



Bayesian spatiotemporal modeling for inverse problems

Shiwei Lan¹ · Shuyi Li¹ · Mirjeta Pasha²

Received: 22 April 2022 / Accepted: 2 May 2023 / Published online: 10 June 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Inverse problems with spatiotemporal observations are ubiquitous in scientific studies and engineering applications. In these spatiotemporal inverse problems, observed multivariate time series are used to infer parameters of physical or biological interests. Traditional solutions for these problems often ignore the spatial or temporal correlations in the data (static model), or simply model the data summarized over time (time-averaged model). In either case, the data information that contains the spatiotemporal interactions is not fully utilized for parameter learning, which leads to insufficient modeling in these problems. In this paper, we apply Bayesian models based on spatiotemporal Gaussian processes (STGP) to inverse problems with spatiotemporal data and show that the spatial and temporal information provides more effective parameter estimation and uncertainty quantification (UQ). We demonstrate the merit of Bayesian spatiotemporal modeling for inverse problems compared with traditional static and time-averaged approaches using a time-dependent advection–diffusion partial differential equation (PDE) and three chaotic ordinary differential equations (ODE). We also provide theoretic justification for the superiority of spatiotemporal modeling to fit the trajectories even if it appears cumbersome (e.g. for chaotic dynamics).

Keywords Spatiotemporal inverse problems · Spatiotemporal Gaussian process · Chaotic dynamics · Trajectory fitting · Uncertainty quantification

1 Introduction

Many inverse problems in science and engineering involve large scale spatiotemporal data, typically recorded as multivariate time series. There are examples in fluid dynamics that describe the flow of liquid (e.g. petroleum) or gas (e.g. flame jet) (Baukal Jr 2000). Other examples include dynamical systems with chaotic behavior prevalent in weather prediction (Lorenz 1963), biology (Liz and Ruiz-Herrera 2012), economics (Brooks 1998) etc. where small perturbation of the initial condition could lead to large deviation from what is observed/calculated in time. The goal of such inverse problems is to recover the parameters from given observations

and knowledge of the underlying physics. The spatiotemporal information is crucial and should be respected when considering proper statistical models for parameter learning. This is not only of interest in statistics, but also beneficial for practical applications of physics and biology to obtain inverse solutions and UQ more effectively.

Traditional methods for these spatiotemporal inverse problems often ignore the time dependence in the data for a simplified solution (Villa et al. 2021; Cleary et al. 2021; Lan et al. 2022). They either treat the observed time series statically as independent identically distributed (i.i.d.) observations across time (Villa et al. 2021; Lan et al. 2022) (hence named as “static” model), or summarize them by taking time average or higher order moments (Morzfeld et al. 2018; Cleary et al. 2021; Huang et al. 2022) (referred as “time-averaged” approach). The former is prevalent in Bayesian inverse problems with time series observations (Lan et al. 2022). The latter is especially common in parameter learning of chaotic dynamics, e.g. Lorenz systems (Lorenz 1963; Cleary et al. 2021), due to their sensitivity to the initial conditions and the system parameters (and the nature of the time-averaging procedure), which in turn cause a rough landscape of the objective function. In both scenarios, the

✉ Shiwei Lan
slan@asu.edu

Shuyi Li
shuyili3@asu.edu

Mirjeta Pasha
mirjeta.pasha@tufts.edu

¹ School of Mathematical and Statistical Sciences, Arizona State University, 901 S. Palm Walk, Tempe, AZ 85287, USA

² Department of Mathematics, Tufts University, 177 College Ave, Medford, MA 02155, USA

spatiotemporal information is not fully integrated into the statistical modeling.

In this paper, we propose to apply Bayesian methods based on Gaussian Process (GP) (Sacks et al. 1989; Oakley and O'Hagan 2002, 2004; Kennedy and O'Hagan 2001; Higdon et al. 2004; O'Hagan 2006) to the inverse problems with spatiotemporal data to account for the space-time interdependence. This leads to fitting the whole trajectories of the observed data, rather than their statistical summaries, with elaborated models. More specifically, we use the STGP model (Cressie and Wikle 2011; Gneiting 2002; Paciorek et al. 2003; Wang et al. 2020; Lan 2022) to fit the observed multivariate time series in comparison with the static or the time-averaged (for summarized data) models. Theoretically, we justify why the STGP model should be preferred to by investigating their Fisher information, which can be used as a measurement of convexity: STGP renders a more convex likelihood than the other two models and leads to an easier learning of the parameters. We also demonstrate in numerical experiments (Sect. 4) that the STGP model yields parameter estimates closer to the truth with smaller observation window required, and also provides more reasonable UQ results. Note this implies faster convergence (future work) by the STGP model, which is computationally important because complex ODE/PDE systems are usually expensive to solve.

Spatiotemporal reasoning/modeling was introduced to inverse problems. However, it was either applied to specific domains such as functional magnetic resonance imaging (fMRI) (Woolrich et al. 2004), electroencephalography (EEG) (Siregar and Sinteff 1996) and electrocardiography (ECG) (Shcherbakova et al. 2021), or to a simplified Gauss-linear problem (Long 2011; Ojeda et al. 2019; Conjard and Omre 2021; Yang 2017). Spatiotemporal information was also used to construct prior (Zhang et al. 2005) and regularization (Yao and Yang 2016; Pasha et al. 2021), or to reduce the number of parameters (Echeverría Ciaurri and Mukerji 2009). However, none of them formulates the spatiotemporal modeling in the general framework of Bayesian inverse problems with spatiotemporal observations. We summarize the main contributions of this work as follows:

- It formulates a Bayesian modeling framework for inverse problems with spatiotemporal data that includes traditional static and time-averaged methods;
- It provides a theoretical justification on why the STGP model is preferable in the spatiotemporal inverse problems;
- It numerically demonstrates the advantage of the STGP model in parameter learning and UQ.

The rest of the paper is organized as follows: Sect. 2 reviews the background of Bayesian UQ for inverse problems, with a particular framework named *Calibrate-Emulate-*

Sample (CES) (Cleary et al. 2021; Lan et al. 2022). In Sect. 3 we generalize the problem setup to include spatiotemporal observations and compare the STGP model (Sect. 3.3) with the static model (Sect. 3.1) and the time-averaged model (Sect. 3.2). We prove in Theorems 3.1 and 3.2 that the STGP model can have more convex likelihood than the static and the time-averaged models. Then in Sect. 4 we demonstrate the advantage of the STGP model over the other two traditional approaches with inverse problems involving an advection–diffusion equation and three chaotic dynamics. Finally we conclude with some discussions on future directions in Sect. 5.

2 Background: Bayesian UQ for inverse problems

In many inverse problems, we are interested in finding an unknown parameter, u (which could be a function or a vector), given the observed data, y . The parameter u usually appears as a quantity of interest in the inverse problem, e.g. the initial condition of a time-dependent advection–diffusion problem (Sect. 4.1) or the coefficient vector in the chaotic dynamics (Sect. 4.2). Let $\mathbb{X}$ and $\mathbb{Y}$ be two separable Hilbert spaces. A forward mapping $\mathcal{G} : \mathbb{X} \rightarrow \mathbb{Y}$ from the parameter space $\mathbb{X}$ to the data space $\mathbb{Y}$ (e.g. $\mathbb{Y} = \mathbb{R}^m$ for $m \geq 1$) connects $u \in \mathbb{X}$ to $y \in \mathbb{Y}$ as follows:

$$y = \mathcal{G}(u) + \eta, \quad \eta \sim \mathcal{N}(0, \Gamma) \quad (1)$$

We can define the potential function (negative log-likelihood), $\Phi : \mathbb{X} \times \mathbb{Y} \rightarrow \mathbb{R}$, often with $\Gamma = \sigma^2 I$:

$$\Phi(u; y) = \frac{1}{2} \|y - \mathcal{G}(u)\|_{\Gamma}^2 = \frac{1}{2} \langle y - \mathcal{G}(u), \Gamma^{-1}(y - \mathcal{G}(u)) \rangle \quad (2)$$

The forward mapping $\mathcal{G}$ represents physical laws usually expressed as large and complex ODE/PDE systems that could be highly non-linear. Therefore repeated evaluations of $\mathcal{G}(u)$ (which require solving ODE/PDE) and hence $\Phi(u; y)$ are expensive for different u 's.

In the Bayesian setting, a prior measure μ_0 is imposed on u , independent of η . For example, we could assume a Gaussian prior $\mu_0 = \mathcal{N}(0, \mathcal{C})$ with the covariance $\mathcal{C}$ being a positive, self-adjoint and trace-class operator on $\mathbb{X}$. Then we can obtain the posterior of u , denoted as $\mu(u|y)$, using Bayes' theorem (Stuart 2010; Dashti and Stuart 2017) if $0 < Z := \int_{\mathbb{X}} \exp(-\Phi(u; y)) \mu_0(du) < +\infty$:

$$\frac{d\mu}{d\mu_0}(u) = \frac{1}{Z} \exp(-\Phi(u; y)) \quad (3)$$

Bayesian UQ for inverse problems involves learning the posterior distribution $\mu(du)$ which often exhibits strongly

non-Gaussian behavior, posing significant challenges for efficient inference methods such as Markov Chain Monte Carlo (MCMC).

There are three urging computational challenges in the Bayesian UQ for inverse problems: (1) intensive computation for likelihood evaluations, which require expensive solving of forward problems; (2) complex (non-Gaussian) posterior distributions; and (3) high dimensionality of the discretized parameter (still denoted as u when there is no confusion from the context). The last one makes the first two more difficult in the sense that high dimensionality not only demands efficient forward solvers, but also challenges the robustness of sampling algorithms. To address these challenges, an approximate inference framework named Calibrate-Emulate-Sample (CES) has recently been proposed by Cleary et al. (2021) whose GP emulator is replaced with neural network (NN) by Lan et al. (2022) for high dimensional approximate UQ. It consists of the following three stages:

1. *Calibrate* using optimization-based (ensemble Kalman) algorithms to obtain parameter estimation and collect expensive forward evaluations for the emulation step;
2. *Emulate* recycling forward evaluations from calibration to build an emulator for sampling;
3. *Sample* sampling the posterior approximately based on the emulator, which is much cheaper than the original forward mapping.

CES calibrates the model with ensemble Kalman (EnK) methods (Evensen 1994; Evensen and van Leeuwen 1996). Two algorithms, ensemble Kalman inversion (EKI) (Schillings and Stuart 2017a; Garbuno-Inigo et al. 2020a) and ensemble Kalman sampler (EKS) (Garbuno-Inigo et al. 2020a, b), evolve K ensemble particles $\{u^{(k)}\}_{k=1}^K$ according to the following equations respectively:

$$\begin{aligned} \text{EKI: } \frac{du^{(k)}}{dt} &= \frac{1}{K} \sum_{k'=1}^K \left(\mathcal{G}(u^{(k')}) - \bar{\mathcal{G}}, y - \mathcal{G}(u^{(k)}) \right. \\ &\quad \left. + \sqrt{\Sigma} \frac{dW^{(k)}}{dt} \right)_{\Gamma} (u^{(k')} - \bar{u}) \\ \text{EKS: } \frac{du^{(k)}}{dt} &= \frac{1}{K} \sum_{k'=1}^K \left(\mathcal{G}(u^{(k')}) - \bar{\mathcal{G}}, y - \mathcal{G}(u^{(k)}) \right)_{\Gamma} \\ &\quad (u^{(k')} - \bar{u}) - C(u)C^{-1}u^{(k)} + \sqrt{2C(u)} \frac{dW^{(k)}}{dt} \end{aligned}$$

where $\bar{u} := \frac{1}{K} \sum_{k=1}^K u^{(k)}$, $\bar{\mathcal{G}} := \frac{1}{K} \sum_{k=1}^K \mathcal{G}(u^{(k)})$, $\Sigma = 0$ or Γ , $\{W^{(k)}\}$ are independent cylindrical Brownian motions on $\mathbb{Y}$, and $C(u) := \frac{1}{K} \sum_{k=1}^K (u^{(k)} - \bar{u}) \otimes (u^{(k)} - \bar{u})$. Implemented in parallel, EnK algorithms converge quickly to the

optimal parameter with a few (usually hundreds of) ensembles without explicit calculation of gradients. However, due to the collapse of ensembles (Schillings and Stuart 2017a, b; de Wiljes et al. 2018; Chada et al. 2020), the sample variance given by $\{u^{(k)}\}_{k=1}^K$ tends to underestimate the actual uncertainty (see Figure 1 in Lan et al. 2022,).

CES recovers the proper uncertainty by running sampling algorithms based on an emulator $\mathcal{G}^e : \mathbb{X} \rightarrow \mathbb{Y}$ trained on data $\{u_n^{(k)}, \mathcal{G}(u_n^{(k)})\}_{k=1, n=0}^{K, N}$ that have been collected in the calibration stage. The emulator can be GP (Cleary et al. 2021; Sacks et al. 1989; Kennedy and O'Hagan 2001; Higdon et al. 2004; O'Hagan 2006) or NN, e.g. convolutional NN (Lan et al. 2022; Krizhevsky et al. 2012; Goodfellow et al. 2016; Zammit-Mangion et al. 2020).

Once the emulator is built, CES approximately samples from the posterior with dimension-independent MCMC algorithms based on the emulated likelihood Φ^e and its gradient $D\Phi^e$ (defined by substituting $\mathcal{G}$ with $\mathcal{G}^e$ in (2)) at much lower computational cost. Note, a class of dimension-independent algorithms—including *preconditioned Crank-Nicolson (pCN)* (Cotter et al. 2013), *infinite-dimensional MALA* (∞ -MALA) (Beskos et al. 2008), *infinite-dimensional HMC* (∞ -HMC) (Beskos et al. 2011), and *infinite-dimensional manifold MALA* (∞ -mMALA) (Beskos 2014) and *HMC* (∞ -mHMC) (Beskos et al. 2017)—are used in place of traditional Metropolis–Hastings algorithms to avoid the deteriorating mixing time as the dimension of parameter space increases.

3 Spatiotemporal inverse problems (STIP)

When the observations are taken from a spatiotemporal process, $y(\mathbf{x}, t)$, simple Gaussian likelihood function as (2) with $\Gamma = \sigma^2 I$, for example, may not be sufficient to describe the space-time interactions. To address this issue, we propose to rewrite the data model (1) in terms of a GP with spatiotemporal kernel $\Gamma(\mathbf{x}, t)$:

$$\begin{aligned} y(\mathbf{x}, t) &= \mathcal{G}(u)(\mathbf{x}, t) + \eta(\mathbf{x}, t), \\ \eta(\mathbf{x}, t) &\sim \mathcal{GP}(0, \Gamma(\mathbf{x}, t)) \end{aligned} \quad (5)$$

In practice, the forward model often involves time-dependent PDE, e.g. heat equation and Navier–Stokes equations. Therefore, it is crucial to allow for the spatiotemporal correlations in the statistical analysis of such inverse problems. Compared to (1), the model (5) offers a more appropriate definition of the likelihood by incorporating the spatiotemporal structures in the data.

Note, the proposed general framework (5) also includes many existing statistical models as special cases. For example, if we define the forward map based on some covariates, $\mathbf{X}(\mathbf{x}, t)$, $\mathcal{G}(\beta)(\mathbf{x}, t) = \mathbf{X}(\mathbf{x}, t)\beta(\mathbf{x}, t)$, then (5) is simply a

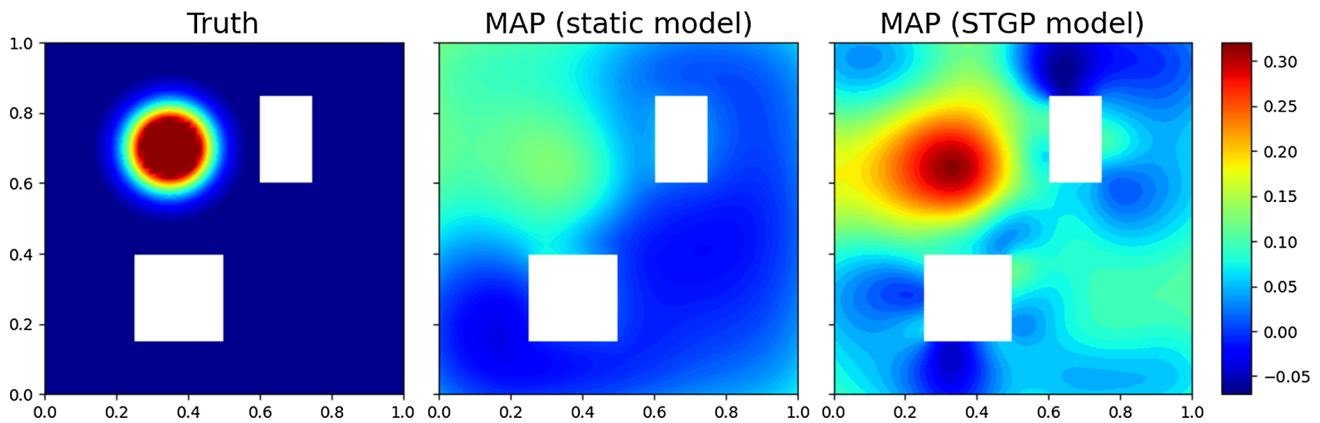


Fig. 1 Advection–diffusion inverse problem: comparing maximum *a posteriori* (MAP) estimates of parameter $u_0 = u(\mathbf{x}, 0)$ by the static model (middle) and the STGP model (right) with the truth $u_0^\dagger$ (left)

regression model. If we set $\mathcal{G}(u)(\mathbf{x}, t) = \mathbf{L}(\mathbf{x}, t)u(\mathbf{x}, t)$ with loading matrix $\mathbf{L}(\mathbf{x}, t)$, then (5) becomes a latent factor model (Agarwal and Chen 2009; Webb 2011; Jenatton et al. 2012).

In the following, we will review the static (Sect. 3.1) and the time-averaged (Sect. 3.2) models and unify them in the framework of STGP model (Sect. 3.3). For the convenience of exposition, we fix some notations in the following. Denote $\mathbf{X} := \{\mathbf{x}_i\}_{i=1}^I$, $\mathbf{t} := \{t_j\}_{j=0}^{J-1}$, and $\mathbf{Y} := y(\mathbf{X}, \mathbf{t}) = \{y(\mathbf{x}_i, t_j)\}_{i=1, j=0}^{I, J-1}$, data observed at I locations and J time points. $\mathbf{C}_*$ is the covariance matrix of the covariance kernel C_* restricted on the finite-dimensional discrete space for $*$ being either $\mathbf{x}$ or t .

3.1 Static model

In the literature of Bayesian inverse problems (Dashti and Stuart 2017; Cleary et al. 2021), the noise η is often assumed i.i.d. over time in (5), i.e. $\eta(\mathbf{x}, t_j) \stackrel{iid}{\sim} \mathcal{N}(0, C_{\mathbf{x}})$. This leads to the following static model where the temporal correlation is ignored:

$$\begin{aligned} y(\mathbf{x}, t)|u, \Gamma &\sim \mathcal{GP}(\mathcal{G}(u)(\mathbf{x}, t), \Gamma(\mathbf{x}, t)) \\ \text{static : } \Gamma(\mathbf{x}, t) &= C_{\mathbf{x}} \otimes \mathcal{I}_t \end{aligned} \quad (6)$$

where $\mathcal{I}_t$ is the Dirac operator such that $\mathcal{I}_t(t, t') = 1$ only if $t = t'$. When the spatial dependence is also suppressed (as in the advection–diffusion example of Sect. 4.1 and in Villa et al. (2021); Lan et al. (2022)), we have $C_{\mathbf{x}} = \sigma_{\epsilon}^2 \mathcal{I}_{\mathbf{x}}$.

The temporal correlation is disregarded in the static model (6). When there is (spatio-)temporal effect in the residual η , the static model (6) may be insufficient to account for the spatiotemporal relationships contained in the data. For illustration, we consider an inverse problem involving advection–diffusion (Sect. 4.1) equation (Villa et al. 2021; Lan et al. 2022) of an evolving concentration field $u(\mathbf{x}, t)$, e.g., temperature for heat transfer, and seek the solution to

the initial condition, $u_0 = u(\mathbf{x}, 0)$, based on spatiotemporal solutions observed (through an observation operator $\mathcal{O}$) on the boundaries of two boxes (Fig. 1, left panel) for a given time period, i.e. $y = \mathcal{O}u(\mathbf{x}, t) + \eta$, $\eta \sim \mathcal{N}(0, \sigma_{\eta}^2)$. As shown in Fig. 1, the simple static model (6) used in Lan et al. (2022) does not account for space-time interactions hence yields the result underestimating the true function $u_0^\dagger$ (left panel). On the contrary, the estimate by the spatiotemporal model (15) (right panel) is much closer to the truth.

