scientific reports

Check for updates

OPEN

Exact Gaussian processes for massive datasets via non-stationary sparsity-discovering kernels

Marcus M. Noack[1]✉, Harinarayan Krishnan[1], Mark D. Risser[2] & Kristofer G. Reyes[3]

A Gaussian Process (GP) is a prominent mathematical framework for stochastic function approximation in science and engineering applications. Its success is largely attributed to the GP's analytical tractability, robustness, and natural inclusion of uncertainty quantification. Unfortunately, the use of exact GPs is prohibitively expensive for large datasets due to their unfavorable numerical complexity of $O(N^3)$ in computation and $O(N^2)$ in storage. All existing methods addressing this issue utilize some form of approximation—usually considering subsets of the full dataset or finding representative pseudo-points that render the covariance matrix well-structured and sparse. These approximate methods can lead to inaccuracies in function approximations and often limit the user's flexibility in designing expressive kernels. Instead of inducing sparsity via data-point geometry and structure, we propose to take advantage of naturally-occurring sparsity by allowing the kernel to discover—instead of induce—sparse structure. The premise of this paper is that the data sets and physical processes modeled by GPs often exhibit natural or implicit sparsities, but commonly-used kernels do not allow us to exploit such sparsity. The core concept of exact, and at the same time sparse GPs relies on kernel definitions that provide enough flexibility to learn and encode not only non-zero but also zero covariances. This principle of ultra-flexible, compactly-supported, and non-stationary kernels, combined with HPC and constrained optimization, lets us scale exact GPs well beyond 5 million data points.

A Gaussian Process (GP) is the most prominent member of the larger family of stochastic processes and provides a powerful and flexible framework for stochastic function approximation in the form of Gaussian Process Regression (GPR). This is because a GP is characterized as a Gaussian probability distribution over a function space $\left\{f : f(\mathbf{x}) = \sum_i^N \alpha_i k(\mathbf{x}, \mathbf{x}_i; h)\ \forall \mathbf{x} \in \mathcal{X}\right\}$, where $k(\mathbf{x}, \mathbf{x}_i; h)$ is the kernel function and h is a set of hyperparameters. The mean and the covariance of the Gaussian probability distribution can be learned by constrained function optimization from data $\mathcal{D} = \{\mathbf{x}_i, y_i\}$ and conditioned on the observations y_i to yield a posterior probability density function. Throughout this paper, we will refer to this optimization often as training or learning to emphasize the link to machine learning (ML). GPs assume a model of the form $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x})$, where $f(\mathbf{x})$ is the unknown latent function, $y(\mathbf{x})$ is the noisy function evaluation (the measurement), $\epsilon(\mathbf{x})$ is the noise term, and $\mathbf{x}$ is an element of the input space (or index set) $\mathcal{X}$. Learning the hyperparameters h of a GP and subsequent conditioning leads to a stochastic representation of a model function which can be used for decision-making, visualizations, and interpretations. This paper deals with the example of regression (GPR), but the proposed methodology can easily be applied to other GP-related tasks; we will therefore simply use the more-general acronym "GP" throughout this paper.

In comparison to neural networks, GPs can scale better with the dimensionality of the input space—since the number of weights or parameters does not depend on it—and provide more exact function approximations[1]. Additionally, GPs provide highly-coveted Bayesian uncertainty quantification on top of such function approximations. While some neural-network-based methods can estimate errors, these are most often not the result of rigorous Bayesian uncertainty quantification. Even so, GPs come with one difficult-to-circumvent problem:

[1]Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [2]Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [3]Department of Materials Design and Innovation, University at Buffalo, Buffalo, NY 14260, USA. ✉email: MarcusNoack@lbl.gov

due to their unfavorable scaling of $O(N^3)$ in computation and $O(N^2)$ in storage[2], the applicability of GPs has largely been limited to small and moderate dataset sizes (N), which prevents the method from being used in many fields where large datasets are common and is a major disadvantage compared to other ML methods, e.g., neural networks. Those fields include many machine learning applications, earth, environmental, climate and materials sciences, and engineering. The numerical complexity of GPs stems from the need to store and invert a typically-dense covariance matrix[2]. While the direct inversion can be replaced by iterative linear system solves, the speedup is rather modest for dense covariance matrices.

Methods to alleviate the GP's scaling issues exist but are largely based on approximations. These workarounds fall into a few broad categories:

- A set of local GP experts: The dataset is divided into subsets, each of which serves as input into separate GPs, and the resulting posteriors are then combined[3–5]. This can also be interpreted as one large GP with a sparse (block-diagonal) covariance matrix. It is common to divide the dataset by locality, leading to the name "local GP experts".
- Inducing-points methods: Instead of inducing a sparse covariance matrix by picking subsets of the dataset, inducing-points methods place new points inside the domain, inducing a favorable data structure that translates into sparsity. The function values at those points are calculated via standard interpolation techniques. Popular examples of this approach include KISS-GP[6], the predictive process[7,8], and fixed-rank Kriging[9,10]. Generally, inducing-points methods are not agnostic to the kernel definition and therefore limit which kernels can be used. This limitation is a major drawback given that recent applications are increasingly using flexible non-stationary kernel functions (for instance[11]), which are generally incompatible with inducing-points methods.
- Structure-exploiting methods: These methods are a special kind of inducing-points method that places pseudo-points on a grid so that the covariance matrix has Toeplitz algebra, which leads to fast linear algebra needed to train and condition the GP. Again, the success of those methods is not agnostic to the kernel definition.
- Vecchia approximations: Instead of calculating the full conditional probability density function of a GP prior, the Vecchia approximation[12,13] is used to pick a subset of the data to condition on. This method is also kernel-dependent and has largely been applied using stationary kernels.

The statistics literature contains a variety of other related approaches; see[14] for a recent summary of both traditional and state-of-the-art approaches with a direct comparison of the methods on a common dataset. Another outstanding review is by[15]. All of the existing methods introduced above have one thing in common: sparsity or exploitable structure in the covariance matrix is introduced by operating on the data points—either by considering subsets of the full dataset or by utilizing representative pseudo-points that allow for a favorable structure (e.g. Toeplitz) in the covariance, and sparsity. This commonality leads to one major issue of all existing methods: they are approximations of exact GPs[16], which leads to poor prediction performance for highly non-linear functions—i.e. functions exhibiting large first and second-order derivatives with frequently changing signs. For high-fidelity approximations, the number of sub-selected data points or pseudo points must approach the size of the original dataset[3], which eliminates the methods' advantages. More fundamentally, the sparsity and structure of the covariance matrix should be dictated by the nature of the problem and the data, not by our computational constraints. This leads us to consider kernels that can take advantage of naturally occurring—problem and data dictated—sparsity.

Instead of operating on the input points—by selecting subsets or pseudo-points—an alternative approach is to let the kernel find the most expressive and sparse structure of the covariance matrix. In principle, a very flexible kernel could discover—not induce—naturally occurring sparse structure in the covariance matrix without acting on the data points at all. In that case, there is no approximation taking place (compared to inducing-points, local-experts, and Vecchia methods) and no ad-hoc point selection is required. Additionally, we shall see that there are no restrictions on the used problem-specific kernel functions as long as they are combined with our proposed sparsity-enabling, and therefore, sparsity-discovering kernels. An added advantage is that the kernel-discovered sparsity is entirely independent of spatial relationships of data points, meaning, very distant data points can be discovered to have high covariances while points in close proximity might be independent; there is no ad-hoc dependency of covariances on Euclidean point distance in $\mathcal{X}$—in contrast to local GP experts for instance.

