Bayesian optimization for parameter estimation of a local particle filter

Shoichi AKAMI[1]

Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan
Meteorological Research Institute, Japan Meteorological Agency, Tsukuba, Japan

and

Keiichi KONDO

Meteorological Research Institute, Japan Meteorological Agency, Tsukuba, Japan

and

Hiroshi L. TANAKA

Organization of Volcanic Disaster Mitigation, Shinjuku, Japan

and

Mizuo KAJINO

Meteorological Research Institute, Japan Meteorological Agency, Tsukuba, Japan
Faculty of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan

August 9th, 2025

---

1) Corresponding author: Shoichi AKAMI, Graduate School of Life and Environmental Sciences, University of Tsukuba, 1-1, Tennoudai, Tsukuba, Ibaraki, Japan, 305-8572.
Email: akami@mri-jma.go.jp
Tel: +81-29-853-8695

Abstract

Particle filter (PF) is a powerful data assimilation method that does not assume the linearity in the time evolution of errors or Gaussian error distributions. However, the number of particles required increases exponentially with the dimensions of the dynamical system, which is a bottleneck when applying the PF to numerical weather prediction (NWP) models. Local particle filter (LPF) realizes the PF in high-dimensional systems by the localization, but it has high parameter sensitivity and is challenging to operate stably. On the other hand, when using a nonlinear observation operator, it is possible to estimate the analysis with higher accuracy than the local ensemble transformation Kalman filter (LETKF) by setting the weight inflation factor, which smooths the weights among particles, and the localization scale, to the optima. Therefore, an efficient parameter estimation method is required.

Bayesian optimization (BO) is a method for efficiently solving optimization problems of black box functions with high computational costs, and is used for parameter optimization of neural networks. Therefore, we estimated the weight inflation factor and localization scale that minimize the root mean square error between the observations and the forecasts (RMSE(o vs. f)) in the LPF using the BO in the Lorenz-96 40-variable model (L96). As a result, the BO was able to model the response surface with high accuracy and estimate the weight inflation factor and localization scale with accuracy equal to or better than random sampling (RS). In addition, this result was robust to changes in the observation set. However, as the number of parameters to be estimated increased, the BO did not always obtain

estimations close to the optima, depending on the observation set.

This study has clarified that the BO contributes to improving the practicality of the LPF, and it has also provided suggestions on how the BO should be developed in the future. Since the LPF can estimate high-precision analysis even in strongly nonlinear phenomena, the development of the technology in this study is expected to improve the accuracy of heavy rainfall prediction in the future. The BO will be helpful in atmospheric model experiments for the practical application of the LPF.

Keywords: Local particle filter; Parameter estimation; Bayesian optimization; Gaussian process regression

1. Introduction

In high-dimensional chaotic dynamical systems, such as numerical weather prediction (NWP) models, even small errors in the initial conditions can develop over time and become large errors. Data assimilation is a technique for estimating the analysis closer to the truth from forecasts and observations, and forecast errors can be improved by using the high-precision analyses as the initial conditions. The ensemble Kalman filter (EnKF; Evensen, 1994) and 4D-Var (Dimet and Talagrand, 1986), which are currently the mainstream data assimilation methods, can estimate the optimum analysis when the errors develop linearly over time and the error distribution follows a Gaussian distribution. On the other hand, when these conditions are not satisfied—around cumulus convection and storm tracks—it may not be possible to estimate the optimum analyses (Kondo and Miyoshi, 2019).

Particle filter (PF; Gordon et al., 1993) does not assume linearity or Gaussianity, and therefore can be an appropriate data assimilation method for dynamical systems with strong nonlinearity. However, the PF obtains the analyses by resampling ensembles (particles) based on weights obtained from the likelihood of observations; therefore, "weight collapse" may occur in high-dimensional systems. The PF requires an exponential increase in the number of particles necessary for the dimensions of the dynamical system (Snyder et al., 2008), and this problem is a bottleneck when applying the PF to the NWP systems.

