\frac{r}{\theta_2})$$

- Example, rank distribution on $2k \times 2k$ matrix where $nb = 500$, 2D problem.

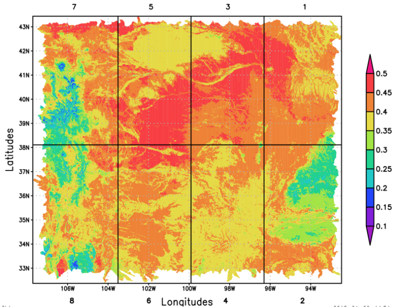


Real Datasets Examples

- Soil Moisture data at the top layer of the Mississippi River Basin in the United States, on January 1st, 2004.

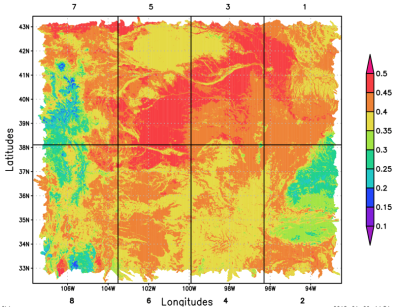
Real Datasets Examples

- Soil Moisture data at the top layer of the Mississippi River Basin in the United States, on January 1st, 2004.
- $\sim 2M$ Locations.



Real Datasets Examples

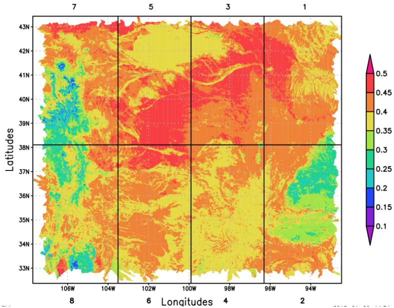
- Soil Moisture data at the top layer of the Mississippi River Basin in the United States, on January 1st, 2004.
- $\sim 2M$ Locations.



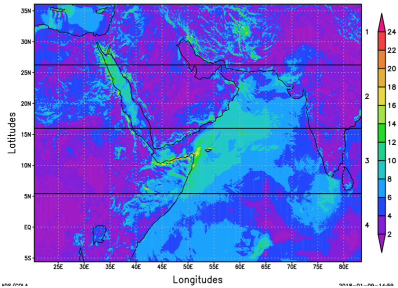
- Wind Speed data at Middle East, on September 1st, 2017.

Real Datasets Examples

- Soil Moisture data at the top layer of the Mississippi River Basin in the United States, on January 1st, 2004.
- $\sim 2M$ Locations.

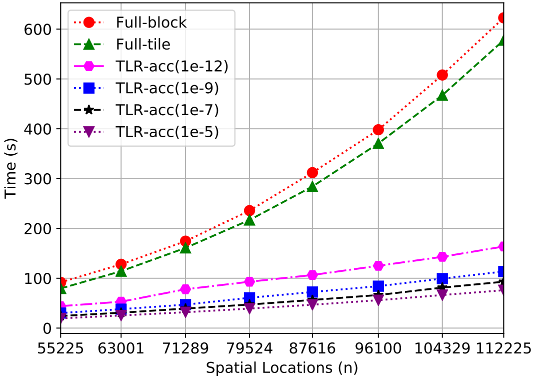


- Wind Speed data at Middle East, on September 1st, 2017.
- $\sim 1M$ Locations.



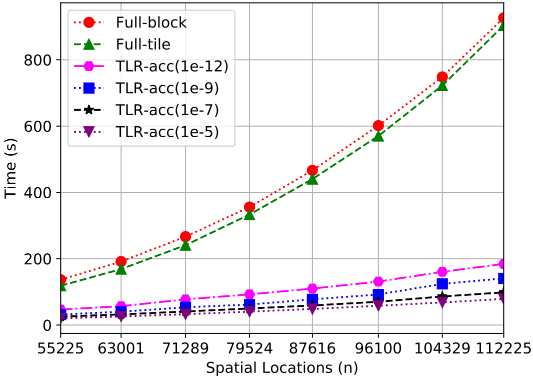
Performance on Shared Memory

- Intel Haswell,
  - Intel(R) Xeon(R) CPU E5-2699 v3 @ 2.30GHz.
  - Dual-socket 18-core.
  - Memory: 256 GB.
  - Around $6.21X$ speedup compared to Full-tile with accuracy 10^{-5}.



