ROBUST, ONLINE, AND ADAPTIVE DECENTRALIZED GAUSSIAN PROCESSES

Fernando Llorente[1], *Daniel Waxman*[2], *Sanket Jantre*[1], *Nathan M. Urban*[1], *Susan E. Minkoff*[1]

[1] Applied Mathematics Department, Brookhaven National Laboratory
[2] Department of Electrical and Computer Engineering, Stony Brook University

ABSTRACT

Gaussian processes (GPs) offer a flexible, uncertainty-aware framework for modeling complex signals, but scale cubically with data, assume static targets, and are brittle to outliers, limiting their applicability in large-scale problems with dynamic and noisy environments. Recent work introduced decentralized random Fourier feature Gaussian processes (DRFGP), an online and distributed algorithm that casts GPs in an information-filter form, enabling exact sequential inference and fully distributed computation without reliance on a fusion center. In this paper, we extend DRFGP along two key directions: first, by introducing a robust-filtering update that downweights the impact of atypical observations; and second, by incorporating a dynamic adaptation mechanism that adapts to time-varying functions. The resulting algorithm retains the recursive information-filter structure while enhancing stability and accuracy. We demonstrate its effectiveness on a large-scale Earth system application, underscoring its potential for in-situ modeling.

***Index Terms*—** Gaussian processes, random features, robust learning, decentralized inference, Earth system modeling.

1. INTRODUCTION

Gaussian processes (GPs) are widely used in signal processing and machine learning due to their ability to capture complex dependencies with principled uncertainty quantification [2]. However, standard GP inference incurs cubic cost in the number of observations, assumes static functions, and is sensitive to outliers. These limitations are further amplified in modern applications where data arrive sequentially, are distributed across several sensors, and both computation and communication resources are limited [3]. In such settings, the model must learn online, adapt to time-varying environments, and facilitate robust inference.

In recent years, researchers have adapted GPs to use online approximations to enable sequential updates [4, 5], distributed algorithms based on product-of-experts fusion to scale to large datasets [6, 7], and decentralized approaches to training and inference in multi-agent systems [8–10]. Robust filtering [11–14] and dynamic kernels [1, 15–17] further extend GPs to handle outliers and temporal evolution. Online kernel machines such as random–feature methods and distributed consensus updates–address similar goals of scalable, streaming learning but lack principled, calibrated uncertainty quantification [18–20]. At the application level, decentralized and distributed GPs have been used for cooperative control of multi-agent systems [21], distributed environmental mapping and sensing [22], collective online learning in large-scale systems [10], Earth system modeling [23, 24], and spatiotemporal traffic modeling and mobility-on-demand systems [25].

Despite this progress, existing decentralized GP algorithms fail to meet at least one of the following desiderata: (a) approximation is
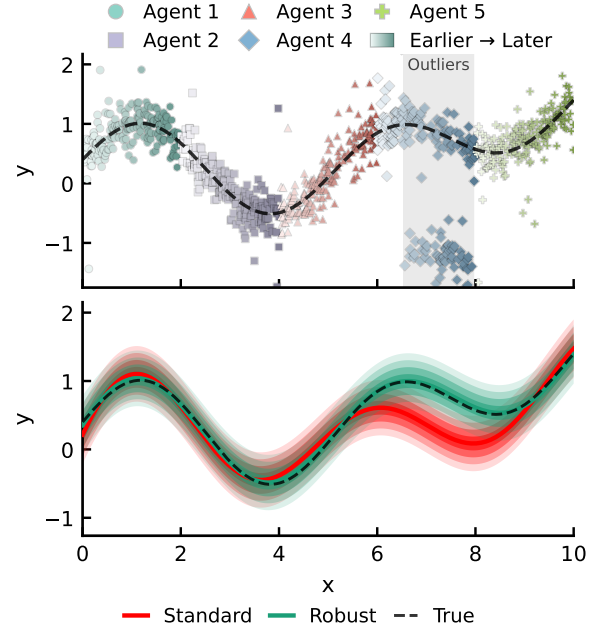

Fig. 1: An illustration of our proposed online, decentralized, and robust GP inference approach. **Top:** Five agents receive data sequentially (with a data point's shade representing its temporal proximity). Agent 4's data is contaminated by localized random outliers. **Bottom:** The standard online decentralized RF-GP [1] becomes inaccurate due to these outliers, but our robust method, using tools from robust filtering theory, maintains superior performance.

clearly controlled via hyperparameters; (b) online inference is possible; (c) they are easily adaptable to time-varying or spatiotemporal environments; and (d) they are robust to contamination via outliers.

