Massively Parallel Likelihood Function Optimization to Accelerate Air Pollution Prediction on Large-Scale Systems

MARY LAI O. SALVANA*, SAMEH ABDULAH, HATEM LTAIEF, YING SUN, MARC G. GENTON, and DAVID E. KEYES, Extreme Computing Research Center (ECRC), King Abdullah University of Science and Technology, Saudi Arabia

Gaussian geostatistical space-time modeling is an effective tool for performing statistical inference of field data evolving in space and time, generalizing spatial modeling alone at the cost of greater complexity of operations and storage, and pushing geostatistical modeling even further into the arms of high performance computing. It makes inferences for missing data by leveraging space-time measurements of one or more fields. We propose a high-performance implementation of a widely applied space-time modeling method for large-scale systems, using a two-level parallelization technique. At the inner level, we rely on state-of-the-art dense linear algebra libraries and parallel runtime systems to perform complex matrix operations required in modeling and prediction operations using maximum likelihood estimation (MLE), i.e., the Cholesky factorization of the Gaussian space-time covariance matrix. At the outer level, we parallelize the optimization process using a distributed implementation of the particle swarm optimization (PSO) algorithm. At this level, parallelization is accomplished using MPI sub-communicators, where nodes in each sub-communicator perform a single MLE iteration at a time. We evaluate the performance and the accuracy of the proposed implementation using synthetic datasets and a real particulate matter (PM) dataset illustrating the application of the technique to air pollution. We achieve up to 24.45, 49.70, 100.06, 189.67, 369.22, and 757.16 TFLOPS/s using 32, 64, 128, 256, 512, and 1024 nodes, respectively, of a Cray XC40 system, with an average of 60% of the peak performance on 1024 nodes with 490K problem size.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Mary Lai O. Salvana, Sameh Abdulah, Hatem Ltaief, Ying Sun, Marc G. Genton, and David E. Keyes. XX. Massively Parallel Likelihood Function Optimization to Accelerate Air Pollution Prediction on Large-Scale Systems. In *XX*. ACM, New York, NY, USA, 19 pages. https://doi.org/XX

1 INTRODUCTION

Geostatistical inference has emerged as an important complement to physics-based modeling of spatio-temporal phenomena through, e.g., discretized partial differential equations, because it is widely applicable and straightforward to use with sufficient data. Gaussian models embed correlations arising from causality and are adequate for many phenomena. Pollutant dispersion models are representative of the complex conservation law models that require many assumed inputs, including physical parameters, initial conditions, and boundary conditions, that non-expert users may find complex to supply. Meanwhile, observational data are becoming more available, especially with the availability of massively deployed sensors, drones, and satellites. The challenges to geostatistical models are the cubic growth of computational complexity and the quadratic growth of storage complexity in the number of observations correlated through the likelihood evaluation process. This complexity is linked to the Cholesky factorization operation required to compute the inverse of the application-associated covariance matrix. The model fitting (MLE) is an iterative

procedure that requires the log-likelihood evaluation at each iteration with a cubic complexity. As shown herein, these are addressable through data sparsity and multi-level parallelism.

The dangers of high levels of air pollution, in particular, the mixture of solid particles and liquid droplets known as particulate matter or PM, have been revealed in a broad range of studies. High concentrations of PM can cause cardiovascular, respiratory, and Parkinson's diseases and decrease life expectancy [20, 30, 33, 51]. PM also has devastating effects on the environment, including disturbances in nutrient cycling [19], cloud formation, solar radiation, the occurrence of acid rain, and acceleration of climate change [14]. Through its long-range transport patterns, PM can travel to other locations and spread infectious diseases [21]. Understanding and predicting the evolution of PM is a key to mitigating its effects. The modeling of PM concentrations is an active field of research with computational models involving particle concentration, momentum, and turbulence equations. These models need to be fed with various model input parameters [24]. Lack of domain expert knowledge impedes the adoption of these models by non-experts, who may lack the knowledge required to supply modeling requirements. Thus, the development of simpler and less expensive data-driven models, such as linear regression [27], time-series [26], Kalman filter [31], and machine learning [54] is paramount. With the advent of advanced data collection technologies, modelers are provided with a continuous stream of data. These data are often geographically referenced. By nature, geographically referenced data may be expected to exhibit some degree of space-time dependence which can be described by a space-time covariance function.

Gaussian geostatistical space-time models offer a way to model space-time data by considering them as realizations of a space-time random field. Denote the measurement obtained at space-time location $(\mathbf{s}, t) \in \mathbb{R}^d \times \mathbb{R}, d \geq 1$, as $Z(\mathbf{s}, t)$, such that the elements in the vector of measurements $\mathbf{Z} = \{Z(\mathbf{s}_1, t_1), Z(\mathbf{s}_2, t_2), \ldots, Z(\mathbf{s}_n, t_n)\}^\top$, where $n \in \mathbb{Z}^+$ is the number of space-time locations, come from the distribution of $Z(\mathbf{s}, t)$ with mean, $\mathrm{E}\{Z(\mathbf{s}, t)\} = \mu(\mathbf{s}, t)$, and covariance, $\mathrm{cov}\{Z(\mathbf{s}_1, t_1), Z(\mathbf{s}_2, t_2)\} = C(\mathbf{s}_1, \mathbf{s}_2; t_1, t_2|\boldsymbol{\theta})$. Here $\mu(\mathbf{s}, t)$ is the mean function that models the deterministic part of the space-time random field and $C(\mathbf{s}_1, \mathbf{s}_2; t_1, t_2|\boldsymbol{\theta})$ is a space-time covariance function, parameterized by $\boldsymbol{\theta} \in \mathbb{R}^q, q \geq 1$, that describes the strength of dependence of the measurements at any two space-time locations $(\mathbf{s}_1, t_1)$ and $(\mathbf{s}_2, t_2)$. A space-time random field is second-order stationary in space and time when the covariance depends only on the space-time lag $(\mathbf{h}, u)$. That is, suppose $Z(\mathbf{s}, t)$ is a second-order stationary space-time random field. Then its covariance is given by $C(\mathbf{s}_1, \mathbf{s}_2; t_1, t_2|\boldsymbol{\theta}) = C(\mathbf{h}, u|\boldsymbol{\theta})$, where $\mathbf{h} = \mathbf{s}_1 - \mathbf{s}_2$ and $u = t_1 - t_2$. The parametric space-time covariance function C can be chosen among many established models in the Gaussian geostatistical literature [9, 17]. A typical workflow in Gaussian geostatistical space-time modeling involves estimating the parameter vector $\boldsymbol{\theta}$ given space-time measurements $\mathbf{Z}$. This is often done via the maximum likelihood estimation (MLE). Other approaches involve moment-based estimation techniques, which require the following empirical covariance function:

$$\hat{C}(\mathbf{h}, u) = \frac{1}{|\mathcal{S}_\mathbf{h}|} \frac{1}{|\mathcal{T}_u|} \sum_{(\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{S}_\mathbf{h}) \cap (t_1, t_2 \in \mathcal{T}_u)} Z(\mathbf{s}_1, t_1) Z(\mathbf{s}_2, t_2),$$

where $\mathcal{S}_\mathbf{h} = \{\mathbf{x}, \mathbf{y} | \mathbf{x} - \mathbf{y} = \mathbf{h}\}$, $\mathcal{T}_u = \{a, b | a - b = u\}$, and $|\mathcal{S}_\mathbf{h}|$ and $|\mathcal{T}_u|$ are the cardinalities of the sets $\mathcal{S}_\mathbf{h}$ and $\mathcal{T}_u$. The parameter estimates under moment-based estimation are then obtained by minimizing the objective function:

$$Q(\boldsymbol{\theta}) = \sum_{\mathbf{h} \in \mathcal{H}} \sum_{u \in \mathcal{U}} \{\hat{C}(\mathbf{h}, u) - C(\mathbf{h}, u|\boldsymbol{\theta})\}^2,$$

with respect to $\boldsymbol{\theta}$, where $\mathcal{H}$ and $\mathcal{U}$ are the sets of the spatial and temporal lags. However, in literature, the moment-based estimation techniques show less efficiency compared to a more reliable method like the MLE and its parameter estimates are highly sensitive to the lags included in $\mathcal{H}$ and $\mathcal{U}$. MLE remains the gold standard for statistical inference owing to its desirable properties, namely, efficiency, consistency, and asymptotic normality, under general conditions [35, 36, 39].

An extensive simulation study comparing both approaches can be found in [53], where MLE is shown to outperform moment-based estimation.

In MLE, the Gaussian log-likelihood function

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

is maximized with respect to $\boldsymbol{\theta}$. Here $\Sigma(\boldsymbol{\theta}) = \{C(\mathbf{s}_i, \mathbf{s}_j; t_i, t_j|\boldsymbol{\theta})\}_{i,j=1}^{n}$ is the $n \times n$ covariance matrix of $\mathbf{Z}$ and $|\Sigma(\boldsymbol{\theta})|$ is the determinant of $\Sigma(\boldsymbol{\theta})$. The covariance matrix $\Sigma(\boldsymbol{\theta})$ represents the correlation between different locations based on a given geostatistical covariance function, herein, the space-time covariance function.