As we outline below, creating an exact GP that learns and utilizes naturally-occurring sparsity shall require three main building blocks: (1) ultra-flexible, non-stationary, compactly-supported kernel functions, specially customized to learn and encode zero-covariances, (2) a high-performing implementation that can compute sub-matrices of the covariance matrix in a distributed, parallel fashion, and (3) a constrained or augmented optimization routine to ensure the learned covariance matrix is sparse (or at least enforce a preference for sparsity). This last point is important for large problems to protect the computing system from over-utilization. In the extreme case, in which naturally-occurring sparsity is insufficient or non-existent, having a sparsity-inducing optimization routine would seamlessly result in an optimal approximate GP. The contributions of this paper can be summarized as follows. We show that, by combining tailored kernel designs, HPC implementation, and constrained optimization, exact GPs can be scaled to datasets of any size, under the assumption of naturally-occurring sparsity. The core idea, that allows such scaling, is a sparsity-discovering kernel design and an optimization that learns which data points are not correlated, independent of their respective locality in the input space $\mathcal{X}$. The sparse structure is not artificially "induced" as in all state-of-the-art methods; instead, we allow the GP to discover the natural sparsity in the dataset. This principle is visualized in Figure 2. For reference we have included a table comparing and contrasting different existing methods and our proposed method (see

Figure 1). While an in-depth quantitative comparison of our proposed method versus existing state-of-the-art approaches would be illuminating, we argue that such an exercise is beyond the scope of this manuscript due to the fact that performance depends on a variety of subjective choices: data application, kernel functions, computing architecture, and prior mean functions, among other things.

Contributions of this Paper at a Glance: (1) We propose a new non-stationary, flexible, and compactly-supported kernel design that allows a Gaussian process to discover sparsity; (2) We show how to use the new kernel design in concert with distributed computing to scale GPs to millions of data points; and (3) We draw attention to the hyperparameter optimization process so that solutions that allow sparsity are preferred.

Method

Basics. A Gaussian Process (GP) is characterized by a Gaussian probability density function over function values $\mathbf{f}$

$$p(\mathbf{f}) = \frac{1}{\sqrt{(2\pi)^{\dim}|\mathbf{K}|}} \exp\left[-\frac{1}{2}(\mathbf{f}-\mathbf{m})^T\mathbf{K}^{-1}(\mathbf{f}-\mathbf{m})\right], \tag{1}$$

and a Gaussian likelihood

$$p(\mathbf{y}|\mathbf{f}) = \frac{1}{\sqrt{(2\pi)^{\dim}|\mathbf{V}|}} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{f})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{f})\right], \tag{2}$$

where $\mathbf{V}$ is the observation-noise matrix, which is most often diagonal, $\mathbf{K}$ is the covariance matrix defined by the kernel function $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, and $\mathbf{m}$ is the prior-mean vector. Training the GP is done by maximizing the marginal log-likelihood (ignoring an additive constant)

$$\ln(L(h)) = -\frac{1}{2}(\mathbf{y}-\mathbf{m}(h))^T\mathbf{K}(h)^{-1}(\mathbf{y}-\mathbf{m}(h)) - \frac{1}{2}\ln(|\mathbf{K}(h)|) \tag{3}$$

with respect to the hyperparameters h. After the hyperparameters are found, the posterior is defined as

$$\begin{aligned} p(\mathbf{f}_0|\mathbf{y}) &= \int_{\mathbb{R}^N} p(\mathbf{f}_0|\mathbf{f},\mathbf{y})\, p(\mathbf{f}|\mathbf{y})\, d\mathbf{f} \\ &\propto \mathcal{N}\Big(\mathbf{m}_0 + \boldsymbol{\kappa}^T (\mathbf{K}+\mathbf{V})^{-1} (\mathbf{y}-\mathbf{m}_0), \boldsymbol{\mathcal{K}} - \boldsymbol{\kappa}^T (\mathbf{K}+\mathbf{V})^{-1} \boldsymbol{\kappa}\Big), \end{aligned} \tag{4}$$

	Dimensionality of the Input Space	Point Geometry	Allowed Kernel Designs
Local GP Experts	no restrictions	best if points naturally occur in clusters	stationary
Inducing-Points Methods	often limited to low-dimensional spaces	point geometry is adjusted, sometimes to a grid	mostly stationary
Structure-Exploiting Methods	no restrictions	points on grid	stationary
Vecchia Approximations	no restrictions	no restrictions	only derived for stationary kernels
Proposed Method	no restrictions	no restrictions	no restrictions

Figure 1. Table comparing the existing approximate methods for large-scale Gaussian processes and the proposed method. The proposed method is the only one with no restrictions on the dimensionality of the input space, kernel design, or data-point geometry.

where $\boldsymbol{\kappa} = k(\mathbf{x}_0, \mathbf{x}_j)$, and $\boldsymbol{\mathcal{K}} = k(\mathbf{x}_0, \mathbf{x}_0)$. This basic framework can be extended by ever-more flexible mean, noise and kernel functions. Our proposed method is entirely agnostic and even symbiotic—in the sense that there is mutual support—to those extensions and we will therefore omit the dependencies thereof.

The bottleneck of training GPs, that is, optimizing (3) with respect to h, is the $O(N^3)$ numerical complexity of calculating $\mathbf{K}(h)^{-1}\mathbf{y}$—or equivalently solving a linear system—and $\ln(|\mathbf{K}(h)|)$, and the storage of $\mathbf{K}$, which scales $O(N^2)$. However, if $\mathbf{K}$ is very sparse, both problems would be avoided. This is the goal of all approximate methods, which work by synthetically inducing this sparsity. In contrast to approximate techniques, we propose to achieve sparsity purely by flexible kernel design, and not through approximations, leading to a sparse but exact GP. The sparsity, in this case, is discovered, not induced. However, if the problem does not have natural sparsity, the constrained optimization described below used to optimize (3) shall guarantee the minimal approximations needed to satisfy system-dictated minimum-sparsity constraints.

Consider, as a simple example, the squared exponential kernel

$$k(\mathbf{x}_1, \mathbf{x}_2) = \sigma_s^2 \ \exp\left(-0.5\ ||\mathbf{x}_1 - \mathbf{x}_2||^2/l^2\right), \tag{5}$$

which is used in approximately 90% of GP applications[17]; even if data points were naturally uncorrelated, the squared exponential kernel would not be able to learn this independence because its global support will always return covariances > 0. This is true for all commonly used stationary kernels and most non-stationary kernels. Instead of formulating kernels that learn well what points are dependent, we propose to consider kernels that are tailored to be capable of learning independence. Such a non-stationary, flexible, and compactly-supported kernel is the first building block of the proposed framework. Even if such a kernel can be defined, the covariance matrix still has to be computed and stored which is time-consuming and often prohibitive due to storage requirements. Distributed computing on HPC compute architecture—as the second building block—can help by splitting up the computational and storage burden. The third building block, augmented and constrained optimization can guarantee that sparse solutions are given preference, or are even a requirement.

Building Block 1: Non-stationary, ultra-flexible and compactly-supported kernel functions. For natural sparsity to be discovered, a kernel function $k(\mathbf{x}_1, \mathbf{x}_2)$ should be designed such that it can flexibly encode correlations between data points, including instances where no correlations exist[18]. The kernel has to meet three requirements:

1. Compact Support: This is the most obvious necessary property. Since we are attempting to discover zero covariances, the kernel has to be compactly supported.
2. Non-Stationarity: Compactly-supported kernels have been used before, but mostly in the stationary case. However, in the stationary case, sparsity is only taken advantage of in an entirely local way—i.e. only if a point happens to be far away from all other points can the covariance be zero. Such kernels are not able to learn more complicated distance-unrelated sparsity-exploiting dependencies.
3. Flexibility: To pick up on sparsity across geometries and distances, a kernel has to be flexible to recognize that neighboring points may be correlated and some points in the distance are not, and vice versa.

Combining compact support, non-stationarity and flexibility yields kernels that are tailored to learn existing and non-existing covariances. Below we examine a few examples to solidify this idea. The kernel

$$k_s(\mathbf{x}_1, \mathbf{x}_2) = \tilde{k}(\mathbf{x}_1, \mathbf{x}_2)\ g(\mathbf{x}_1) g(\mathbf{x}_2), \tag{6}$$

is a rather well-known example of a non-stationary kernel. The subscript "s" stands for "sparsity" since this is the kernel that will allow us later to discover sparsity. The kernel $\tilde{k}$ is assumed to be compactly-supported and stationary (for instance[18]); the non-stationarity is produced by the term $g(\mathbf{x}_1)g(\mathbf{x}_2)$. The flexibility of this kernel depends entirely on the parameterization of g. In the most flexible case, g could be a sum of Kronecker-δ functions centered at a subset the data points $\hat{\mathcal{D}} \subseteq \{\mathbf{x}_i\}_{i=1}^N$, i.e.,

$$g(\mathbf{x}) = \sum_{\mathbf{x}_i \in \hat{\mathcal{D}}} h_i \delta(\mathbf{x}, \mathbf{x}_i),$$

where the $|\hat{\mathcal{D}}|$ binary coefficients $h_i \in \{0, 1\}$ are hyperparameters that may be optimized during training. If we allowed $\hat{\mathcal{D}}$ to include *all* data points, we would obtain a GP that has learned which such points can safely be ignored ($h_i = 0$) without impacting the marginal log-likelihood.