A step toward these goals is the *decentralized random Fourier feature Gaussian process* (DRFGP) [26] which uses random Fourier features to map inputs into a finite-dimensional embedding, approximating a GP with a stationary kernel by a Bayesian linear model [27, 28]. Exact and recursive updates are then available via Kalman filtering. Using natural parameters, additive consensus provides a decentralized solution, where agents learn single, shared GP model, in contrast to product-of-experts schemes that do not preserve a coherent global GP [6, 7]. We further discuss the DRFGP in Section 2. The DRFGP can be viewed both as a decentralization of online GPs via basis expansions, [29], or as a kernelized analogue of distributed Bayesian linear regression that aggregates sufficient statistics by consensus [30].

In this paper, we extend DRFGP with two simple yet effective

enhancements: (i) per-agent, per-sample robust updates that downweight atypical observations, inspired by the robust filtering literature; and (ii) an explicit dynamic model that adapts to time-varying functions. Together, these yield the *robust, online, and adaptive decentralized GP (ROAD-GP)*. The enhancements are straightforward to integrate and, in practice, substantially improve stability and accuracy. We demonstrate the approach in an Earth system application that assimilates large volumes of streaming data, showing that ROAD-GP enables effective sequential, distributed inference.

2. DECENTRALIZED RF-BASED GAUSSIAN PROCESSES

We first review the DRFGP [26], recalling how random Fourier feature GPs admit a Bayesian linear model form with an analytical posterior. We then present the information-filter formulation, which naturally enables online and distributed inference and underpins our robust and dynamic extensions.

2.1. Fusion Center Solution

We begin with the centralized setting, in which all data are available at a single location. This setting serves as the reference solution that decentralized algorithms aim to approximate. We want to learn a function $f(\mathbf{x}) \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^d$ using a GP with a stationary kernel $k(\mathbf{x}, \mathbf{x}')$. In the random feature GP (RF-GP), the stationary kernel of a GP is approximated using random Fourier features [27], i.e., $k(\mathbf{x}, \mathbf{x}') \approx \phi(\mathbf{x})^\top \phi(\mathbf{x}')$ where $\phi(\mathbf{x}) \in \mathbb{R}^{2J}$ is defined as

$$\phi(\mathbf{x}) = \frac{1}{\sqrt{J}} \left[\sin(\mathbf{x}^\top \mathbf{v}_1), \cos(\mathbf{x}^\top \mathbf{v}_1), \ldots, \sin(\mathbf{x}^\top \mathbf{v}_J), \cos(\mathbf{x}^\top \mathbf{v}_J) \right]^\top$$

This mapping is computed using samples from the spectral density of $k(\mathbf{x}, \mathbf{x}')$ [27,28]. Then the RF-GP approximation of $f(\mathbf{x})$ reduces to a Bayesian linear model with parameters $\boldsymbol{\theta} \in \mathbb{R}^{2J}$, $f(\mathbf{x}) = \phi(\mathbf{x})^\top \boldsymbol{\theta}$, with prior $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \sigma_\theta^2 \mathbf{I}_{2J})$. Given a dataset of observations of $f(\mathbf{x})$ corrupted by Gaussian noise with variance σ_{obs}^2, such that $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ with $\mathbf{X} \in \mathbb{R}^{T \times d}$ and $\mathbf{y} \in \mathbb{R}^T$. The posterior of $f(\mathbf{x})$ is determined by the posterior of $\boldsymbol{\theta}$, which is also Gaussian. Denoting $\boldsymbol{\Phi} = [\phi(\mathbf{x}_1) \ldots \phi(\mathbf{x}_T)] \in \mathbb{R}^{2J \times T}$, the posterior of $\boldsymbol{\theta}$ is $p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}) = \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, with

$$\boldsymbol{\mu}_c = \frac{1}{\sigma_{\text{obs}}^2} \boldsymbol{\Sigma}_c \boldsymbol{\Phi} \mathbf{y}, \qquad \boldsymbol{\Sigma}_c = \left(\frac{1}{\sigma_{\text{obs}}^2} \boldsymbol{\Phi} \boldsymbol{\Phi}^\top + \frac{1}{\sigma_\theta^2} \mathbf{I} \right)^{-1}. \quad (1)$$