Once the maximum likelihood estimates, $\hat{\boldsymbol{\theta}}$, are obtained, prediction, also known as kriging, can be done by recognizing that the vector of known space-time measurements $\mathbf{Z}_1$ and the vector of missing space-time measurements $\mathbf{Z}_2$ are jointly Gaussian, i.e.,

$$\begin{bmatrix}\mathbf{Z}_1\\ \mathbf{Z}_2\end{bmatrix} \sim \mathcal{N}_{n_1+n_2}\left(\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11}(\boldsymbol{\theta}) & \Sigma_{12}(\boldsymbol{\theta})\\ \Sigma_{21}(\boldsymbol{\theta}) & \Sigma_{22}(\boldsymbol{\theta})\end{bmatrix}\right), \tag{2}$$

where n_1 and n_2 are the number of given and missing space-time locations, respectively, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors of $\mathbf{Z}_1$ and $\mathbf{Z}_2$, respectively, Σ_{11} and Σ_{22} are the covariance matrices of $\mathbf{Z}_1$ and $\mathbf{Z}_2$, respectively, and $\Sigma_{12} = \Sigma_{21}^\top$ is the cross-covariance matrix of $\mathbf{Z}_1$ and $\mathbf{Z}_2$. The conditional distribution of the missing space-time measurements $\mathbf{Z}_2$ is:

$$\mathbf{Z}_2|\mathbf{Z}_1 \sim \mathcal{N}_{n_2}\{\boldsymbol{\mu}_2 + \Sigma_{21}(\boldsymbol{\theta})\Sigma_{11}^{-1}(\boldsymbol{\theta})(\mathbf{Z}_1 - \boldsymbol{\mu}_1), \Sigma_{22}(\boldsymbol{\theta}) - \Sigma_{21}(\boldsymbol{\theta})\Sigma_{11}^{-1}(\boldsymbol{\theta})\Sigma_{12}(\boldsymbol{\theta})\}. \tag{3}$$

Under the zero-mean assumption, the best linear unbiased predictor of $\mathbf{Z}_2$ is:

$$\hat{\mathbf{Z}}_2 = \Sigma_{21}(\hat{\boldsymbol{\theta}})\Sigma_{11}^{-1}(\hat{\boldsymbol{\theta}})\mathbf{Z}_1, \tag{4}$$

with prediction uncertainty

$$\hat{\mathbf{V}}_2 = \mathrm{diag}\{\Sigma_{22}(\hat{\boldsymbol{\theta}}) - \Sigma_{21}(\hat{\boldsymbol{\theta}})\Sigma_{11}^{-1}(\hat{\boldsymbol{\theta}})\Sigma_{12}(\hat{\boldsymbol{\theta}})\}. \tag{5}$$

The grand challenge in large-scale Gaussian geostatistical modeling lies in the inversion of $\Sigma(\boldsymbol{\theta})$. The MLE operation also involves an iterative optimization process to evaluate the likelihood function and find the optimum set of parameters that satisfies the global maximum likelihood value. We adopt a parallel implementation of the particle swarm optimization algorithm (PSO) [50], which belongs to the evolutionary computing algorithms. The PSO algorithm depends on the interaction between independent particles to scan the search space of the problem to find the globally optimum solutions [57]. It has several advantages, including its low computational cost and amenability to parallelization on large-scale systems. The implementation in [50] incorporates the pattern search method to the PSO algorithm to generate a more aggressive version of the optimization process that allows more accurate estimations of the required optimum solutions.

In this work, we use a two-level parallelization technique to perform the MLE operation in space-time dimensions. The inner level involves matrix operations of the likelihood function using parallel task-based linear solvers. The higher level involves the parallel PSO (PPSO) algorithm to distribute the optimization process between different numbers of nodes with the aid of the Message Passing Interface (MPI) API.

The remainder of the paper is structured as follows. Section 2 articulates the contributions of this paper. Section 3 gives a brief background on space-time covariance functions, the PSO algorithm, and MPI communicators. Section 4 highlights related works. Section 5 contains a detailed description of the proposed implementation. Section 6 reports the performance results in terms of accuracy/time-to-solution and we conclude in Section 7.

2 CONTRIBUTIONS

This work takes existing Gaussian geostatistical space-time models to the realm of large-scale space-time modeling using a two-level parallelization technique. The five-fold contributions of the paper are as follows:

- We present a high-performance implementation of the Gaussian geostatistical space-time modeling on large-scale systems involving the log-likelihood function.
- We incorporate the Parallel PSO algorithm to the MLE operation to utilize the execution performance on distributed environments.
- We exploit the MPI communicators paradigm to create a set of independent executions for the log-likelihood function that allows massively parallel execution of the MLE operation while relying on task-based parallel linear solvers to perform the log-likelihood matrix operations.
- We are able to achieve up to 24.45, 49.70, 100.06, 189.67, 369.22, and 757.16 TFLOPS/s using 32, 64, 128, 256, 512, and 1024 nodes, respectively, on a Cray XC40 system with an average of 60% of the peak performance on these numbers of nodes.
- We perform a set of experiments to assess accuracy in terms of inference and prediction on both synthetic datasets and real datasets.
- We apply our new implementation on two air pollution datasets from Middle East and US regions. The evaluation shows promising results of the proposed approach in satisfying high prediction accuracy in such applications.

3 BACKGROUND OF THE STUDY

In this section, we give a brief overview of space-time covariance functions, the PSO algorithm, and MPI communicators.

3.1 Space-Time Covariance Functions

Research in developing geostatistical space-time covariance function models is intensive, given its crucial role in prediction and uncertainty quantification. The space-time covariance function is used to generate a positive definite covariance matrix whose parameters can be calibrated using the log-likelihood or MLE function with a memory complexity of $O(n^2)$ and computation complexity of $O(n^3)$, where n represents the dimension of the covariance matrix. A popular space-time covariance function model proposed in [16] has the form:

$$C(\mathbf{h}, u) = \frac{\sigma^2}{|u|^{2\alpha}/a_t + 1} \mathcal{M}_\nu \left\{ \frac{\|\mathbf{h}\|/a_s}{(|u|^{2\alpha}/a_t + 1)^{\beta/2}} \right\}, \tag{6}$$

where $\mathcal{M}_\nu$ is the univariate Matérn correlation function parameterized with ν, $\sigma^2 > 0$ is the variance parameter, $\nu > 0$ and $\alpha \in (0, 1]$ are the smoothness parameters in space and time, respectively, $a_s, a_t > 0$ are the range parameters in space and time, respectively, and $\beta \in [0, 1]$ is the space-time interaction parameter. When $\beta = 0$, the model is classified as a separable model such that it factors into its purely spatial and purely temporal components, which implies independence of the space and time aspects. This means that the dependence structure in space is modeled separately from dependence structure in time. Nonzero values of β imply that the dependence structure in space relates to some degree with the dependence structure in time. The resulting models when $\beta > 0$ are termed nonseparable. In Figure 1, we show two simulated space-time random fields springing from the nonseparable ($\beta = 1$) and separable ($\beta = 0$) versions of the model in (6). The nonseparable and separable models have the same parameter values except for β. From the figure, one can see that the random fields are identical at $t = 1$, however, they evolve in different ways through time.

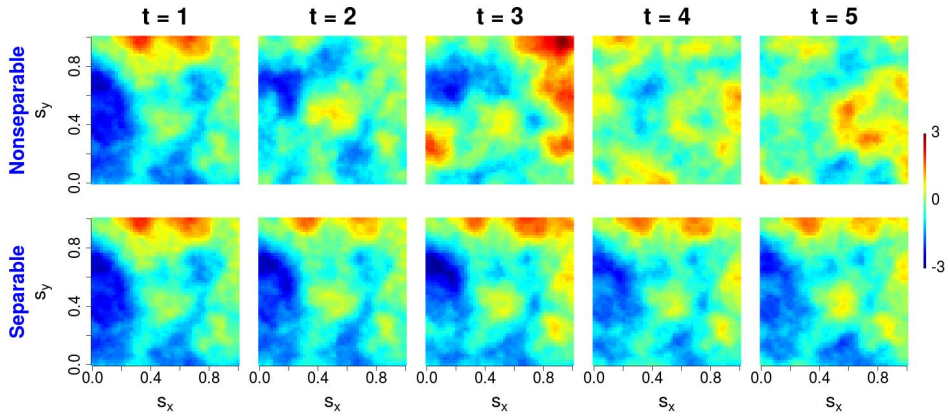


Fig. 1. Simulated space-time realizations from the model in (6) with different values of the space-time interaction parameter, β. The nonseparable and separable models correspond to $\beta = 1$ and $\beta = 0$, respectively.