3.2 Time-averaged model

In many chaotic dynamics, we observe the trajectories as multivariate time series that are very sensitive to the initial condition and the parameters. This usually results in a complex objective function with multiple local minima (Abarbanel 2013). They in turn form a rough landscape of the objective and pose extreme difficulties on parameter learning (Cleary et al. 2021) (See also Fig. 6). The time-averaged approach is commonly used in the spirit of extracting sufficient statistics from the raw data (Fisher and Russell 1922).

We consider the same data model as in (5) with $\mathcal{G}(u)$ being the observed solution $\mathbf{x}(t; u, \mathbf{x}_0)$ of the following chaotic dynamics (r th order ODE) for a given parameter $u \in \mathbb{R}^P$:

$$\dot{\mathbf{x}} := \frac{d\mathbf{x}}{dt} = f(t, \mathbf{x}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(r)}; u), \quad \mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^I \quad (7)$$

That is, $\mathcal{G}(u) = \mathcal{O}\mathbf{x}(t; u, \mathbf{x}_0)$ with an observation operator $\mathcal{O}$. At each time t , the observed vector could include components of $\mathbf{x}$ and up to their k th order interactions for $k \geq 1$. For example, if $\mathbf{x} = [x_1, \dots, x_I]$, we could include all the first and second order moments in the observation vector, $\mathcal{O}\mathbf{x} = [x_1, \dots, x_I, x_1^2, x_1x_2, \dots, x_ix_j, \dots, x_I^2]$. Because the trajectories of $\mathcal{G}(u)$ are usually complex, it is often to average them over time and consider the following

forward mapping instead:

$$\mathcal{G}_T(u; \mathbf{x}_0) := \frac{1}{T} \int_{t_0}^{t_0+T} \mathcal{O}\mathbf{x}(t; u, \mathbf{x}_0) dt \quad (8)$$

where t_0 is the spin-up time and T is the window length for averaging the observed trajectories of the dynamics.

Following Cleary et al. (2021), we make the same assumption regarding the dynamical system (7):

Assumption 1 1. For $u \in \mathbb{X}$, (7) has a compact attractor $\mathcal{A}$, supporting an invariant measure $\mu(d\mathbf{x}; u)$. The system is ergodic, and the following limit of Law of Large Numbers (LLN) is satisfied: for $\mathbf{x}_0 \sim \mu(\cdot; u)$ fixed, with probability one,

$$\lim_{T \rightarrow \infty} \mathcal{G}_T(u; \mathbf{x}_0) = \mathcal{G}_\infty(u) := \int_{\mathcal{A}} \mathcal{O}\mathbf{x}(t; u, \mathbf{x}_0) \mu(d\mathbf{x}; u) \quad (9)$$

2. The Central Limit Theorem (CLT) holds quantifying the ergodicity: for $\mathbf{x}_0 \sim \mu(\cdot; u)$,

$$\mathcal{G}_T(u; \mathbf{x}_0) \sim \mathcal{N}(\mathcal{G}_\infty(u), T^{-1} \Sigma(u)) \quad (10)$$

The limit $\mathcal{G}_\infty(u)$ becomes independent of the initial condition $\mathbf{x}_0$. However, the finite-time truncation, $\mathcal{G}_T(u; \mathbf{x}_0)$, with different random initializations $\mathbf{x}_0$, generates random errors from the limit $\mathcal{G}_\infty(u)$, which are assumed approximately Gaussian. Assume the data y can be observed with a true parameter $u^\dagger$, i.e. $y = \mathcal{G}_T(u^\dagger; \mathbf{x}_0)$. The following time-averaged model is usually adopted for the inverse problems involving chaotic dynamics (Cleary et al. 2021):

$$y|u, \Sigma(u) \sim \mathcal{N}(\mathcal{G}_\infty(u), T^{-1} \Sigma(u)) \quad (11)$$

time-average : $T^{-1} \Sigma(u) \approx \Gamma_{\text{obs}}$

where the empirical covariance Γ_{obs} can be estimated with $\mathcal{G}_\tau(u; \mathbf{x}_0)$ for $\tau \gg T$.

In practice, we follow Cleary et al. (2021) to observe from $\mathcal{G}_\tau(u; \mathbf{x}_0)$ long enough (τ) to estimate Γ_{obs} . This is done once to reduce the influence of the initial condition $\mathbf{x}_0$. The inference is conducted on $[t_0, t_0 + T]$ where we need to choose t_0 and T appropriately to trade off between the stability and the computational feasibility (See more studies in Sect. 4.2 for Figs. 7, 11, 15). We also replace $\mathcal{G}_\infty(u)$ with $\mathcal{G}_T(u; \mathbf{x}_0)$ in (11) and define the potential $\Phi_\tau(u)$ of parameter u for the time-averaged model (11) as follows:

$$\Phi_\tau(u) = \frac{1}{2} \|y - \mathcal{G}_T(u; \mathbf{x}_0)\|_{\Gamma_{\text{obs}}}^2 \quad (12)$$

If we observe the trajectories (all the first order moments, i.e. $\mathcal{O}\mathbf{x} = \mathbf{x}$) at discrete time points $\mathbf{t}$ with $t_{J-1} = t_0 +$

T , then $\mathcal{O}\mathbf{x}(t; u)$ yields multivariate time series, denoted as $\mathbf{X}(u)_{I \times J} = \mathbf{x}(\mathbf{t}; u) = [\mathbf{x}(t_0; u), \dots, \mathbf{x}(t_{J-1}; u)]$. Then we have

$$\begin{aligned} \mathcal{G}_T(u; \mathbf{x}_0) &= \bar{\mathbf{X}}(u) := \mathbf{X}(u) \frac{\mathbf{1}_J}{J}, \quad y = \mathbf{X}(u^\dagger) \frac{\mathbf{1}_J}{J}, \\ \Gamma_{\text{obs}} &= \mathbf{X}(u^\dagger) \left[\mathbf{I}_J - \frac{\mathbf{1}_J \mathbf{1}_J^\top}{J} \right] \mathbf{X}(u^\dagger)^\top \end{aligned} \quad (13)$$

Denote $\mathbf{X}_0 = \mathbf{X}(u) - \mathbf{X}(u^\dagger)$. Therefore the potential Φ_τ becomes

$$\Phi_\tau(u) = \frac{1}{2} \frac{\mathbf{1}_J^\top}{J} \mathbf{X}_0^\top \Gamma_{\text{obs}}^{-1} \mathbf{X}_0 \frac{\mathbf{1}_J}{J} = \frac{1}{2} \text{tr} \left[\frac{\mathbf{1}_J \mathbf{1}_J^\top}{J^2} \mathbf{X}_0^\top \Gamma_{\text{obs}}^{-1} \mathbf{X}_0 \right] \quad (14)$$

Note, averaging the trajectories over time does not ease the difficulty of rough landscape, see for instance Fig. 6 for an illustration. However, the potential function for the following STGP model (15) is more convex around the true values $u^\dagger$ compared with the time-averaged approach (11).

The aforementioned two approaches, the static model (6) and the time-averaged model (11), can be recognized as special cases of a more general framework of spatiotemporal modeling based on STGP, to be discussed in the following section.

3.3 Spatiotemporal GP model

For the spatiotemporal data $y(\mathbf{x}, t)$ in the inverse problems, we consider the following likelihood model based on STGP:

$$\begin{aligned} y(\mathbf{x}, t)|u, \Gamma &\sim \mathcal{GP}(\mathcal{G}(u)(\mathbf{x}, t), \Gamma(\mathbf{x}, t)) \\ \text{STGP : } \Gamma(\mathbf{x}, t) &= \mathcal{C}_{\mathbf{x}} \otimes \mathcal{C}_t \end{aligned} \quad (15)$$

where $\mathcal{C}_{\mathbf{x}}$ and $\mathcal{C}_t$ are spatial and temporal kernel respectively.

If we observe the process $y(\mathbf{x}, t)$ according to (15), the resulted data matrix $\mathbf{Y} = \mathcal{G}(u)(\mathbf{X}, \mathbf{t}) + \boldsymbol{\eta}$ follows the matrix normal distribution (denoted as ' $\mathcal{MN}$ ') (Gupta and Nagar 2018) for which we can also specify three above-mentioned models

$$\begin{aligned} \mathbf{Y} | \mathbf{M}, \mathbf{U}, \mathbf{V} &\sim \mathcal{MN}(\mathbf{M}, \mathbf{U}, \mathbf{V}), \quad \mathbf{M} = \mathcal{G}(u^\dagger)(\mathbf{X}, \mathbf{t}) \\ \text{static : } \mathbf{U}_s &= \sigma_\epsilon^2 \mathbf{I}_{\mathbf{x}}, \quad \mathbf{V}_s = \mathbf{I}_t \end{aligned} \quad (16a)$$

$$\text{time-average : } \mathbf{U}_T = \Gamma_{\text{obs}}, \quad \mathbf{V}_T = J^2 (\mathbf{1}_J \mathbf{1}_J^\top)^- \quad (16b)$$

$$\text{STGP : } \mathbf{U}_{\text{ST}} = \mathbf{C}_{\mathbf{x}}, \quad \mathbf{V}_{\text{ST}} = \mathbf{C}_t \quad (16c)$$

where $\mathbf{U}(\mathbf{V})$ refers to the covariance matrix characterizing the row (column) wise dependence in data $\mathbf{Y}$, $\mathbf{Y} = \mathcal{O}\mathbf{x}(t; u) = \mathbf{X}(u)$ for the static model and M^- is the pseudo-inverse of M .

In all the above three models (16), we assume $\mathbf{Y}$ i.i.d. over u 's. Denote Φ_* and $\mathcal{I}_*$ as potential function and Fisher

information matrix with $*$ being ‘S’ for the static model (16a), ‘T’ for the time-averaged model (16b) and ‘ST’ for the STGP model (16c) respectively. The following theorem compares the convexity of their likelihoods and indicates that the STGP model (16c) with proper configuration has the advantage of parameter learning with the most convex likelihood among the three models.

Theorem 3.1 *If we set the maximal eigenvalues of $\mathbf{C}_x$ and $\mathbf{C}_t$ such that $\lambda_{\max}(\mathbf{C}_x)\lambda_{\max}(\mathbf{C}_t) \leq \sigma_\epsilon^2$, then the following inequality holds regarding the Fisher information matrices, $\mathcal{I}_s$ and $\mathcal{I}_{st}$, of the static model and the STGP model respectively:*

$$\mathcal{I}_{st}(u) \geq \mathcal{I}_s(u) \quad (17)$$

If we control the maximal eigenvalues of $\mathbf{C}_x$ and $\mathbf{C}_t$ such that $\lambda_{\max}(\mathbf{C}_x)\lambda_{\max}(\mathbf{C}_t) \leq J\lambda_{\min}(\Gamma_{obs})$, then the following inequality holds regarding the Fisher information matrices, $\mathcal{I}_t$ and $\mathcal{I}_{st}$, of the time-averaged model and the STGP model respectively:

$$\mathcal{I}_{st}(u) \geq \mathcal{I}_t(u) \quad (18)$$

Proof See “Appendix A”. $\square$

Remark 1 In practice, we can choose proper parameters, e.g. the magnitude σ^2 , of the covariance kernels $\mathcal{C}_x$ and $\mathcal{C}_t$ to make their maximal eigenvalues upper-bounded. For example, we can set σ^2 small enough so that inequalities in the condition can be satisfied.

The following theorem considers a special case, $\mathbf{C}_x = \Gamma_{obs}$, under a milder condition (straightforward to check e.g. in Sect. 4.2) in comparing the likelihood convexity of the time-averaged model and the STGP model.

Theorem 3.2 *If we choose $\mathbf{C}_x = \Gamma_{obs}$ and require the maximal eigenvalue of $\mathbf{C}_t$, $\lambda_{\max}(\mathbf{C}_t) \leq J$, then the following inequality holds regarding the Fisher information matrices, $\mathcal{I}_t$ and $\mathcal{I}_{st}$, of the time-averaged model and the STGP model respectively:*

$$\mathcal{I}_{st}(u) \geq \mathcal{I}_t(u) \quad (19)$$

Proof See “Appendix A”. $\square$

Remark 2 In general, $\Phi_*(u)$ is not the potential of a Gaussian distribution because of the possible non-linearity of $\mathcal{G}(u)$. Theorems 3.1 and 3.2 indicate that for each $u \in \mathbb{X}$, the STGP model can have a more convex Gaussian proxy in the Laplace approximation.

Often we are interested in predicting the underlying process $y(\mathbf{x}, t)$ at future time t_* given the spatiotemporal

observations $\mathbf{Y}$. Based on the STGP model (15), we could use the following posterior predictive distribution

$$p(y(\mathbf{x}, t_*)|\mathbf{Y}) = \int p(y(\mathbf{x}, t_*)|u, \mathbf{Y})p(u|\mathbf{Y})du \quad (20)$$

Denote the conditional prediction $E[y(\mathbf{x}, t_*)|u, \mathbf{Y}]$ as

$$\mathcal{G}^*(u)(\mathbf{x}, t_*) = \underbrace{\mathcal{G}(u)(\mathbf{x}, t_*)}_{\text{Physical}} + \underbrace{\Gamma_{t_*\mathbf{t}}\Gamma_{\mathbf{t}\mathbf{t}}^{-1}(\mathbf{Y} - \mathcal{G}(u)(\mathbf{X}, \mathbf{t}))}_{\text{Statistical}} \quad (21)$$

Then we predict $y(\mathbf{x}, t_*)$ with the following predictive mean

$$\begin{aligned} E[y(\mathbf{x}, t_*)|\mathbf{Y}] &= E_{u|\mathbf{Y}}[\mathcal{G}^*(u)(\mathbf{x}, t_*)] \\ &\approx \bar{\mathcal{G}}(\mathbf{x}, t_*) + \Gamma_{t_*\mathbf{t}}\Gamma_{\mathbf{t}\mathbf{t}}^{-1}(\mathbf{Y} - \bar{\mathcal{G}}(\mathbf{X}, \mathbf{t})) \end{aligned} \quad (22)$$

where $\bar{\mathcal{G}}(\mathbf{x}, t_*) := \frac{1}{S} \sum_{s=1}^S \mathcal{G}(u^{(s)})(\mathbf{x}, t_*)$ with $u^{(s)} \sim p(u|\mathbf{Y})$. And we can quantify the uncertainty using the law of total conditional variance:

$$\begin{aligned} \text{Var}[y(\mathbf{x}, t_*)|\mathbf{Y}] &= \Gamma_{t_*t_*} - \Gamma_{t_*\mathbf{t}}\Gamma_{\mathbf{t}\mathbf{t}}^{-1}\Gamma_{\mathbf{t}t_*} + \text{Var}_{u|\mathbf{Y}}[\mathcal{G}^*(u)(\mathbf{x}, t_*)] \\ &\approx \Gamma_{t_*t_*} - \Gamma_{t_*\mathbf{t}}\Gamma_{\mathbf{t}\mathbf{t}}^{-1}\Gamma_{\mathbf{t}t_*} + s_{\mathcal{G}^*}^2(\mathbf{x}, t_*) \end{aligned} \quad (23)$$

where $s_{\mathcal{G}^*}^2(\mathbf{x}, t_*) := \frac{1}{S} \sum_{s=1}^S [\mathcal{G}^*(u^{(s)})(\mathbf{x}, t_*) - \bar{\mathcal{G}}(\mathbf{x}, t_*)]^2$ with $u^{(s)} \sim p(u|\mathbf{Y})$.

Assume $t_* \notin \mathbf{t}$. For the static model (6), we have $\Gamma_{t_*\mathbf{t}} = 0$ thus $\mathcal{G}^*(u)(\mathbf{x}, t_*) = \mathcal{G}(u)(\mathbf{x}, t_*)$. Therefore we have the simplified results

$$\begin{aligned} E[y(\mathbf{x}, t_*)|\mathbf{Y}] &\approx \bar{\mathcal{G}}(\mathbf{x}, t_*), \\ \text{Var}[y(\mathbf{x}, t_*)|\mathbf{Y}] &\approx \sigma_\epsilon^2 + s_{\mathcal{G}}^2(\mathbf{x}, t_*) \end{aligned} \quad (24)$$

This may underestimate the uncertainty compared with the more general STGP model (15). If we are only interested in predicting the forward map $\mathcal{G}(u)$ to new time $t = t_*$, we actually have similar results

$$\begin{aligned} E[\mathcal{G}(u)(\mathbf{x}, t_*)|\mathbf{Y}] &\approx \bar{\mathcal{G}}(\mathbf{x}, t_*), \\ \text{Var}[\mathcal{G}(u)(\mathbf{x}, t_*)|\mathbf{Y}] &\approx s_{\mathcal{G}}^2(\mathbf{x}, t_*) \end{aligned} \quad (25)$$

Note all the above prediction is feasible only if we are able to solve ODE/PDE systems to time t_* , i.e. we can evaluate $\mathcal{G}(u^{(s)})(\mathbf{x}, t)$ at $t = t_*$. When we do not have the computer codes available for doing so, we could model $\mathcal{G}(u)(\mathbf{x}, t)$ with another GP $\mathcal{GP}(0, \Gamma^{\mathcal{G}})$ and further predict the forward mapping:

$$\begin{aligned} \mathcal{G}(u)(\mathbf{x}, t_*)|\mathcal{G}(u)(\mathbf{X}, \mathbf{t}) &\sim \\ \mathcal{N}(\Gamma_{t_*\mathbf{t}}^{\mathcal{G}}(\Gamma_{\mathbf{t}\mathbf{t}}^{\mathcal{G}})^{-1}\mathcal{G}(u)(\mathbf{X}, \mathbf{t}), \Gamma_{t_*t_*}^{\mathcal{G}} - \Gamma_{t_*\mathbf{t}}^{\mathcal{G}}(\Gamma_{\mathbf{t}\mathbf{t}}^{\mathcal{G}})^{-1}\Gamma_{\mathbf{t}t_*}^{\mathcal{G}}) \end{aligned} \quad (26)$$

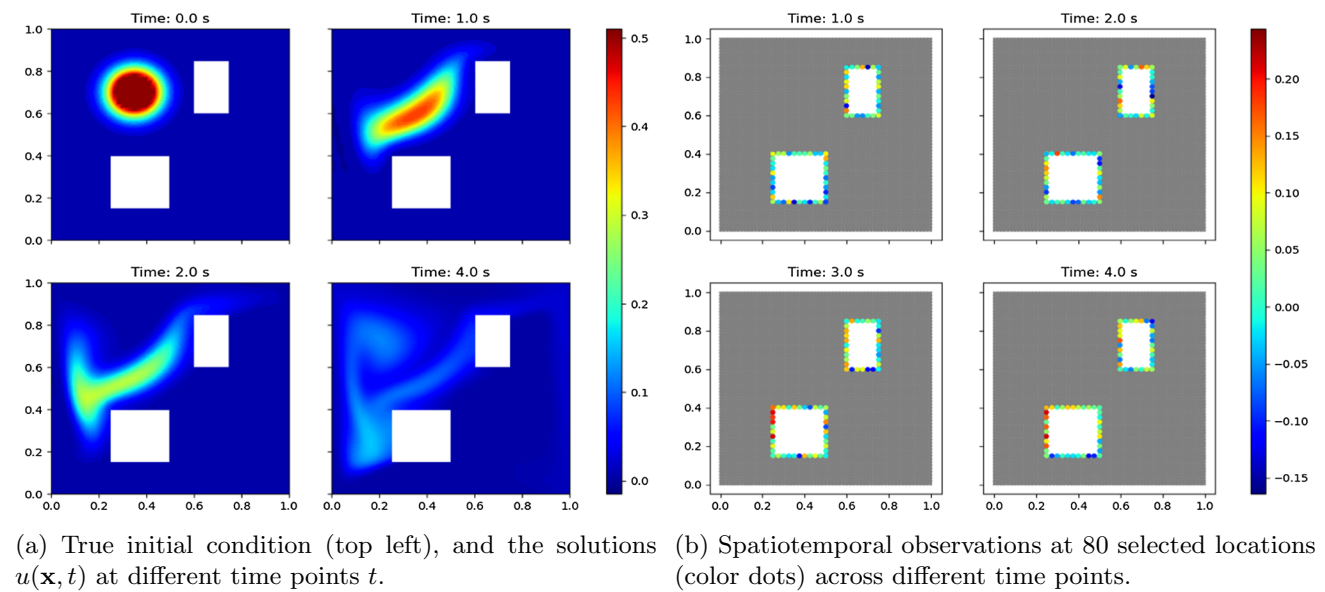


Fig. 2 Advection–diffusion inverse problem

Before concluding this section, we comment on the computational cost of these models in (16). Given the mean (forward output), the complexity of the static model (16a) is $\mathcal{O}(I^3 J)$ in general, or $\mathcal{O}(IJ)$ if $\mathbf{C}_{\mathbf{x}} = \mathbf{I}_{\mathbf{x}}$, the time-averaged model (16b) has complexity $\mathcal{O}(I^3 J)$ and the STGP model (16c) has complexity $\mathcal{O}(I^3 J + IJ^3)$. However, compared with the cost of fitting these statistical models, it is usually more computationally demanding to evaluate the physical model $\mathcal{G}$, whose cost can be reduced by solving the system for less time with better statistical (STGP) model for the inverse solution at the same level of accuracy (See Sect. 4.2). There is in general a trade-off between the model complexity (hence computational expense) and the quality of parameter estimation and UQ.