It will be convenient later to express the above in an equivalent information form. To this end, let $\mathbf{D}_c = \boldsymbol{\Sigma}_c^{-1}$ and $\boldsymbol{\eta}_c = \sigma_{\text{obs}}^{-2} \boldsymbol{\Phi} \mathbf{y}$, then Eq. (1) can be rewritten as

$$\boldsymbol{\mu}_c = \mathbf{D}_c^{-1} \boldsymbol{\eta}_c, \qquad \mathbf{D}_c = \frac{1}{\sigma_{\text{obs}}^2} \boldsymbol{\Phi} \boldsymbol{\Phi}^\top + \frac{1}{\sigma_\theta^2} \mathbf{I}. \quad (2)$$

This approximation provides the baseline centralized solution, with complexity $\mathcal{O}(TJ^2 + J^3)$, compared with the $\mathcal{O}(T^3)$ scaling of vanilla GPs [2], thereby reducing computational cost when $T \gg J$.

While the closed-form solution is exact, it assumes the data are centralized. To make this approach both online and distributed, we exploit its additive structure. The *information* quantities $\boldsymbol{\eta}_c$ and $\mathbf{D}_c$ admit an additive form in the case of conditionally independent Gaussian observations:

$$\mathbf{D}_c = \sum_{t=1}^T \mathbf{P}_t = \frac{1}{\sigma_\theta^2} \mathbf{I} + \sum_{t=1}^T \left(\frac{1}{\sigma_{\text{obs}}^2} \boldsymbol{\Phi}_t \boldsymbol{\Phi}_t^\top \right), \quad (3)$$

$$\boldsymbol{\eta}_c = \sum_{t=1}^T \mathbf{s}_t = \sum_{t=1}^T \frac{1}{\sigma_{\text{obs}}^2} \boldsymbol{\Phi}_t \mathbf{y}_t, \quad (4)$$
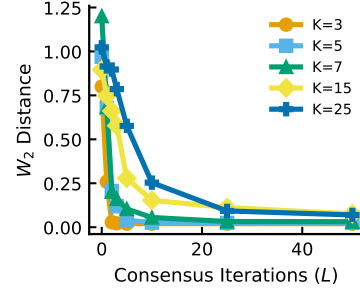

Fig. 2: An illustration of consensus building: $K \in \{3, 5, 7, 15, 25\}$ agents are connected in a "ring" topology and perform L consensus iterations. The figure shows the 2-Wasserstein distance to the centralized posterior, which decreases rapidly as L increases.

where $\boldsymbol{\Phi}_t \in \mathbb{R}^{2J \times N_t}$, $\mathbf{y}_t \in \mathbb{R}^{N_t}$ and N_t denotes the number of observations in the t-th batch. For $N_t = 1$, we have $\boldsymbol{\Phi}_t = \phi(\mathbf{x}_t) \in \mathbb{R}^{2J}$ and $\mathbf{y}_t = y_t \in \mathbb{R}$. This decomposition opens the path for online and distributed learning. In the online setting, new data arrive at each time instant t, updating the quantities as follows:

$$\mathbf{D}_{c,t} = \mathbf{D}_{c,t-1} + \mathbf{P}_t \quad (5)$$

$$\boldsymbol{\eta}_{c,t} = \boldsymbol{\eta}_{c,t-1} + \mathbf{s}_t, \quad (6)$$

where we define the incremental quantities $\mathbf{P}_t$ and $\mathbf{s}_t$ as

$$\mathbf{P}_t = \frac{1}{\sigma_{\text{obs}}^2} \boldsymbol{\Phi}_t \boldsymbol{\Phi}_t^\top, \quad \mathbf{P}_0 = \frac{1}{\sigma_\theta^2} \mathbf{I}. \quad (7)$$

$$\mathbf{s}_t = \frac{1}{\sigma_{\text{obs}}^2} \boldsymbol{\Phi}_t \mathbf{y}_t, \quad \mathbf{s}_0 = \mathbf{0}. \quad (8)$$

In summary, rewriting the posterior in information form reveals its recursive structure. Specifically, every new batch of observations contributes an additive update to $\boldsymbol{\eta}_t$ and $\mathbf{D}_t$, as per Eqs. (5)-(6), enabling online learning without storing all past data.