Because of the flexibility of the space-time model in (6), we implement this for large-scale space-time modeling for air pollution.

3.2 Particle Swarm Optimization

The Particle Swarm Optimization (PSO) algorithm belongs to the class of evolutionary computing algorithms. The evolutionary computing paradigm, inspired by the functionality and structure of living entities from the same taxonomy, has been successfully applied to many scientific fields. Although single living entities have no perception of a high-level goal to achieve, non-organized low-level goals achieved through different entities lead to a global solution for a complex problem. There are several examples of applying the evolutionary computing paradigm to solve scientific challenges such as genetic algorithms [46], simulated annealing [1, 13], differential evolution [37, 42], and PSO [22, 41]. PSO is a heuristic global optimization method that mimics biological swarms' social learning, for instance, a shoal of fish or flock of birds. The search space includes a population of candidate solutions, i.e., a set of candidate parameter vectors. The PSO algorithm considers every single solution as a particle that aims to move with a certain speed to scan the global minimum solution's search space. The particle is associated with a velocity vector that directs the next step of the particle based on the current position and the direction of the best particle position. The PSO algorithm is conceptually simple and compatible with parallelization. This makes PSO a prevalent choice for scientific applications. However, some limitations should be emphasized, such as local optima convergence and stagnation, where swarms stagnate before reaching the global minimum, and the swarms remain highly diverse. Gaussian geostatistical space-time modeling involves an optimization process to solve a nonconvex estimation problem by maximizing the log-likelihood function. Herein, the PSO algorithm is applied to solve the optimization process in parallel by relying on an existing parallel implementation of the PSO algorithm, i.e., PPSwarm [50]. This implementation incorporates the pattern search algorithm [25] into the PSO algorithm to generate a more aggressive search method that can remove the limitations of the original algorithm.

3.3 MPI Communicators

The MPI API is the predominant programming method in developing HPC applications on large-scale systems [38]. It relies on exchanging data between distributed-memory processing units through a message passing. Distributed parallel applications usually require collective operations that involve a group of processes communicating in an isolated context. Thus, MPI defines a set of collective routines to perform group communications over distributed architectures, for instance, *MPI_Bcast*, *MPI_Reduce*, and *MPI_Gather*. However, to perform such collective routines between a set of processes, they should share the same MPI communicator. The MPI communicator manages the intra-communicator operations messages between different nodes within the same MPI session. The MPI default communicator is *MPI_COMM_WORLD*, which involves all the initiated MPI processes defined by the user. This default communicator can be split into x sub-communicators where processes in each sub-communicator (x_i) can exchange data through the collective operations associated with this MPI sub-communicator. The *MPI_Comm_split* MPI function is a common way to partition the *MPI_COMM_WORLD* communicator into disjoint subgroups associated with different sub-communicators. Each group of processes has one value of color so that each sub-communicator contains all processes of the same color. In this work, we exploit the MPI communicator/sub-communicators to perform a high-level parallelization, to optimize the MLE process, and to effectively utilize the underlying hardware resources.

4 RELATED WORK

Although space-time models are constantly being developed, research on scaling these models to handle large volumes of space-time data is lagging behind. Large-scale Gaussian geostatistical modeling has been attempted in the past but mostly in the purely spatial setting, i.e., not including the temporal dimension. GPUs have been widely used to accelerate the kriging operation in spatial applications. However, most of the contributions were in reducing the amount of data transfer time between the GPU, and the hosted CPU [10, 11, 29, 52]. Due to the GPU memory limits, all their experiments were performed using datasets with only a few thousand spatial locations. In [56], another GPU-based implementation was proposed with the main contribution of enabling stream-processing. In [2], a unified exascale geospatial statistics software was proposed. The software is supported with advanced linear algebra solvers and dynamic runtime systems to enable high-performance spatial data processing. The parallelization is performed in the MLE process in the purely spatial context. Because of the added temporal dimension, extending these to space-time is not trivial. A few existing space-time implementations include [55] using OpenCL language with a separable space-time covariance function. In [12], daily pollen levels were reconstructed using the space-time inverse distance weighted (ST-IDW) interpolation approach parallelized on concurrent CPUs in a shared-nothing manner.

Air pollution prediction has been performed in the literature through a variety of means. Often, aerodynamic analyses are employed, wherein a computational model is initiated to model the atmospheric pollution concentrations. This process requires intensive computations and an in-depth understanding of physical models. The Community Multi-scale Air Quality (CMAQ) [43], the Nested Air Quality Prediction Modeling System (NAQPMS) [15], and the WRFChem [45] are some examples of these computational models. Although these models are invaluable tools, they have several drawbacks. For instance, they require defining many domain-specific parameters that only domain scientists can obtain and understand. Statistical methods, which thrive in finding associations and correlations within the dataset, have an advantage compared to the physics-based modeling approaches because practitioners need not first become experts about the process being modeled [23]. Several statistical tools have been used in literature to perform weather prediction with high accuracy [3, 18, 32, 44, 48]. Several works in literature use statistical methods to perform prediction on the

air pollution problem [7, 28, 34, 40]. This work provides a framework upon which Gaussian geostatistical space-time models can be used to perform large-scale air pollution predictions.

5 PARALLEL MODELING FRAMEWORK

This section provides a detailed description of the proposed parallel framework through a two-level parallelization technique. At the inner level, parallel linear solvers and dynamic runtime systems are used through a set of task-based algorithms to parallelize the MLE operation. At the outer level, the PPSwarm optimization algorithm is used to distribute the optimization process between a set of MPI sub-communicators. The following two subsections describe in detail our proposed parallel MLE framework.

5.1 Parallel Likelihood Function Estimation

High-performance geostatistics modeling software has been proposed in [2] with three main components: a synthetic data generator, an MLE modeling tool, and a predictor. This software mainly relies on `Chameleon` state-of-the-art linear algebra library [8] and `StarPU` dynamic runtime system [6]. Through a set of task-based linear solvers, a set of high-level geostatistical operations were implemented, mainly synthetic spatial data generation, the log-likelihood maximization, and spatial prediction functions. The underlying linear solvers are supported by tile-based algorithms, where the covariance matrix is split into a set of small tiles and operate directly on these tiles [4]. Compared to the block-based algorithms, which were widely used in parallel linear algebra software, tile-based algorithms enable a fine-grained look-ahead mechanism to the trailing submatrix and relaxes artifactual loop-ordered and subroutine boundary synchronization points encountered.

Task-based linear solvers rely on dependence analyses derived directly from the serial code and represented in Directed Acyclic Graphs (DAGs). Each DAG defines the tasks through nodes and the data dependencies through arrows. An example of a 4×4 matrix Cholesky factorization DAG is shown in Figure 2 with four tile-based operations: POTRF, TRSM, SYRK, and GEMM. DAGs can be sequentially coded and passed to the underlying dynamic runtime system to enable fine-grain execution on the underlying hardware architecture. In [2], `StarPU` dynamic runtime system is used to support the tile-based linear solvers and is preferred due to its broad hardware architecture support (shared-memory, GPUs, and distributed memory systems) [5]. `StarPU` can directly manage sequential coded DAGs to schedule tasks to the available hardware resources. The `StarPU` internal scheduler ensures the utilization of the resources and the tasks' ordering to prevent possible access violation. `StarPU` provides a high level of abstraction to improve user productivity and creativity. We adopt task-based parallelism in the MLE operation through the `Chameleon` parallel linear algebra library and `StarPU` runtime system in the proposed implementation.

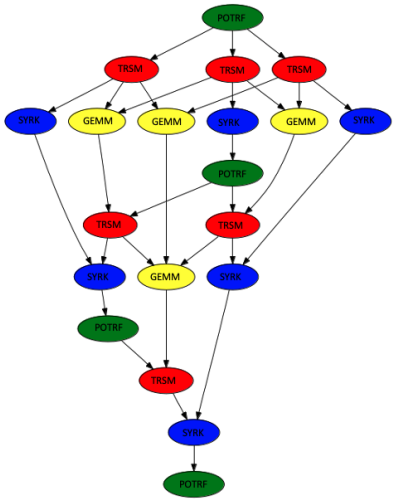
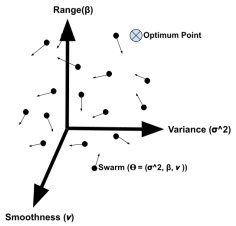
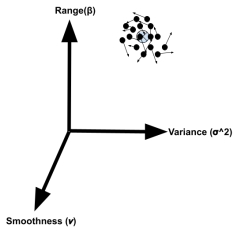


Fig. 2. 4×4 Cholesky factorization DAG with four operations, namely, positive-definite triangular Cholesky factorization (POTRF), triangular solve matrix (TRSM), symmetric rank-K update (SYRK), and general matrix-matrix multiply (GEMM).