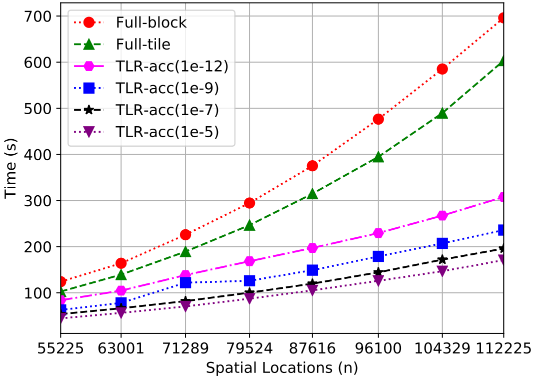
Performance on Shared Memory

- Intel Broadwell,
  - Intel(R) Xeon(R) CPU E5-2680 v4@ 2.40GHz.
  - Dual-socket 14-core.
  - Memory: 128 GB.
  - Around 9.16X speedup compared to full-tile with accuracy 10^{-5}.



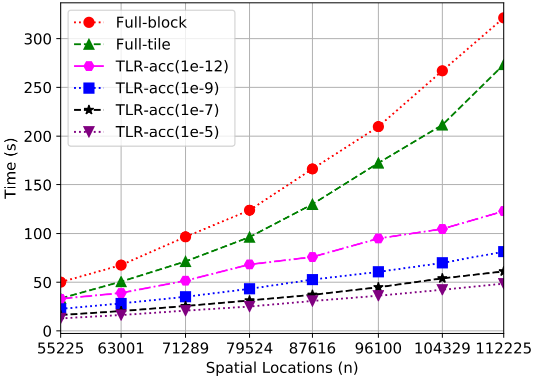
Performance on Shared Memory

- Intel Knights Landing,
  - Intel(R) Xeon Phi(TM) CPU 7210 @ 1.30GHz.
  - Single socket 64-core.
  - Memory: 112 GB.
  - Around 13X speedup compared to full-tile with accuracy 10^{-5}.



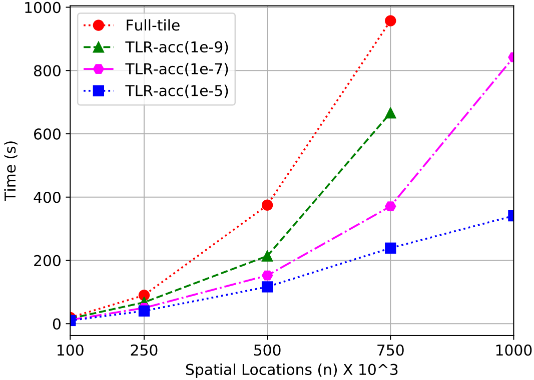
Performance on Shared Memory

- Intel Skylake,
  - Intel(R) Xeon(R) Platinum 8176 CPU @ 2.10GHz.
  - Dual-socket 28-core.
  - Memory: 256 GB.
  - Around $4.48X$ speedup compared to full-tile with accuracy 10^{-5}



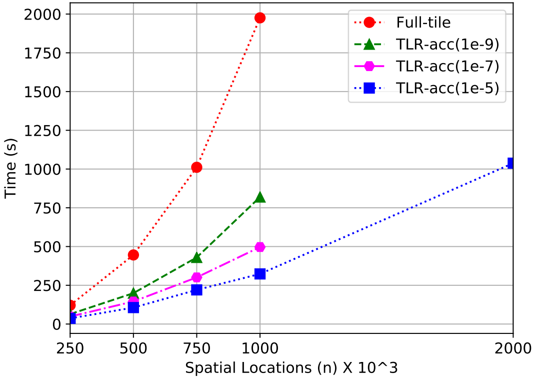
Performance on Distributed Memory

- Shaheen-2,
  - 6174 Intel Haswell Processors.
  - Each processor: dual-socket 16-core.
  - 790 TB of aggregate memory.
  - Around $5X$ speedup compared to Full-tile with accuracy 10^{-5} on 256 nodes.