2.2. Decentralized multi-agent systems

The additive formulation also applies when data are spread across N agents at each time instant t. Introducing a subscript k to denote the k-th agent, the update equations become

$$\mathbf{D}_{c,t} = \mathbf{D}_{c,t-1} + \mathbf{P}_t = \mathbf{D}_{c,t-1} + \sum_{k=1}^K \mathbf{P}_{k,t}, \quad (9)$$

$$\boldsymbol{\eta}_{c,t} = \boldsymbol{\eta}_{c,t-1} + \mathbf{s}_t = \boldsymbol{\eta}_{c,t-1} + \sum_{k=1}^K \mathbf{s}_{k,t}. \quad (10)$$

This update scheme corresponds to a distributed information filter, in which each agent at time t computes the local quantities $\mathbf{P}_{k,t}$ and $\mathbf{s}_{k,t}$. The exact solution can be recovered by aggregating the agents' quantities to obtain $\mathbf{P}_t = \sum_{k=1}^K \mathbf{P}_{k,t}$ and $\mathbf{s}_t = \sum_{k=1}^K \mathbf{s}_{k,t}$. However, in a fully distributed scenario, no fusion center exists, and agents can only communicate with their neighbors.

If we want a *fully* decentralized implementation, the goal is thus to approximate the sums $\mathbf{P}_t = \sum_{k=1}^K \mathbf{P}_{k,t}$ and $\mathbf{s}_t = \sum_{k=1}^K \mathbf{s}_{k,t}$ that a fusion center would compute, while sharing information only with neighbors. Under mild conditions [31], this goal can be achieved via additive consensus algorithms, in which each agent approximates $\widetilde{\mathbf{P}}_{k,t}^{(L)} \approx \mathbf{P}_t$ and $\widetilde{\mathbf{s}}_{k,t}^{(L)} \approx \mathbf{s}_t$ for all k. Here, a tilde sign denotes the

consensus approximation at agent k after L communication rounds. Each agent then acts approximately as a fusion center that computes the full posterior up to time t. The convergence of GP predictions to the centralized posterior under additive consensus is shown in Fig. 2.

2.3. Ensembles

Thus far, we have described decentralized inference for a single RF-GP. In practice, performance depends on kernel hyperparameters such as length scales. To handle hyperparameter tuning in DRFGP, we use an ensemble of M RF-GPs at each agent (e.g. by sampling length scales from a prior). The agent network then aims to approximate the mixture of RF-GPs at a fusion center. In this case, even with a fusion center, we do not recover the exact Bayesian model averaging (BMA) weights unless the data batches are independent–a common assumption in distributed GPs [7]. Nevertheless, ensembles provide a simple way to further improve robustness and predictive performance of the final model.

3. NOVEL SCHEMES

Having introduced DRFGP, we now present two extensions that make it suitable for dynamic environments and robust to outliers. These extensions are easy to incorporate and improve stability and accuracy in challenging scenarios.

3.1. Time-varying DRFGP

The DRFGP introduced in Section 2 assumes a static regression function $f(\mathbf{x})$, and hence does not handle time-varying functions. In the centralized setting, one solution is to impose a state transition on the model parameters $\boldsymbol{\theta}$ (e.g., a random walk), a popular way to introduce temporal drift. This dynamical process turns the original Bayesian model into a state-space model (SSM),

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \mathbf{u}_t \tag{11}$$

$$y_t = \boldsymbol{\phi}(\mathbf{x}_t)^\top \boldsymbol{\theta}_t + \mathbf{n}_t, \quad t = 1, 2, \ldots, \tag{12}$$

where $\mathbf{u}_t$ and $\mathbf{n}_t$ are independent Gaussian random variables for all t. The back-to-prior (B2P) and uncertainty-injection (UI) can be derived from Eq. (11) [1, 17]. In practice, B2P and UI forgetting corresponds to downweighting the (per-agent) quantities $\mathbf{D}_{t-1}$ and $\boldsymbol{\eta}_{t-1}$ using a forgetting coefficient $\nu \in [0, 1]$ before updating with the new batch of observations.