Local particle filter (LPF; Penny and Miyoshi, 2016; Potthast et al., 2019; Kotsuki et al., 2022) is a method for realizing the PF in high-dimensional systems by reducing the

dimensions of the dynamical system through localization. Spatial localization is justified by the fact that long-range correlations are weaker than spurious or nearby correlations (Hamill et al., 2001). In fact, the LPF can estimate the more accurate analyses than the EnKF in cases of non-Gaussian observation errors, nonlinear observation operators, and sparse observation networks (Poterjoy and Anderson, 2016; Poterjoy, 2016; Penny and Miyoshi, 2016). This advantage of the LPF is extremely important for improving the prediction accuracy of heavy rainfall. This is because, for example, radar reflectivity and precipitation intensity have a nonlinear relationship, and when assimilating radar reflectivity, it is necessary to introduce a nonlinear observation operator.

However, in the LPF, it is necessary to set the localization scale and the inflation factor that smooths the weights among particles to the optima. In addition, excessive resampling causes "weight collapse," so adjusting the resampling frequency based on the number of effective particles is also an effective approach. Another method for implementing the PF is to approximate the prior distribution by a combination of Gaussian kernels centered on the values of each particle. In this method, the amplitude of the Gaussian kernel is an important parameter (Stordal et al., 2011). In the LPF, the filter diverges unless these various parameters are optimized (Kotsuki et al., 2022).

With the improvement of methods and the advancement of models applied to the practical implementation of the LPF, the number of parameters to be optimized is expected to increase, and the computational cost of numerical experiments is anticipated to become

substantial. The simplest method for searching for the optima of these parameters is grid search (GS), which divides the parameter space into equal parts and comprehensively searches all combinations. The GS always obtains the optima within the set range, but it requires experiments that increase exponentially with the number of parameters. Random sampling (RS) is a method that randomly selects combinations from the parameter space. It works more efficiently than the GS in high-dimensional spaces, but if the number of samples is insufficient, it may not obtain the optima. Therefore, a more efficient optimization method is required.

One method for reducing computational costs is to replace the system response to parameters with a surrogate model (e.g., Sawada, 2020). Bayesian optimization (BO; Mockus, 1989) is a method for estimating parameters that minimize an objective function and is used for parameter optimization of neural networks (Snoek et al., 2012). In this method, Gaussian process regression (GPR) is employed to emulate the objective function. Consequently, even if the shape of the response surface of the input-output data is unknown or the function is a multi-peaked function that cannot be differentiated, it is still possible to efficiently search for a global optimal solution. In addition, the BO and other systems are independent and easy to implement.

The effectiveness of using the BO for parameter estimation within the EnKF framework has already been demonstrated (Lunderman et al., 2021). Therefore, this study builds on this research line to investigate whether parameter estimation using the BO can improve the

practicality of the LPF. In addition, since the BO has been used as a tool in previous studies and there is a lack of knowledge that contributes to an essential understanding, we conducted a survey focusing on the GPR prediction distributions and Lipschitz constants to inform future technological developments. This survey was conducted using a data assimilation experiment using the Lorenz-96 40-variable model (L96: Lorenz and Emanuel, 1998).

This paper is organized as follows. Section 2 introduces the methodology, Section 3 describes the experimental setup, Section 4 compares the estimation accuracy of the RS and the BO, and investigates the estimation results of the BO in detail from the perspectives of the GPR prediction distribution and Lipschitz constant. Section 5 presents prospects and conclusions.

2. Method

a. Local particle filter

The PF estimates the posterior distribution using the Monte Carlo method and Bayes' theorem:

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) = \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t)\, p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})}{p(\boldsymbol{y}_t|\boldsymbol{y}_{1:t-1})}, \tag{1}$$

where p represents the probability distribution; $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ denotes the posterior distribution of state variable $\boldsymbol{x}$ at time t $(t = 1, \dots, T)$ given all observations $\boldsymbol{y}$ up to time t; $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ is the likelihood of $\boldsymbol{x}$ given $\boldsymbol{y}$; $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1})$ is the prior distribution given all $\boldsymbol{y}$ up to one time before analysis time; and $p(\boldsymbol{y}_t|\boldsymbol{y}_{1:y-1})$ denotes the marginal likelihood of $\boldsymbol{y}$, which can be expressed as a constant computed by climate data in the NWP. The prior distribution can be approximated using particles (or ensemble members) of the numerical forecast:

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1}) \approx \frac{1}{M}\sum_{m=1}^{M} \delta\big(x_t - \mathcal{F}(x_{t-1}^m)\big), \tag{2}$$

where the subscripts m $(m = 1, \dots, M)$ denote the indices of the particle, δ is the Dirac delta function, and $\mathcal{F}$ is the numerical model. In this study assumes a Gaussian likelihood function, given by

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t) = \frac{1}{\sqrt{(2\pi)^o|\boldsymbol{R}|}} \exp\left[-\frac{1}{2}\big(\boldsymbol{y}_t - h(\boldsymbol{x}_t)\big)^{\mathsf{T}}\boldsymbol{R}^{-1}\big(\boldsymbol{y}_t - h(\boldsymbol{x}_t)\big)\right]. \tag{3}$$

where o represents the dimension of $\boldsymbol{y}$. In addition, $\boldsymbol{R}$ denotes the observation error covariance matrix, and $|\boldsymbol{R}|$ is its determinant. h denotes the observation operator. The weight of each particle is the normalized likelihood, computed for all particles as follows:

$$w_t^m = \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t^m)}{\sum_{m'=1}^{M} p\left(\boldsymbol{y}_t|\boldsymbol{x}_t^{m'}\right)}, \tag{4}$$

where the subscripts m' denote the indices of the particles for summation. The posterior distribution is obtained by resampling each particle of the prior distribution in proportion to its weight:

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) \approx \sum_{m=1}^{M} w_t^m \delta\left(x_t - \mathcal{F}(x_{t-1}^m)\right). \tag{5}$$

The resampling method is also arbitrary. This study defined the analysis particles as the sum of the transformation for perturbations of forecast particles and the mean of the forecast particles:

$$\boldsymbol{X}^a = \overline{\boldsymbol{X}}^f + \delta\boldsymbol{X}^f\,\boldsymbol{T}\,, \tag{6}$$

where $\boldsymbol{X}^a$ denotes the analysis particles; $\overline{\boldsymbol{X}}^f$ represent the mean of forecast particles; and $\delta\boldsymbol{X}^f$ denotes the perturbation of forecast particles, where the row and column of $\boldsymbol{X}^a$, $\overline{\boldsymbol{X}}^f$, and $\delta\boldsymbol{X}^f$ indicate the particle size and dimension of the NWP model, respectively. $\boldsymbol{T}$ denotes the ensemble transform matrix, defined as a square matrix of order M. As resampling is performed using the ensemble transform matrix in the LPF, the matrix markedly affects filter performance (Farchi and Bocquet, 2018; Kotsuki et al., 2022). When the particle size is sufficiently large, the ratio of resampled particle sizes will closely match the ratio of weights; otherwise, the sampling error may become substantial.

In addition, the weights among grid points differ because varying observations are assimilated at each grid point through localization. As the pronounced weight difference may

cause spatial discontinuity, the ensemble transform matrix should satisfy a spatially smooth transition. Addressing the smoothing issue presents an interesting challenge. For example, Kotsuki et al. (2022) addressed this problem by sorting the particles and creating an ensemble transform matrix close to an identity matrix (see also Potthast et al., 2019). Our resampling method is based on Algorithm 1 of Kotsuki et al. (2022) and additionally uses stochastic universal resampling (SUR) instead of probabilistic resampling to reduce sampling error. The SUR is implemented as follows. Create a normalized cumulative probability distribution divided by the weight of each particle, and select a random starting point in the range $[0, 1/M]$. Set M pointers at equal intervals between the starting point and $1/M$, and sample the particles corresponding to the cumulative probabilities pointed to by each pointer.

Furthermore, we used localization to limit the impact of observations within the local domain to avoid "weight collapse" (Penny and Miyoshi, 2016; Kotsuki et al., 2022). This localization method is applied by independently computing the analysis at every grid point, similar to the local ensemble transform Kalman filter (LETKF; Hunt et al., 2007). Specifically, it is implemented by computing the product of the inverse of observation error covariance matrix $\boldsymbol{R}$ in Eq. (3) and the inverse of localization function $\boldsymbol{L}(r)$:

$$\exp\left[-\frac{1}{2}\left(\boldsymbol{y}_t - h(\boldsymbol{x}_t)\right)^{\top}\boldsymbol{R}^{-1}\{\boldsymbol{L}(r)\}^{-1}\left(\boldsymbol{y}_t - h(\boldsymbol{x}_t)\right)\right]. \tag{7}$$