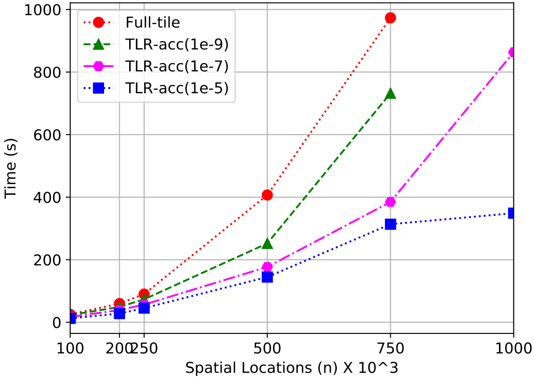
Performance on Distributed Memory

- Shaheen-2,
  - 6174 Intel Haswell Processors.
  - Each processor: dual-socket 16-core.
  - 790 TB of aggregate memory.
  - Around $6X$ speedup compared to Full-tile with accuracy 10^{-5} on 1024 nodes



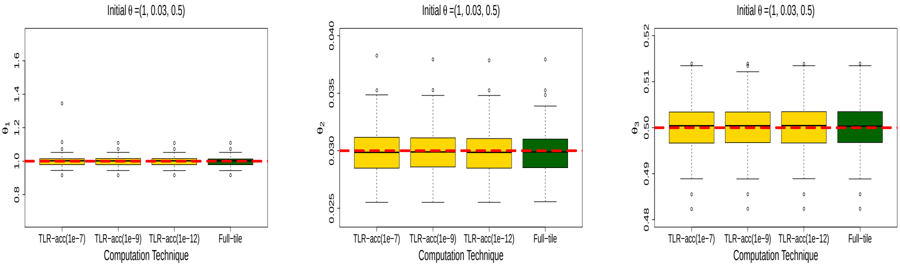
Performance on Distributed Memory (Prediction)

- Shaheen-2,
  - 6174 Intel Haswell Processors.
  - Each processor: dual-socket 16-core.
  - 790 TB of aggregate memory.
  - Around $5X$ speedup compared to Full-tile with accuracy 10^{-5} on 256 nodes



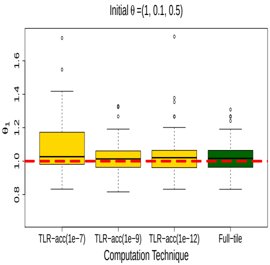
Qualitative Results (Data Correlation Impact)

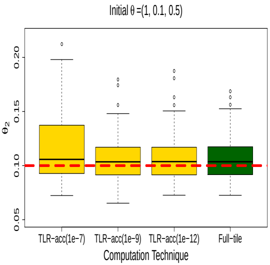
- Synthetic Datasets (40k)
  - Initial (θ)=(1, 0.03, 0.5).
    - Full-tile, TLR w acc. $= 10^{-12}$, TLR w acc. $= 10^{-9}$, and TLR w acc. $= 10^{-7}$

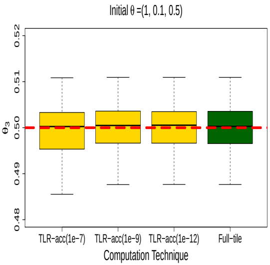


Qualitative Results (Data Correlation Impact)

- Synthetic Datasets (40k)
  - Initial (θ)=(1, 0.01, 0.5).
    - Full-tile, TLR w acc. $= 10^{-12}$, TLR w acc. $= 10^{-9}$, and TLR w acc. $= 10^{-7}$

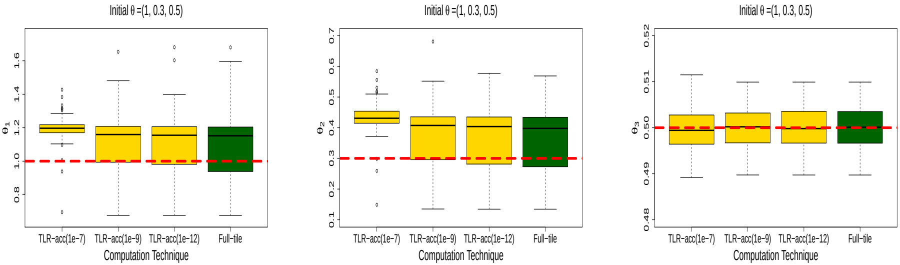






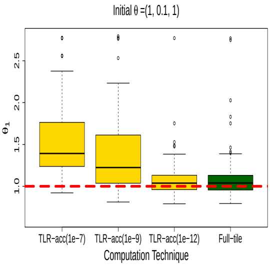
Qualitative Results (Data Correlation Impact)