4 Numerical experiments

In this section, we demonstrate the numerical advantage of spatiotemporal modeling in parameter estimation and UQ. More specifically, we compare the STGP model (15) with the static model (6) using an advection–diffusion inverse problem (Sect. 4.1) previously considered in Villa et al. (2021); Lan et al. (2022) with the static method. Then we compare the STGP model (15) with the time-averaged model (11) using three chaotic dynamical inverse problems (Sect. 4.2) of which the Lorenz problem (Sect. 4.2.1) was studied by Cleary et al. (2021) with the time-averaged approach. Numerical evidences are presented to support that the STGP model (15) is preferable to the other two models. All the computer codes are publicly available at <https://github.com/lanzithinking/Spatiotemporal-inverse-problem>.

4.1 Advection–diffusion inverse problem

First, we consider an inverse problem governed by a parabolic PDE within the Bayesian inference framework. The underlying PDE is a time-dependent advection–diffusion equation that can be applied to heat transfer, air pollution, etc. The inverse problem involves inferring an unknown initial condition $u_0 \in L^2(\Omega)$ from spatiotemporal point measurements $\{y(\mathbf{x}_i, t_j)\}$.

The parameter-to-observable forward mapping $\mathcal{G} : u_0 \rightarrow \mathcal{O}u$ maps the initial condition u_0 to pointwise spatiotemporal observations of the concentration field $u(\mathbf{x}, t)$ through the solution of the following advection–diffusion equation (Petra and Stadler 2011; Villa et al. 2021):

$$\begin{aligned} u_t - \kappa \Delta u + \mathbf{v} \cdot \nabla u &= 0 \quad \text{in } \Omega \times (0, T) \\ u(\cdot, 0) &= u_0 \quad \text{in } \Omega \\ \kappa \nabla u \cdot \mathbf{n} &= 0, \quad \text{on } \partial\Omega \times (0, T) \end{aligned} \quad (27)$$

where $\Omega \subset [0, 1]^2$ is a bounded domain shown in Fig. 2a, $\kappa = 10^{-3}$ is the diffusion coefficient, and $T > 0$ is the final time. The velocity field $\mathbf{v}$ is computed by solving the following steady-state Navier–Stokes equation with the side walls driving the flow (Petra and Stadler 2011):

$$\begin{aligned} -\frac{1}{\text{Re}} \Delta \mathbf{v} + \nabla q + \mathbf{v} \cdot \nabla \mathbf{v} &= 0 \quad \text{in } \Omega \\ \nabla \cdot \mathbf{v} &= 0 \quad \text{in } \Omega \\ \mathbf{v} &= \mathbf{g}, \quad \text{on } \partial\Omega \end{aligned} \quad (28)$$

Here, q is the pressure, and Re is the Reynolds number, which is set to 100 in this example. The Dirichlet boundary data

$\mathbf{g} \in \mathbb{R}^2$ is given by $\mathbf{g} = \mathbf{e}_2 = (0, 1)$ on the left wall of the domain, $\mathbf{g} = -\mathbf{e}_2$ on the right wall, and $\mathbf{g} = \mathbf{0}$ everywhere else.

We set the true initial condition $u_0^\dagger = 0.5 \wedge \exp\{-100[(x - 0.35)^2 + (y - 0.7)^2]\}$, illustrated in the top left panel of Fig. 2a, which also shows a few snapshots of the solutions $u(\mathbf{x}, t)$ at other time points on a regular grid mesh of size 61×61 . To obtain spatiotemporal observations $\{y(\mathbf{x}_i, t_j)\}$, we collect solutions $u(\mathbf{x}, t)$ solved on a refined mesh at $I = 80$ selected locations $\{\mathbf{x}_i\}_{i=1}^I$ across $J = 16$ time points $\{t_j\}_{j=1}^J$ evenly distributed between 1 and 4 seconds (thus denoted as $\mathcal{O}u$) and inject some Gaussian noise $\mathcal{N}(0, \sigma_\eta^2)$ such that the noise standard deviation is $\sigma_\eta = 0.5 \max \mathcal{O}u$, i.e.,

$$y = \mathcal{G}(u_0^\dagger) = \mathcal{O}u(\mathbf{x}, t; u_0^\dagger) + \eta, \quad \eta \sim \mathcal{N}(0, \sigma_\eta^2 I_{1280})$$

Fig. 2b plots 4 snapshots of these observations at 80 locations along the inner boundary. In the Bayesian setting, we adopt a GP prior for $u_0 \sim \mu_0 = \mathcal{GP}(0, \mathcal{C})$ with the covariance kernel $\mathcal{C} = (\delta \mathcal{I} - \gamma \Delta)^{-2}$ defined through the Laplace operator Δ , where δ governs the variance of the prior and γ/δ controls the correlation length. We set $\gamma = 2$ and $\delta = 10$ in this example.

The Bayesian inverse problem involves obtaining an estimate of the initial condition u_0 and quantifying its uncertainty based on the 80×16 spatiotemporal observations. The Bayesian UQ in this example is especially challenging not only because of its large dimensionality (3413) of spatially discretized u (Lagrange degree 1) at each time t , but also due to the spatiotemporal correlations in these observations.

We compare two likelihood models (6) and (15). The static model (6) is commonly used in the literature of Bayesian inverse problems (Lan 2019; Villa et al. 2021; Lan et al. 2022). Here the STGP model (15) is considered to better account for the spatiotemporal relationships in the data. We estimate the variance parameter of the joint kernel from data. The correlation length parameters are determined ($\ell_x = 0.5$ and $\ell_t = 0.2$) by investigating their autocorrelations as in Fig. 18. Figure 1 compares the maximum a posteriori (MAP) estimates of the parameter u_0 by the two likelihood models (right two panels) with the true parameter $u_0^\dagger$ (left panel). The STGP model yields a better MAP estimate closer to the truth compared with the static model.

We also run MCMC algorithms (pCN, ∞ -MALA, and ∞ -HMC) to estimate u_0 . For each algorithm, we run 6000 iterations and burn in the first 1000. The remaining 5000 samples are used to obtain the posterior mean estimate $\hat{u}_0$ (Fig. 3a) and the posterior standard deviation (Fig. 3b). The STGP model (15) consistently generates estimates closer to the true values (refer to Fig. 1) with smaller posterior standard deviation than the static model (6) using various MCMC algorithms. Such improvement of parameter estimation by the STGP model (15) is also verified by smaller relative error

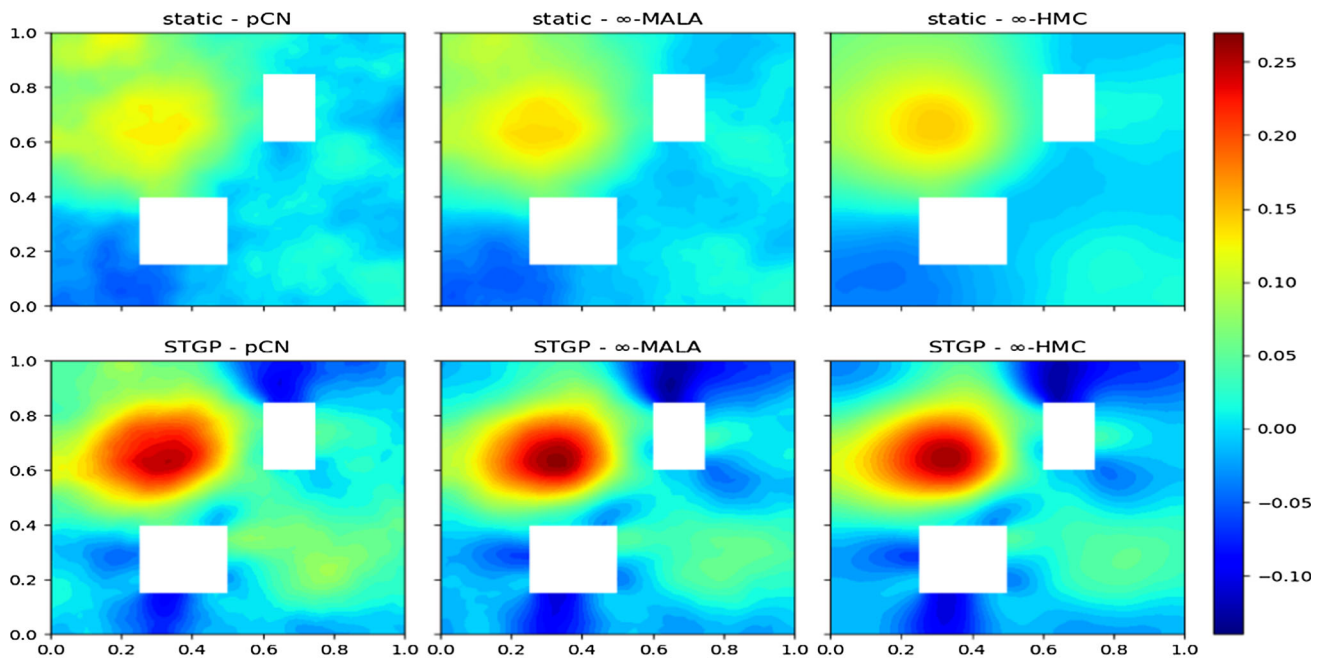
of mean estimates, $\text{REM} = \frac{\|\hat{u}_0 - u_0^\dagger\|}{\|u_0^\dagger\|}$, reported in Table 1, which summarizes the results of 10 repeated experiments with their standard deviations in the brackets.

Finally, we consider the forward prediction (25) over the time interval $[0, 5]$. We substitute each of the 5000 samples $\{u^{(s)}\}_{s=1}^{5000}$ generated by ∞ -HMC into $\mathcal{G}(u^{(s)})(\mathbf{x}, t_*)$ to solve the advection–diffusion equation (27) for $t_* \in [0, 5]$. We observe each of these 5000 solutions at the 80 locations (Fig. 2b) for 50 points equally spaced in $[0, 5]$. Then we obtain the prediction by $\bar{\mathcal{G}}(\mathbf{x}, t_*)_{80 \times 50} = \frac{1}{5000} \sum_{s=1}^{5000} \mathcal{G}(u^{(s)})(\mathbf{x}, t_*)$, and compute the relative errors in terms of the Frobenius norm of the difference between the prediction and the true solution $\mathcal{G}(u_0^\dagger)(\mathbf{x}, t_*)$: $\frac{\|\bar{\mathcal{G}}(\mathbf{x}, t_*) - \mathcal{G}(u_0^\dagger)(\mathbf{x}, t_*)\|}{\|\mathcal{G}(u_0^\dagger)(\mathbf{x}, t_*)\|}$.

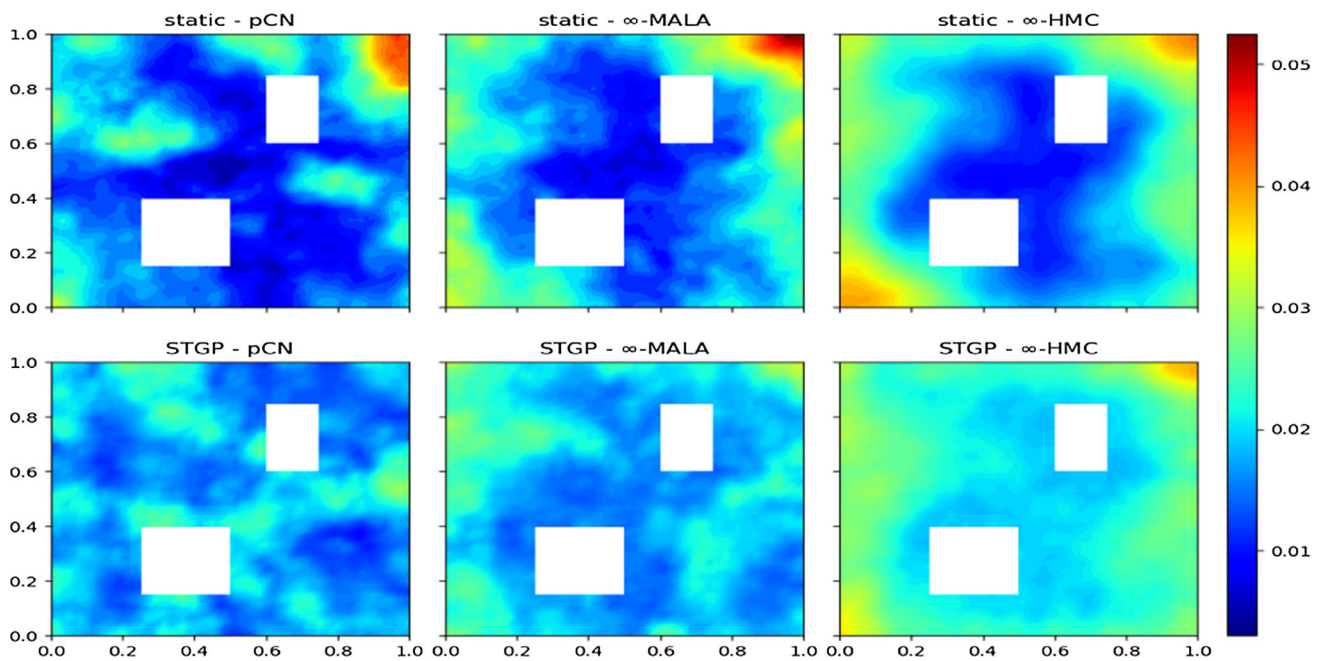
Table 1 shows the STGP model (15) provides more accurate predictions with smaller errors compared with the static model (6). Figure 4 depicts the predicted time series $\bar{\mathcal{G}}(\mathbf{x}, t_*)$ at two selective locations based on the static (blue dashed lines) and the STGP (orange dot-dashed lines) models along with their credible bands (shaded regions) compared with the truth (red solid lines) in the right two panels. Note that with smaller credible bands, the static model is more certain about its prediction that is further away from the truth. While the STGP model provides wider credible bands that cover more of the true trajectories, indicating a more appropriate uncertainty being quantified. Therefore, on the left panel of Fig. 4, the STGP model has a higher rate for its credible intervals among these 80 locations to cover the truth, $\mathcal{G}(u_0^\dagger)(\mathbf{x}, t_*)$, on most of $t_* \in [0, 5]$. Note these models are trained on $t \in [1, 4]$, so the STGP model does not show much advantage at the beginning but quickly outperforms the static model after $t_* = 1$.

4.2 Chaotic dynamical inverse problems

Chaos, refers to the behavior of a dynamical system that appears to be random in long term even its evolution is fully determined by the initial condition. Many physical systems are characterized by the presence of chaos that has been extensively demonstrated (Lorenz 1963; Ivancevic and Ivancevic 2008; Bishop 2017). The main challenges of analyzing chaotic dynamical systems include the stability, the transitivity, and the sensitivity to the initial conditions (which contributes to the seeming randomness) (Effah-Poku et al. 2018). In the study of chaotic dynamical systems, one of the interests is determining the essential system parameters given the observed data. In this section, we will investigate three chaotic dynamical systems, Lorenz63 (Lorenz 1963), Agiza and Yassen (2001) and Yassen (2003), that can be summarized as the first-order ODE: $\dot{\mathbf{x}} = f(\mathbf{x}; u)$. We will apply the CES framework (Sect. 2) to learn the system parameter u and quantify its associated uncertainty based on the observed



(a) Posterior mean estimates of the initial concentration field $u_0(\mathbf{x})$.



(b) Posterior standard deviation estimates of the initial concentration field $u_0(\mathbf{x})$.

Fig. 3 Advection–diffusion inverse problem: comparing posterior estimates of parameter u_0 in the static model (upper row) and the STGP model (lower row) based on 5000 samples by various MCMC algorithms

Table 1 Advection–diffusion inverse problem: comparing (i) posterior estimates of parameter u_0 in terms of relative error of mean REM = $\frac{\|\hat{u}_0 - u_0^\dagger\|}{\|u_0^\dagger\|}$ and (ii) the forward predictions $\mathcal{G}(u)(\mathbf{x}, t_*)$ in terms of relative error $\frac{\|\tilde{\mathcal{G}}(\mathbf{x}, t_*) - \mathcal{G}(u_0^\dagger)(\mathbf{x}, t_*)\|}{\|\mathcal{G}(u_0^\dagger)(\mathbf{x}, t_*)\|}$ by two likelihood models (static and STGP)

Models	Estimation			Prediction		
	pCN	∞ -MALA	∞ -HMC	pCN	∞ -MALA	∞ -HMC
static	0.83 (0.023)	0.81 (0.011)	0.79 (0.005)	0.43 (0.013)	0.4 (0.006)	0.4 (0.003)
STGP	0.74 (0.021)	0.73 (0.012)	0.73 (0.003)	0.44 (0.068)	0.32 (0.016)	0.31 (0.005)

Each experiment is repeated for 10 runs of MCMC (pCN, ∞ -MALA, and ∞ -HMC respectively) and the numbers in the bracket are standard deviations of these repeated experiments

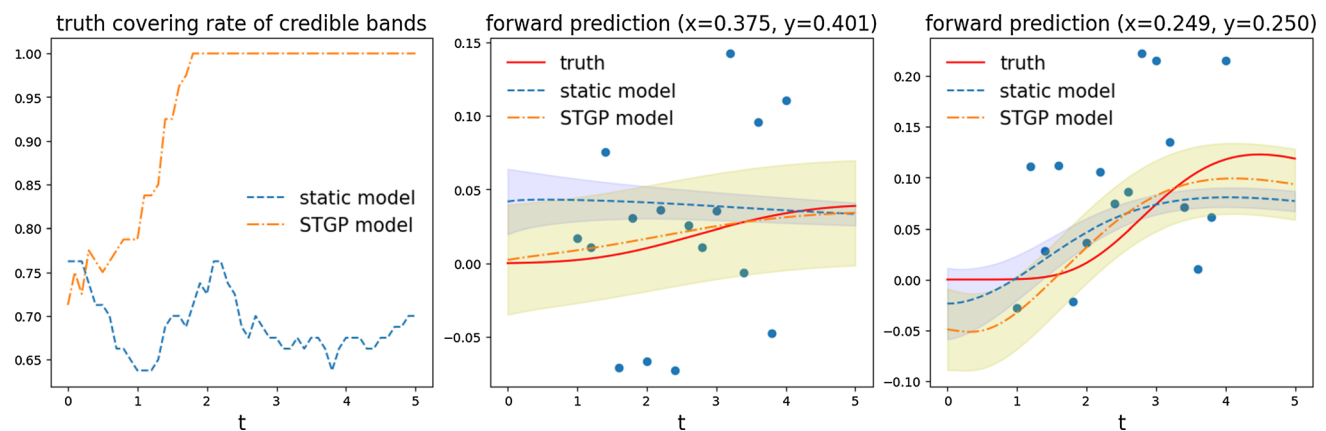


Fig. 4 Advection–diffusion inverse problem: comparing forward predictions, $\tilde{\mathcal{G}}(\mathbf{x}, t_*)$, based on the static model and the STGP model. The left panel plots the curves representing the percentage of 80 (corresponding to the selected locations) credible bands that cover the true solution

$\mathcal{G}(u_0^\dagger)(\mathbf{x}, t_*)$ at each time $t_* \in [0, 5]$. The right two panels show the predicted time series (blue dashed and orange dot-dashed lines) along with the credible bands (shaded regions) by the two models compared with the truth (red solid lines) at two selective locations $\mathbf{x} = (0.375, 0.401)$ and $\mathbf{x} = (0.249, 0.250)$. Blues dots are observations

trajectories. We find the spatiotemporal models numerically more advantageous by fitting the whole trajectories than the common approach by averaging the trajectories over time (Schneider et al. 2017; Cleary et al. 2021; Huang et al. 2022).