(a) Iteration 0



(b) Iteration N

Fig. 3. PSO-based MLE optimization of a univariate purely spatial covariance function.

5.2 Proposed Parallel Optimization Strategy

Although the state-of-the-art parallel linear algebra solvers provide a high level of parallelization and scalability on large systems, enough workload should be guaranteed for all processing units, which cannot be satisfied because of the memory limits. Thus, we propose to pick the required number of nodes based on the problem size. This should satisfy the higher performance requirement and avoid memory size restrictions. Further scalability can be achieved by parallelizing the optimization process on a larger number of nodes. In MLE, $\widehat{\boldsymbol{\theta}} = (\widehat{\theta_1}, ..., \widehat{\theta_q})^\top$ represents the parameter vector of the maximum likelihood estimates which maximizes the log-likelihood function. Finding the maximum value of the log-likelihood function requires several optimization iterations over a pool of parameter vectors. In the proposed framework, we rely on the PPSwarm software, a parallel implementation of the PSO algorithm mentioned in Section 3.2 [50]. The parallel implementation utilizes several particles to scan the search space simultaneously, such that each particle can find the maximum value of the log-likelihood function using a single parameter vector on one MPI node. Thus, a single MPI communicator is used to organizing the communication between different nodes, i.e., *MPI_COMM_WORLD*. We incorporate the PPSwarm algorithm with the runtime-based MLE implementation. Assuming a number of MPI nodes under the default MPI communicator, we split the *MPI_COMM_WORLD* communicator into a set of sub-communicators where each sub-communicator can evaluate a single log-likelihood function solution. In this way, we can evaluate several log-likelihood functions in parallel while keeping the MLE operation's low-level parallelization. Figure ?? shows how different particles move from iteration 0 to iteration N to find the optimum point in the search space. Each particle exclusively evaluates one log-likelihood function.

Figure 4 depicts the two-level node splitting operation. Assuming that a set of 32 MPI nodes is used, the *MPI_Comm_split* function can be used to split the default *MPI_COMM_WORLD* communicator into a set of sub-communicators, as shown by the figure. Each sub-communicator includes the nodes that will internally communicate to evaluate a single MLE solution with the aid of the underlying parallel linear algebra library and the dynamic runtime system.

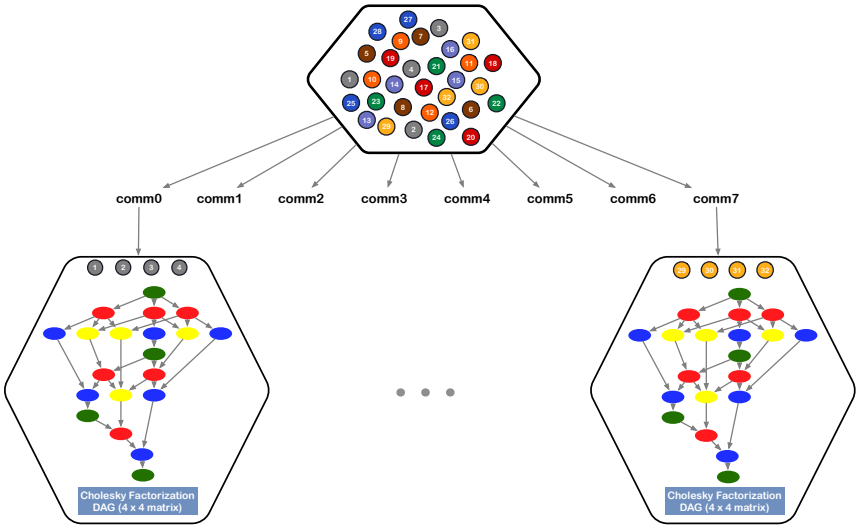


Fig. 4. The proposed framework using 32 nodes and 8 MPI sub-communicators. Each sub-communicator includes 4 nodes that estimate the log-likelihood function with a certain set of parameters in parallel using the StarPU runtime system.

6 PERFORMANCE RESULTS AND ANALYSIS

In this section, the performance and the correctness of the proposed two-level parallelization technique are assessed using both synthetic and real space-time datasets. We generated synthetic datasets with different spatial and temporal dependence profiles which illustrate a variety of cases that can appear in real data. We also obtained two real PM2.5 datasets each with more than 400, 000 space-time locations to validate the prediction accuracy of our implementation. The performance is tested on an Intel-based Cray XC40 system with 6,174 compute nodes, each of which has two 16-core Intel Haswell CPUs at 2.30 GHz and 128 GB of memory.

6.1 Qualitative Analysis Using Synthetic Data

In this section, we illustrate the merits of our proposed framework on two fronts: 1) we show the advantage of using nonseparable models over separable models, and 2) we test the accuracy of our implementation in recovering the parameters of the nonseparable model in (6).

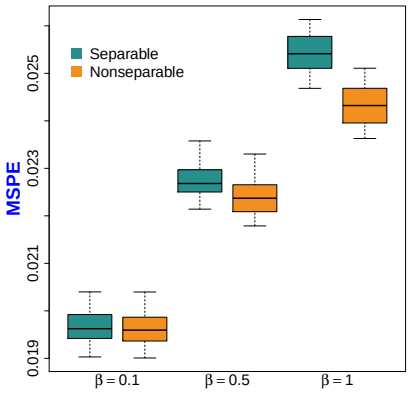


Fig. 5. Boxplots of the prediction errors when fitting a separable and nonseparable model on space-time data with varying degrees of space-time interaction. Weak, moderate, and strong space-time interactions are represented by $\beta = 0.1, 0.5, 0.9$, respectively.

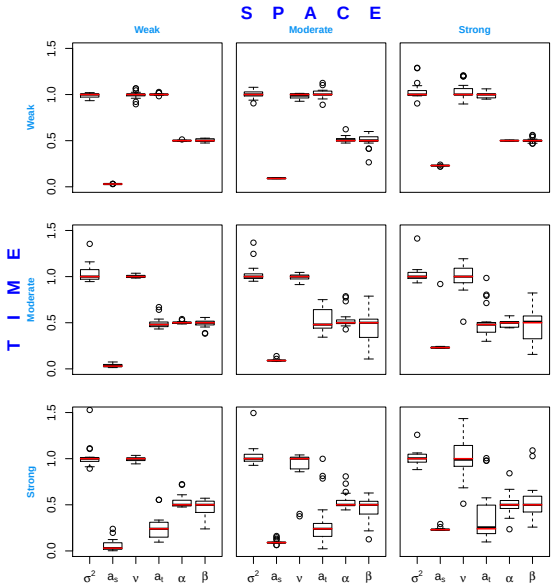


Fig. 6. Boxplots of parameter estimates under varying degrees of space-time dependence. Weak, moderate, and strong dependence in space and time corresponds to $a_s = \{0.03, 0.09, 0.23\}$ and $a_t = \{1, 0.48, 0.24\}$, respectively. The true parameters are highlighted in red.

Even though separable models have a computational advantage over nonseparable models, such that the full covariance matrix resulting from a separable model is simply a Kronecker product of smaller covariance matrices in space and time, they are limited in the scope of space-time dynamics they can model. On the other hand, nonseparable models introduce some sophistication in the dynamics between space and time components through the space-time interaction parameter, β. We perform some numerical experiments to illustrate the strength of nonseparable models clearly. We simulate space-time realizations at 400 spatial locations and 100 temporal locations from the model in (6) for each values of $\beta \in \{0.1, 0.5, 0.9\}$. These datasets are representative of space-time interactions, from weak to strong. To each simulated dataset, we fit both a separable ($\beta = 0$) and a nonseparable model ($\beta \geq 0$) on the training set (90%) using local optimization and perform predictions on the testing set (10%). We perform this simulation, estimation, and prediction for 100 rounds and plot the resulting prediction errors, computed using the mean squared prediction error (MSPE) metric, as boxplots for each scenario. Figure 5 shows the results of the experiments. From the results, one can see that both separable and nonseparable models are at par in prediction performance when the interaction in space-time is weak, i.e., $\beta = 0.1$. The difference between the two models appears when they are tested in moderate and strong scenarios. As expected, the separable model cannot adequately capture non-negligible space-time interaction.

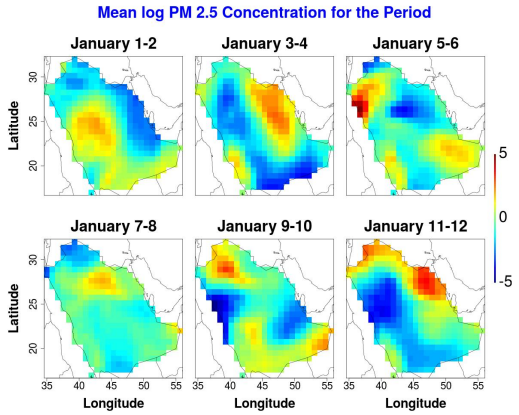


Fig. 7. Visualization of the log PM2.5 dataset after space-time mean removal at the first six time points in 2016.