The kernel (6) is very flexible but has two issues. First, it explicitly depends on potentially millions of binary hyperparameters. Second, it is unable to encode varying covariances between data points; points are either turned "on" or "off". An even more flexible kernel, that can in fact turn on and off selected covariances instead of just points, can be defined as

$$k_s(\mathbf{x}_1, \mathbf{x}_2) = \tilde{k}(\mathbf{x}_1, \mathbf{x}_2)\ \left(g_1(\mathbf{x}_1)g_1(\mathbf{x}_2) + g_2(\mathbf{x}_1)g_2(\mathbf{x}_2)\right), \tag{7}$$

where

$$g_1(\mathbf{x}) = \sum_{\mathbf{x}_i \in \hat{\mathcal{D}}} h_i^{g_1} \delta(\mathbf{x}, \mathbf{x}_i) \tag{8}$$

$$g_2(\mathbf{x}) = \sum_{\mathbf{x}_i \in \hat{\mathcal{D}}} h_i^{g_2} \delta(\mathbf{x}, \mathbf{x}_i) \tag{9}$$

where the $h_i^{g_1}$ and $h_i^{g_2} \in \{0, 1\}$ or $\in [0, \infty]$. This kernel can effectively discover that certain covariances (perhaps most) are zero. If we include all data points in $\hat{\mathcal{D}}$, then this kernel has $2N$ hyperparameters to optimize, which can be an overwhelming optimization if N is large.

To alleviate the challenge of a large number of hyperparameters, we can trade some of the flexibility and therefore sparsity for a parameterization with fewer hyperparameters. For this purpose, we propose the kernel function

$$k_s(\mathbf{x}_1, \mathbf{x}_2) = \tilde{k}(\mathbf{x}_1, \mathbf{x}_2) \sum_i^{n_1} g_i(\mathbf{x}_1) g_i(\mathbf{x}_2), \tag{10}$$

where

$$g_i(\mathbf{x}) = \sum_{j=1}^{n_2} a_{ij} \exp\left[\frac{-\beta_{ij}}{1 - \frac{||\mathbf{x} - \mathbf{x}_0^{ij}||_2^2}{r_{ij}^2}} + \beta_{ij}\right] \chi\left(r_{ij} > ||\mathbf{x} - \mathbf{x}_0^{ij}||_2\right). \tag{11}$$

Equation (11) is a sum of, so-called, bump functions, where χ is the indicator function which is 1 if $r_{ij} > ||\mathbf{x} - \mathbf{x}_0^{ij}||_2$ and 0 otherwise, $\mathbf{x}_0^{ij}$ are the bump function locations, r_{ij} are the radii, and β_{ij} are shape parameters. Bump functions are $\in C^\infty$ and compactly supported; precisely the properties we need to create sparsity-discovering kernel functions. The kernel function (10) allows us to seamlessly choose between flexibility, which directly impacts the ability to discover sparsity, and the number of hyperparameters (compare (10) with (7) for $n_1 = 2$ and $n_2 = 1$). See Fig. 3 for a visualization of this kernel. For our test, we will combine the above kernel with a compactly-supported stationary kernel given by

$$\tilde{k}(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} \frac{\sqrt{2}}{3\sqrt{\pi}} \left(\left(3\left(\frac{d}{r}\right)^2\right) \log\left(\frac{\frac{d}{r}}{1 + \sqrt{1 - \left(\frac{d}{r}\right)^2}} \right) + \left(2\left(\frac{d}{r}\right)^2 + 1\right) \sqrt{1 - \left(\frac{d}{r}\right)^2} \right) & \text{if } d < r, \\ 0 \text{ else} \end{cases} \tag{12}$$

where $d = ||\mathbf{x}_1 - \mathbf{x}_2||_2$, and r is the radius of support. Kernel function (12) is a rather well-known compactly-supported stationary kernel. Since kernels can be multiplied, we can combine our sparsity-discovering kernel k_s (10) with any other kernel ($k_c \cdot k_s$), leading to no restrictions on the core kernel k_c. The bump functions in k_s can be normalized and shaped in order to equal one within its support and zero otherwise, which can then be understood as a mask that leaves the core kernel k_c untouched in areas of support. Since the bump function only appears in the kernel in g, any shape will lead to positive semi-definiteness of the kernel. The kernel k_s also gives us the opportunity to estimate the sparsity of the covariance matrix. In the limit of adding infinitely many, uniformly distributed data points inside the fixed domain, the discrete covariance matrix becomes the covariance operator (the kernel) and the number of non-zero entries becomes an integral. In that case, the sparsity s of the covariance matrix is bounded from above such that

$$s = \frac{\text{number of non-zero covariances}}{N^2} \leq \frac{\int_{S_k} d\mathbf{x} d\mathbf{x}}{\int_{\mathcal{X} \times \mathcal{X}} d\mathbf{x} d\mathbf{x}}, \tag{13}$$

which we can use to formulate objective functions that allow us to give preference to sparse solutions, or to formulate sparsity constraints; note that a small s here means high sparsity. S_k in (13) is the set of support of the kernel, i.e., $S_k \subset \mathcal{X} \times \mathcal{X}$, which, in our case is the Cartesian product of two balls $\mathcal{B} \subset \mathcal{X}$—the volume of the Cartesian-product set of two balls embedded in $\mathbb{R}^n$, $\mathcal{B}_1 \times \mathcal{B}_2$ is the product of their respective volumes. Therefore, for kernel (10), the sparsity can be bounded from above—assuming entirely disjoint support since any overlap increases sparsity (lowers s)—so that

$$\sup\left\{ \int_{S_k} d\mathbf{x} d\mathbf{x} \right\} = \sum_i^{n_1} \sum_j^{n_2} \sum_k^{n_2} Vol_s(dim, r_{ij}) Vol_s(dim, r_{ik}), \tag{14}$$

where $Vol_s(dim, r)$ is the volume of a dim-dimensional sphere with radius r, defined as

$$Vol_s(dim, r) = \frac{\pi^{dim/2}}{\Gamma\left(\frac{dim}{2} + 1\right)} r^{dim}, \tag{15}$$

where Γ is the gamma function, and dim is the dimensionality of $\mathcal{X}$. As $\beta \to 0$ in Eq. (11), the kernel's effect on k_c in regions of support vanishes (see Fig. 3 for an example). β can therefore be seen as a shape parameter. As $\beta \to \infty$, the bump functions become delta functions and kernel (7) is obtained.

Building Block 2: High-performance computing to take advantage of sparse kernels. While flexible non-stationary and compactly-supported kernels are the core building block of our algorithm for extreme-scale GPs, the covariance matrix has to be computed in a dense format first to take full advantage

of multi-threading, however, this could violate RAM restrictions for large datasets; computing the covariance matrix in a sparse format in place would be prohibitively inefficient. To avoid slow computations or going beyond the RAM limit, we define a "host" covariance matrix (on one host machine) as sparse in the first place, compute dense sub-matrices in a distributed way and cast them into a sparse format, and only communicate sparse sub-matrices back to the host machine, where they are inserted into the host covariance matrix. Through this strategy, we address RAM limitations by distributing the covariance matrix across many computing resources—and could even exploit out-of-core methodologies such as utilizing disk storage if needed. Additionally, the computation time is sped up by leveraging heterogeneous architectures such as GPUs, efficient at data-parallel operations, and threading-task-parallel CPU operations. The combination of distributing memory and exploiting parallelism across cores allows our algorithm to operate on datasets of practically-unlimited size—given enough distributed workers and sufficient natural sparsity. The procedure is illustrated in Fig. 4 and shown in pseudo-code 1.