4.2.1 Lorenz system

The most popular example of chaotic dynamics is the Lorenz63 system (named after the author and the year it was proposed in Lorenz 1963) that represents a simplified model of atmospheric convection for the chaotic behavior of the weather. The governing equations of the Lorenz system are given by the following ODE

$$\begin{cases} \dot{x} = \sigma(y - x), \\ \dot{y} = x(\rho - z) - y, \\ \dot{z} = xy - \beta z, \end{cases} \quad (29)$$

where x , y , and z denote variables proportional to the convective intensity, horizontal and vertical temperature differences and $u := (\sigma, \rho, \beta)$ represents the model parameters known as the Prandtl number (σ), the Rayleigh number (ρ), and an

unnamed parameter (β) used for physical proportions of the regions (Ott 1981).

The behavior of Lorenz63 system (29) strongly relies on these parameters. In many studies, the parameter ρ varies in $(0, \infty)$ and the other parameters σ and β are held constant. In particular, (29) has a stable equilibrium point at the origin for $\rho \in (0, 1)$. For $\rho \in (1, \gamma)$ with $\gamma = \sigma \frac{\sigma + \beta + 3}{\sigma - \beta - 1}$, (29) has three equilibrium points, one unstable equilibrium point at the origin and two stable equilibrium points at $(\sqrt{\beta(\rho - 1)}, \sqrt{\beta(\rho - 1)}, \rho - 1)^\top$ and $(-\sqrt{\beta(\rho - 1)}, -\sqrt{\beta(\rho - 1)}, \rho - 1)^\top$. When $\rho > \gamma$, the equilibrium points become unstable and it results in unstable spiral shaped trajectories. One classical configuration for the parameters in (29) is $\sigma = 10$, $\beta = \frac{8}{3}$, $\rho = 28$ when the system exhibits two-lobe orbits (Yang et al. 2002) (See the left panel of Fig. 5). In this example, we seek to infer such parameter $u^\dagger = (\sigma^\dagger, \beta^\dagger, \rho^\dagger) = (10, 8/3, 28)$ based on the observed chaotic trajectories demonstrated in the middle panel of Fig. 5. Note the solutions $(x(t), y(t), z(t))$ highly depend on the initial conditions $(x(0), y(0), z(0))$, we hence fix $(x(0), y(0), z(0))$ in the following.

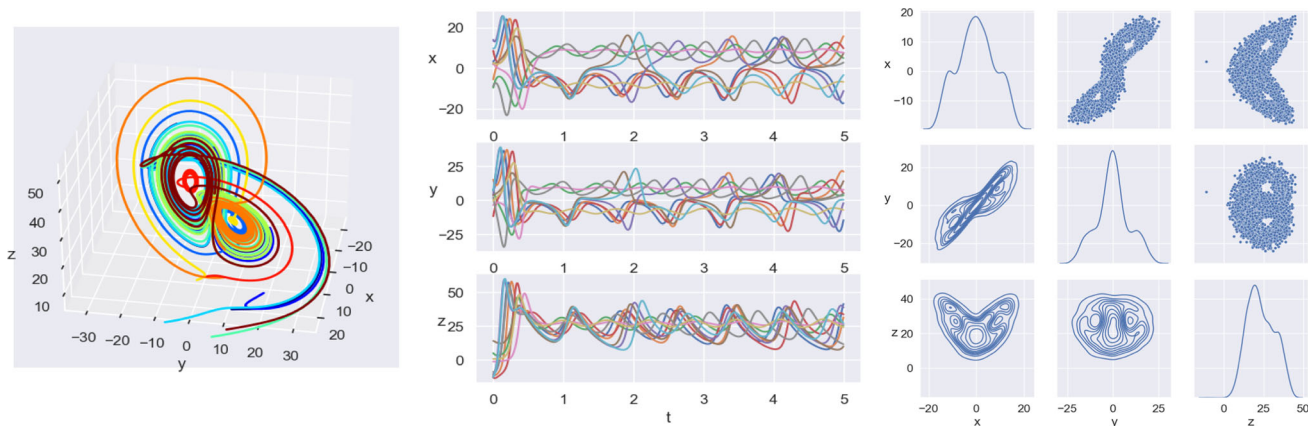


Fig. 5 Lorenz63 dynamics: two-lobe orbits (left), chaotic solutions (middle) and coordinates' distributions (right)

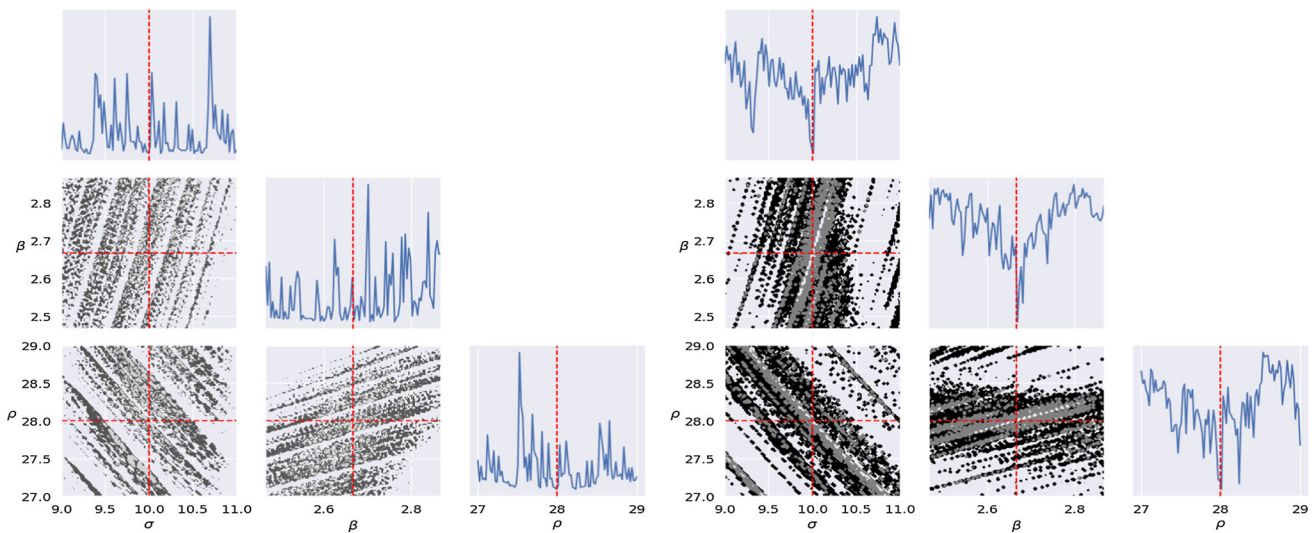


Fig. 6 Lorenz inverse problem: marginal (diagonal) and pairwise (lower triangle) sections of the joint density $p(u)$ by the time-averaged model (left) and the STGP model (right) respectively. Red dashed lines indicate the true parameter values

Due to the chaotic nature of the states $\{(x(t), y(t), z(t)) : t \in [0, \tau]\}$, we can treat these coordinates as random variables. In the right panel of Fig. 5, we demonstrate their marginal and pairwise distributions (diagonal and lower triangle) estimated by a collection of states (upper triangle) along a long-time trajectory solved with $u^\dagger$. For a given parameter $u = (\sigma, \beta, \rho)$, we have the trajectory $\mathcal{G}(u)$ as the following map:

$$\mathcal{G}(u) : \mathbb{R}_+ \rightarrow \mathbb{R}^3, \quad t \mapsto (x(t; u), y(t; u), z(t; u)) \quad (30)$$

where $(x(t; u), y(t; u), z(t; u))$ is the solution of (29) for a given parameter u . We generate spatiotemporal data from the chaotic dynamics (29) with $u^\dagger = (\sigma^\dagger, \beta^\dagger, \rho^\dagger)$ by observing its trajectory on $J = 100$ equally spaced time points $t_j \in [t_0, t_0 + T]$: $\mathbf{X}(u^\dagger)_{3 \times 100} := \{\mathcal{G}(u^\dagger)(t_j) = (x(t_j; u^\dagger), y(t_j; u^\dagger), z(t_j; u^\dagger))\}_{j=1}^J$. These observations can be viewed as a 3-dimensional time series that estimate the

empirical covariance Γ_{obs} as in Cleary et al. (2021). The inverse problem involves learning the parameter u given these observations, also known as parameter identification (Negrini et al. 2021).

Following Cleary et al. (2021), we endow a log-Normal prior on u : $\log u \sim \mathcal{N}(\mu_0, \sigma_0^2)$ with $\mu_0 = (2.0, 1.2, 3.3)$ and $\sigma_0 = (0.2, 0.5, 0.15)$. We compare the two likelihood models (11) and (15) for this dynamical inverse problem. For the time-averaged model (11), we substitute $\bar{\mathbf{X}}(u)_{3 \times 1}$ with $\bar{\mathbf{X}}^*(u)_{9 \times 1} = \mathcal{OG}^*(u)(t)$ by averaging the following augmented trajectory $\mathcal{G}^*(u)(t)$ in time (Cleary et al. 2021):

$$\mathcal{G}^*(u)(t) = (x(t), y(t), z(t), x^2(t), y^2(t), z^2(t), x(t)y(t), x(z)z(t), y(t)z(t))$$

For the spatiotemporal likelihood model STGP (15), we set the correlation length $\ell_x = 0.4$ and $\ell_t = 0.1$ for the spatial

kernel $\mathcal{C}_x$ and the temporal kernel $\mathcal{C}_t$ respectively. They are chosen to reflect the spatial and temporal resolutions.

We first notice that the spatiotemporal modeling facilitates the learning of the true parameter $u^\dagger$. As illustrated in Fig. 6, despite of the rough landscape, the marginal (e.g. $p(\sigma, \beta^\dagger, \rho^\dagger)$) and pairwise (e.g. $p(\sigma, \beta, \rho^\dagger)$) sections of the joint density $p(u)$ by the STGP model (15) are more convex in the neighbourhood of $u^\dagger$ compared with the time-averaged model (11). This verifies the implication of Theorem 3.2 on their difference in convexity. Therefore, particle based algorithms such as EnK methods have higher chance of concentrating their ensemble particles around the true parameter value $u^\dagger$, leading to better estimates. Here, the roughness of the posterior creates barrier for the direct application of MCMC algorithms. Therefore, we apply more robust EnK methods for the parameter estimation.

We run each of the two EnK algorithms for $N = 50$ iterations and choose the ensembles (of size K) when its ensemble mean attains the minimal error in estimating the parameter u with reference to its true value $u^\dagger$. In practice, EnK algorithms usually converge quickly within a few iterations so $N = 50$ suffices the need for most applications.

To investigate the roles of spin-up length t_0 and observation window size T , we run EnK multiple times varying each of the two quantities one at a time. Seen from Fig. 7, we observe consistently smaller errors by the STGP model (15) compared with the time-averaged model (11). More specifically, the upper row indicates that the estimation errors, measured by $\text{REM} = \frac{\|\hat{u} - u^\dagger\|}{\|u^\dagger\|}$, are not every sensitive to the spin-up t_0 given sufficient window size $T = 10$. On the other hand, for fixed spin-up $t_0 = 100$, both models decrease errors with increasing window size T as they aggregate more information. However, the STGP model requires only about $\frac{1}{4}$ time length as the time-averaged model to attain accuracy at the same level ($T = 1$ vs. $T = 4$). This supports that the STGP is preferable to the time-average approach as the former may add a small overhead for the statistical inference but could save much more in resolving the physics (solving ODE/PDE), which is usually more expensive.

Now we set spin-up $t_0 = 100$ long enough to ignore the effect of initial condition in the dynamics and choose the observation window size $T = 10$. We compare the two models (11) (15) using EnK algorithms with different ensemble sizes (K) to obtain an estimate $\hat{u}$ of the parameter u . Figure 19 shows that the STGP model performs better than the time-averaged model in generating smaller errors (REM) for almost all cases. In general more ensembles help reduce the errors except for the time-averaged model using EKI algorithm. Note, the STGP model with EKS algorithm yields parameter estimates with the lowest errors. Table 2 summarizes the REM's by different combinations of the two

likelihood models (time-averaged and STGP) and two EnK algorithms (EKI and EKS). Again we can see consistent advantage of the spatiotemporal likelihood model STGP (15) over the simple time-averaged model (11) in producing more accurate parameter estimates.

Next, we apply CES (Sect. 2) (Cleary et al. 2021; Lan et al. 2022) to quantify the uncertainty of the estimate $\hat{u}$. Direct application of MCMC suffers from the extremely low acceptance rate because of the rough density landscape (Fig. 6). Ensemble particles from EnK algorithm cannot provide rigorous systematic UQ due to the ensemble collapse (Schillings and Stuart 2017a, b; de Wiljes et al. 2018; Chada et al. 2020) (See red dots in Fig. 8). Therefore, we run approximate MCMC based on NN emulators built on the EnK outputs $\{u_n^{(k)}, \mathcal{G}(u_n^{(k)})\}_{k=1, n=0}^{K, N}$. Note, we have different structures for the observed data in the two models (11) (15): 9-dimensional summary of time series for the time-averaged model (11) and 3×100 time series for the STGP model (15). Therefore we build densely connected NN (DNN) $\mathcal{G}^e: \mathbb{R}^3 \rightarrow \mathbb{R}^9$ for the former and DNN-RNN (recurrent NN) type of network $\mathcal{G}^e: \mathbb{R}^3 \rightarrow \mathbb{R}^{3 \times 100}$ for the latter to account for their different data structures in the forward output. Figure 8 compares the marginal (diagonal) and pairwise (lower triangle) posterior densities of u estimated by 10,000 samples (upper triangle) of the pCN algorithm based on the corresponding NN emulators for the two models. The spatiotemporal model STGP (15) estimates the posterior densities peaked around the true parameter $u^\dagger$ in contrast to the estimates by the time-averaged model (11) that are far off from the truth. This is because the DNN-RNN network adopted by the STGP model (15) has an RNN layer for the output to respect the time series nature of the data thus provides a more informative surrogate for the underlying posterior than DNN does for the time-averaged model (11).

Finally, we consider the forward prediction $\bar{\mathcal{G}}(\mathbf{x}, t_*)$ (25) for $t_* \in [t_0, t_0 + 1.5T]$ with $K = 500$ EKS ensembles corresponding to the lowest error. Figure 9 compares the prediction results given by these two models. The result by the STGP model is very close to the truth till $t = 113$ while the prediction by the time-averaged model quickly departs from the truth only after $t = 102$. The STGP model predicts the future of this challenging chaotic dynamics significantly better than the time-averaged model.

4.2.2 Rössler system

Next we consider the Rössler dynamics (Hegazi et al. 2001) governed by the following system of autonomous differential equations:

$$\begin{cases} \dot{x} = -y - z, \\ \dot{y} = x + ay, \\ \dot{z} = b + z(x - c). \end{cases} \quad (31)$$

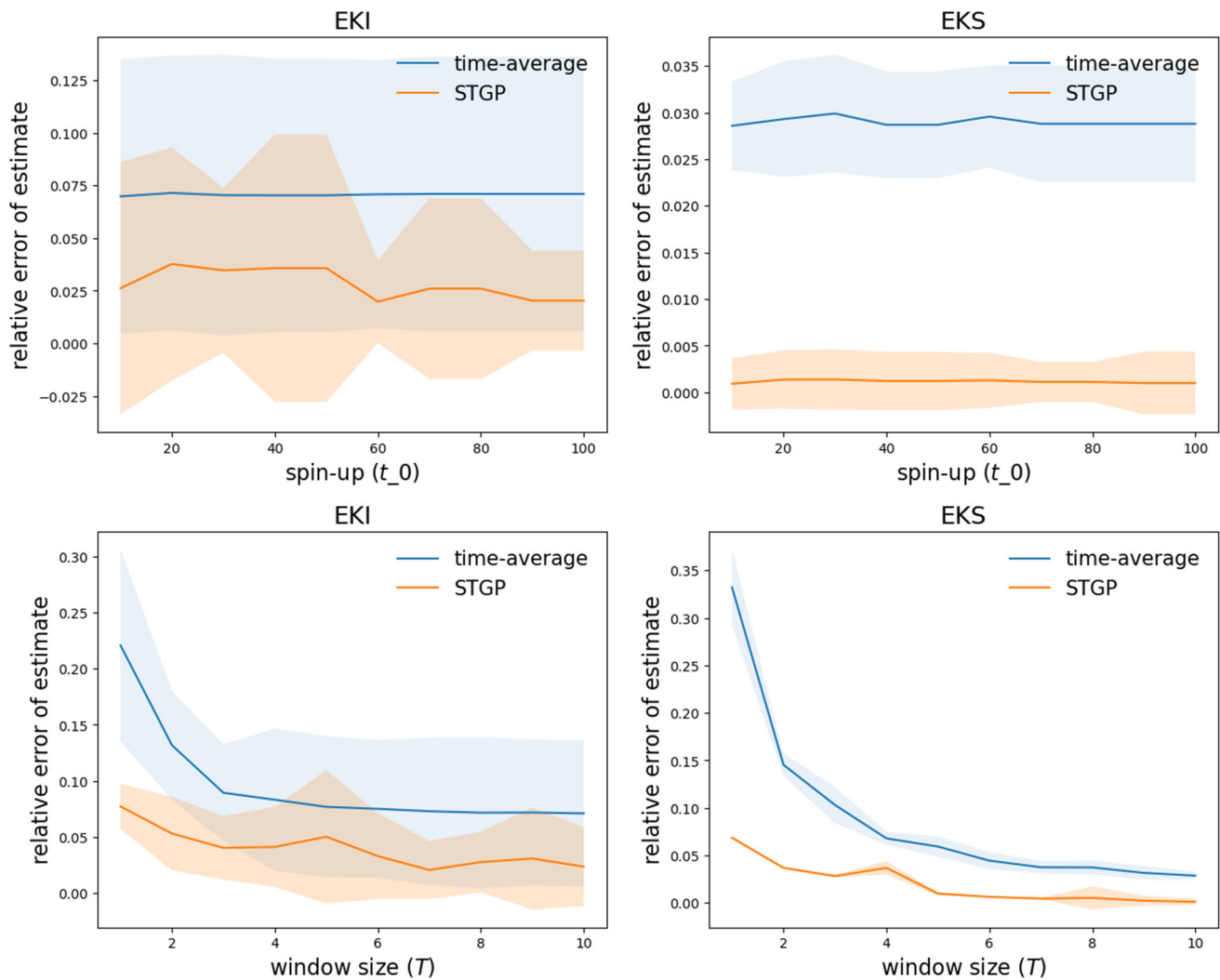


Fig. 7 Lorenz inverse problem: comparing posterior estimates of parameter u for two models (time-average and STGP) in terms of relative error of mean $REM = \frac{\|\hat{u} - u^\dagger\|}{\|u^\dagger\|}$. The upper row shows the results by varying the spin-up t_0 and fixing $T = 10$. The lower row

shows the results by varying the observation window size T and fixing $t_0 = 100$. Each experiment is repeated for 10 runs of EnK (EKI and EKS respectively) with $K = 500$ ensembles and the shaded regions indicate standard deviations of such repeated experiments

Table 2 Lorenz inverse problem: comparing posterior estimates of parameter u for two models, time-average (Tavg) and STGP, in terms of relative error of median $REM = \frac{\|\hat{u} - u^\dagger\|}{\|u^\dagger\|}$

Model-Algorithms	K = 50	K = 100	K = 200	K = 500	K = 1000
Tavg-EKI	0.06 (0.03)	0.09 (0.03)	0.09 (0.01)	0.06 (0.04)	0.07 (0.02)
Tavg-EKS	0.10 (0.02)	0.07 (4.62e−03)	0.05 (2.60e−03)	0.03 (3.04e−03)	0.03 (8.56e−04)
STGP-EKI	0.07 (0.03)	0.04 (0.03)	0.03 (0.02)	0.02 (0.03)	0.02 (0.01)
STGP-EKS	0.09 (0.03)	0.05 (0.03)	0.03 (0.02)	3.97e−04 (1.06e−03)	5.52e−04 (6.37e−04)