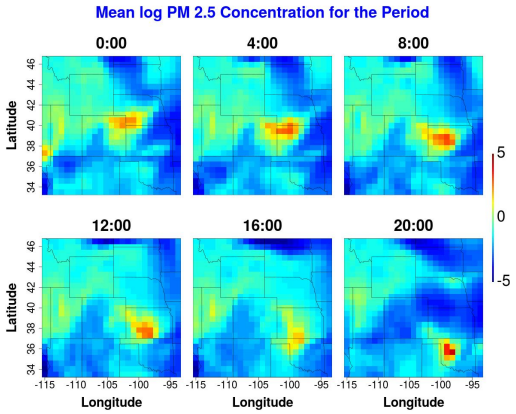


Fig. 8. Visualization of the log PM2.5 dataset after space-time mean removal at four hour intervals on January 1, 2016 over the Midwest US.

Since a nonseparable model is preferable due to its flexibility, we check whether the parallel implementation can recover all the parameters of such a model. In this set of experiments, we simulate realizations from the nonseparable model in (6) with varying degrees of space-time dependence. We generate space-time random fields from combinations of weak, moderate, and strong dependence in space and time using different values of a_s and a_t, the parameters responsible for the long-range dependence in space-time, respectively. Figure 6 presents the values of the parameter estimates for every dependence scenario. The median values of the parameter estimates match with the true parameter values. This shows that the PSO can satisfactorily perform in any combination of space-time dependence. The boxplots being wider in scenarios with strong dependence in either space and/or time may be due to having fewer data points with separation distance in space and/or time that are larger than the spatial or temporal effective ranges, which are the distances at which the spatial or temporal correlation drops to approximately 0.05. Adding more spatial and temporal locations can resolve this.

6.2 Qualitative Analysis Using Real PM2.5 Data

In this section, we apply the proposed framework on real space-time datasets downloaded from the NASA Earthdata website. The datasets are Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) reanalysis datasets comprised of hourly PM2.5 measurements. PM2.5 are PMs with an aerodynamic diameter of $2.5\mu m$ or less. They usually come from roads, agricultural production, construction activities, and open burning of biomass [49]. The first dataset includes the residuals of the log PM2.5 averaged over 48 hours in Saudi Arabia from 2016 to 2019. Such data transformation ensures that the measurements satisfy the zero-mean, Gaussianity, and stationarity in the space-time second-order structure assumptions; see Section 1, where these assumptions were laid out. The dataset contains 550 spatial locations, each with 730 temporal measurements. The resulting space-time covariance matrix for this dataset is of size $401,500 \times 401,500$. Figure 7 shows the measurements at every spatial location at the first six time points. We expect some substantial correlation in space from the data as there are recognizable blobs of yellow, red, and blue, indicating that the measurement at any spatial location is similar to its neighbors. We may expect weak

Table 1. A summary of the estimated parameters of the nonseparable (NS) and separable (S) models and their corresponding errors (MSPE) and prediction uncertainty (PU) for the Saudi Arabia and US datasets. MSPE1 and PU1 correspond to Testing Dataset 1, while MSPE2 and PU2 point to Testing Dataset 2. The best model reports the lower MSPE and PU.

	Model	$\hat{\sigma}^2$	$\hat{a}_s$	$\hat{\nu}$	$\hat{a}_t$	$\hat{\alpha}$	$\hat{\beta}$	MSPE1 / PU1	MSPE2 / PU2
Saudi Arabia	NS	1.2942	1.3461	2.1568	1.1284	0.1401	0.7548	0.00179877 / 76.91	1.08391 / 875.85
	S	2.6194	1.2737	2.1544	2.0466	0.0308	0	0.00180453 / 78.63	1.14705 / 1066.00
US	NS	0.4757	1.3383	1.1266	6.7722	0.7262	0.1451	0.00280177 / 155.27	0.05822582 / 322.09
	S	2.1237	1.5409	1.4709	7.9975	0.4867	0	0.00318483 / 134.52	0.06386783 / 1118.20

correlation in time as the spatial images are substantially different from one time point to the next. The second dataset contains the residuals of the hourly log PM2.5 in Midwest US where the terrains are relatively flat. Over this certain region, the hourly sampled log PM2.5 already satisfies the stationarity assumption, hence, no averaging is performed. This other dataset contains 945 spatial locations, each with 500 temporal measurements, with a resulting space-time covariance matrix of size $472,500 \times 472,500$. Figure 8 shows the measurements of the second dataset at every spatial location at four hour intervals. Similarly, moderate to strong correlation in space can be seen. However, unlike the first dataset, a strong correlation in time can be hypothesized as the spatial image structure from the first hour tends to persist even until after 20 hours. Furthermore, Figure 8 illustrates a special phenomenon called transport phenomenon where objects such as PM2.5 particles get transported from one place to another by some media such as the wind. Such behavior can be detected by following the red blob at 0:00 located over Colorado and Nebraska. At 20:00, the red blob has traveled to Oklahoma. This transport phenomenon and many other environmental phenomena are more appropriately modeled by nonseparable space-time models [16, 47].

For both datasets, we split the full space-time dataset into training (90%) and testing (Testing Dataset 1) (10%) and fit the separable ($\beta = 0$) and nonseparable ($\beta > 0$) versions of the model in (6). After obtaining the estimated parameters, we predict (Equation 4) the measurements at the space-time locations included in Testing Dataset 1 and compute the prediction errors measured by the MSPE and the associated prediction uncertainty (PU) (Equation 5). Furthermore, for both Saudi Arabia and US datasets, we also predict the full map forward in time (two-time steps ahead); that is, we forecast the values on all spatial locations at some future time points. We compare these predictions to the true measurements, which we label as Testing Dataset 2. The results are reported in Table 1.

The parameter estimates in Table 1 imply several properties regarding the two space-time random fields. First, the nonseparability parameter, β, for Saudi Arabia has an estimated value of 0.7548. This is higher than the estimated β value of 0.1451 for the US. Therefore, the space-time interaction for log PM2.5 is much stronger in Saudi Arabia than in the US. Second, the range parameter in time, a_t, has an estimated value that is a lot higher in the US than that of Saudi Arabia. This suggests that log PM2.5 values at time points that are far apart remain strongly correlated. Based on the parameter values, for the Saudi Arabia case, the correlation drops to 0.05 after 8-time points, while the correlation drops to 0.05 after 150-time points for the US one. Third, some parameters in the separable model differ a lot from their counterparts in the nonseparable model. This can be expected as the parameters in the separable model need to compensate for the lack of the β parameter. In particular, the variance parameter, σ^2, is artificially inflated and is the most impacted parameter. Moreover, when we compute the sample variances of the training dataset for Saudi Arabia and the US, we obtain the values 0.9979 and 0.805, respectively. The variance parameter estimates under the nonseparable model are closer to these values obtained from the real data. This is one disadvantage of the separable model, i.e., we obtain erroneous parameter estimates that are not supported by the real data. Fourth, the errors obtained

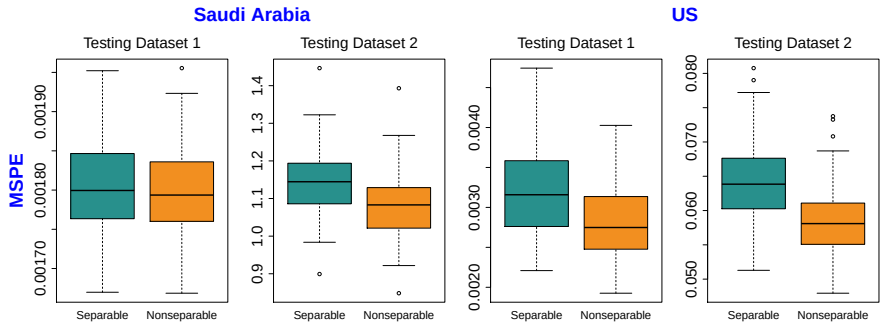


Fig. 9. Boxplots of the prediction errors for the real data pseudo cross-validation study.