We split up the dataset of length $|\mathcal{D}|$ into batches of size b. For large $|\mathcal{D}|$, the only way the covariance matrix can be computed is by distributing the computational burden by dividing the host covariance matrix into block sub-matrices, each representative of a unique data-batch pair (see Fig. 4). The batch pairs are transmitted to different workers (often a few per node) via the Python library DASK; we denote the number of parallel-executed tasks by n (one task per worker). In each task, the exact batch-covariance is computed. Because of the specifically-designed kernel, many elements of each sub-matrix will be zero. That way, theoretically, any-size covariance matrix can be computed and stored in a distributed way. As the sub-matrices are transferred back, they will be translated into a sparse representation and injected into the sparse host covariance matrix on the host machine. While this matrix is $|\mathcal{D}| \times |\mathcal{D}|$ in size, its sparsity avoids problems with storing or computations. The computation of a batch of the covariance matrix can be accelerated by taking advantage of the many parallel threads a GPU or CPU has to offer. Future work will compare the compute performance of different implementations and architectures.

The proposed algorithm, given more resources, is able to compute solutions faster, exhibiting the strong scaling properties inherent in the design (see Fig. 5). Furthermore, as the problem size increases, the algorithm matches the set of resources also highlighting weak scaling. In summary, our formulation speeds up computation, reduces memory burden, and provides an ability to exploit heterogeneous architectures (CPUs/GPUs/TPUs), providing future compatibility of the proposed framework since future architectures can be leveraged.

The theoretical computing time of the covariance matrix can be calculated as

$$T_c = \frac{|\mathcal{D}|}{2nb}\left(\frac{|\mathcal{D}|}{b} + 1\right) t_b, \tag{16}$$

where t_b is the compute time for one sub-matrix, whose scaling depends on the exact implementation, and availability and number of parallel CPU or GPU threads. Equation (16) suggests that, as the number of tasks n, the number of parallel workers, approaches $\frac{|\mathcal{D}|}{b}$, the scaling becomes linear in $|\mathcal{D}|$, i.e. complexity $O(|\mathcal{D}|)$. By extension, as the number of workers approaches the total batch number, the scaling becomes constant. The linear-system solution can be accomplished by the conjugate gradient method which has numerical complexity $O\left(m\sqrt{k}\right)$, where m is the number of non-zero entries in the covariance matrix and k is the condition number. The log-determinant computation can be done via Cholesky factorization whose scaling depends on the exact structure of the matrix. Furthermore, since for most intents and purposes $\frac{|\mathcal{D}|}{b} >> 1$, we can approximate

$$T_c \approx \frac{|\mathcal{D}|^2 t_b}{2nb^2}, \tag{17}$$

which can help estimate the optimal batch size given a particular architecture. For sequential computations, t_b scales $O\left(b^2\right)$ and the batch size drops out of the equation. For the other extreme, perfect parallelization, t_b scales $O(b)$ and we, therefore, want to maximize the batch size up to the point where the linear scaling stops. That number depends on the particular architecture.

Algorithm 1 Distributed Covariance Computation

```
procedure COMPUTE COVARIANCE
    SparseCovariance = sparse matrix(dataset size,dataset size)
    tasks = []
    for i = 0:number of batches do in parallel
        for j = i:number of batches do in parallel
            select worker
            tasks.append(dask.distributed.client.submit(kernel,batches,batch_coordinates,hyperparameters))
            SparseCovariance, tasks = collect_submatrices(SparseCovariance,tasks)
        end for
    end for
    client.gather(tasks)
end procedure

procedure COLLECT_SUBMATRICES(SparseCovariance,tasks)
    new_tasks = []
    for task in tasks do
        if task is finished then
            submatrix = task.result() #only sparse matrices are communicated
            SparseCovariance = insert(submatrix) #insert the sub-matrix into host covariance matrix
        else
            new_tasks.append(task)
        end if
    end for
    return SparseCovariance, new_tasks
end procedure
```

Building Block 3: Augmented and constrained optimization. The proposed kernel definition (10) can be paired with constrained

$$\begin{aligned} &\underset{h}{\operatorname{argmax}} \ln(L) \\ &\text{subject to } s < \text{ sparsity requirement} \end{aligned} \tag{18}$$

or augmented optimization

$$\underset{h,s}{\operatorname{argmax}}(\ln(L) + (1-s)\ln(L)), \tag{19}$$

where s is the estimated sparsity (Eq. 13) (not a Lagrange multiplier). The constraint means that this formulation will only be exact up until the RAM restriction is hit, then the GP will turn itself into an approximate GP, but without the need for the user to make decisions on which points are being considered. The augmented (or biased) optimization will always prefer sparse covariance matrices. However, caution has to be exercised to ensure the optimization is not dominated by the need for sparsity. The formulation in Eq. (19) gives priority to the likelihood, since $s \in [0,1]$, which means the objective function is bounded by $[\ln(L), 2\ln(L)]$.

A note on solving linear systems, log-determinants, and optimization strategies. After computing the sparse covariance matrix in a distributed fashion, all that is left to do to enable GP training is to optimize the marginal log-likelihood. For this, we need to solve

$$\mathbf{Kx} = \mathbf{y} \tag{20}$$

and compute

$$\log(|\mathbf{K}|). \tag{21}$$

A common approach in the dense and the sparse case is to use Cholesky or LU factorization. Given the factorization, both the linear-system solution and the log-determinant computation is trivial. However, even for a sparse input matrix, both Cholesky and LU might have large memory requirements depending on fill-in and pivoting options. In addition, for those decomposition methods to be successful, the matrices have to be extremely sparse with only a handful of non-diagonal non-zero entries; a level of sparsity we might not be able to guarantee for matrices originating from a GP. In our experience, it is better to use iterative methods (e.g. conjugate gradients) to solve the linear system. This leaves us with the problem of estimating the log-determinant accurately. For this work, we have employed random linear algebra (RLA). More specifically, we have implemented the method presented in[19]. Since we are training the GP via Markov-Chain Monte-Carlo the random noise induced by RLA

won't affect the training. As we move to more deterministic optimizers, especially derivative-based optimizers, this discussion will have to be revisited.

A moderately-sized example to verify error convergence

To demonstrate the functionality of the method, we investigate the error convergence of the GP-predicted model. Our proposed method is only viable if the ground truth can be recovered. The data we use is the United States topography. Of the 25000 points, we choose 24000 points as the training dataset and 1000 points as the test dataset $\{\mathbf{x}_i^{test}, y_i^{test}\}$. While this dataset is only moderately large, it is outside of the capabilities of most exact-GP algorithms. We chose a smaller dataset to be able to calculate the root-mean-square error (RMSE) in each iteration of the training somewhat efficiently. For this example, we used kernel (10) with $n_1 = n_2 = 1$. We are employing a Markov-Chain-Monte-Carlo (MCMC) algorithm for the training. The RMSE is defined as

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(y_i^{test} - f_i\right)^2}, \tag{22}$$

where y_i^{test} are the test measurements (elevations) and f_i are the calculated posterior means (Eq. (4)). The result is presented in Fig. 6. From this result, we conclude that the error converges toward zero as the hyperparameter search progresses.

A climate example with over 5 million data points

We demonstrate the proposed methodology on a temporal extension of the dataset shown in Fig. 2. The dataset contains daily maximum temperatures from 1990 to 2019 from circa 7500 gauge-based weather stations across the continental United States[20,21]; after accounting for missing daily measurements, these stations yield over 51.6 million data points across this approximately 10000-day (30-year) period. Due to computing constraints, while writing this paper, we randomly extracted a dataset of 5165718 points to use for our example. For reproducibility purposes, the dataset can be found online at

ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/. For our tests, we used the *gp2Scale* library that is part of the *fvGP* Python package, available from GitHub (*https://github.com/lbl-camera/fvGP*) and pypi (*pip install fvgp*).

A Gaussian process is needed to analyze these data for a variety of reasons. First, while in situ measurements of daily weather variables provide the most realistic data source for understanding historical climate, users of such data often require geospatially-interpolated datasets that account for irregular sampling density and provide a complete picture of how temperatures vary over space. Daily maximum temperatures furthermore exhibit strong spatial autocorrelations due to their driving physical mechanisms in the ocean and atmosphere, which GPs are particularly well-suited to model statistically. In the temporal domain, autocorrelations are also generally quite strong in daily temperatures due to, e.g., seasonality imposed by the solar cycle, and as such GPs are needed to appropriately impute missing measurements at the gauge locations.