Each experiment is repeated for 10 runs of EnK (EKI and EKS respectively) and the numbers in the bracket are standard deviations of such repeated experiments

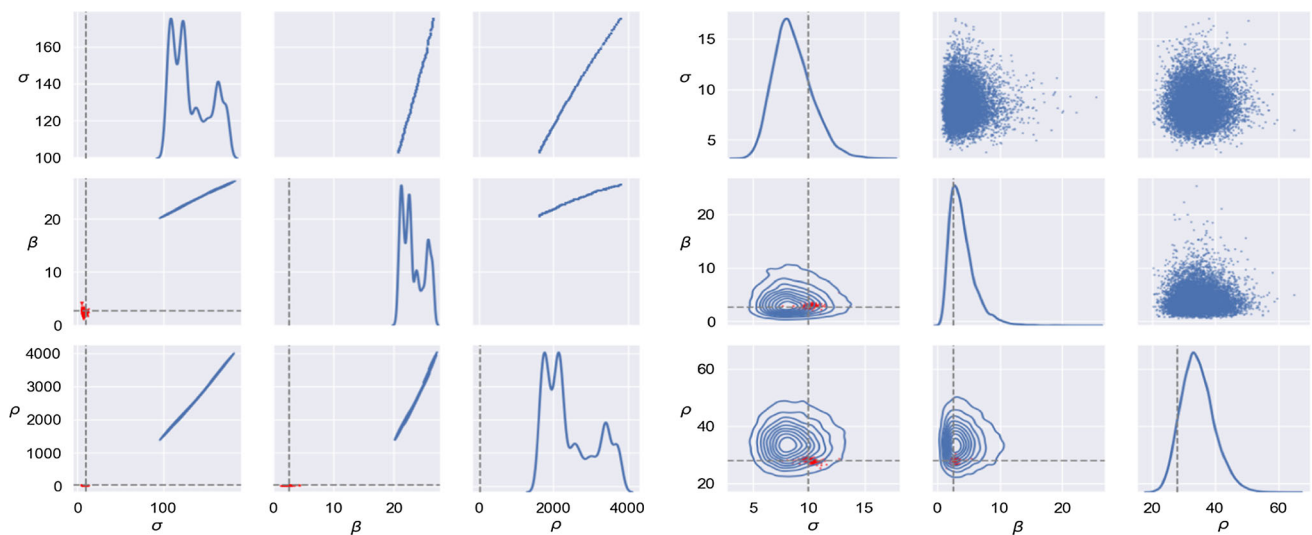


Fig. 8 Lorenz inverse problem: marginal (diagonal) and pairwise (lower triangle) distributions estimated with 10,000 samples (upper triangle) by the pCN algorithm based on NN emulators for the time-averaged model (left) and the STGP model (right) respectively. Red

dots (lower triangle) are selective 10,000 ensemble particles from running the EKS algorithm. Gray dashed lines indicate the true parameter values

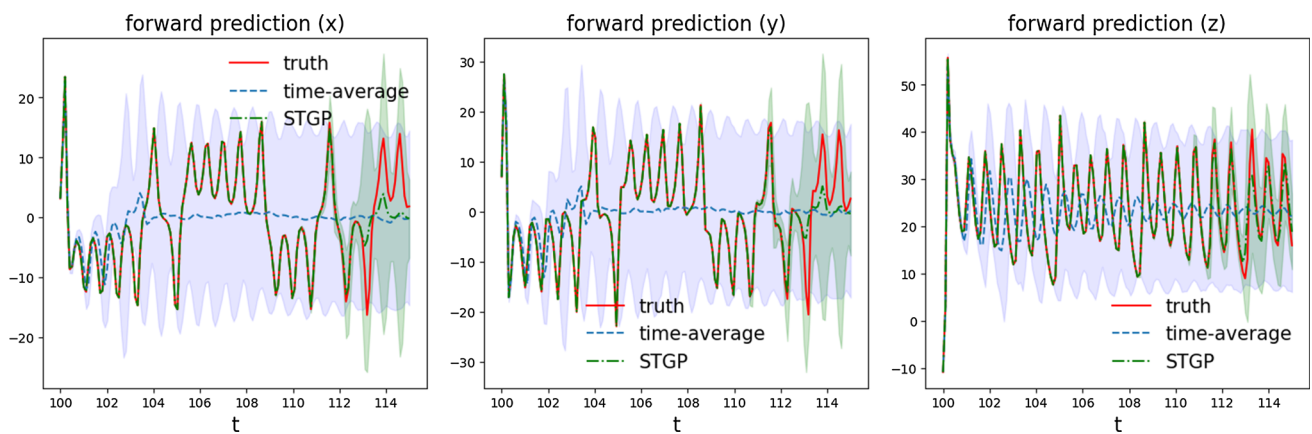


Fig. 9 Lorenz inverse problem: comparing forward predictions $\tilde{\mathcal{G}}(\mathbf{x}, t_*)$ based on the time-averaged model and the STGP model

where $a, b, c > 0$ are parameters determining the behavior of the system. The Rössler attractor was originally discovered by German biochemist Otto Eberhard (Rössler 1976; Rossler 1979). When $c^2 > 4ab$, the system (31) exhibits continuous-time chaos and has two unstable equilibrium points $(a\gamma_-, -\gamma_-, \gamma_-)$ and $(a\gamma_+, -\gamma_+, \gamma_+)$ with $\gamma_+ = \frac{c + \sqrt{c^2 - 4ab}}{2a}$, $\gamma_- = \frac{c - \sqrt{c^2 - 4ab}}{2a}$. Note that the Rössler attractor has similarities to the Lorenz attractor, nevertheless it has a single lobe and offers more flexibility in the qualitative analysis. The true parameter that we try to infer is $u^\dagger = (a^\dagger, b^\dagger, c^\dagger) = (0.2, 0.2, 5.7)$. Figure 10 illustrates the single-lobe orbits (left), the chaotic solutions (middle) and their marginal and pairwise distributions (right) of their coordinates viewed as random variables.

Note, the Rössler dynamics evolve at a lower rate compared with the Lorenz63 dynamics (compare the middle panels of Figs. 10, 5). Therefore, we adopt a longer spin-up length ($t_0 = 1000$) and a larger window size ($T = 100$). For the STGP model (15), spatiotemporal data are generated by observing the trajectory (30) of the chaotic dynamics (31) with $u^\dagger = (a^\dagger, b^\dagger, c^\dagger)$ for $J = 100$ time points in $[t_0, t_0 + T]$. We also augment the time-averaged data with second-order moments for the time-averaged model (11). In this Bayesian inverse problem, we adopt a log-Normal prior on u : $\log u \sim \mathcal{N}(\mu_0, \sigma_0^2)$ with $\mu_0 = (-1.5, -1.5, 2.0)$ and $\sigma_0 = (0.15, 0.15, 0.2)$. Once again, with spatiotemporal likelihood model STGP (15), learning the true parameter value $u^\dagger$ becomes easier because the posterior density $p(u)$ concentrates more on $u^\dagger$ compared with the time-averaged

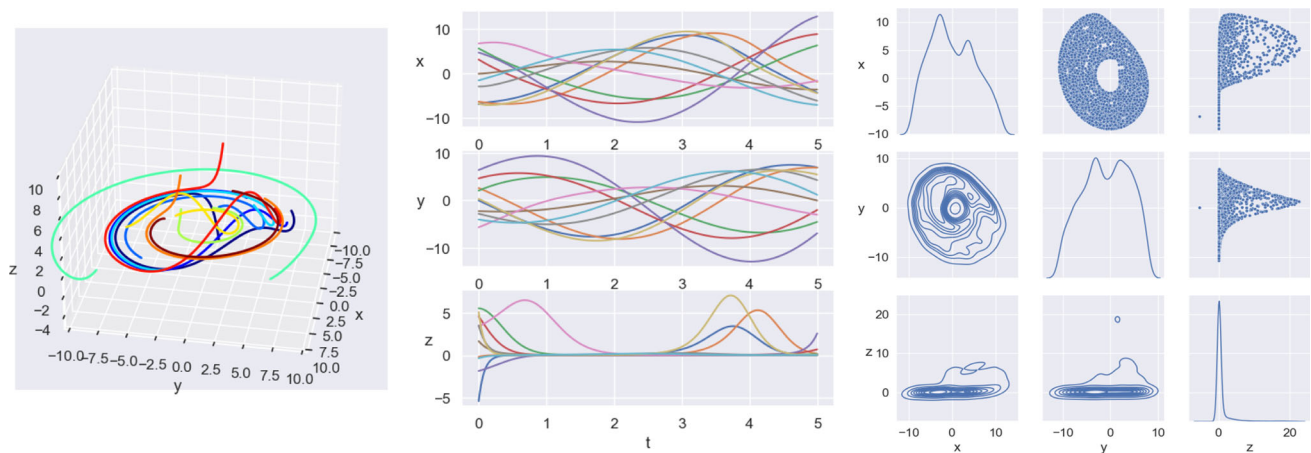


Fig. 10 Rössler dynamics: single-lobe orbits (left), chaotic solutions (middle) and coordinates' distributions (right)

model (11), as indicated by Theorem 3.2. See Fig. 20 for the comparison on their marginal and pairwise sections of the joint density $p(u)$.

We also compare the two models (11) (15) when investigating the roles of spin-up length t_0 and observation window size T in Fig. 11. Despite of the consistent smaller errors (expressed in terms of REM) by the STGP model, REM is not every sensitive to the spin-up t_0 given sufficient window size $T = 100$. However, for fixed spin-up $t_0 = 100$, the STGP model (15) is superior than the time-averaged approach (11) in reducing the estimation error using smaller observation time window T : the former requires only half time length as the latter to attain the same level of accuracy ($T = 30$ vs. $T = 60$ with EKI and $T = 20$ vs. $T = 40$ with EKS).

Now we fix $t_0 = 1000$ and $T = 100$. Figure 21 compares these two models (11) (15) in terms of REM's of the parameter estimation by EnK algorithms with different ensemble sizes (K). The STGP model (15) shows universal advantage over the time-averaged model (11) in generating smaller REM's. Note, the time-averaged model becomes over-fitting if running EKS more than 10 iterations, a phenomenon also reported in Iglesias et al. (2013) and Iglesias (2016). Table 3 summarizes the REM's by different combinations of likelihood models and EnK algorithms and confirms the consistent advantage of the STGP model over the time-averaged model in rendering more accurate parameter estimation.

We apply CES (Sect. 2) (Cleary et al. 2021; Lan et al. 2022) for the UQ. Based on the EKS ($K = 500$) outputs, we build DNN $\mathcal{G}^e : \mathbb{R}^3 \rightarrow \mathbb{R}^9$ for the time-averaged model (11) and DNN-RNN $\mathcal{G}^e : \mathbb{R}^3 \rightarrow \mathbb{R}^{3 \times 100}$ for the STGP model (15) to account for their different data structures. Figure 12 compares the marginal and pairwise posterior densities of u estimated by 10,000 samples of the pCN algorithm based on the corresponding NN emulators for the two models. The STGP model (15) generates more appropriate UQ results than the time-averaged model (11) does. Finally, we consider

the forward prediction $\bar{\mathcal{G}}(\mathbf{x}, t_*)$ (25) for $t_* \in [t_0, t_0 + 1.5T]$ with $K = 500$ EKS ensembles corresponding to the lowest error. Figure 13 shows that the STGP model provides better prediction consistent with the truth throughout the whole time window while the result by the time-averaged model deviates from the truth quickly after $t = 1020$.

4.2.3 Chen system

Yet another chaotic dynamical system we consider is the Chen system (Chen and Ueta 1999) described by the following ODE:

$$\begin{cases} \dot{x} = a(y - x), \\ \dot{y} = (c - a)x - xz + cy, \\ \dot{z} = xy - bz. \end{cases} \quad (32)$$

where $a, b, c > 0$ are parameters. When $a = 35, b = 3, c = 28$, the system (32) has a double-scroll chaotic attractor often observed from a physical, electronic chaotic circuit. The true parameter that we will infer is $u^\dagger = (a^\dagger, b^\dagger, c^\dagger) = (35, 3, 28)$. With $u^\dagger$, the system has three unstable equilibrium states given by $(0, 0, 0)$, $(\gamma, \gamma, 2c - a)$, and $(-\gamma, -\gamma, 2c - a)$ where $\gamma = \sqrt{b(2c - a)}$ (Yassen 2003). Figure 14 illustrates the two-scroll attractor (left), the chaotic trajectories (middle) and their marginal and pairwise distributions (right) of their coordinates viewed as random variables.

The Chen dynamics has trajectories changing rapidly as the Lorenz63 dynamics (compare the middle panels of Fig. 14 and Fig. 5). Therefore we adopt the same spin-up length ($t_0 = 100$) and observation window size ($T = 10$) as in the Lorenz inverse problem (Sect. 4.2.1). We generate the spatiotemporal data and the augmented time-averaged summary data by observing the trajectory of (32) over $[t_0, t_0 + T]$ solved with $u^\dagger$ similarly as in the previous sections. A log-Nomral prior is

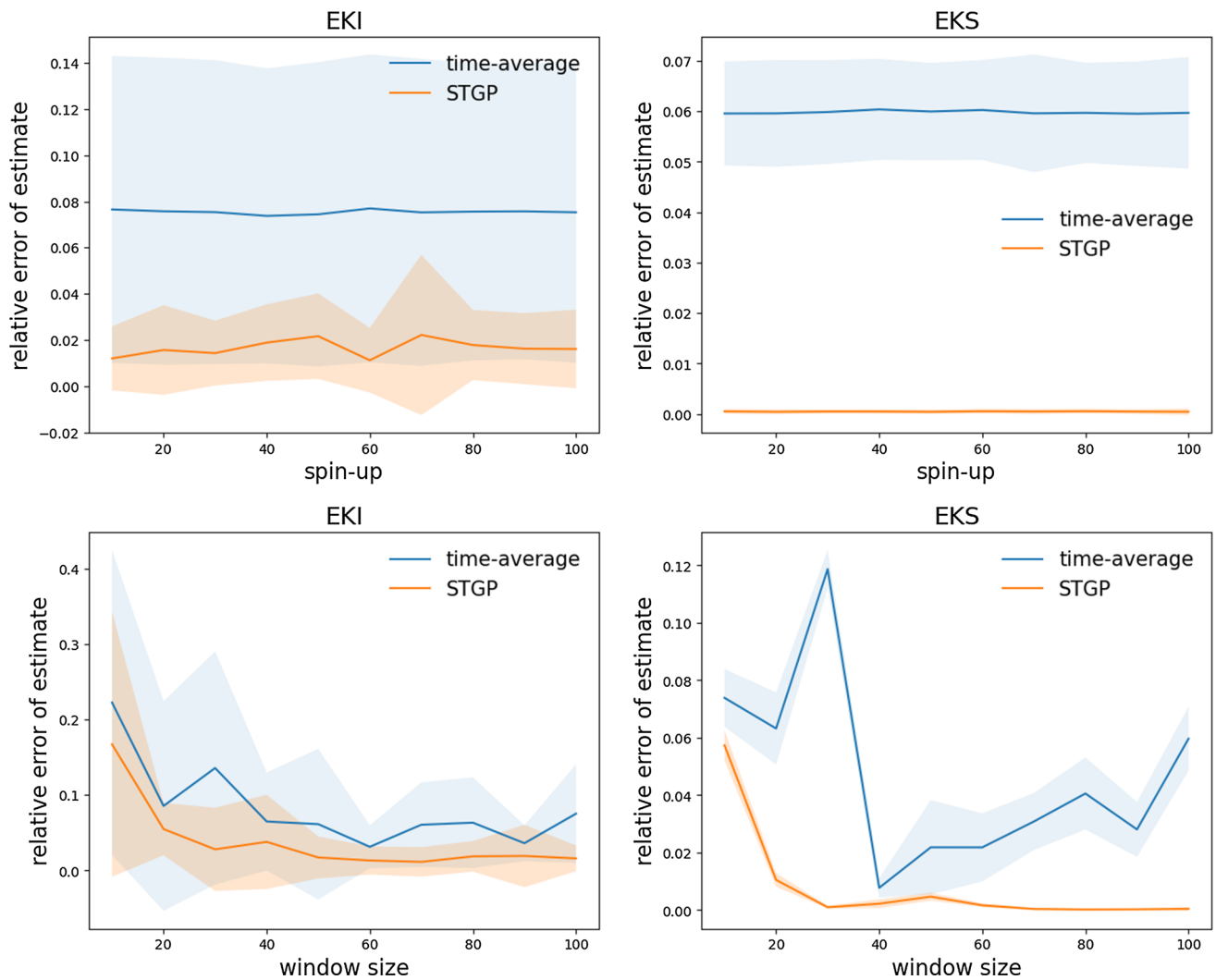


Fig. 11 Rössler inverse problem: comparing posterior estimates of parameter u for two models (time-average and STGP) in terms of relative error of mean $\text{REM} = \frac{\|\hat{u} - u^\dagger\|}{\|u^\dagger\|}$. The upper row shows the results by varying the spin-up t_0 and fixing $T = 100$. The lower row shows

the results by varying the window size T and fixing $t_0 = 100$. Each experiment is repeated for 10 runs of EnK (EKI and EKS respectively) with $K = 500$ ensembles and the shaded regions indicate standard deviations of such repeated experiments

Table 3 Rössler inverse problem: comparing posterior estimates of parameter u for two models (time-average and STGP) in terms of relative error of median $\text{REM} = \frac{\|\hat{u} - u^\dagger\|}{\|u^\dagger\|}$

Model- Algo	K = 50	K = 100	K = 200	K = 500	K = 1000
Tavg-EKI	0.16 (0.09)	0.11 (0.06)	0.10 (0.07)	0.07 (0.04)	0.11 (0.07)
Tavg-EKS	0.06 (0.02)	0.06 (7.61e−03)	0.06 (6.20e−03)	0.06 (5.37e−03)	0.06 (2.53e−03)
STGP-EKI	0.02 (0.02)	0.01 (0.01)	0.02 (0.02)	0.01 (9.09e−03)	0.01 (0.02)
STGP-EKS	0.02 (0.01)	2.47e−03 (0.02)	7.63e−04 (2.86e−03)	4.23e−04 (2.45e−04)	3.62e−04 (1.19e−04)

Each experiment is repeated for 10 runs of EnK (EKI and EKS respectively) and the numbers in the bracket are standard deviations of such repeated experiments

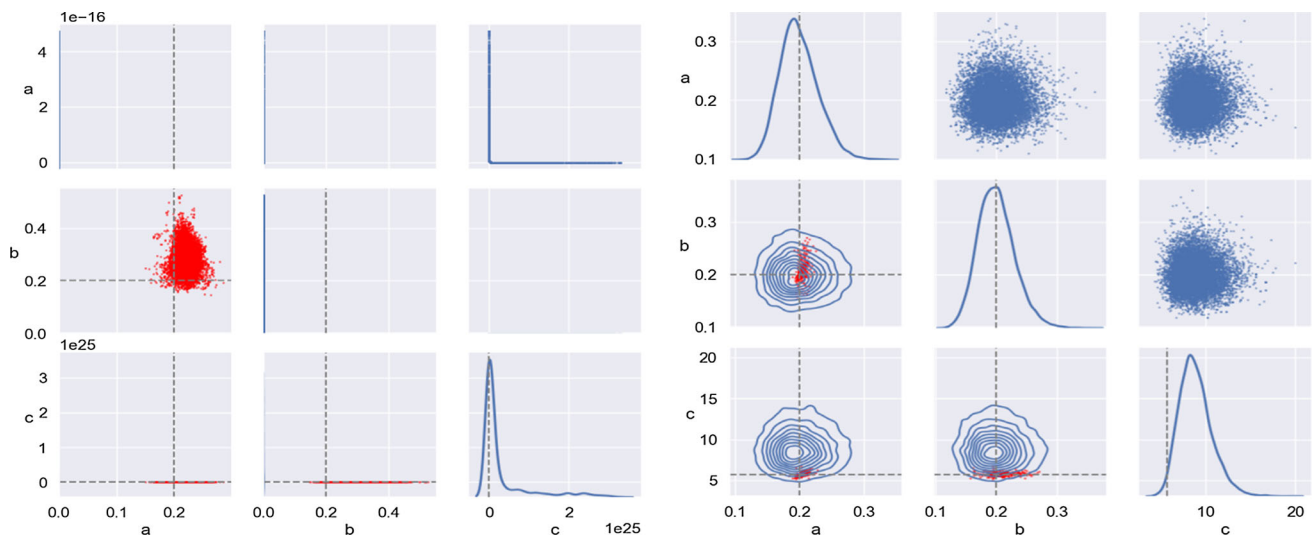


Fig. 12 Rössler inverse problem: marginal (diagonal) and pairwise (lower triangle) distributions estimated with 10,000 samples (upper triangle) by the pCN algorithm based on NN emulators for the time-averaged model (left) and the STGP model (right) respectively. Red

dots (lower triangle) are selective 10,000 ensemble particles from running the EKS algorithm. Gray dashed lines indicate the true parameter values