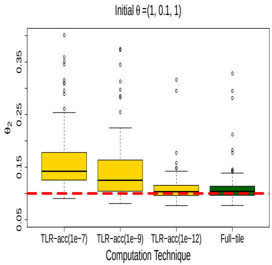
- Synthetic Datasets (40k)
  - Initial (θ)=(1, 0.3, 0.5).
    - Full-tile, TLR w acc. $= 10^{-12}$, TLR w acc. $= 10^{-9}$, and TLR w acc. $= 10^{-7}$

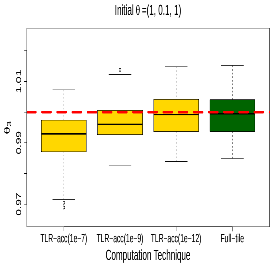


Qualitative Results (Data Smoothness Impact)

- Synthetic Datasets (40k)
  - Initial (θ)=(1, 0.1, 1).
    - Full-tile, TLR w acc. $= 10^{-12}$, TLR w acc. $= 10^{-9}$, and TLR w acc. $= 10^{-7}$

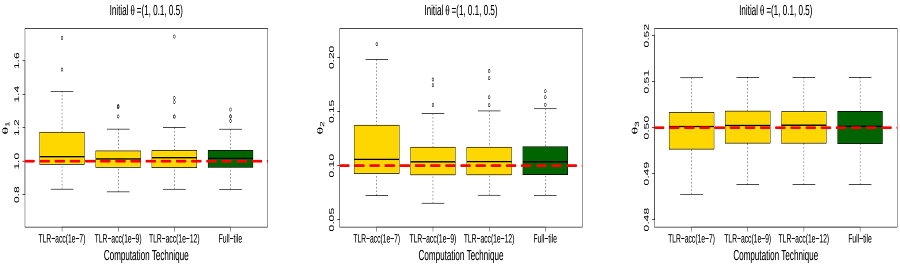






Qualitative Results (Data Smoothness Impact)

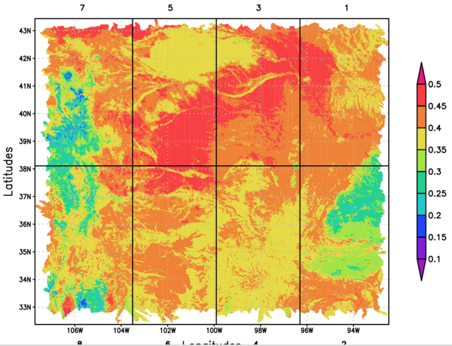
- Synthetic Datasets (40k)
  - Initial (θ)=(1, 0.1, 0.5).
    - Full-tile, TLR w acc. $= 10^{-12}$, TLR w acc. $= 10^{-9}$, and TLR w acc. $= 10^{-7}$



Qualitative Results (Soil Moisture Dataset)

R	Matérn Covariance: Variance (θ_1), TLR Accuracy					Spatial Range (θ_2), TLR Accuracy					Smoothness (θ_3), TLR Accuracy				
	10^{-5}	10^{-7}	10^{-9}	10^{-12}	*Full-tile*	10^{-5}	10^{-7}	10^{-9}	10^{-12}	*Full-tile*	10^{-5}	10^{-7}	10^{-9}	10^{-12}	*Full-tile*
R1	0.855	0.855	0.855	0.855	0.852	6.039	6.034	6.034	6.033	5.994	0.559	0.559	0.559	0.559	0.559
R2	0.383	0.378	0.378	0.378	0.380	10.457	10.307	10.307	10.307	10.434	0.491	0.491	0.491	0.491	0.490
R3	0.282	0.283	0.283	0.283	0.277	11.037	11.064	11.066	11.066	10.878	0.509	0.509	0.509	0.509	0.507
R4	0.382	0.38	0.38	0.38	0.41	7.105	7.042	7.042	7.042	7.77	0.532	0.533	0.533	0.533	0.527
R5	0.832	0.837	0.837	0.837	0.836	9.172	9.225	9.225	9.225	9.213	0.497	0.497	0.497	0.497	0.496
R6	0.646	0.615	0.621	0.621	0.619	10.886	10.21	10.317	10.317	10.323	0.521	0.524	0.524	0.524	0.523
R7	0.430	0.452	0.452	0.452	0.553	14.101	15.057	15.075	15.075	19.203	0.519	0.516	0.516	0.516	0.508
R8	0.661	1.194	0.769	0.769	0.906	18.603	37.315	22.168	22.168	27.861	0.469	0.462	0.467	0.467	0.461

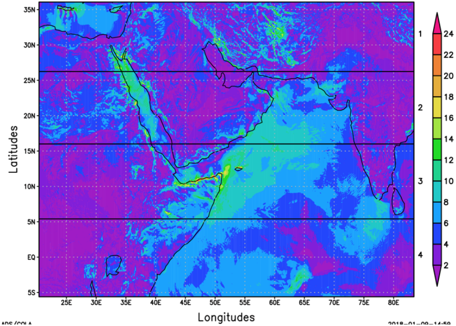
- Highly correlated regions require higher TLR accuracy (ex., regions 7 and 8).