For this example, we defined the kernel as

$$k(\mathbf{x}_1, \mathbf{x}_2) = \tilde{k}(\mathbf{x}_1, \mathbf{x}_2) \cdot \left(g_1(\mathbf{x}_1)g_1(\mathbf{x}_2) + g_2(\mathbf{x}_1)g_2(\mathbf{x}_2)\right), \tag{23}$$

where $\tilde{k}$ is defined in Eq. (12), and g_1 and g_2 are defined in Eq. (11) with $n_2 = 4$, giving rise to 42 hyperparameters.

To deliver a proof-of-concept of the proposed strategy, we are again employing a Markov-Chain-Monte-Carlo (MCMC) training up to 160 function evaluations. Since the total compute time scales linearly with the number of function evaluations, it is straightforward to estimate the compute time for many other training strategies. For this test, we chose two different architectures, namely Nersc's Cori Haswell Nodes (*https://www.nersc.gov/systems/cori/*) and Perlmutter's GPU nodes (Perlmutter Phase 1: *https://www.nersc.gov/systems/perlmutter/*). Due to challenges with allocating DASK workers on Cori, the result shown was calculated on 256 of Perlmutter's A100 Nvidia GPUs. Computing a batch of size 10000 can be accomplished in circa 0.6 seconds on each GPU node. See Fig. 7 for the visualization of the result.

Due to early-access constraints, we split up this run into 4 separate runs, storing the hyperparameters and therefore the state of the training. Therefore, the total run time of 24 h contains 4 initializations. Each iteration of the MCMC took circa 460 sec, leading to a total estimated run time of 72384 sec. We also included information about error convergence in the figure using a smaller subset of the full dataset.

Summary, discussion, and conclusion

In this paper, we have proposed a new methodology and algorithm for extreme-scale exact Gaussian processes (GPs) based on flexible, non-stationary, and compactly-supported kernels, and distributed computing. Our method is not another approximate GP but is designed to discover—not induce—naturally occurring sparsity and use it to alleviate challenges with numerical complexity in compute time and storage. It is our strong belief that this natural sparsity is very common in many modern datasets. The fundamental assumption in this work is that GPs often give rise to sparse covariance matrices naturally if given enough flexibility, through non-stationary kernel designs, to discover the sparsity. This can only be achieved for kernels that are very flexible, non-stationary, and compactly supported. For efficiency reasons, the covariance still has to be computed in a dense format first which is accomplished by distributing the workload over many CPU or GPU nodes. Constrained or augmented optimization is used to give sparse solutions priority or to constrain sparsity. These constraints only take effect when RAM or computing restrictions of the system are exceeded and would then turn the exact GP into an optimal sparse GP.

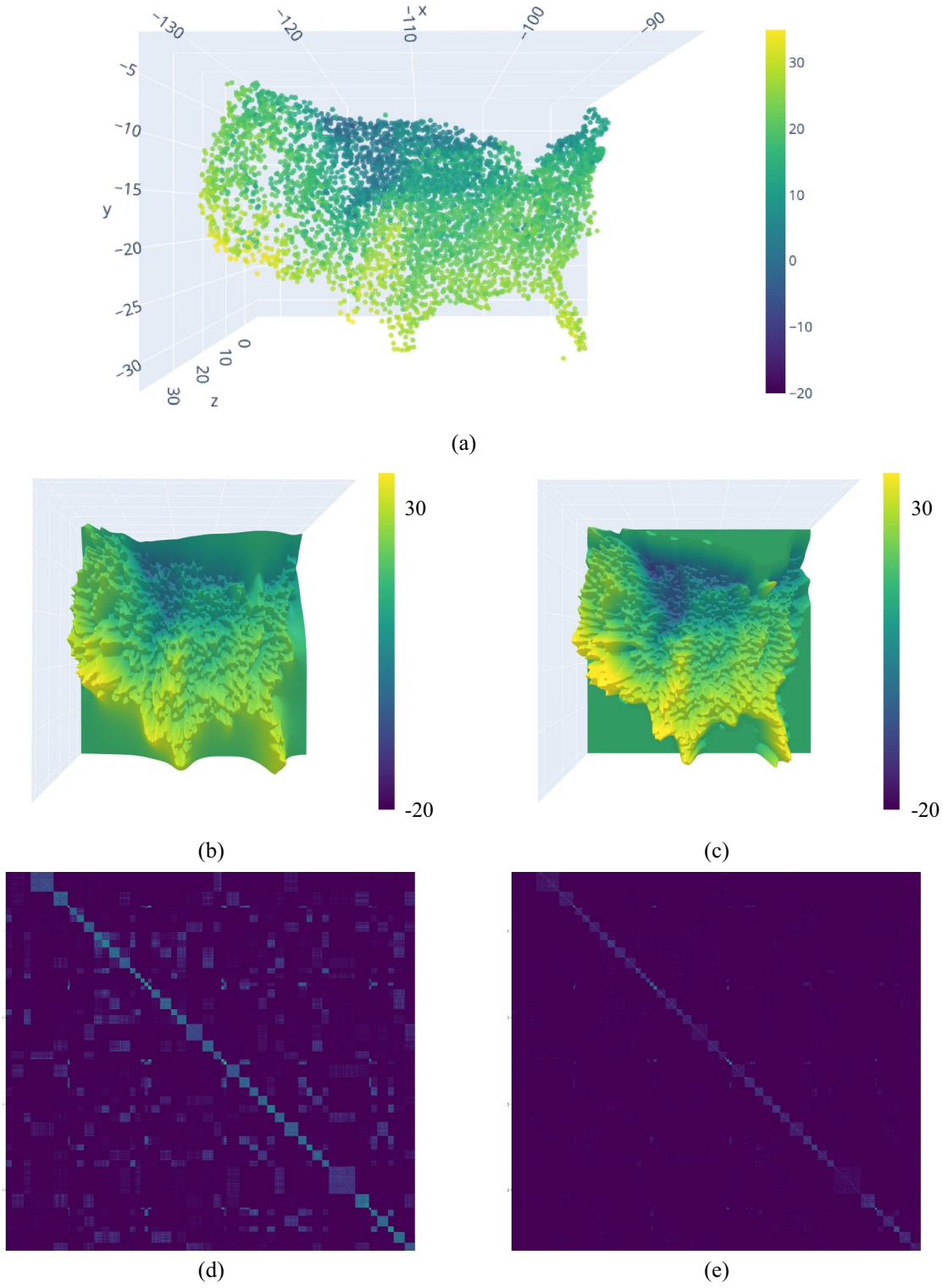


Figure 2. Figure illustrating the premise of our proposed algorithm. Panel (**a**) shows the test data, measured daily maximum temperatures (°C) from April 10th, 1990 across the United States ($N = 4718$). This problem size is still well within the capabilities of a standard GP, whose posterior mean is shown in (**b**). If we employ a flexible, non-stationary, and compactly-supported kernel, we can learn through optimization of the marginal log-likelihood that only a few covariances are of essence for the prediction. Our sparse result is shown in (**c**). Panels (**d**) and (**e**) show the covariance matrix of the dense and sparse GP, respectively, where the sparse covariance only has 1.5% of the non-zero entries of the full dense matrix. The sparsity in this problem is discovered, not induced, leading to an exact GP. This principle, in combination with HPC, for truly large covariance matrices, and constrained function optimization enables GPs to be scaled to tens of millions of data points.