Here, a Gaussian function with compact support was used as the localization function (Gaspari and Cohn, 1999):

$$L(r) = \begin{cases} \exp\left(-\frac{q^2}{2r^2}\right) \text{ if } q < 2\sqrt{\frac{10}{3}}\, r \\ 0 \text{ else} \end{cases}, \quad (8)$$

where q denotes the distance between the analysis grid point and the observation point and r represents the standard deviation of the Gaussian function, defining the localization scale. Observations beyond the radius of influence where the localization function is zero are not assimilated, while those within the localization scale are weighted based on the localization function. Therefore, r is the parameter that determines the localization scale, and It is necessary to set the appropriate value.

In addition, to avoid filter divergence, it is necessary to maintain particle diversity. Therefore, we smoothed the weights among particles to prevent a few particles from occupying most of the weights. We refer to this approach as weight inflation in this study:

$$w_t^m \leftarrow \tau w_t^m + \frac{1-\tau}{M}, (0 \le \tau \le 1), \quad (9)$$

where τ represents the inflation factor in the LPF. If τ is not 1, the weights w_t^m are smoothed, and all particles have equal weights when τ equals 0. On the other hand, if the original weights are used, the LPF tends to diverge due to "weight collapse." As τ becomes smaller, the LPF deviates from the theoretical PF but works stably. Thus, the relationship between mathematical rigor and stability is a trade-off on the inflation factor τ. Note that this approach is mathematically equivalent to Eq. (23) in Kotsuki et al. (2022). However, while Kotsuki et al. (2022) smoothed the weights in the time direction, we smoothed the weights among particles.

221

222 *b. Local ensemble transform Kalman filter*

223 The LETKF is a computationally efficient data assimilation method that combines the local
224 ensemble Kalman filter (LEKF; Ott et al., 2004) and the ensemble transform Kalman filter
225 (ETKF, Bishop et al., 2001). The analysis ensemble is obtained by the following equation:

226 $$\boldsymbol{X}^a = \overline{\boldsymbol{X}}^f + \delta\boldsymbol{X}^f \left[\widetilde{\boldsymbol{P}}^a \left(h(\delta\boldsymbol{X}^f) \right)^T \boldsymbol{R}^{-1} \left(\boldsymbol{y} - h(\overline{\boldsymbol{X}}^f) \right) + \sqrt{M-1} \left(\widetilde{\boldsymbol{P}}^a \right)^{\frac{1}{2}} \right]. \quad (10)$$

227 Here, $\widetilde{\boldsymbol{P}}^a$ is the analysis error covariance matrix in the ensemble space. It is given by the
228 following equation:

229 $$\widetilde{\boldsymbol{P}}^a = \left[\frac{(M-1)}{\alpha} \boldsymbol{I} + \left(h(\delta\boldsymbol{X}^f) \right)^T \boldsymbol{R}^{-1} \left(h(\delta\boldsymbol{X}^f) \right) \right]^{-1}. \quad (11)$$

230 Here, α represents the covariance inflation factor in the LETKF, $\boldsymbol{I}$ is the identity matrix. With
231 a finite ensemble size, the forecast error covariance matrix is generally underestimated,
232 which leads to filter divergence due to a decrease in ensemble spread. In multiplicative
233 inflation, the underestimation is prevented by inflating the forecast error covariance matrix
234 by a constant α.

235

236 *c. Bayesian optimization*

237 The BO estimates input data that minimizes the objective function by modeling response
238 surface using the GPR and evaluating using an acquisition function. The GPR assumes that
239 a joint distribution $p(\boldsymbol{g})$ of input data $\boldsymbol{z} = \{z_1, z_2, \dots, z_S\}$ and corresponding output data $\boldsymbol{g} =$
240 $\{g(z_1), g(z_2), \dots, g(z_S)\}$ follow the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K})$. Here, The

input data and output data subscripts s $(s = 1, \ldots, S)$ denote the indices of the data. This assumption is written as follows:

$$\boldsymbol{g} \sim \mathcal{GP}\left(\boldsymbol{\mu}(\boldsymbol{z}), \boldsymbol{K}(\boldsymbol{z}, \boldsymbol{z}')\right), \tag{12}$$