Alternatively, we can keep the Bayesian model static, and enlarge our input vector $\mathbf{x}$ with a variable t so that we learn the dynamic function directly, $f(\mathbf{x}, t) = \phi(\mathbf{x}, t)^\top \boldsymbol{\theta}$. Let $\widetilde{\mathbf{x}} = [\mathbf{x}, t] \in \mathbb{R}^{d+1}$, where $\mathbf{x}$ denotes the *spatial*[1] variable and t denotes time. The mapping $\boldsymbol{\phi}(\widetilde{\mathbf{x}})$ is equivalent to using a stationary *spatio-temporal* kernel,

$$k_{\mathrm{st}}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{x}}') = k_{\mathrm{s}}(\mathbf{x}, \mathbf{x}') \times k_{\mathrm{t}}(t, t') \approx \boldsymbol{\phi}(\widetilde{\mathbf{x}})^\top \boldsymbol{\phi}(\widetilde{\mathbf{x}}'). \tag{13}$$

There is a natural connection between the forgetting mechanism and the use of temporal kernels. For instance, B2P forgetting corresponds to using an Ornstein-Uhlenbeck temporal kernel [17]. In practice, we find the spatio-temporal kernel approach simpler and more effective, allowing us to proceed exactly as in the static case while implicitly capturing temporal correlations.

---

[1]$\mathbf{x}$ is not required to be a spatial variable in the strict physical sense; $\mathbf{x}$ are the parameters of the function we aim to learn, but we assume the function itself changes with time.

3.2. Robust DRFGP

Another practical challenge is robustness. The RF-GP with Gaussian likelihood can be highly sensitive to anomalies. This is particularly important in decentralized systems because a single agent receiving biased observations can contaminate the entire network. To improve robustness, we adapt tools from robust filtering theory and M-estimation, in particular the unified form of [13], which is similar to the generalized Bayes updates of [12]. Similar robust filters have recently demonstrated success in the centralized sequential GP literature [11, 14].

In a *robust* distributed information filter, each local agent computes a diagonal weight matrix $\mathbf{W}_{k,t} = \mathrm{diag}(w_{k,t}^{(i)})$. The weight $w_{k,t}^{(i)}$ is a function of the standardized residual $e_{n,t}^{(i)} = (y_{k,t}^{(i)} - \widehat{y}_{k,t}^{(i)})/\sigma_{y,k,t}$, where $\widehat{y}_{k,t}$ and $\sigma_{y,k,t}$ denote the mean prediction and its standard deviation computed at agent k for $i = 1, \ldots, K_t$. The update equations are now

$$\mathbf{D}_t \leftarrow \mathbf{D}_{t-1} + \sum_k \frac{\boldsymbol{\Phi}_{k,t} \mathbf{W}_{k,t} \boldsymbol{\Phi}_{k,t}^\top}{\sigma_{\mathrm{obs}}^2} \tag{14}$$

$$\boldsymbol{\eta}_t \leftarrow \boldsymbol{\eta}_{t-1} + \sum_k \frac{\boldsymbol{\Phi}_{k,t}^\top \mathbf{W}_{k,t}^\top \mathbf{y}_{k,t}}{\sigma_{\mathrm{obs}}^2} \tag{15}$$

Weights $w_{k,t}^{(i)}$ are derived from robust weighting functions. For instance, the Huber and Hampel strategies are defined as

$$w_{\mathrm{Huber}}^{(i)}(e) = \begin{cases} 1, & |e| \leq \delta, \\ \delta/|e|, & |e| > \delta, \end{cases} \tag{16}$$

$$w_{\mathrm{Hampel}}^{(i)}(e) = \begin{cases} 1, & |e| \leq a, \\ a/|e|, & a < |e| \leq b, \\ c - |e|\,/\,(c-b), & b < |e| \leq c, \\ 0, & |e| > c, \end{cases} \tag{17}$$

These updates are equivalent to carrying out inference under a tempered likelihood that downweights atypical observations.

4. EXPERIMENTS

We now illustrate the proposed extensions in an Earth-system application. The goal is not to solve a specific scientific problem but to empirically validate ROAD-GP in a realistic scenario where (i) there is large-scale streaming data, (ii) there may be data contamination and non-stationarity, and (iii) distributed computation and communication efficiency are desirable. We focus on two main questions. First, do the robust modules prevent the network from being contaminated when an agent receives extreme observations? Second, does consensus allow the network to converge to a common approximation in decentralized settings?