Qualitative Results (Wind Speed Dataset)

R	Matérn Covariance											
	Variance (θ_1)				Spatial Range (θ_2)				Smoothness (θ_3)			
	TLR Accuracy				TLR Accuracy				TLR Accuracy			
	10^{-5}	10^{-7}	10^{-9}	*Full-tile*	10^{-5}	10^{-7}	10^{-9}	*Full-tile*	10^{-5}	10^{-7}	10^{-9}	*Full-tile*
R1	7.406	9.407	12.247	8.715	29.576	33.886	39.573	32.083	1.214	1.196	1.175	1.210
R2	11.920	13.159	13.550	12.517	26.011	28.083	28.707	27.237	1.290	1.267	1.260	1.274
R3	10.588	10.944	11.232	10.819	18.423	18.783	19.114	18.634	1.418	1.413	1.407	1.416
R4	12.408	17.112	12.388	12.270	17.264	17.112	17.247	17.112	1.168	1.170	1.168	1.170

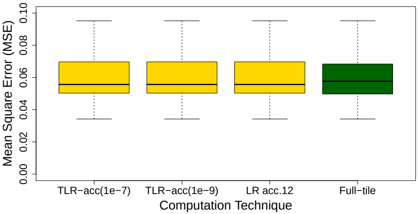
- Highly correlated regions require higher TLR accuracy (ex., regions 1, 2, and 3).



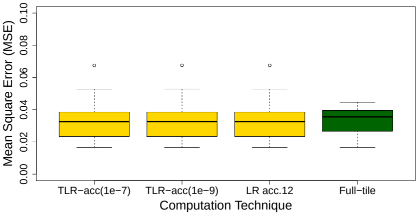
Qualitative Results (Soil Moisture Dataset)

- Prediction Accuracy

- Soil Moisture (Region 1)



- Soil Moisture (Region 3)



ExaGeoStat Under the Microscope

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

ExaGeoStat Under the Microscope

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

ExaGeoStat Under the Microscope

- ExaGeoStat is an open-source software which is available at https://github.com/ecrc/exageostat.
- ExaGeoStatR is available at https://github.com/ecrc/exageostatR.
- ExaGeoStat 0.1.0 (Nov. 9th 2017)

ExaGeoStat Under the Microscope

- ExaGeoStat is an open-source software which is available at https://github.com/ecrc/exageostat.
- ExaGeoStatR is available at https://github.com/ecrc/exageostatR.
- ExaGeoStat 0.1.0 (Nov. 9th 2017)
    - Support exact computation using Chameleon dense Linear algebra library and StarPU runtime system.

ExaGeoStat Under the Microscope

- ExaGeoStat is an open-source software which is available at https://github.com/ecrc/exageostat.
- ExaGeoStatR is available at https://github.com/ecrc/exageostatR.
- ExaGeoStat 0.1.0 (Nov. 9th 2017)
  - Support exact computation using Chameleon dense Linear algebra library and StarPU runtime system.
  - Support real and synthetic geospatial datasets.

ExaGeoStat Under the Microscope

- ExaGeoStat is an open-source software which is available at https://github.com/ecrc/exageostat.
- ExaGeoStatR is available at https://github.com/ecrc/exageostatR.
- ExaGeoStat 0.1.0 (Nov. 9th 2017)
  - Support exact computation using Chameleon dense Linear algebra library and StarPU runtime system.
  - Support real and synthetic geospatial datasets.
  - Soil moisture dataset at Mississippi basin area.