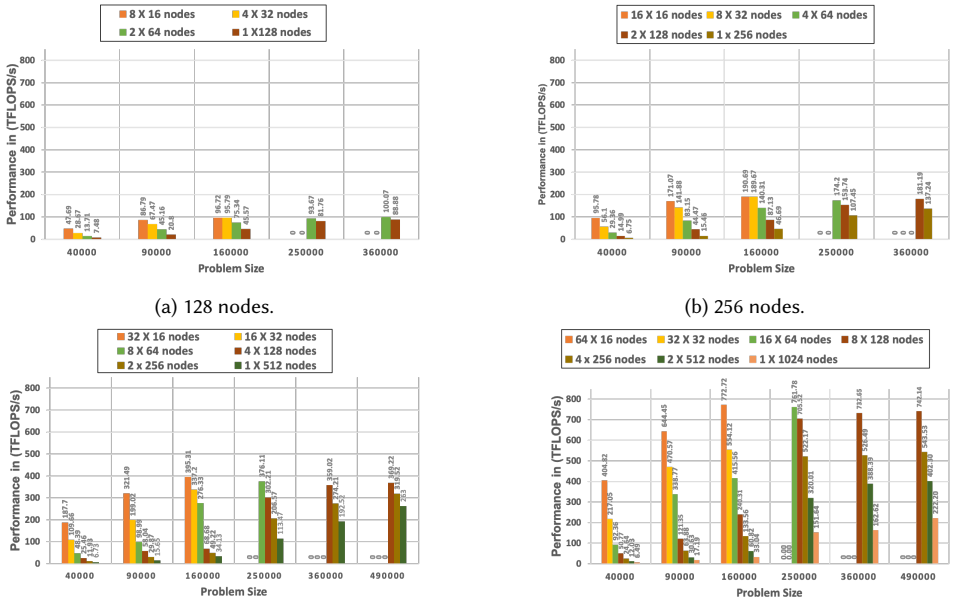
in Testing Dataset 1 are smaller than those obtained in Testing Dataset 2. This is expected as predicting on Testing Dataset 1 involves filling the missing values at some locations in space for each time point. On the other hand, predicting on Testing Dataset 2 entails forecasting forward in time where all values on the map are missing for several time points ahead. Prediction on datasets such as Testing Dataset 2 is much more common in practice wherein full maps of measurements in the next few time points are desired. Lastly, the separable model returns higher prediction errors than the nonseparable model. In particular, the percentages of improvements of the nonseparable over the separable model for MSPE in Testing Dataset 2 is 6% in Saudi Arabia and 9% in the US and for the prediction uncertainty (PU) is 18% in Saudi Arabia and 71% in the US.

The difference in MSPEs between the nonseparable and separable models may not be huge because the values are small. Hence, to further show that the nonseparable model is superior, we perform a pseudo cross-validation study wherein for 100 rounds, we take a subset of each of the testing datasets and compute the resulting MSPE. We collect the MSPE values and illustrate the distribution as boxplots in Figure 9. The boxplots have shown that indeed the nonseparable model returns lower errors in both weak correlation (Saudi Arabia) and strong correlation (US) scenarios. Note that these real data results mirror the synthetic data experiments results in Figure 5.

6.3 Performance Evaluation at Scale

We aim in this section to evaluate the performance of the proposed two-level parallel implementation. Parallelization of the MLE operation through a set of MPI sub-communicators requires determining the number of nodes in each MPI sub-communicator that satisfies a decent performance (execution time and FLOPS/s). Assuming a number of compute nodes, several combinations of the number of sub-communicators and the number of nodes in each MPI sub-communicator can be applied. We use $x \times y$ nodes notation to define a specific distribution where x represents the number of MPI sub-communicators, and y represents the number of nodes in each MPI sub-communicator. Herein, we evaluate these combinations for a set of given nodes, i.e., 32, 64, 128, 256, 512, and 1024 on a Cray XC40 system. We assume a load-balanced workload distribution since all computations are performed in dense format (full-double precision).

Figure 10 shows the performance of different combinations ($x \times y$ nodes) using different number of nodes. The missing data come with the memory limitation of the small number of nodes and the large problem size. In this case, it is mandatory to increase the number of nodes in each sub-communicator to be able to handle the generated covariance



(a) 128 nodes.

(b) 256 nodes.

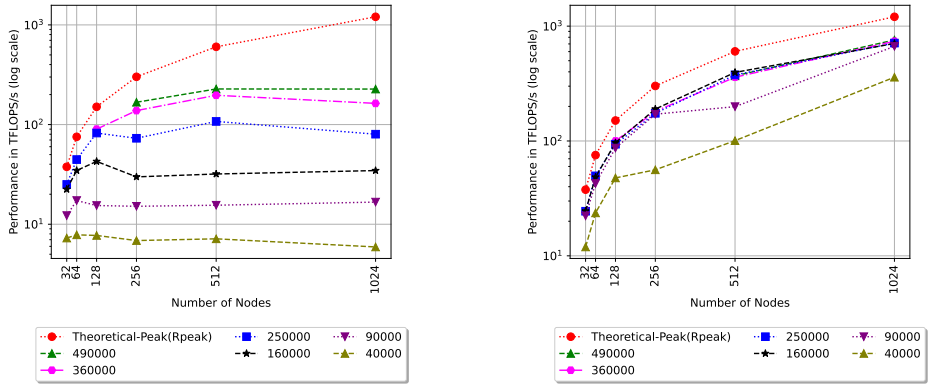
(c) 512 nodes.

(d) 1024 nodes.

Fig. 10. Performance of a single MLE optimization step using a different number of nodes on a Cray XC40 Supercomputer. The legends show how many MPI sub-communicators x are used and how many y nodes exist in each MPI sub-communicator.

matrix. For the different number of nodes, the same remarks hold. First, fewer nodes per MPI sub-communicator y (or distributing the workload through more MPI sub-communicators) is the main option to achieve a decent performance. Of course, this is restricted by the size of the problem, which cannot be executed on a small number of nodes. The best performance can be obtained by the smallest number of nodes per sub-communicator that can handle a given problem size. Second, the performance improvement can be clearly observed in all the subfigures with small matrix sizes since the workload is small and cannot saturate a large number of nodes. For instance, in Figure 10a, using 128 nodes, the performance improvement of using 8 sub-communicators compared to the single communicator with all the 128 nodes is about 6.38X with a problem size of 40K. However, in the case of 160K, only 2.12X performance improvement has been achieved. Another example can be captured from Figure 10c, using 512 nodes, the achieved performance improvements are 27.89X and 11.58X with 40K and 160K problem sizes, respectively. Third, the proposed framework shows higher performance improvements with a larger number of nodes. For instance, using 160K problem size, the achieved performance improvements are 2.12X, 4.08X, 11.58X, and 23.39X using 128, 256, 512, and 1024 nodes, respectively.

In the following subsection, we compare our proposed two-level parallelization scalability with the single MPI communicator implementation. We used the tuned number of sub-communicators x for each number of nodes and problem sizes to achieve the highest performance. The tuning setting relies on compute-bound driven by the kernel performance of matrix multiplication (GEMM) as shown by the DAG in Figure 2. We select the number of nodes per MPI sub-communicator that satisfies a decent performance percentage with a single GEMM operation.



(a) One-level MLE parallelization.

(b) Two-level MLE parallelization.

Fig. 11. Performance of an MLE optimization step using single and n MPI communicators on a Cray XC40 Supercomputers. In (b) x MPI sub-communicators is used where x is tuned for performance.

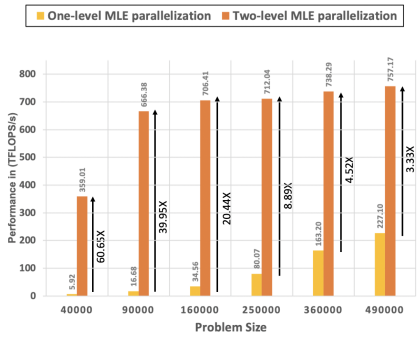
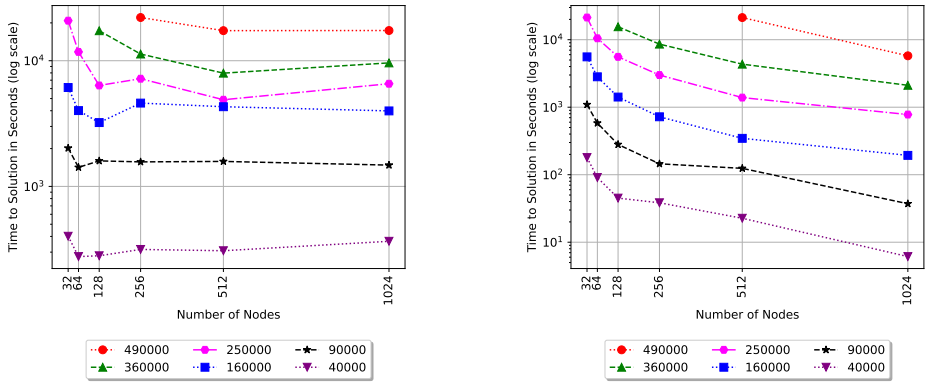


Fig. 12. One-level versus two-level MLE parallelization performance using 1024 nodes on a Cray XC40 system.

6.4 Performance Comparison

Herein, we aim at comparing the proposed two-level parallel MLE implementation with the one-level parallel implementation shown in [2]. We use both performance (FLOPS/s) and time-to-solution to conduct this comparison. The number of MPI sub-communicators has been tuned for each number of nodes and problem size to achieve the highest performance.