4.1. Experimental setup

We use gridded weather data for a region of central Asia from the CRU dataset [32], processing monthly temperatures in streaming fashion. We aim to highlight ROAD-GP's ability to ingest large volumes of streaming data and summarize them with fixed-size statistics, while yielding an approximation of the centralized solution with low communication overhead. We include ensembles with spatial lengthscales $\ell_s \in \{0.01, 0.05, 0.1\}$, a temporal lengthscale $\ell_t = 4.0$, observation variance $\sigma_{\mathrm{obs}}^2 = 0.05$, and process variance $\sigma_\theta^2 \in$
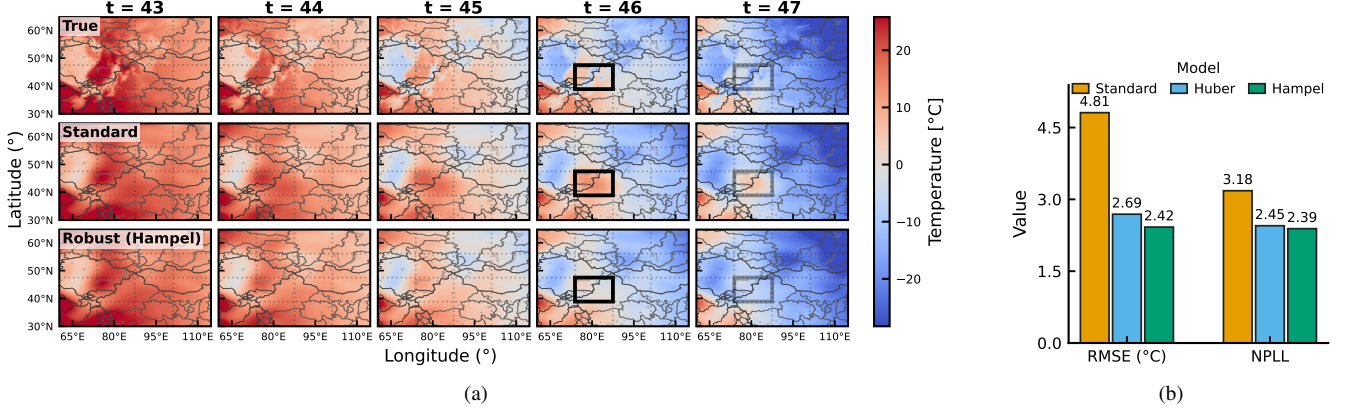
Fig. 3: **(a)** Visualizations of spatiotemporal predictions in the "robustness to outliers" experiment. Grid lines denote agent boundaries, and squares indicate the region where outliers are injected at $t = 46$. Note the effect on predictions at $t = 47$. **(b)** RMSE and negative predictive log-likelihood (NPLL) computed over the final time instant ($t = 47$).
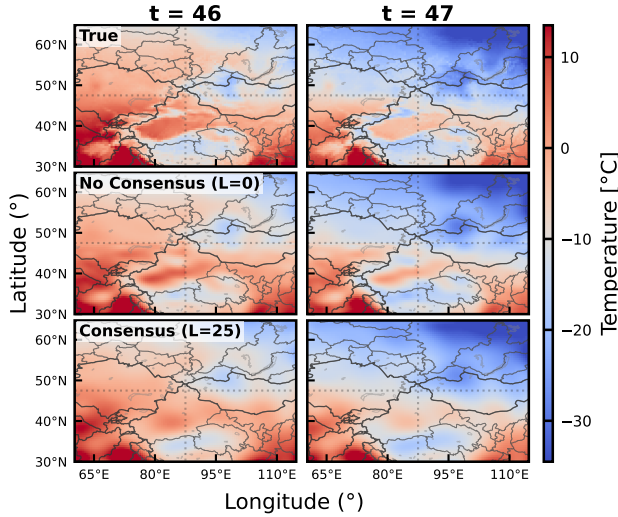