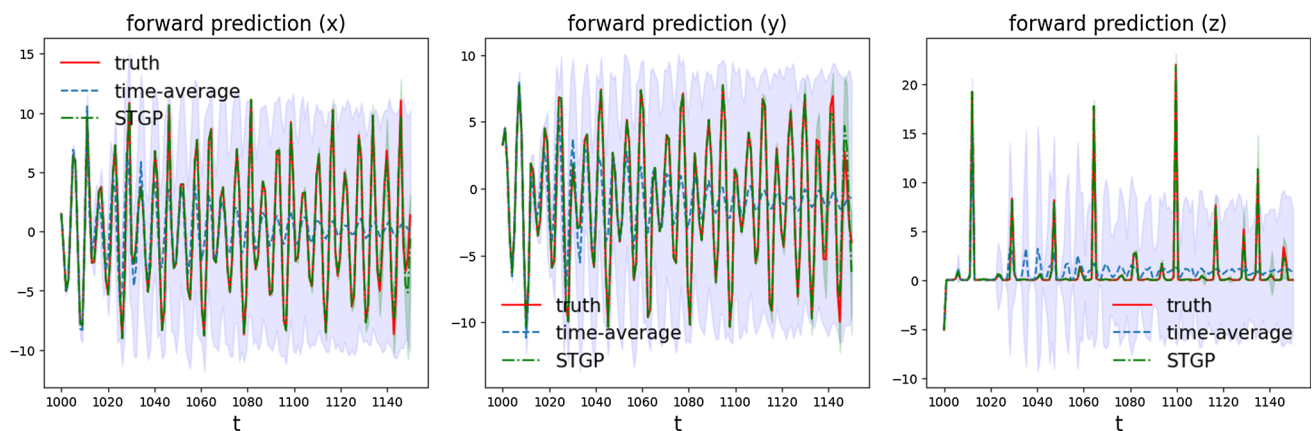


Fig. 13 Rössler inverse problem: comparing forward predictions $\tilde{G}(\mathbf{x}, t_*)$ based on the time-averaged model and the STGP model

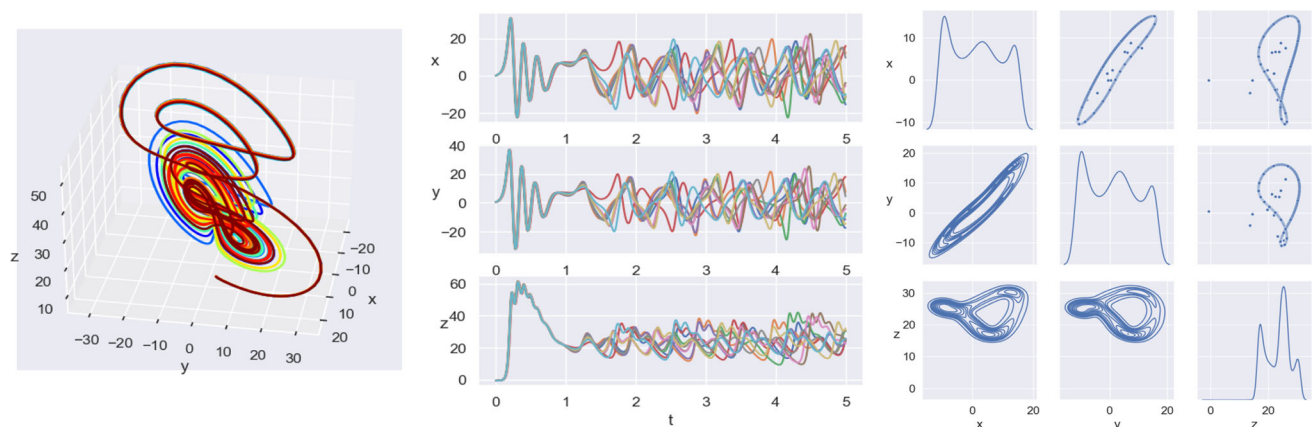


Fig. 14 Chen dynamics: double-scroll attractor (left), chaotic solutions (middle) and coordinates' distributions (right)

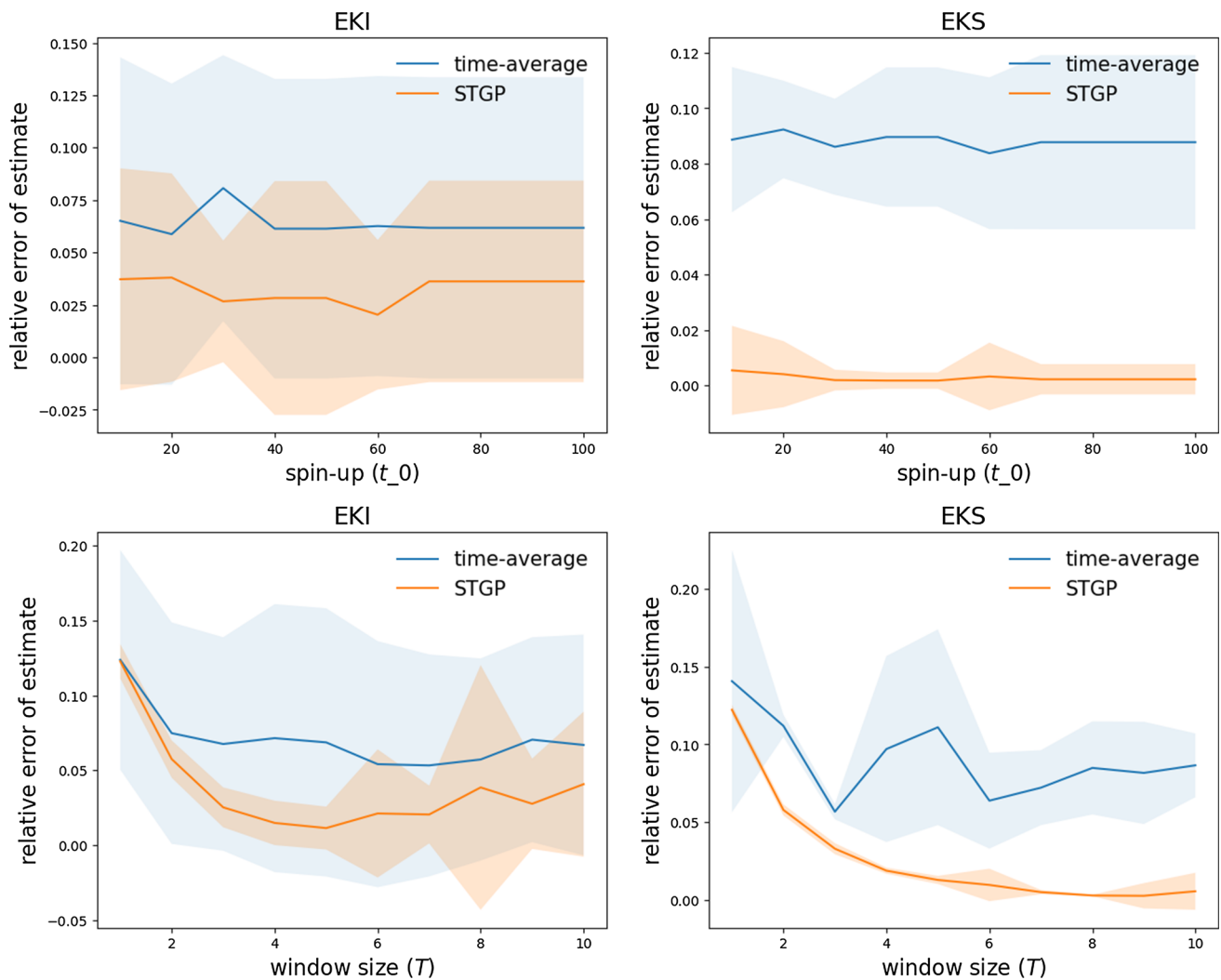


Fig. 15 Chen inverse problem: comparing posterior estimates of parameter u for two models (time-average and STGP) in terms of relative error of mean $\text{REM} = \frac{\|\hat{u} - u^\dagger\|}{\|u^\dagger\|}$. The upper row shows the results by varying the spin-up t_0 and fixing $T = 10$. The lower row shows the results

by varying the observation window size T and fixing $t_0 = 100$. Each experiment is repeated for 10 runs of EnK (EKI and EKS respectively) with $K = 500$ ensembles and the shaded regions indicate standard deviations of such repeated experiments

Table 4 Chen inverse problem: comparing posterior estimates of parameter u for two models (time-average and STGP) in terms of relative error of median $\text{REM} = \frac{\|\hat{u} - u^\dagger\|}{\|u^\dagger\|}$

Model-Algo	K = 50	K = 100	K = 200	K = 500	K = 1000
Tavg-EKI	0.07 (0.03)	0.04 (0.04)	0.04 (0.04)	0.05 (0.04)	0.04 (0.04)
Tavg-EKS	0.12 (0.03)	0.10 (0.02)	0.09 (0.02)	0.09 (0.01)	0.09 (0.01)
STGP-EKI	0.14 (0.09)	0.09 (0.08)	0.09 (0.08)	0.03 (0.03)	0.01 (9.87e−03)
STGP-EKS	0.07 (0.04)	0.05 (0.04)	0.01 (0.01)	2.89e−03 (6.07e−03)	3.32e−04 (4.66e−04)

Each experiment is repeated for 10 runs of EnK (EKI and EKS respectively) and the numbers in the bracket are standard deviations of such repeated experiments

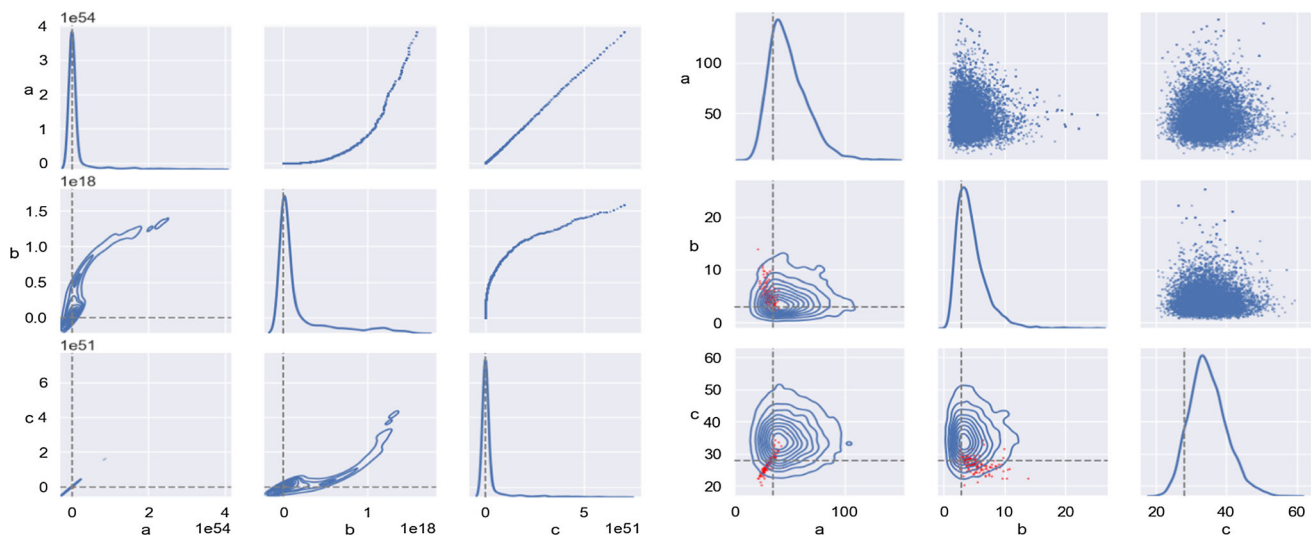


Fig. 16 Chen inverse problem: marginal (diagonal) and pairwise (lower triangle) distributions estimated with 10,000 samples (upper triangle) by the pCN algorithm based on NN emulators for the time-averaged

model (left) and the STGP model (right) respectively. Red dots (lower triangle) are selective 10,000 ensemble particles from running the EKS algorithm. Gray dashed lines indicate the true parameter values

adopted for u : $\log u \sim \mathcal{N}(\mu_0, \sigma_0^2)$ with $\mu_0 = (3.5, 1.2, 3.3)$ and $\sigma_0 = (0.35, 0.5, 0.15)$. The STGP model (15) still possesses more convex posterior density $p(u)$ than the time-averaged model (11) as illustrated by its marginal and pairwise sections plotted in Fig. 22.

Varying the spin-up length t_0 and the observation window size T one at a time in Fig. 15, we observe similar advantage of the STGP model compared with the time-averaged model regardless of the insensitivity of errors with respect to t_0 . Similarly, the STGP model demands a smaller observation window than the time-averaged model ($T = 2$ vs. $T = 6$ with EKI and $T = 2$ vs. $T = 3$ with EKS) to reach the same level of accuracy.

Again we see the merit of the STGP model (15) in reducing the error (REM) of parameter estimation compared with

the time-averaged model (11) in various combinations of EnK algorithms with different ensemble sizes (K) in Fig. 23 and Table 4. As in the previous problem (Sect. 4.2.2), similar over-fitting (bottom left of Fig. 23) by the time-averaged model occurs if running EKS algorithms more than 5 iterations (or earlier).

UQ results (Fig. 16) by CES show the STGP model estimates the uncertainty of parameter u more appropriately than the time-averaged model. Finally, though the prediction is challenging to the Chen dynamics (32), the STGP model still performs much better than the time-averaged model by predicting more accurate trajectory for longer time ($t = 111$ vs $t = 101$) as shown in Fig. 17.

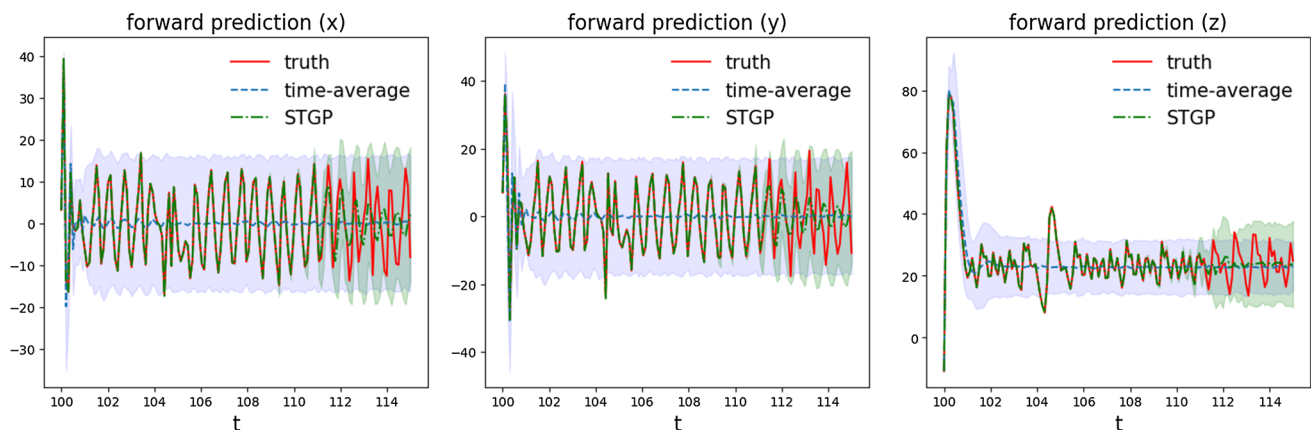


Fig. 17 Chen inverse problem: comparing forward predictions $\bar{G}(\mathbf{x}, t_*)$ based on the time-averaged model and the STGP model

5 Conclusion

In this paper, we investigate the inverse problems with spatiotemporal data. We compare the Bayesian models based on STGP with traditional static and time-averaged models that do not fully integrate the spatiotemporal information. By fitting the trajectories of the observed data, the STGP model provides more effective parameter estimation and more appropriate UQ. We explain the superiority of the STGP model in theorems showing that it renders more convex likelihood that facilitates the parameter learning. We demonstrate the advantage of the spatiotemporal modeling using an inverse problem constrained by an advection–diffusion PDE and three inverse problems involving chaotic dynamics.

Theorems 3.1 and 3.2 compare the STGP model with the static and the time-averaged models regarding their statistical convexity. These novel qualitative results imply that the parameter learning (based on EnK methods) with the STGP model converges faster than the other two traditional methods. In the future work, we will explore a quantitative characterization on their convergence rates particularly in terms of covariance properties.

The STGP model (15) considered in this paper has a classical separation structure in their joint kernel. This may not be sufficient to characterize complex spatiotemporal relationships, e.g. the temporal evolution of spatial dependence (TESD) (Lan 2022). We will expand this work by considering non-stationary and/or non-separable STGP models (Cressie and Wikle 2011; Zhang and Cressie 2020; Wang et al. 2020; Lan 2022) to account for more complicated space-time interactions in these spatiotemporal inverse problems.

Acknowledgements SL is supported by NSF grant DMS-2134256.

Appendix A Proofs

Theorem (3.1) *If we set the maximal eigenvalues of $\mathbf{C}_x$ and $\mathbf{C}_t$ such that $\lambda_{\max}(\mathbf{C}_x)\lambda_{\max}(\mathbf{C}_t) \leq \sigma_\epsilon^2$, then the following inequality holds regarding the Fisher information matrices, $\mathcal{I}_s$ and $\mathcal{I}_{st}$, of the static model and the STGP model respectively:*

$$\mathcal{I}_{st}(u) \geq \mathcal{I}_s(u) \quad (\text{A1})$$

If we control the maximal eigenvalues of $\mathbf{C}_x$ and $\mathbf{C}_t$ such that $\lambda_{\max}(\mathbf{C}_x)\lambda_{\max}(\mathbf{C}_t) \leq J\lambda_{\min}(\Gamma_{obs})$, then the following inequality holds regarding the Fisher information matrices, $\mathcal{I}_t$ and $\mathcal{I}_{st}$, of the time-averaged model and the STGP model respectively:

$$\mathcal{I}_{st}(u) \geq \mathcal{I}_t(u) \quad (\text{A2})$$

Proof Denote $\mathbf{Y}_0 = \mathbf{Y} - \mathbf{M}$. We have $\Phi_*(u) = \frac{1}{2} \text{tr} [\mathbf{V}_*^{-1} \mathbf{Y}_0^T \mathbf{U}_*^{-1} \mathbf{Y}_0]$ with $*$ being S or ST. $\mathbf{U}_s, \mathbf{V}_s, \mathbf{U}_{st}$ and $\mathbf{V}_{st}$ are specified in (16). We notice that both $\mathbf{U}_*$ and $\mathbf{V}_*$ are symmetric, then we have

$$\begin{aligned} \frac{\partial \Phi_*}{\partial u_i} &= \frac{1}{2} \left\{ \text{tr} \left[\mathbf{V}_*^{-1} \frac{\partial \mathbf{Y}_0^T}{\partial u_i} \mathbf{U}_*^{-1} \mathbf{Y}_0 \right] + \text{tr} \left[\mathbf{V}_*^{-1} \mathbf{Y}_0^T \mathbf{U}_*^{-1} \frac{\partial \mathbf{Y}_0}{\partial u_i} \right] \right\} \\ &= \text{tr} \left[\mathbf{V}_*^{-1} \mathbf{Y}_0^T \mathbf{U}_*^{-1} \frac{\partial \mathbf{Y}_0}{\partial u_i} \right] \\ \frac{\partial^2 \Phi_*}{\partial u_i \partial u_j} &= \text{tr} \left[\mathbf{V}_*^{-1} \mathbf{Y}_0^T \mathbf{U}_*^{-1} \frac{\partial^2 \mathbf{Y}_0}{\partial u_i \partial u_j} \right] + \text{tr} \left[\mathbf{V}_*^{-1} \frac{\partial \mathbf{Y}_0^T}{\partial u_i} \mathbf{U}_*^{-1} \frac{\partial \mathbf{Y}_0}{\partial u_j} \right] \end{aligned}$$

Due to the i.i.d. assumption in both models, $\mathbf{Y}_0$ is independent of either $\frac{\partial \mathbf{Y}_0}{\partial u_i}$ or $\frac{\partial^2 \mathbf{Y}_0}{\partial u_i \partial u_j}$. Therefore

$$\begin{aligned} (\mathcal{I}_*)_{ij} &= \mathbb{E} \left[\frac{\partial^2 \Phi_*}{\partial u_i \partial u_j} \right] = \mathbb{E} \left[\text{tr} \left(\mathbf{V}_*^{-1} \frac{\partial \mathbf{Y}_0^T}{\partial u_i} \mathbf{U}_*^{-1} \frac{\partial \mathbf{Y}_0}{\partial u_j} \right) \right] \\ &= \mathbb{E} \left[\text{vec} \left(\frac{\partial \mathbf{Y}_0}{\partial u_i} \right)^T (\mathbf{V}_*^{-1} \otimes \mathbf{U}_*^{-1}) \text{vec} \left(\frac{\partial \mathbf{Y}_0}{\partial u_j} \right) \right] \quad (\text{A3}) \end{aligned}$$

For any $\mathbf{w} = (w_1, \dots, w_p) \in \mathbb{R}^p$ and $\mathbf{w} \neq \mathbf{0}$, denote $\tilde{\mathbf{w}} := \sum_{i=1}^p w_i \text{vec} \left(\frac{\partial \mathbf{Y}_0}{\partial u_i} \right)$. To prove $\mathcal{I}_{st}(u) \geq \mathcal{I}_s(u)$, it suffices to show $\tilde{\mathbf{w}}^T (\mathbf{V}_{st} \otimes \mathbf{U}_{st})^{-1} \tilde{\mathbf{w}} \geq \tilde{\mathbf{w}}^T (\mathbf{V}_s \otimes \mathbf{U}_s)^{-1} \tilde{\mathbf{w}}$.