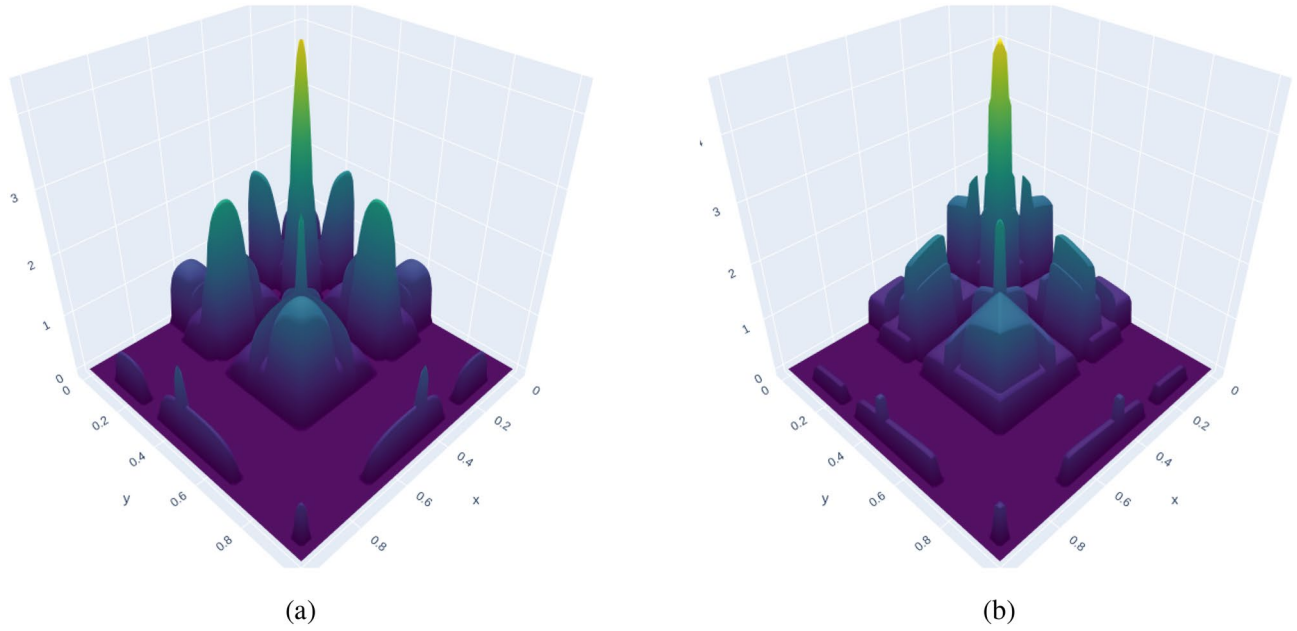


Figure 3. Figure showing a very flexible, non-stationary and compactly-supported kernel function $k(x, y)$ (panel **a**, k_s Eq. (10)), and a "sparsified" squared-exponential core kernel k_c (panel **b**). Only for a one-dimensional input domain, we can visualize the kernel as a function over $\mathcal{X} \times \mathcal{X} \subset \mathbb{R} \times \mathbb{R}$. The kernel uses a set of compactly-supported bump functions to naturally discover sparsity through optimization of the bump functions' positions, heights, radii, and shapes.Since any multiplication of kernels is a valid kernel, our sparsity-discovering kernel k_s (panel **a**) can be combined with any kernel; therefore, compared to most approximate methods, it does not limit the user's ability to design and employ arbitrary kernel functions. Panel (**b**) shows that concept; where the kernel k_s has support, the covariance function becomes the squared-exponential kernel.

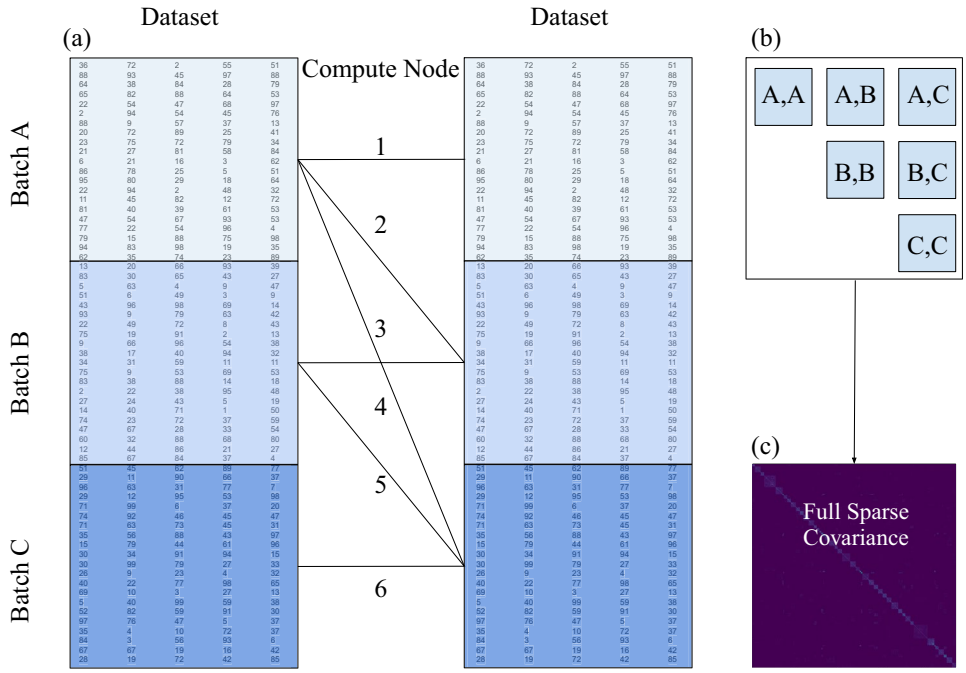
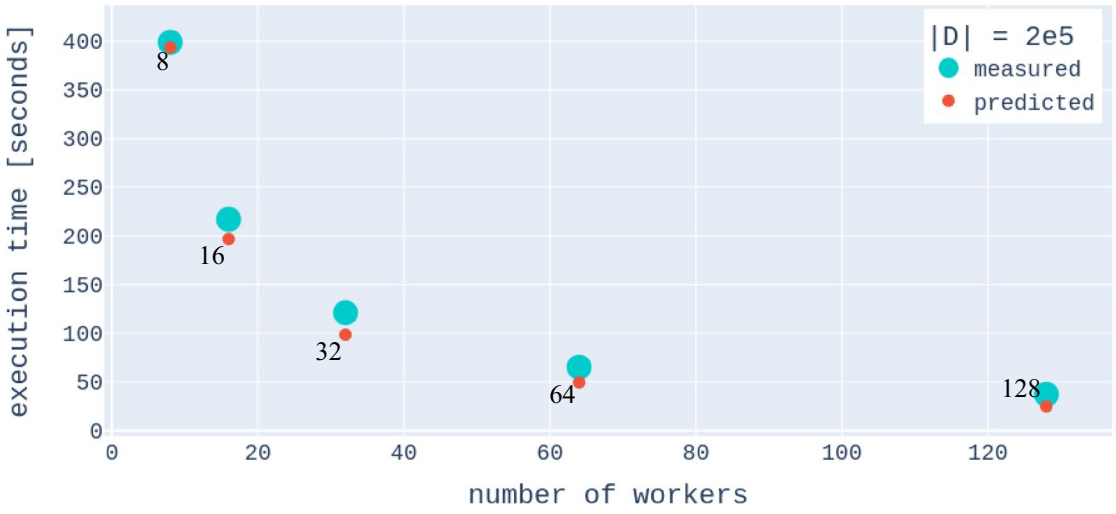
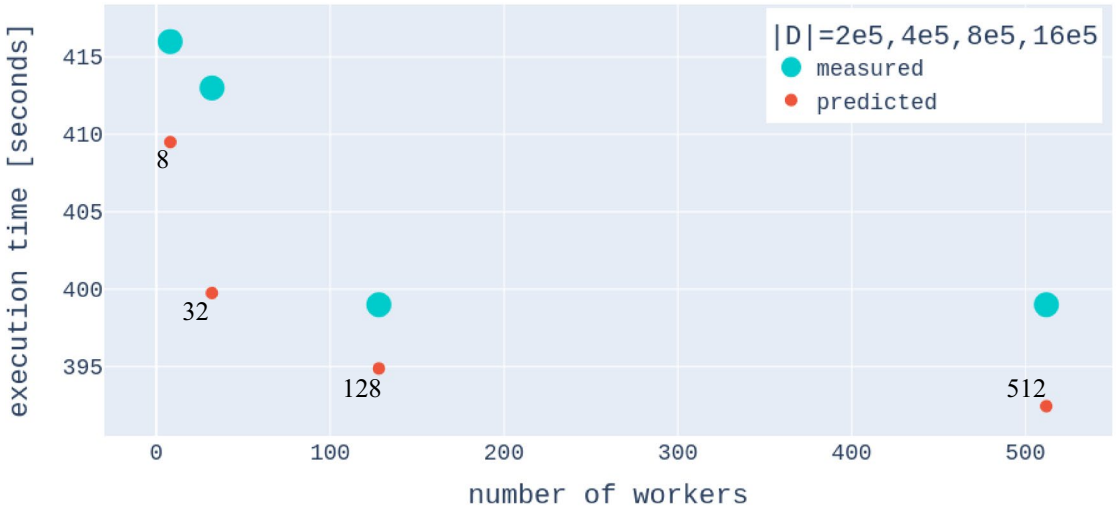


Figure 4. Figure illustrating the computational building block of the proposed algorithm. The dataset is divided into batches (panel **a**). Pairs of batches are sent to the compute nodes where the associated sub-matrices of the covariance matrix are calculated using the presented sparse kernels (Eq. (10), panel **b**). The sub-matrices are cast into a sparse format on the compute nodes before being sent back to the host. There, they get assembled to obtain the full sparse master covariance matrix (panel **c**). All subsequent mathematical operations needed for a GP, namely calculating the log-determinant and solving a linear system, are performed efficiently on the sparse covariance matrix.