Fig. 4: Predictions of the spatiotemporal field with and without the use of additive consensus. In "No Consensus," each agent's local GP is stitched together; Fig. 5 visualizes each agent's global predictions.
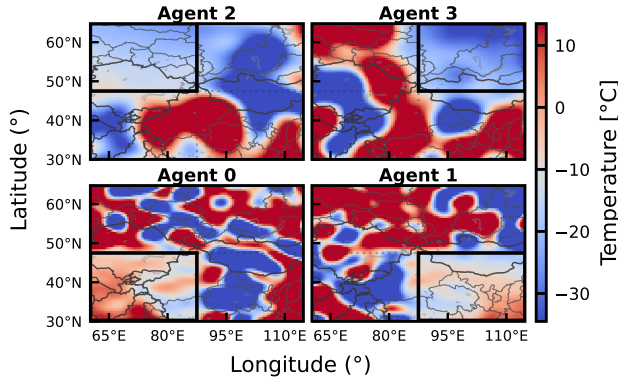

Fig. 5: Global estimates of each local agent in Fig. 4. Black rectangles denote the area in which each agent receives data.

$\{1.0, 25.0\}$. We normalize the spatial and output dimensions, but not the temporal dimension.

4.2. Robustness to outliers

We first assess the robustness of ROAD-GP. We consider $K = 4$ agents, with $J = 1000$ random frequencies, each receiving $N_{\text{batch}} = 459$ spatial observations over $T = 48$ successive months, for a total of $\approx 325\,000$ data points. Inspired by [14], we inject extreme outliers at time $t = 46$, biasing 30% of the observations to random values 8 standard deviations warmer than recorded – similar to extremely anomalous events or sensor failure. We illustrate predictions for $t = 43, \cdots, 47$ for the standard DRFGP with a spatiotemporal kernel and ROAD-GP with Hampel weights in Fig. 3a. ROAD-GP safely ignores the influence of these outliers, which is evident both visually and in the reported metrics of Fig. 3b.

4.3. Consensus and approximation to the centralized solution

We next evaluate the effect of consensus, using four agents on the same dataset for better visualization. With $L = 0$ communication rounds, the model reduces to a collection of independent (local) RF-GPs, one per agent. With sufficiently large $L > 0$, each agent's posterior then converges to the global, centralized solution. Fig. 5 shows that local solutions ($L = 0$) can produce crisper predictions, but at the cost of potentially poor global estimates.

5. CONCLUSIONS

We have extended the decentralized random Fourier feature Gaussian process (DRFGP) framework with two lightweight yet powerful modules: robust updates that mitigate the effect of outliers, and dynamic mechanisms that adapt to time-varying functions. These modifications preserve the recursive and decentralized structure of DR-FGP, allowing agents to maintain constant-size statistics and collectively approximate the centralized solution. Using curated weather datasets, we have validated our approach and shown that the new modules can be vital for obtaining stable solutions under corrupted data streams. These results highlight the potential for robust, dynamic in-situ modeling of large-scale streaming data.

6. REFERENCES

[1] F. Llorente and P. M. Djurić, "Dynamic Random Feature Gaussian Processes for Bayesian Optimization of Time-Varying Functions," in *IEEE ICASSP*, 2024.

[2] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006, vol. 1.

[3] P. M. Djurić and C. Richard, *Cooperative and Graph Signal Processing: Principles and Applications*. Academic Press, 2018.

[4] T. D. Bui, C. Nguyen, and R. E. Turner, "Streaming sparse Gaussian process approximations," in *NIPS*, 2017.

[5] W. J. Maddox, S. Stanton, and A. G. Wilson, "Conditioning sparse variational Gaussian processes for online decision-making," *NeurIPS*, 2021.

[6] M. Deisenroth and J. W. Ng, "Distributed Gaussian processes," in *ICML*, 2015.

[7] H. Liu, J. Cai, Y. Wang, and Y. S. Ong, "Generalized robust Bayesian committee machine for large-scale Gaussian process regression," in *ICML*, 2018.

[8] G. P. Kontoudis and D. J. Stilwell, "Decentralized nested Gaussian processes for multi-robot systems," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[9] ——, "Scalable, federated Gaussian process training for decentralized multi-agent systems," *IEEE Access*, 2024.

[10] T. N. Hoang, Q. M. Hoang, K. H. Low, and J. How, "Collective online learning of Gaussian processes in massive multi-agent systems," in *AAAI*, 2019.