By [Theorem 4.2.12 in Horn and Johnson (1991)], we know that any eigenvalue of $\mathbf{V}_* \otimes \mathbf{U}_*$ has the format as a product of eigenvalues of $\mathbf{V}_*$ and $\mathbf{U}_*$ respectively, i.e. $\lambda_k(\mathbf{V}_* \otimes \mathbf{U}_*) = \lambda_i(\mathbf{V}_*)\lambda_j(\mathbf{U}_*)$, where where $\{\lambda_j(M)\}$ are the ordered eigenvalues of M , i.e. $\lambda_1(M) \geq \dots \geq \lambda_d(M)$. By the given condition we have

$$\begin{aligned} \lambda_{IJ}((\mathbf{V}_{st} \otimes \mathbf{U}_{st})^{-1}) &= \lambda_1^{-1}(\mathbf{V}_{st} \otimes \mathbf{U}_{st}) \\ &= \lambda_1^{-1}(\mathbf{C}_t)\lambda_1^{-1}(\mathbf{C}_x) \geq \sigma_\epsilon^{-2} = \lambda_1((\mathbf{V}_s \otimes \mathbf{U}_s)^{-1}) \quad (\text{A4}) \end{aligned}$$

Thus it completes the proof of the first inequality.

Similarly by the second condition, we have

$$\begin{aligned} \lambda_{IJ}((\mathbf{V}_{st} \otimes \mathbf{U}_{st})^{-1}) &= \lambda_1^{-1}(\mathbf{C}_t)\lambda_1^{-1}(\mathbf{C}_x) \\ &\geq J^{-1}\lambda_{\min}^{-1}(\Gamma_{obs}) = \lambda_1(\mathbf{V}_s^- \otimes \mathbf{U}_s^{-1}) \quad (\text{A5}) \end{aligned}$$

and complete the proof of the second inequality. $\square$

Theorem (3.2) *If we choose $\mathbf{C}_x = \Gamma_{obs}$ and require the maximal eigenvalue of $\mathbf{C}_t$, $\lambda_{\max}(\mathbf{C}_t) \leq J$, then the following inequality holds regarding the Fisher information matrices, $\mathcal{I}_t$ and $\mathcal{I}_{st}$, of the time-averaged model and the STGP model respectively:*

$$\mathcal{I}_{st}(u) \geq \mathcal{I}_t(u) \quad (\text{A6})$$

Proof Denote $\mathbf{Y}_0 = \mathbf{Y} - \mathbf{M}$. We have $\Phi_*(u) = \frac{1}{2} \text{tr} [\mathbf{V}_*^{-1} \mathbf{Y}_0^T \mathbf{U}_*^{-1} \mathbf{Y}_0]$ with $*$ being T or ST. $\mathbf{U}_t, \mathbf{V}_t, \mathbf{U}_{st}$ and $\mathbf{V}_{st}$ are specified in (16).

By the similar argument of the proof in Theorem 3.1, we have

$$(\mathcal{I}_*)_{ij} = \mathbb{E} \left[\frac{\partial^2 \Phi_*}{\partial u_i \partial u_j} \right] = \text{tr} \left[\mathbf{V}_*^{-1} \mathbb{E} \left(\frac{\partial \mathbf{Y}_0^\top}{\partial u_i} \mathbf{U}_*^{-1} \frac{\partial \mathbf{Y}_0}{\partial u_j} \right) \right] \quad (\text{A7})$$

For any $\mathbf{w} = (w_1, \dots, w_p) \in \mathbb{R}^p$ and $\mathbf{w} \neq \mathbf{0}$, denote $\mathbf{W} := \sum_{i,j=1}^p w_i \mathbb{E} \left(\frac{\partial \mathbf{Y}_0^\top}{\partial u_i} \mathbf{U}_*^{-1} \frac{\partial \mathbf{Y}_0}{\partial u_j} \right) w_j$. We know $\mathbf{W} \geq \mathbf{0}_{J \times J}$. It suffices to show $\text{tr}[\mathbf{V}_{\text{ST}}^{-1} \mathbf{W}] \geq \text{tr}[\mathbf{V}_T^{-1} \mathbf{W}]$.

By the corollary (Marshall et al. 2011) of Von Neumann's trace inequality (Mirsky 1975), we have

$$\begin{aligned} \sum_{j=1}^J \lambda_j(\mathbf{V}_*^{-1}) \lambda_{J-j+1}(\mathbf{W}) &\leq \text{tr}(\mathbf{V}_*^{-1} \mathbf{W}) \\ &\leq \sum_{j=1}^J \lambda_j(\mathbf{V}_*^{-1}) \lambda_j(\mathbf{W}) \end{aligned} \quad (\text{A8})$$

where $\{\lambda_j(M)\}$ are the ordered eigenvalues of M , i.e. $\lambda_1(M) \geq \dots \geq \lambda_d(M)$. The only non-zero eigenvalue of $\mathbf{V}_T^- = J^{-2}(\mathbf{1}_J \mathbf{1}_J^\top)$ is $\lambda_1(\mathbf{V}_T^-) = J^{-1}$. Therefore, we have

$$\begin{aligned} \text{tr}[\mathbf{V}_T^- \mathbf{W}] &\leq J^{-1} \lambda_1(\mathbf{W}) \leq \lambda_J(\mathbf{V}_{\text{ST}}^{-1}) \lambda_1(\mathbf{W}) + \\ &\sum_{j=1}^{J-1} \lambda_j(\mathbf{V}_{\text{ST}}^{-1}) \lambda_{J-j+1}(\mathbf{W}) \leq \text{tr}[\mathbf{V}_{\text{ST}}^{-1} \mathbf{W}] \end{aligned} \quad (\text{A9})$$

where $\lambda_J(\mathbf{V}_{\text{ST}}^{-1}) = \lambda_1^{-1}(\mathbf{C}_t) \geq J^{-1}$ and $\lambda_j(\mathbf{V}_{\text{ST}}^{-1}), \lambda_j(\mathbf{W}) \geq 0$. $\square$

Appendix B More numerical results

See Figs. 18, 19, 20, 21, 22 and 23.

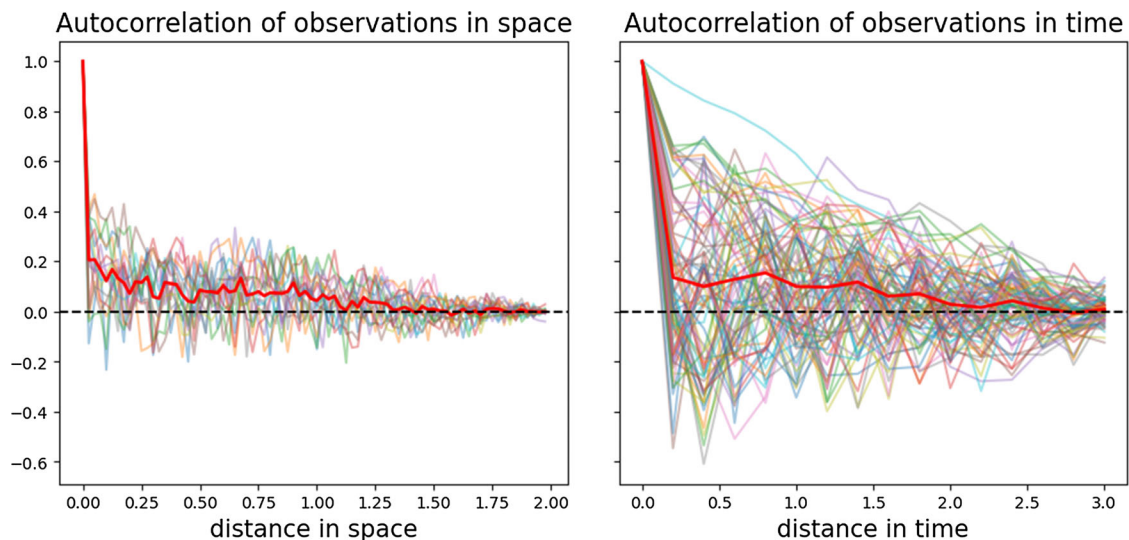


Fig. 18 Advection–diffusion inverse problem: auto-correlations of observations in space (left) and time (right) respectively

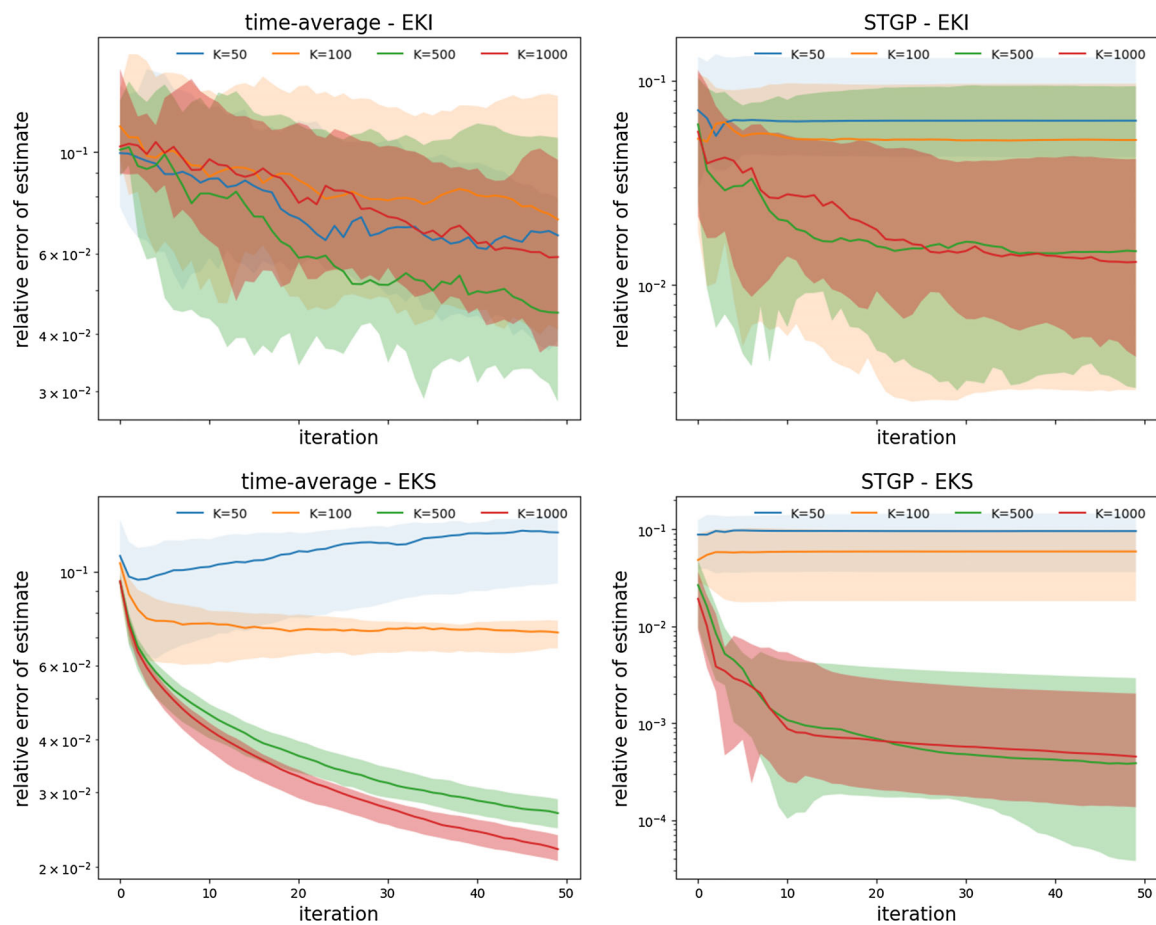


Fig. 19 Lorenz inverse problem: comparing posterior estimates of parameter u for two models (time-average and STGP) in terms of rela-

tive error of median REM = $\frac{\|\hat{u} - u^\dagger\|}{\|u^\dagger\|}$. Each experiment is repeated for 10 runs of EnK (EKI and EKS respectively) and shaded regions indicate 5–95% quantiles of such repeated results

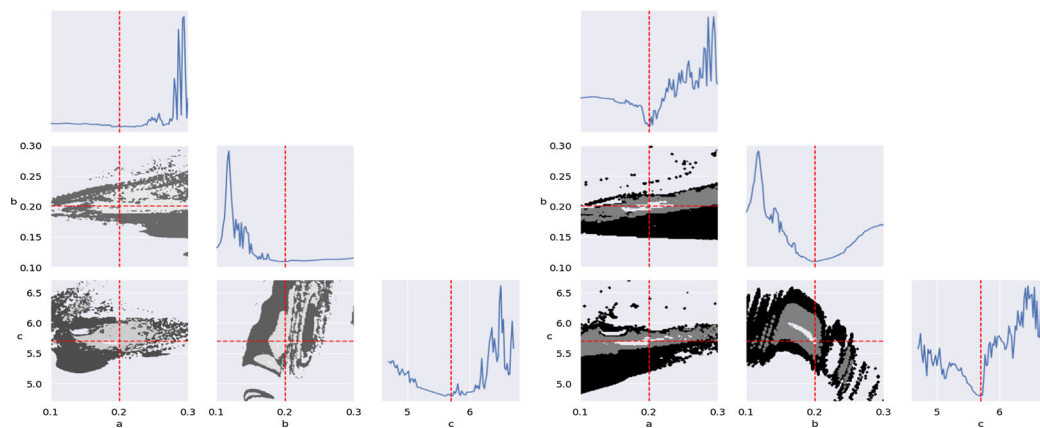


Fig. 20 Rössler inverse problem: marginal (diagonal) and pairwise (lower triangle) sections of the joint density $p(u)$ by the time-averaged model (left) and the STGP model (right) respectively. Red dashed lines indicate the true parameter values

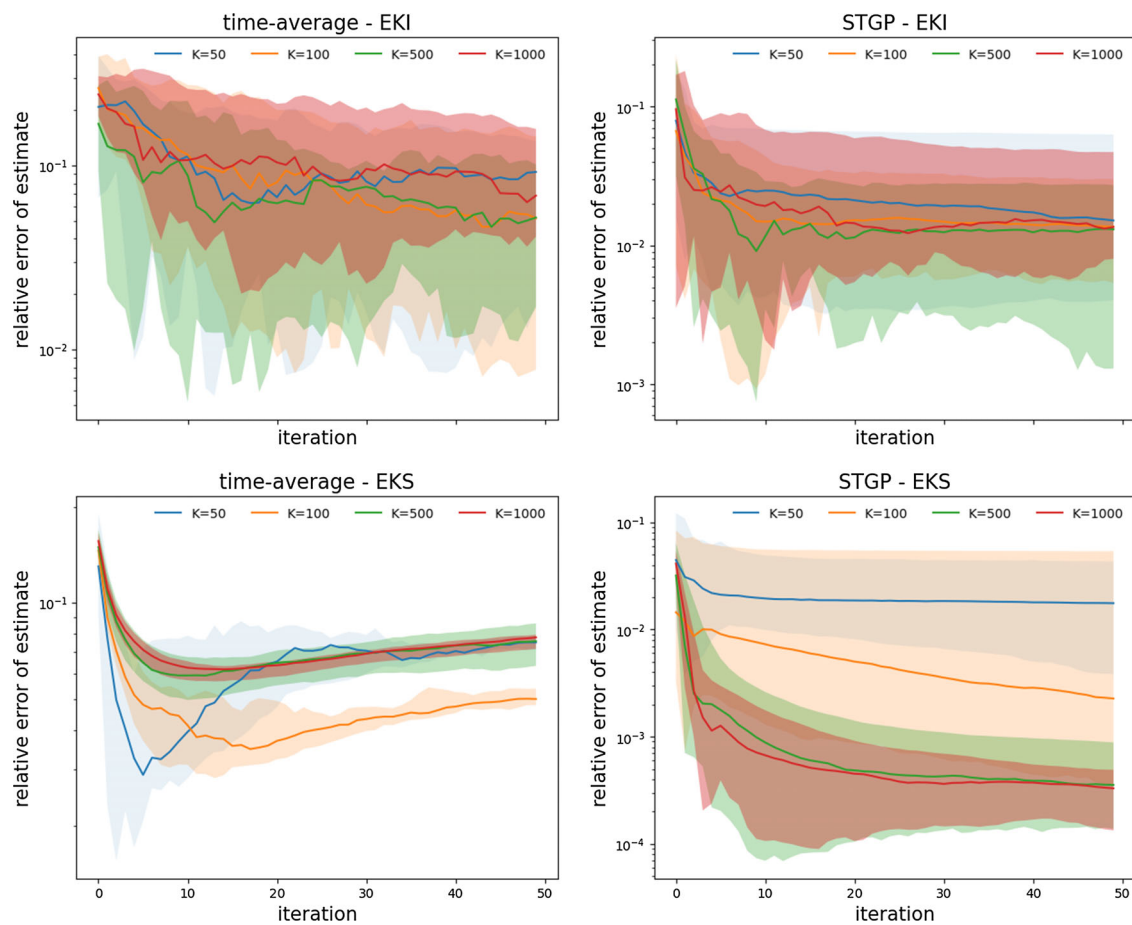


Fig. 21 Rössler inverse problem: comparing posterior estimates of parameter u for two models (time-average and STGP) in terms of rel-

ative error of median REM = $\frac{\|\hat{u} - u^*\|}{\|u^*\|}$. Each experiment is repeated for 10 runs of EnK (EKI and EKS respectively) and shaded regions indicate 5–95% quantiles of such repeated results

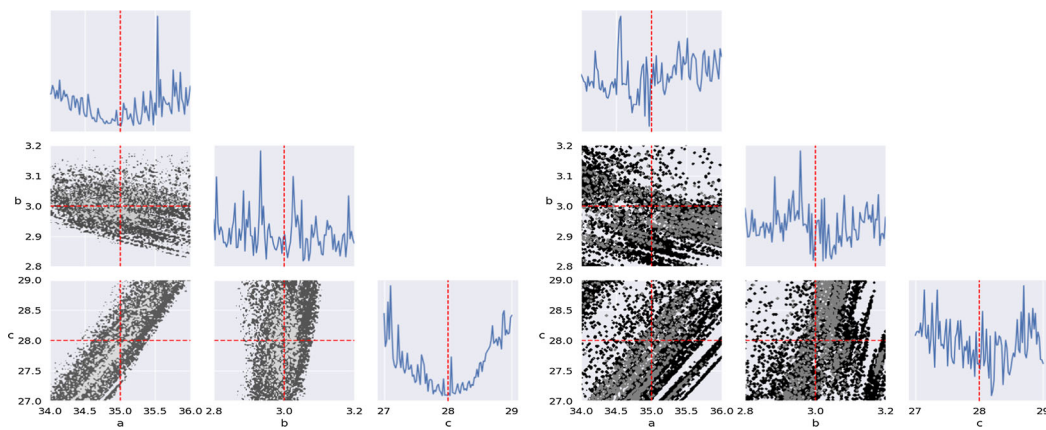


Fig. 22 Chen inverse problem: marginal (diagonal) and pairwise (lower triangle) sections of the joint density $p(u)$ by the time-averaged model (left) and the STGP model (right) respectively. Red dashed lines indicate the true parameter values