Figure 11 compares the performance of one-level parallelism using a single MPI communicator (a single PSO algorithm step using one particle) with the performance of two-level parallelism through x MPI sub-communicators (a single PSO algorithm step using x particles) with the same number of nodes. The x- and y-axes show the number of nodes and the performance (TFLOPS/s) in log-scale, respectively. Six different matrix sizes are used for the evaluation, i.e., 40K, 90K, 160K, 250K, 360K, and 490K.



(a) One-level MLE parallelization.

(b) Two-level MLE parallelization.

Fig. 13. Time-to-solution of full MLE operation using 100 optimization iterations on a Cray XC40 Supercomputers. In (b) x MPI sub-communicators is used where x is tuned for performance.

Figure 11a reports the one-level implementation performance where the parallel linear solvers are applied to the log-likelihood function. The red curve shows the peak performance (Rpeak) of each number of nodes as a reference to the total performance percentage that each run can achieve.

Theoretically, doubling the number of nodes should halve the execution time and double the performance simultaneously (i.e., linear scalability). However, in small workloads, the performance degrades compared to what is expected because there is not enough work to saturate this large number of nodes.

Moreover, compared to the peak performance (Rpeak) of each number of nodes (in the red curve), the one-level parallelization can achieve up to 67%, 59%, 60%, 56%, 47%, and 28% of Rpeak using 32, 64, 128, 256, 512, and 1024 respectively. It is clear that the performance scalability deteriorates with smaller problem sizes since there is not enough work to keep all the nodes busy. Thus, large problem sizes scale better than small problem sizes.

Figure 11b shows performance using two-level parallelization where the number of particles equals the number of MPI sub-communicators. The figure confirms the scalability of the proposed implementation when x MPI sub-communicators are used simultaneously. We report the best performance of different numbers nodes and different problem sizes with a tuned x value, i.e., the number of sub-communicators. The achieved performance compared to the peak performance can reach up to 65%, 66%, 66%, 60%, 59%, and 58% using 32, 64, 128, 256, 512, and 1024, respectively. Compared to Figure 11a, the proposed implementation satisfies stable performance with a different number of nodes and with different problem sizes.

Figure 12 highlights the gained performance in using the proposed two-level parallelization compare to the one-level parallelization using 1024 nodes on a Cray XC40 system. The figure shows that the proposed framework can achieve 3.3X more TFLOPS/s than the baseline framework using a 490K dataset. Smaller matrix sizes can achieve higher performance increase, but we think it might be overkill to highlight this speedup since the scalability of the one-level parallelization degraded so much with a small problem size and a large number of nodes.

Of course, performance analysis addresses processing capability; however, time-to-solution is also essential to assess the proposed method's effectiveness. Figure 13 shows the time-to-solution, where the number of optimization iterations is fixed to 100. When one-level parallism is used (Figure 13a), smaller workloads take the same execution time even

when the number of nodes is increased. Thus, at a certain point, increasing the number of nodes will not improve the time-to-solution anymore. For instance, with a 360K problem size, using 1024 nodes instead of 512 nodes will not accelerate the execution time. On the contrary, the proposed implementation keeps the execution time speedup gains with a larger number of nodes. For instance, with a 360K problem size, the total execution time of the MLE operation is 4343.42 and 2111.48 seconds with 256 and 512 nodes, respectively, with a speedup of 2.06X.

7 CONCLUSION AND FUTURE WORK

We propose a massively parallel implementation of spatio-temporal inference using statistical modeling that can predict air pollution using observations in a specific space-time domain, illustrating the importance of relaxing the assumption of independence of space and time. The proposed two-level framework relies on MPI sub-communicators at the upper level to perform independent log-likelihood function evaluation with a different set of parameters. These independent computations result from applying parallel optimization through the parallel PSO algorithm, where each swarm represents a single log-likelihood operation. A task-based parallel technique is applied at the lower level to perform linear solver operations on a given set of nodes representing a single MPI sub-communicator. Each linear solver is represented as a Directed Acyclic Graph (DAG), representing a set of tasks that can be distributed to different hardware resources using the StarPU runtime system. The performance and the accuracy of the proposed implementation have been assessed using synthetic datasets and real PM2.5 atmospheric pollution datasets from two geographical regions, i.e., Middle East and the US, on a Cray XC40 system with up to 1024 nodes.

In future work, we intend to represent all the linear algebraic operations as a single DAG graph that allows a batch of asynchronous executions of several MLE iterations on all the available hardware resources simultaneously. These asynchronous executions may permit to mitigate the load imbalance overhead seen from contentions of shared network interconnect resources, promote situations of overlapping communications with computations, and increase on-node data reuse, especially in the presence of GPU hardware accelerators. Moreover, computations approximation such as low-rank approximation and covariance tapering that suffer from the load imbalance problems can benefit from the proposed approach.

REFERENCES

[1] Emile Aarts, Jan Korst, and Wil Michiels. 2005. Simulated annealing. In *Search methodologies*. Springer, 187–210.

[2] Sameh Abdulah, Hatem Ltaief, Ying Sun, Marc G Genton, and David E Keyes. 2018. ExaGeoStat: A high performance unified software for geostatistics on manycore systems. *IEEE Transactions on Parallel and Distributed Systems* 29, 12 (2018), 2771–2784.

[3] Sameh Abdulah, Hatem Ltaief, Ying Sun, Marc G Genton, and David E Keyes. 2019. Geostatistical modeling and prediction using mixed precision tile cholesky factorization. In *2019 IEEE 26th International Conference on High Performance Computing, Data, and Analytics (HiPC)*. IEEE, 152–162.

[4] Emmanuel Agullo, Jim Demmel, Jack Dongarra, Bilel Hadri, Jakub Kurzak, Julien Langou, Hatem Ltaief, Piotr Luszczek, and Stanimire Tomov. 2009. Numerical linear algebra on emerging architectures: The PLASMA and MAGMA projects. In *Journal of Physics: Conference Series*, Vol. 180. IOP Publishing, 012037.

[5] Cédric Augonnet, Samuel Thibault, Raymond Namyst, and Pierre-André Wacrenier. 2009. StarPU: A unified platform for task scheduling on heterogeneous multicore architectures. In *European Conference on Parallel Processing*. Springer, 863–874.

[6] Cédric Augonnet, Samuel Thibault, Raymond Namyst, and Pierre-André Wacrenier. 2011. StarPU: a unified platform for task scheduling on heterogeneous multicore architectures. *Concurrency and Computation: Practice and Experience* 23, 2 (2011), 187–198.

[7] Omer Saud Azeez, Biswajeet Pradhan, and Helmi ZM Shafri. 2018. Vehicular CO emission prediction using support vector regression model and GIS. *Sustainability* 10, 10 (2018), 3434.

[8] CHAMELEON 2021. The Chameleon Project: A dense linear algebra software for heterogeneous architectures. Available at https://project.inria.fr/chameleon/.

[9] Wanfang Chen, Marc G Genton, and Ying Sun. 2021. Space-time covariance structures and models. *Annual Review of Statistics and Its Application* 8 (2021).

[10] Tangpei Cheng. 2013. Accelerating universal Kriging interpolation algorithm using CUDA-enabled GPU. *Computers & Geosciences* 54 (2013), 178–183.

[11] E Gutiérrez De Ravé, Francisco J Jiménez-Hornero, Ana B Ariza-Villaverde, and JM Gómez-López. 2014. Using general-purpose computing on graphics processing units (GPGPU) to accelerate the ordinary kriging algorithm. *Computers & Geosciences* 64 (2014), 1–6.

[12] Michael R Desjardins, Alexander Hohl, Adam Griffith, and Eric Delmelle. 2019. A space–time parallel framework for fine-scale visualization of pollen levels across the Eastern United States. *Cartography and Geographic Information Science* 46, 5 (2019), 428–440.

[13] Kathryn Anne Dowsland and Jonathan Thompson. 2012. Simulated annealing. *Handbook of natural computing* (2012), 1623–1655.

[14] Goran Gašparac, Amela Jeričević, Prashant Kumar, and Branko Grisogono. 2020. Regional-scale modelling for the assessment of atmospheric particulate matter concentrations at rural background locations in Europe. *Atmospheric Chemistry and Physics* 20, 11 (2020), 6395–6415.

[15] Baozhu Ge, Zifa Wang, Xiaobin Xu, Jianbin Wu, Xiaolan Yu, and Jian Li. 2014. Wet deposition of acidifying substances in different regions of China and the rest of East Asia: Modeling with updated NAQPMS. *Environmental pollution* 187 (2014), 10–21.

[16] Tilmann Gneiting. 2002. Nonseparable, stationary covariance functions for space–time data. *J. Amer. Statist. Assoc.* 97, 458 (2002), 590–600.

[17] Tilmann Gneiting, Marc G Genton, and Peter Guttorp. 2007. Geostatistical space-time models, stationarity, separability, and full symmetry. *Finkenstaedt, B., Held, L. and Isham, V. (eds.), Statistics of Spatio-Temporal Systems, Chapman & Hall/CRC Press, Monographs in Statistics and Applied Probability* 107 (2007), 151–175.