[11] D. Waxman and P. M. Djurić, "A gaussian Process-based Streaming Algorithm for Prediction of Time Series With Regimes and Outliers," in *International Conference on Information Fusion (FUSION)*, 2024.

[12] G. Duran-Martin, M. Altamirano, A. Y. Shestopaloff, L. Sánchez-Betancourt, J. Knoblauch, M. Jones, F.-X. Briol, and K. Murphy, "Outlier-robust kalman filtering through generalised bayes," *arXiv preprint arXiv:2405.05646*, 2024.

[13] L. Chang and K. Li, "Unified form for the robust Gaussian information filtering based on M-estimate," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 412–416, 2017.

[14] W. Laplante, M. Altamirano, A. Duncan, J. Knoblauch, and F.-X. Briol, "Robust and conjugate spatio-temporal Gaussian processes," *arXiv preprint arXiv:2502.02450*, 2025.

[15] F. M. Nyikosa, M. A. Osborne, and S. J. Roberts, "Bayesian optimization for dynamic problems," *arXiv preprint arXiv:1803.03432*, 2018.

[16] S. Van Vaerenbergh, M. Lázaro-Gredilla, and I. Santamaría, "Kernel recursive least-squares tracker for time-varying regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1313–1326, 2012.

[17] S. Van Vaerenbergh, I. Santamaría, and M. Lázaro-Gredilla, "Estimation of the forgetting factor in kernel recursive least squares," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2012.

[18] Y. Shen, T. Chen, and G. B. Giannakis, "Random feature-based online multi-kernel learning in environments with unknown dynamics," *Journal of Machine Learning Research*, vol. 20, no. 22, pp. 1–36, 2019.

[19] S. Hong and J. Chae, "Distributed online learning with multiple kernels," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 3, pp. 1263–1277, 2021.

[20] ——, "Communication-efficient randomized algorithm for multi-kernel online federated learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9872–9886, 2021.

[21] A. Lederer, Z. Yang, J. Jiao, and S. Hirche, "Cooperative control of uncertain multiagent systems via distributed Gaussian processes," *IEEE Transactions on Automatic Control*, vol. 68, no. 5, pp. 3091–3098, 2022.

[22] J. Di, E. Zobeidi, A. Koppel, and N. Atanasov, "Distributed Gaussian process mapping for robot teams with time-varying communication," in *American Control Conference*, 2022.

[23] K. Rumsey, M. Grosskopf, E. Lawrence, A. Biswas, and N. Urban, "A hierarchical sparse Gaussian process for in situ inference in expensive physics simulations," in *Applications of machine learning*, vol. 12227, 2022, pp. 126–138.

[24] M. Grosskopf, E. Lawrence, A. Biswas, L. Tang, K. Rumsey, L. V. Roekel, and N. Urban, "In-situ spatial inference on climate simulations with sparse gaussian processes," in *Proceedings of the In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization (ISAV)*, 2021.

[25] J. Chen, K. H. Low, Y. Yao, and P. Jaillet, "Gaussian process decentralized data fusion and active sensing for spatiotemporal traffic modeling and prediction in mobility-on-demand systems," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 3, pp. 901–921, 2015.

[26] F. Llorente, D. Waxman, and P. M. Djurić, "Decentralized Online Ensembles of Gaussian Processes for Multi-Agent Systems," in *IEEE ICASSP*, 2025.

[27] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *NIPS*, 2007.

[28] M. Lázaro-Gredilla, J. Quinonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal, "Sparse spectrum Gaussian process regression," *Journal of Machine Learning Research*, vol. 11, pp. 1865–1881, 2010.

[29] D. Waxman and P. M. Djurić, "Dynamic online ensembles of basis expansions," *TMLR*, 2024.

[30] Y. Wang and P. M. Djurić, "Distributed Bayesian estimation of linear models with unknown observation covariances," *IEEE Transactions on Signal Processing*, vol. 64, no. 8, pp. 1962–1971, 2015.

[31] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[32] I. Harris, T. J. Osborn, P. Jones, and D. Lister, "Version 4 of the cru ts monthly high-resolution gridded multivariate climate dataset," *Scientific Data*, vol. 7, no. 1, p. 109, 2020.