(a)

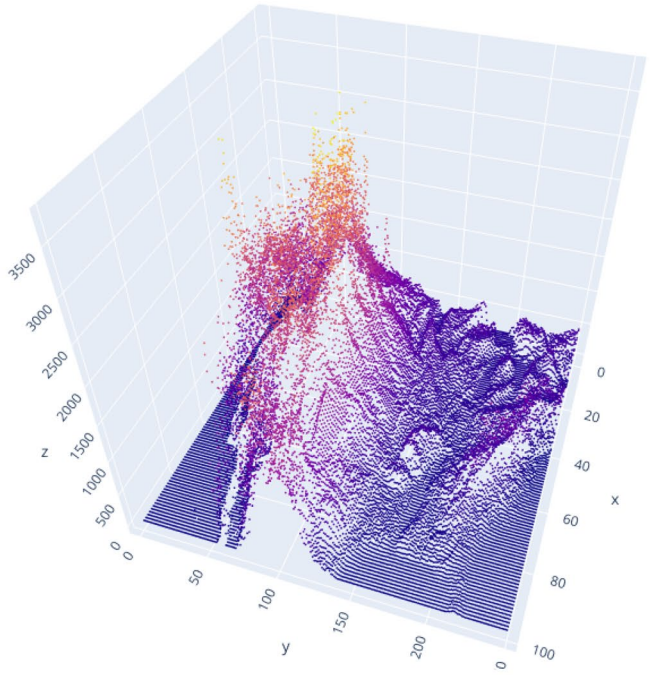


(b)

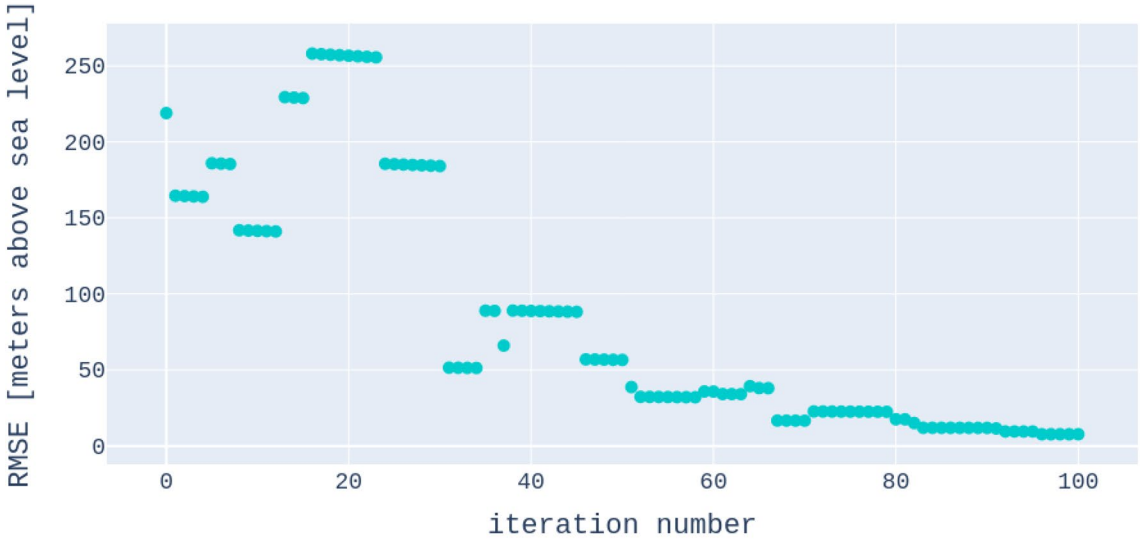
Figure 5. Figure illustrating theoretical and measured strong and weak scaling of the distributed covariance computation. (**a**) The computation time of a problem of fixed size as a function of the number of workers. (**b**) Computation time as a function of the number of workers while the problem dataset size is increased (from the left 2e5, 4e5, 8e5, 16e5, also see label). The figures suggest that there is a strong case to be made for the favorable scalability of exact Gaussian processes. The exact number of workers in each run is indicated as numbers adjacent to the dots.

This work is at a proof-of-concept stage; therefore, there are several challenges with the current form and these will be addressed in future work:

1. The sparsity-discovering kernel for our examples was relatively simple. It has to be shown that much more flexible bump-function-based kernels can be formulated and their hyperparameters can be found robustly. However, there is a trade-off to consider; a more flexible kernel will lead to better detection of sparsity, but a more costly optimization of the hyperparameters. More hyperparameters also mean possible ill-posed optimization problems.
2. We have used MCMC for training, which means only having to evaluate the marginal log-likelihood. The proposed method should be extended for gradient-based optimization of the hyperparameters.
3. While our covariance matrix is computed in a distributed manner, the linear-system solutions and log-determinant computations are serialized even though most workers are idle and should be used for that task. However, the observed sparsity was found to be so substantial that the computations were not a bottleneck.



(a)



(b)

Figure 6. To verify the functionality of the proposed methodology, we present the error convergence between the GP posterior-mean prediction and the test data. Panel (**a**) shows the full data set, the topography of the United States (meters above sea level) evaluated at 25000 points. From that set, we selected 24000 data points for training and 1000 test points randomly and calculated the RMSE as the hyperparameter search via MCMC progressed. Panel (**b**) shows the error convergence. We can confirm that the proposed methodology leads to error convergence as we approach the final hyperparameters.

Despite those shortcomings, the method has shown its strength by training a Gaussian process on more than five million data points. This is, to our knowledge the largest exact GP ever trained. Given the strong and weak scaling shown in Fig. 5 and predicted by Eq. (16), we are confident that exact GPs on 100 million data points are currently possible. The code is available as part of the open-source python packages *fvGP* and *gpCAM*.