ExaGeoStat Under the Microscope

- ExaGeoStat is an open-source software which is available at https://github.com/ecrc/exageostat.
- ExaGeoStatR is available at https://github.com/ecrc/exageostatR.
- ExaGeoStat 0.1.0 (Nov. 9th 2017)
  - Support exact computation using Chameleon dense Linear algebra library and StarPU runtime system.
  - Support real and synthetic geospatial datasets.
  - Soil moisture dataset at Mississippi basin area.
- Today, ExaGeoStat supports,

ExaGeoStat Under the Microscope

- ExaGeoStat is an open-source software which is available at https://github.com/ecrc/exageostat.
- ExaGeoStatR is available at https://github.com/ecrc/exageostatR.
- ExaGeoStat 0.1.0 (Nov. 9th 2017)
  - Support exact computation using Chameleon dense Linear algebra library and StarPU runtime system.
  - Support real and synthetic geospatial datasets.
  - Soil moisture dataset at Mississippi basin area.
- Today, ExaGeoStat supports,
  - Tile-Low Rank approximation (TLR) using HiCMA TLR approximation.

ExaGeoStat Under the Microscope

- ExaGeoStat is an open-source software which is available at https://github.com/ecrc/exageostat.
- ExaGeoStatR is available at https://github.com/ecrc/exageostatR.
- ExaGeoStat 0.1.0 (Nov. 9th 2017)
  - Support exact computation using Chameleon dense Linear algebra library and StarPU runtime system.
  - Support real and synthetic geospatial datasets.
  - Soil moisture dataset at Mississippi basin area.
- Today, ExaGeoStat supports,
  - Tile-Low Rank approximation (TLR) using HiCMA TLR approximation.
  - Super Diagonal Tile (SDT) approximation.

ExaGeoStat Under the Microscope

- ExaGeoStat is an open-source software which is available at https://github.com/ecrc/exageostat.
- ExaGeoStatR is available at https://github.com/ecrc/exageostatR.
- ExaGeoStat 0.1.0 (Nov. 9th 2017)
  - Support exact computation using Chameleon dense Linear algebra library and StarPU runtime system.
  - Support real and synthetic geospatial datasets.
  - Soil moisture dataset at Mississippi basin area.
- Today, ExaGeoStat supports,
  - Tile-Low Rank approximation (TLR) using HiCMA TLR approximation.
  - Super Diagonal Tile (SDT) approximation.
  - Performance results of TLR-based computations on shared and distributed-memory systems attain up to $13X$ and $5X$ speedups.

ExaGeoStat Under the Microscope

- ExaGeoStat is an open-source software which is available at https://github.com/ecrc/exageostat.
- ExaGeoStatR is available at https://github.com/ecrc/exageostatR.
- ExaGeoStat 0.1.0 (Nov. 9th 2017)
  - Support exact computation using Chameleon dense Linear algebra library and StarPU runtime system.
  - Support real and synthetic geospatial datasets.
  - Soil moisture dataset at Mississippi basin area.
- Today, ExaGeoStat supports,
  - Tile-Low Rank approximation (TLR) using HiCMA TLR approximation.
  - Super Diagonal Tile (SDT) approximation.
  - Performance results of TLR-based computations on shared and distributed-memory systems attain up to $13X$ and $5X$ speedups.
  - Support NetCDF Format.

ExaGeoStat Under the Microscope

- ExaGeoStat is an open-source software which is available at https://github.com/ecrc/exageostat.
- ExaGeoStatR is available at https://github.com/ecrc/exageostatR.
- ExaGeoStat 0.1.0 (Nov. 9th 2017)
  - Support exact computation using Chameleon dense Linear algebra library and StarPU runtime system.
  - Support real and synthetic geospatial datasets.
  - Soil moisture dataset at Mississippi basin area.
- Today, ExaGeoStat supports,
  - Tile-Low Rank approximation (TLR) using HiCMA TLR approximation.
  - Super Diagonal Tile (SDT) approximation.
  - Performance results of TLR-based computations on shared and distributed-memory systems attain up to $13X$ and $5X$ speedups.
  - Support NetCDF Format.
  - Support Out-Of-Core (OOC) execution.

ExaGeoStat Under the Microscope

- ExaGeoStat is an open-source software which is available at https://github.com/ecrc/exageostat.
- ExaGeoStatR is available at https://github.com/ecrc/exageostatR.
- ExaGeoStat 0.1.0 (Nov. 9th 2017)
  - Support exact computation using Chameleon dense Linear algebra library and StarPU runtime system.
  - Support real and synthetic geospatial datasets.
  - Soil moisture dataset at Mississippi basin area.
- Today, ExaGeoStat supports,
  - Tile-Low Rank approximation (TLR) using HiCMA TLR approximation.
  - Super Diagonal Tile (SDT) approximation.
  - Performance results of TLR-based computations on shared and distributed-memory systems attain up to $13X$ and $5X$ speedups.
  - Support NetCDF Format.
  - Support Out-Of-Core (OOC) execution.
  - Wind Speed, Temperature datasets at Middle-East.