[18] Nasimul Hasan, Nayan Chandra Nath, and Risul Islam Rasel. 2015. A support vector regression model for forecasting rainfall. In *2015 2nd International Conference on Electrical Information and Communication Technologies (EICT)*. IEEE, 554–559.

[19] Yong H Huang, James E Saiers, Judson W Harvey, Gregory B Noe, and Steven Mylon. 2008. Advection, dispersion, and filtration of fine particles within emergent vegetation of the Florida Everglades. *Water Resources Research* 44, 4 (2008).

[20] Konstantinos E Kakosimos, Marc J Assael, John S Lioumbas, and Anthimos S Spiridis. 2011. Atmospheric dispersion modelling of the fugitive particulate matter from overburden dumps with numerical and integral models. *Atmospheric Pollution Research* 2, 1 (2011), 24–33.

[21] George Kallos, Marina Astitha, Petros Katsafados, and Chris Spyrou. 2007. Long-range transport of anthropogenically and naturally produced particulate matter in the Mediterranean and North Atlantic: Current state of knowledge. *Journal of Applied Meteorology and Climatology* 46, 8 (2007), 1230–1251.

[22] James Kennedy and Russell Eberhart. 1995. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, Vol. 4. IEEE, 1942–1948.

[23] Byeong Soo Kim, Bong Gu Kang, Seon Han Choi, and Tag Gon Kim. 2017. Data modeling versus simulation modeling in the big data era: case study of a greenhouse control system. *Simulation* 93, 7 (2017), 579–594.

[24] Joseph B Knox. 1974. Numerical modeling of the transport diffusion and deposition of pollutants for regions and extended scales. *Journal of the Air Pollution Control Association* 24, 7 (1974), 660–664.

[25] Robert Michael Lewis and Virginia Torczon. 2002. A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Optimization* 12, 4 (2002), 1075–1089.

[26] Jihan Li, Xiaoli Li, and Kang Wang. 2019. Atmospheric PM2. 5 Concentration Prediction Based on Time Series and Interactive Multiple Model Approach. *Advances in Meteorology* 2019 (2019).

[27] Junmin Li and Luping Wang. 2017. The research of PM2.5 concentrations model based on regression calculation model. In *AIP Conference Proceedings*, Vol. 1794. AIP Publishing LLC, 030005.

[28] Bing-Chun Liu, Arihant Binaykia, Pei-Chann Chang, Manoj Kumar Tiwari, and Cheng-Chin Tsao. 2017. Urban air quality forecasting based on multi-dimensional collaborative support vector regression (svr): A case study of beijing-tianjin-shijiazhuang. *PloS one* 12, 7 (2017), e0179763.

[29] Lin Liu, Chonglong Wu, and Zhibo Wang. 2015. Parallelization of the Kriging Algorithm in Stochastic Simulation with GPU Accelerators. In *Geo-Informatics in Resource Management and Sustainable Ecosystem*. Springer, 197–205.

[30] Matthew Loxham, Donna E Davies, and Stephen T Holgate. 2019. The health effects of fine particulate air pollution.

[31] Baolei Lyu, Yuzhong Zhang, and Yongtao Hu. 2017. Improving PM2. 5 air quality model forecasts in China using a bias-correction framework. *Atmosphere* 8, 8 (2017), 147.

[32] Ahmed F Mashaly and AA Alazba. 2016. MLP and MLR models for instantaneous thermal efficiency prediction of solar still under hyper-arid environment. *Computers and Electronics in Agriculture* 122 (2016), 146–155.

[33] Michael C McCarthy, Douglas S Eisinger, Hilary R Hafner, Lyle R Chinkin, Paul T Roberts, Kevin N Black, Nigel N Clark, Peter H McMurry, and Arthur M Winer. 2006. Particulate matter: a strategic vision for transportation-related research. *Environmental science & technology* 40, 18 (2006), 5593–5599.

[34] J Murillo-Escobar, JP Sepulveda-Suescun, MA Correa, and D Orrego-Metaute. 2019. Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization: Case study in Aburrá Valley, Colombia. *Urban Climate* 29 (2019), 100473.

[35] RH Norden. 1972. A survey of maximum likelihood estimation. *International Statistical Review/Revue Internationale de Statistique* (1972), 329–354.

[36] RH Norden. 1973. A survey of maximum likelihood estimation: Part 2. *International Statistical Review/Revue Internationale de Statistique* (1973), 39–58.

[37] Karol R Opara and Jarosław Arabas. 2019. Differential Evolution: A survey of theoretical analyses. *Swarm and evolutionary computation* 44 (2019), 546–558.

[38] Ashwin Pajankar. 2017. Message passing interface. In *Raspberry Pi Supercomputing and Scientific Programming*. Springer, 61–65.

[39] Karl Pearson. 1936. Method of moments and method of maximum likelihood. *Biometrika* 28, 1/2 (1936), 34–59.
[40] Eleni Petrakou and Iasonas Topsis Giotis. 2020. Planetary statistics and forecasting for solar flares. *arXiv preprint arXiv:2006.10694* (2020).
[41] Riccardo Poli, James Kennedy, and Tim Blackwell. 2007. Particle swarm optimization. *Swarm intelligence* 1, 1 (2007), 33–57.
[42] Kenneth V Price. 2013. Differential evolution. In *Handbook of optimization*. Springer, 187–214.
[43] Allison M Ring, Timothy P Canty, Daniel C Anderson, Timothy P Vinciguerra, Hao He, Daniel L Goldberg, Sheryl H Ehrman, Russell R Dickerson, and Ross J Salawitch. 2018. Evaluating commercial marine emissions and their role in air quality policy using observations and the CMAQ model. *Atmospheric Environment* 173 (2018), 96–107.
[44] Mary Salvana, Sameh Abdulah, Huang Huang, Hatem Ltaief, Ying Sun, Marc M Genton, and David Keyes. 2021. High Performance Multivariate Geospatial Statistics on Manycore Systems. *IEEE Transactions on Parallel and Distributed Systems* (2021).
[45] Pierre Sicard, Paola Crippa, Alessandra De Marco, Stefano Castruccio, Paolo Giani, Juan Cuesta, Elena Paoletti, Zhaozhong Feng, and Alessandro Anav. 2021. High spatial resolution WRF-Chem model over Asia: physics and chemistry evaluation. *Atmospheric Environment* 244 (2021), 118004.
[46] SN Sivanandam and SN Deepa. 2008. Genetic algorithms. In *Introduction to genetic algorithms*. Springer, 15–37.
[47] Michael L Stein. 1999. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.
[48] Zhongda Tian, Gang Wang, and Yi Ren. 2020. Short-term wind speed forecasting based on autoregressive moving average with echo state network compensation. *Wind Engineering* 44, 2 (2020), 152–167.
[49] W Gene Tucker. 2000. An overview of PM2.5 sources and control strategies. *Fuel Processing Technology* 65 (2000), 379–392.
[50] A Ismael F Vaz and Luís N Vicente. 2007. A particle swarm pattern search method for bound constrained global optimization. *Journal of Global Optimization* 39, 2 (2007), 197–219.
[51] Jacqueline Whalley and Sara Zandi. 2016. Particulate matter sampling techniques and data modelling methods. In *Air Quality-Measurement and Modeling*. INTECH, 10.
[52] Jianhui Xu, Hongda Hu, Hong Shu, and Zhiyong Hu. 2016. Using compute unified device architecture-enabled graphic processing unit to accelerate fast Fourier transform-based regression Kriging interpolation on a MODIS land surface temperature image. *Journal of Applied Remote Sensing* 10, 2 (2016), 026036.
[53] Yuan Yan and Marc G Genton. 2018. Gaussian likelihood inference on data from trans-Gaussian random fields with Matérn covariance function. *Environmetrics* 29, 5-6 (2018), e2458.
[54] Mehdi Zamani Joharestani, Chunxiang Cao, Xiliang Ni, Barjeece Bashir, and Somayeh Talebiesfandarani. 2019. PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere* 10, 7 (2019), 373.
[55] Yueheng Zhang, Xinqi Zheng, Zhenhua Wang, Gang Ai, and Qing Huang. 2018. Implementation of a parallel GPU-based space-time kriging framework. *ISPRS International Journal of Geo-Information* 7, 5 (2018), 193.
[56] Xu Zhong, Allison Kealy, and Matt Duckham. 2016. Stream Kriging: Incremental and recursive ordinary Kriging over spatiotemporal data streams. *Computers & Geosciences* 90 (2016), 134–143.
[57] Chi Zhou, HB Gao, Liang Gao, and WG Zhang. 2003. Particle Swarm Optimization (PSO) Algorithm [J]. *Application Research of Computers* 12 (2003), 7–11.