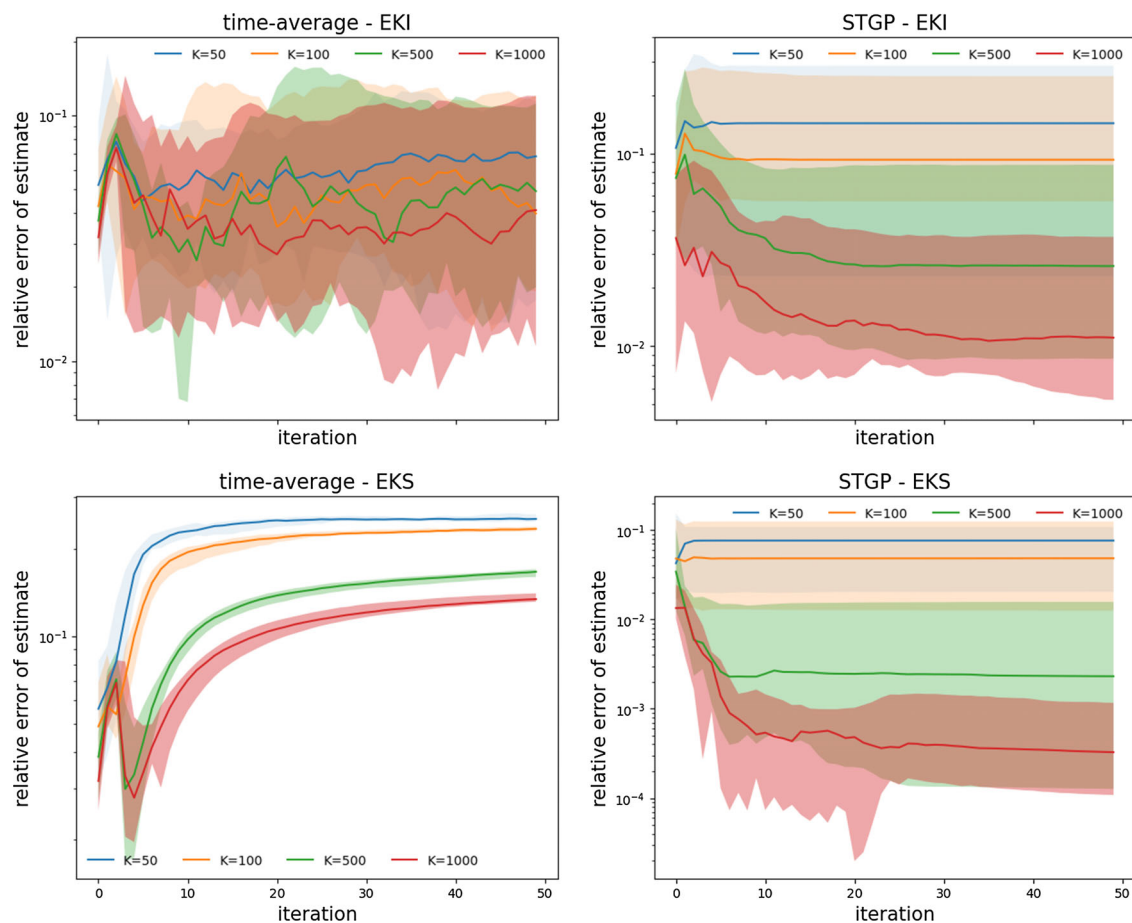


Fig. 23 Chen inverse problem: comparing posterior estimates of parameter u for two models (time-average and STGP) in terms of relative error

of median $REM = \frac{\|\hat{u} - u^{\dagger}\|}{\|u^{\dagger}\|}$. Each experiment is repeated for 10 runs of EnK (EKI and EKS respectively) and shaded regions indicate 5–95% quantiles of such repeated results

References

- Abarbanel, H.: Predicting the Future: Completing Models of Observed Complex Systems, vol. 1. Springer, New York (2013). <https://doi.org/10.1007/978-1-4614-7218-6>
- Agarwal, D., Chen, B.-C.: Regression-based latent factor models. ACM Press, New York (2009). <https://doi.org/10.1145/1557019.1557029>
- Agiza, H., Yassen, M.: Synchronization of Rossler and Chen chaotic dynamical systems using active control. *Phys. Lett. A* **278**(4), 191–197 (2001)
- Baukal, C.E., Jr., Gershtein, V., Li, X.J. (eds.): Computational Fluid Dynamics in Industrial Combustion. CRC Press, Boca Raton (2000). <https://doi.org/10.1201/2F9781482274363>
- Beskos, A.: A stable manifold MCMC method for high dimensions. *Stat. Probab. Lett.* **90**, 46–52 (2014)
- Beskos, A., Roberts, G., Stuart, A., Voss, J.: MCMC methods for diffusion bridges. *Stoch. Dyn.* **8**(03), 319–350 (2008)
- Beskos, A., Pinski, F.J., Sanz-Serna, J.M., Stuart, A.M.: Hybrid Monte-Carlo on Hilbert spaces. *Stoch. Process. Appl.* **121**, 2201–2230 (2011)
- Beskos, A., Girolami, M., Lan, S., Farrell, P.E., Stuart, A.M.: Geometric MCMC for infinite-dimensional inverse problems. *J. Comput. Phys.* **335**, 327–351 (2017)
- Bishop, R.: Chaos. In: Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy*, 2017th edn. Metaphysics Research Lab, Stanford University, Stanford (2017)
- Brooks, C.: Chaos in foreign exchange markets: a sceptical view. *Comput. Econ.* **11**(3), 265–281 (1998). <https://doi.org/10.1023/a:1008650024944>
- Chada, N.K., Stuart, A.M., Tong, X.T.: Tikhonov regularization within ensemble Kalman inversion. *SIAM J. Numer. Anal.* **58**(2), 1263–1294 (2020). <https://doi.org/10.1137/19m1242331>
- Chen, G., Ueta, T.: Yet another chaotic attractor. *Int. J. Bifurc. Chaos* **9**(07), 1465–1466 (1999)
- Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., Stuart, A.M.: Calibrate, emulate, sample. *J. Comput. Phys.* **424**, 109716 (2021). <https://doi.org/10.1016/j.jcp.2020.109716>
- Conjard, M., Omre, H.: Spatio-temporal inversion using the selection Kalman model. *Front. Appl. Math. Stat.* (2021). <https://doi.org/10.3389/fams.2021.636524>
- Cotter, S.L., Roberts, G.O., Stuart, A., White, D.: MCMC methods for functions: modifying old algorithms to make them faster. *Stat. Sci.* **28**(3), 424–446 (2013)
- Cressie, N., Wikle, C.: *Statistics for Spatio-Temporal Data* CourseSmart Series. Wiley (2011). <https://books.google.com/books?id=kOC6D0DiNYC>

- Dashti, M., Stuart, A.M.: The Bayesian Approach to Inverse Problems, pp. 311–428. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-12385-1_7
- de Wiljes, J., Reich, S., Stannat, W.: Long-time stability and accuracy of the ensemble Kalman–Bucy filter for fully observed processes and small measurement noise. *SIAM J. Appl. Dyn. Syst.* **17**(2), 1152–1181 (2018). <https://doi.org/10.1137/17M1119056>
- Echeverría Ciaurri, D., Mukerji, T.: A robust scheme for spatio-temporal inverse modeling of oil reservoirs. Cairns, Australia (2009)
- Effah-Poku, S., Obeng-Denteh, W., Dontwi, I.: A study of chaos in dynamical systems. *J. Math.* **2018**, 1–5 (2018)
- Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**(C5), 10143 (1994). <https://doi.org/10.1029/94jc00572>
- Evensen, G., van Leeuwen, P.J.: Assimilation of Geosat altimeter data for the Agulhas current using the ensemble Kalman filter with a quasigeostrophic model. *Mon. Weather Rev.* **124**(1), 85–96 (1996). [https://doi.org/10.1175/1520-0493\(1996\)124<0085:aogadf>2.0.co;2](https://doi.org/10.1175/1520-0493(1996)124<0085:aogadf>2.0.co;2)
- Fisher, R.A., Russell, E.J.: On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Character* **222**(594–604), 309–368 (1922). <https://doi.org/10.1098/rsta.1922.0009>
- Garbuno-Inigo, A., Hoffmann, F., Li, W., Stuart, A.M.: Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler. *SIAM J. Appl. Dyn. Syst.* **19**(1), 412–441 (2020a). <https://doi.org/10.1137/19M1251655>
- Garbuno-Inigo, A., Nüsken, N., Reich, S.: Affine invariant interacting Langevin dynamics for Bayesian inference. *SIAM J. Appl. Dyn. Syst.* **19**(3), 1633–1658 (2020b). <https://doi.org/10.1137/19m1304891>
- Gneiting, T.: Nonseparable, stationary covariance functions for space-time data. *J. Am. Stat. Assoc.* **97**(458), 590–600 (2002). <https://doi.org/10.1198/016214502760047113>
- Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
- Gupta, A., Nagar, D.: Matrix Variate Distributions. Chapter 2: Matrix Variate Normal Distribution. Chapman and Hall/CRC, New York (2018). <https://doi.org/10.1201/9780203749289>
- Hegazi, A., Agiza, H., El-Dessoky, M.: Controlling chaotic behaviour for spin generator and Rossler dynamical systems with feedback control. *Chaos Solitons Fractals* **12**(4), 631–658 (2001)
- Higdon, D., Kennedy, M., Cavendish, J.C., Cafo, J.A., Ryne, R.D.: Combining field data and computer simulations for calibration and prediction. *SIAM J. Sci. Comput.* **26**(2), 448–466 (2004). <https://doi.org/10.1137/S1064827503426693>
- Horn, R.A., Johnson, C.R.: Topics in Matrix Analysis. Cambridge University Press, Cambridge (1991). <https://doi.org/10.1017/CBO9780511840371>
- Huang, D.Z., Huang, J., Reich, S., Stuart, A.M.: Efficient derivative-free Bayesian inference for large-scale inverse problems. *arXiv:2204.04386* (2022). <https://doi.org/10.48550/ARXIV.2204.04386>
- Iglesias, M.A.: A regularizing iterative ensemble Kalman method for PDE-constrained inverse problems. *Inverse Prob.* **32**(2), 025002 (2016). <https://doi.org/10.1088/0266-5611/32/2/025002>
- Iglesias, M.A., Law, K.J.H., Stuart, A.M.: Ensemble Kalman methods for inverse problems. *Inverse Prob.* **29**(4), 045001 (2013). <https://doi.org/10.1088/0266-5611/29/4/045001>
- Ivancevic, V.G., Ivancevic, T.T.: Complex Nonlinearity. Springer, Berlin (2008). <https://doi.org/10.1007/978-3-540-79357-1>
- Jenatton, R., Roux, N., Bordes, A., Obozinski, G.R., Pereira, F., Burges, C., Bottou, L., Weinberger, K.: A latent factor model for highly multi-relational data. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., Red Hook (2012)
- Kennedy, M.C., O’Hagan, A.: Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B Methodol.* **63**, 425–464 (2001)
- Krizhevsky, A., Sutskever, I., Hinton, G.E., Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105. Curran Associates, Inc., Red Hook (2012)
- Lan, S.: Adaptive dimension reduction to accelerate infinite-dimensional geometric Markov chain Monte Carlo. *J. Comput. Phys.* **392**, 71–95 (2019). <https://doi.org/10.1016/j.jcp.2019.04.043>
- Lan, S.: Learning temporal evolution of spatial dependence with generalized spatiotemporal Gaussian process models. *J. Mach. Learn. Res.* **23**(259), 1–53 (2022)
- Lan, S., Li, S., Shahbaba, B.: Scaling up bayesian uncertainty quantification for inverse problems using deep neural networks. *SIAM/ASA J. Uncertain. Quantif.* (to appear) (2022). [arXiv:2101.03906](https://arxiv.org/abs/2101.03906)
- Liz, E., Ruiz-Herrera, A.: Chaos in discrete structured population models. *SIAM J. Appl. Dyn. Syst.* **11**(4), 1200–1214 (2012). <https://doi.org/10.1137/120868980>
- Long, C.J., et al.: State-space solutions to the dynamic magnetoencephalography inverse problem using high performance computing. *Ann. Appl. Stat.* (2011). <https://doi.org/10.1214/11-aos483>
- Lorenz, E.N.: Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**(2), 130–141 (1963). [https://doi.org/10.1175/1520-0469\(1963\)020<0130:dnf>2.0.co;2](https://doi.org/10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2)
- Marshall, A.W., Olkin, I., Arnold, B.C.: Inequalities: Theory of Majorization and Its Applications, 2nd edn. Springer, New York (2011). <https://doi.org/10.1007/978-0-387-68276-1>
- Mirsky, L.: A trace inequality of john von Neumann. *Monatshefte Math.* **79**, 303–306 (1975)
- Morzfeld, M., Adams, J., Lunderman, S., Orozco, R.: Feature-based data assimilation in geophysics. *Nonlinear Process. Geophys.* **25**(2), 355–374 (2018). <https://doi.org/10.5194/npg-25-355-2018>
- Negrini, E., Citti, G., Capogna, L.: System identification through Lipschitz regularized deep neural networks. *J. Comput. Phys.* **444**, 110549 (2021). <https://doi.org/10.1016/j.jcp.2021.110549>
- O’Hagan, A.: Bayesian analysis of computer code outputs: a tutorial. *Reliab. Eng. Syst. Saf.* **91**(10), 1290–1300 (2006). <https://doi.org/10.1016/j.res.2005.11.025>
- Oakley, J., O’Hagan, A.: Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika* **89**(4), 769–784 (2002). <https://doi.org/10.1093/biomet/89.4.769>
- Oakley, J.E., O’Hagan, A.: Probabilistic sensitivity analysis of complex models: a Bayesian approach. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **66**(3), 751–769 (2004)
- Ojeda, A., Klug, M., Kreutz-Delgado, K., Gramann, K., Mishra, J.: A Bayesian framework for unifying data cleaning, source separation and imaging of electroencephalographic signals. *bioRxiv* (2019)
- Ott, E.: Strange attractors and chaotic motions of dynamical systems. *Rev. Mod. Phys.* **53**(4), 655 (1981)
- Paciorek, C., Schervish, M., Thrun, S., Saul, L., Schölkopf, B.: Non-stationary covariance functions for Gaussian process regression. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) *Advances in Neural Information Processing Systems*, vol. 16. MIT Press, Cambridge (2003)
- Pasha, M., Saibaba, A.K., Gazzola, S., Espanol, M.I., de Sturler, E.: Efficient edge-preserving methods for dynamic inverse problems (2021). [arXiv:2107.05727](https://arxiv.org/abs/2107.05727). <https://doi.org/10.48550/ARXIV.2107.05727>
- Petra, N., Stadler, G.: Model variational inverse problems governed by partial differential equations. Technical Report, The Institute for

- Computational Engineering and Sciences, The University of Texas at Austin (2011)
- Rössler, O.: An equation for continuous chaos. *Phys. Lett. A* **57**(5), 397–398 (1976). [https://doi.org/10.1016/0375-9601\(76\)90101-8](https://doi.org/10.1016/0375-9601(76)90101-8)
- Rossler, O.: An equation for hyperchaos. *Phys. Lett. A* **71**(2), 155–157 (1979). [https://doi.org/10.1016/0375-9601\(79\)90150-6](https://doi.org/10.1016/0375-9601(79)90150-6)
- Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P.: Design and analysis of computer experiments. *Stat. Sci.* **4**(4), 409–423 (1989). <https://doi.org/10.1214/ss/1177012413>
- Schillings, C., Stuart, A.: Analysis of the ensemble Kalman filter for inverse problems. *SIAM J. Numer. Anal.* **55**(3), 1264–1290 (2017a). <https://doi.org/10.1137/16M105959X>
- Schillings, C., Stuart, A.M.: Convergence analysis of ensemble Kalman inversion: the linear, noisy case. *Appl. Anal.* **97**(1), 107–123 (2017b). <https://doi.org/10.1080/00036811.2017.1386784>
- Schneider, T., Lan, S., Stuart, A., Teixeira, J.: Earth system modeling 2.0: a blueprint for models that learn from observations and targeted high-resolution simulations. *Geophys. Res. Lett.* **44**(24), 12396–12417 (2017). <https://doi.org/10.1002/2017GL076101>
- Shcherbakova, A.I., Kupriyanova, Y.A., Zhikhareva, G.V.: Spatio-temporal analysis the results of solving the inverse problem of electrocardiography. *J. Phys. Conf. Ser.* **2091**(1), 012028 (2021). <https://doi.org/10.1088/1742-6596/2091/1/012028>
- Siregar, P., Sinteff, J.-P.: Introducing spatio-temporal reasoning into the inverse problem in electroencephalography. *Artif. Intell. Med.* **8**(2), 97–122 (1996). [https://doi.org/10.1016/0933-3657\(95\)00028-3](https://doi.org/10.1016/0933-3657(95)00028-3)
- Stuart, A.M.: Inverse problems: a Bayesian perspective. *Acta Numer.* **19**, 451–559 (2010)
- Villa, U., Petra, N., Ghattas, O.: hIPPYlib: an extensible software framework for large-scale inverse problems governed by PDEs; part I—deterministic inversion and linearized Bayesian inference. *ACM Trans. Math. Softw.* **47**(2), 1–34 (2021). <https://doi.org/10.1145/3428447>
- Wang, K., Hamelijnck, O., Damoulas, T., Steel, M. III, H.D., Singh, A.: Non-separable non-stationary random fields. In: III, H.D. & Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, pp. 9887–9897. PMLR (2020). <http://proceedings.mlr.press/v119/wang20g.html>
- Webb, G.I., et al.: *Latent Factor Models and Matrix Factorizations*, pp. 571–571. Springer, Berlin (2011). https://doi.org/10.1007/978-0-387-30164-8_887
- Woolrich, M., Jenkinson, M., Brady, J., Smith, S.: Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE Trans. Med. Imaging* **23**(2), 213–231 (2004). <https://doi.org/10.1109/tmi.2003.823065>
- Yang, Y.: Source-space analyses in meg/eeG and applications to explore spatio-temporal neural dynamics in human vision. Ph.D. thesis (CMU) (2017). https://kithub.cmu.edu/articles/Source-Space_Analyses_in_MEG_EEG_and_Applications_to_Explore_Spatio-temporal_Neural_Dynamics_in_Human_Vision/6723065/1. <https://doi.org/10.1184/R1/6723065.V1>
- Yang, S.-K., Chen, C.-L., Yau, H.-T.: Control of chaos in Lorenz system. *Chaos Solitons Fractals* **13**(4), 767–780 (2002)
- Yao, B., Yang, H.: Physics-driven spatiotemporal regularization for high-dimensional predictive modeling: a novel approach to solve the inverse ECG problem. *Sci. Rep.* (2016). <https://doi.org/10.1038/srep39012>
- Yassen, M.: Chaos control of Chen chaotic dynamical system. *Chaos Solitons Fractals* **15**(2), 271–283 (2003)
- Zammit-Mangion, A., Wikle, C.K.: Deep integro-difference equation models for spatio-temporal forecasting. *Spat. Stat.* **37**, 100408 (2020). <https://doi.org/10.1016/j.spasta.2020.100408>
- Zhang, B., Cressie, N.: Bayesian inference of spatio-temporal changes of arctic sea ice. *Bayesian Anal.* **15**(2), 605–631 (2020). <https://doi.org/10.1214/20-ba1209>
- Zhang, Y., Ghodrati, A., Brooks, D.H.: An analytical comparison of three spatio-temporal regularization methods for dynamic linear inverse problems in a common statistical framework. *Inverse Prob.* **21**(1), 357–382 (2005). <https://doi.org/10.1088/0266-5611/21/1/022>

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.