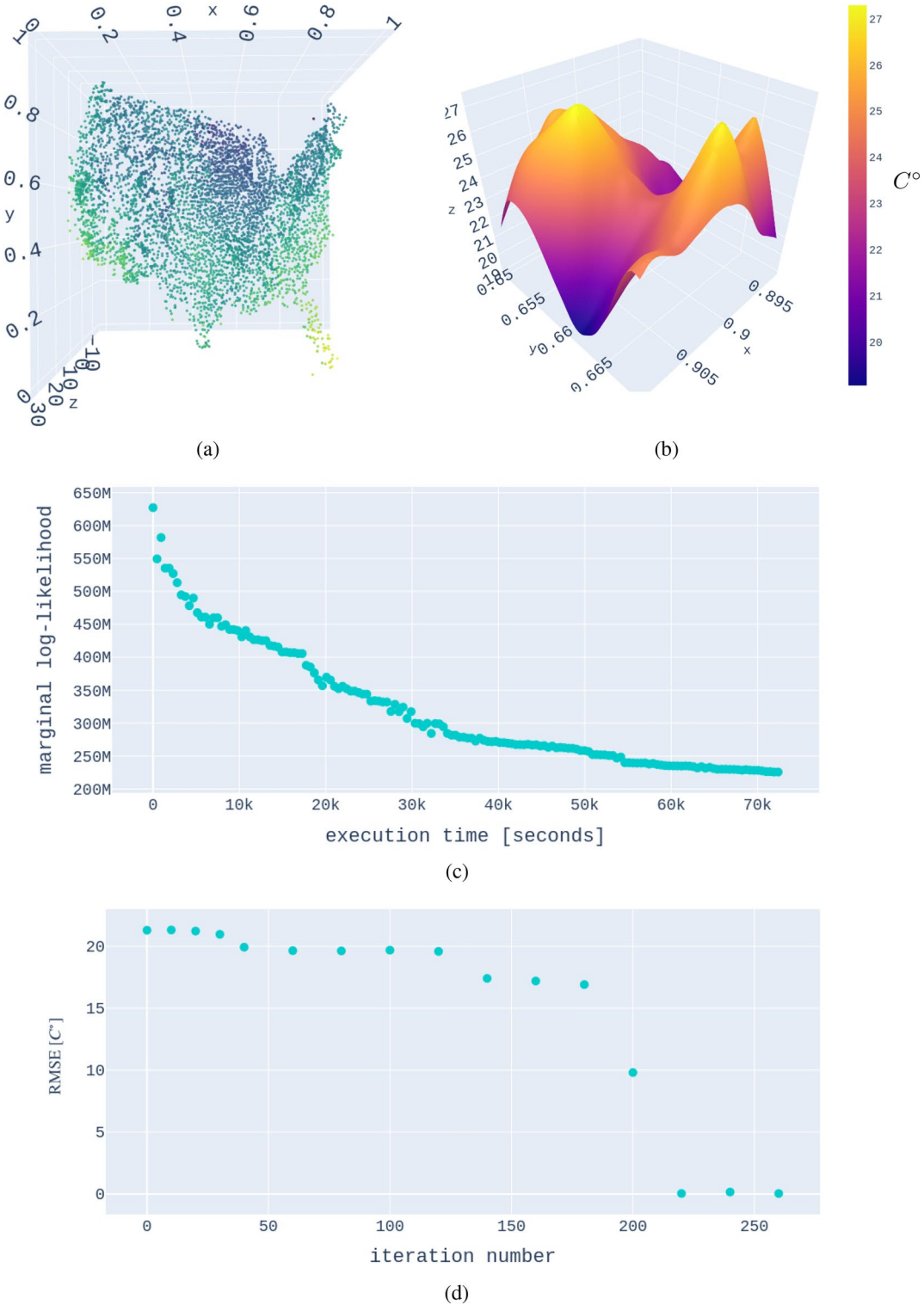


Figure 7. The result of a Gaussian process trained on over 5 million data points. While this paper is best understood as a proof-of-concept, we want to ensure that we show the readers that the resulting model is reasonable by the end of our training (**a**, **b**). (Panel **a**) The distributions of the climate stations with temperatures from the first day of the dataset (Jan 1st, 1990); the axes are normalized. (Panel **b**) The GP interpolation over a subdomain in the northeast at a time slice in June 2004. The noise of the measurement was estimated ad-hoc, which explains the somewhat rough appearance of the posterior-mean function. We trained the GP via MCMC for 160 iterations. While this does not reach convergence, it is enough to demonstrate the feasibility of such an extreme-scale GP. Panel (**c**) shows the marginal log-likelihood as a function of training time. The GP was trained in under 24 h, on 256 GPUs, opening the doors for much larger GPs. To verify error convergence, we also extracted a smaller dataset of 103315 points from the full climate dataset. The RMSE with respect to 1000 test points as a function of MCMC iteration number is visualized in panel (**d**).

Data availability

The topography dataset is available at https://drive.google.com/file/d/1BMNsdv168PoxNCHsNWR_znpDswjdFxXI/view. The climate datasets analyzed during the current study are available from NOAA, https://www.ncei.noaa.gov/data/global-historical-climatology-network-daily/.

Received: 10 June 2022; Accepted: 15 February 2023
Published online: 13 March 2023

References

1. Manzhos, S. & Ihara, M. In *On the optimization of hyperparameters in gaussian process regression*. arXiv preprint arXiv:2112.01374 (2021).
2. Williams, C. K. I. & Rasmussen, C. E. *Gaussian Processes for Machine Learning* Vol. 2 (MIT press, Cambridge, 2006).
3. Cohen, S., Mbuvha, R., Marwala, T. & Deisenroth, M. Healing products of gaussian process experts. In *International Conference on Machine Learning 2068–2077* (PMLR, 2020).
4. Gao, Y., Li, N., Ding, N., Li, Y., Dai, T. & Xia, S.-T. Generalized local aggregation for large scale gaussian process regression. In *2020 International Joint Conference on Neural Networks (IJCNN)* 1–8 (IEEE, 2020).
5. Schürch, M., Azzimonti, D., Benavoli, A. & Zaffalon, M. in*Correlated product of experts for sparse gaussian process regression*. arXiv preprint arXiv:2112.09519 (2021).
6. Wilson, A. & Nickisch, H. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *International Conference on Machine Learning* 1775–1784 (PMLR, 2015).
7. Banerjee, S., Gelfand, A. E., Finley, A. O. & Sang, H. Gaussian predictive process models for large spatial data sets. *J. R. Statist. Soc. Ser. B (Statist. Methodol.)* **70**(4), 825–848 (2008).
8. Finley, A. O., Sang, H., Banerjee, S., Gelfand, A. E. & Alan, E. Improving the performance of predictive process modeling for large datasets. *Comput. Statist. Data Anal.* **53**(8), 2873–2884 (2009).
9. Cressie, N. & Johannesson, G. Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc. Ser. B (Statist. Methodol.)* **70**(1), 209–226 (2008).
10. Stein, M. L. Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statist.* **8**, 1–19 (2014).
11. Remes, S., Heinonen, M., Kaski, S. Non-stationary spectral kernels. *Adv. Neural Inform. Process. Syst.* **30** (2017).
12. Vecchia, A. V. Estimation and model identification for continuous spatial processes. *J. R. Statist. Soc. Ser. B (Methodol.)* **50**(2), 297–312 (1988).
13. Katzfuss, M. & Guinness, J. A general framework for vecchia approximations of gaussian processes. *Stat. Sci.* **36**(1), 124–141 (2021).
14. Heaton, M. J. *et al.* A case study competition among methods for analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* **24**(3), 398–425 (2019).
15. Liu, H., Ong, Y.-S., Shen, X. & Cai, J. When gaussian process meets big data: a review of scalable GPS. *IEEE Trans. Neural Netw. Learn. Syst.* **31**(11), 4405–4423 (2020).
16. Wang, K. *et al.* Exact gaussian processes on a million data points. *Adv. Neural. Inf. Process. Syst.* **32**, 14648–14659 (2019).
17. Pilario, K. E., Shafiee, M., Cao, Y., Lao, L. & Yang, S.-H. A review of kernel methods for feature extraction in nonlinear process monitoring. *Processes* **8**(1), 24 (2020).
18. Melkumyan, A. & Ramos, F. T. A sparse covariance function for exact gaussian process inference in large datasets. In *Twenty-First International Joint Conference on Artificial Intelligence* (2009).
19. Boutsidis, C., Drineas, P., Kambadur, P., Kontopoulou, E.-M. & Zouzias, A. A randomized algorithm for approximating the log determinant of a symmetric positive definite matrix. *Linear Algebra Appl.* **533**, 95–117 (2017).
20. Menne, M. J. *et al.* An overview of the global historical climatology network-daily database. *J. Atmos. Ocean. Technol.* **29**(7), 897–910 (2012).
21. Menne, M. J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R. S., Gleason, B.E. & Houston, T. G. *Global Historical Climatology Network - Daily (GHCN-Daily), Version 3*. NOAA National Climatic Data Center (Accessed 03 August 2020) (2012a).

Acknowledgements

The idea-creation phase of this work and the writing of the first version of the manuscript were funded through the Center for Advanced Mathematics for Energy Research Applications (CAMERA), which is jointly funded by the Advanced Scientific Computing Research (ASCR) and Basic Energy Sciences (BES) within the Department of Energy's Office of Science, under Contract No. DE-AC02-05CH11231. Further development of the kernel designs and testing was supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231. This work was further supported by the Regional and Global Model Analysis Program of the Office of Biological and Environmental Research in the Department of Energy Office of Science under contract number DE-AC02-05CH11231. This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain the correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231 using NERSC award m4055-ERCAP0020612. The authors thank Jeffrey Donatelli from Lawrence Berkeley National Laboratory for reviewing the manuscript.

Author contributions

M.M.N. had the initial idea motivated by discussions with the whole team, derived the mathematics for the kernels with help from M.D.R. and K.G.R., implemented the code with help from H.K., and wrote the first draft of the manuscript. M.D.R. revised the statistical parts of this work and the manuscript. K.G.R. checked and corrected the mathematical derivations. H.K. devised the technical components of the HPC implementation. The whole team revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.M.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.