Conclusion

- We introduce ExaGeoStat, a unified software for computational Geostatistics that exploits recent developments in dense/approximate linear algebra task-based algorithms associated with dynamic runtime systems. https://github.com/ecrc/exageostat.

Conclusion

- We introduce ExaGeoStat, a unified software for computational Geostatistics that exploits recent developments in dense/approximate linear algebra task-based algorithms associated with dynamic runtime systems. https://github.com/ecrc/exageostat.
- We propose an accurate and amenable MLE framework using TLR-based approximation format to reduce the prohibitive complexity of the apparently dense covariance matrix computation.

Conclusion

- We introduce ExaGeoStat, a unified software for computational Geostatistics that exploits recent developments in dense/approximate linear algebra task-based algorithms associated with dynamic runtime systems. https://github.com/ecrc/exageostat.
- We propose an accurate and amenable MLE framework using TLR-based approximation format to reduce the prohibitive complexity of the apparently dense covariance matrix computation.
- ExaGeoStat estimates the statistical model parameters for Geostatistics applications and predict missing measurements.

Conclusion

- We introduce ExaGeoStat, a unified software for computational Geostatistics that exploits recent developments in dense/approximate linear algebra task-based algorithms associated with dynamic runtime systems. https://github.com/ecrc/exageostat.
- We propose an accurate and amenable MLE framework using TLR-based approximation format to reduce the prohibitive complexity of the apparently dense covariance matrix computation.
- ExaGeoStat estimates the statistical model parameters for Geostatistics applications and predict missing measurements.
- ExaGeoStat relies on a single source code to target various hardware resources including shared and distributed-memory systems.

Conclusion

- We propose a synthetic dataset generator that can be used to perform broader scientific experiments related to computational geostatistics applications.

Conclusion

- We propose a synthetic dataset generator that can be used to perform broader scientific experiments related to computational geostatistics applications.
- We propose an R-wrapper functions, ExaGeoStatR, to facilitate the use of our software in the R environment. https://github.com/ecrc/exageostatR

Conclusion

- We propose a synthetic dataset generator that can be used to perform broader scientific experiments related to computational geostatistics applications.
- We propose an R-wrapper functions, ExaGeoStatR, to facilitate the use of our software in the R environment. https://github.com/ecrc/exageostatR
- We evaluate performance using both synthetic and real datasets in terms of elapsed time and number of floating-point operations (Gflop/s) on several hardware systems.

Conclusion

- We propose a synthetic dataset generator that can be used to perform broader scientific experiments related to computational geostatistics applications.
- We propose an R-wrapper functions, ExaGeoStatR, to facilitate the use of our software in the R environment. https://github.com/ecrc/exageostatR
- We evaluate performance using both synthetic and real datasets in terms of elapsed time and number of floating-point operations (Gflop/s) on several hardware systems.
- We assess the quality of the estimation of the Matérn covariance parameters and prediction operation achieved by ExaGeoStat through a quantitative performance analysis and using both exact and approximation techniques.

KAUST Team/Collaborators/Vendors

- Extreme Computing Research Center @ KAUST: **R. Abdelkhalek, S. Abdulah, K. Akbudak, R. AlOmairy, A. Alonazy, T. Alturkestani, W. Boukaram, A. Charara, N. Doucet, M. Genton, E. A. Gonzalez Fisher, D. Keyes, H. Ltaief, A. Mikhalev, D. Sukkari and Y. Sun**
- Tokyo Institute of Technology, Japan: **R. Yokota**
- Innovative Computing Laboratory @ UTK, USA: **PLASMA/MAGMA/PaRSEC Teams**
- INRIA/INP/LaBRI Bordeaux, France: **Runtime/HiePACS Teams**
- Max-Planck Institute@Leipzig, Germany: **R. Kriemann**
- American University of Beirut, Lebanon: **G. Turkiyyah**
- KAUST Supercomputing Lab
- Intel Parallel Computing Center
- Cray Center of Excellence

Intel Parallel Computing Centers

THE SUPERCOMPUTER COMPANY

Thank You!

Questions?

sameh.abdulah@kaust.edu.sa