where $\boldsymbol{z}$ are input data that summarizes the inflation factor τ and the localization scale r into a single vector. The superscript $'$ denotes the another data within the data set. In addition, $\mathcal{GP}$ denotes the Gaussian process with the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{K}$ defined as a square matrix of order S. The elements of covariance matrix $\boldsymbol{K_{ij}}$ is defined as $k(z, z' \mid \boldsymbol{\theta})$. In this study, we used the Gaussian kernel with added white noise as a general choice. Note that the following equation is for the 2-dimensional BO:

$$k(z, z' \mid \boldsymbol{\theta}) = \theta_1 \exp\left(-\frac{(\tau - \tau')^2}{\theta_2} - \frac{(r - r')^2}{\theta_3}\right) + \theta_4 \delta(z, z'). \tag{13}$$

Here, the kernel function $k(z, z')$ defines the correlation between any two data z and z' in the input data $\boldsymbol{z}$. In addition, $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$ denotes the positive hyper-parameters that define the kernel function, while δ represents the Dirac delta function.

When the amplitude parameter θ_1 is small, the variation in the GPR prediction distribution is slight. The GPR prediction distribution becomes smoother when the length scale parameters θ_2 and θ_3 are large. When the noise parameter θ_4 is small, the uncertainties in the GPR prediction distribution near the input data are reduced. Note that when there are two types of input data, using two length scale parameters, θ_2 and θ_3, allows for more flexible modeling tailored to the characteristics of each input data.

In addition, since τ and r have different scales by a factor of 10, we normalized them to

the same scale. Since the Gaussian kernel performs distance-based calculations, the normalization prevents the influence of specific input data from becoming dominant. In our system, this approach markedly contributed to improving the performance of the BO.

When new input data $\boldsymbol{z}^*$ is given, the GPR is updated, and the new joint distribution of output data $\boldsymbol{g}^*$ is expressed as:

$$p(\boldsymbol{g}^*|\boldsymbol{z}^*, \mathcal{D}) = \mathcal{N}(\boldsymbol{k}_*^\top \boldsymbol{K}^{-1}\boldsymbol{g}, k_{**} - \boldsymbol{k}_*^\top \boldsymbol{K}^{-1}\boldsymbol{k}_*), \tag{14}$$

where $\mathcal{D} = (\boldsymbol{z}, \boldsymbol{g})$ denotes the accumulated input data, $\boldsymbol{k}_*$ is the similarity between the new input data $\boldsymbol{z}^*$ and the accumulated input data $\mathcal{D}$. k_{**} represents the similarity of the new input data z^* to themselves.

$$\boldsymbol{k}_* = \big(k(z^*, z_1), k(z^*, z_2), \dots, k(z^*, z_S)\big)^\top, \tag{15}$$

$$k_{**} = k(z^*, z^*). \tag{16}$$

Eq. (14), (17), and (18) are derived under the assumption that $\boldsymbol{\mu}(\boldsymbol{z})$ in Eq. (12) is zero, but in practice, mathematical rigor can be satisfied by subtracting the average from the input data.

In addition, when the covariance matrix $\boldsymbol{K}$ becomes close to a singular matrix due to redundant exploration of the same input data, it may become impossible to calculate the inverse matrix stably (Rasmussen and Nickisch, 2010). There are several techniques to improve numerical stability, but we followed Rasmussen and Williams (2006) and added jitter to the diagonal components of the covariance matrix. However, as far as we have experimented, this technique alone can prevent errors associated with singular matrices,

but cannot prevent the redundant exploration.

The hyper-parameters $\boldsymbol{\theta}$ are optimized by maximizing the negative log marginal likelihood, defined as following equation:

$$\log p(\boldsymbol{g} \mid \boldsymbol{z}, \boldsymbol{\theta}) = -\frac{S}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{K}_{\boldsymbol{\theta}}| - \frac{1}{2}\boldsymbol{g}^{\top}\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{g}, \tag{17}$$

where $\boldsymbol{K}_{\boldsymbol{\theta}}$ denotes the covariance matrix that depends on $\boldsymbol{\theta}$, with elements determined by the kernel function $k(z, z' \mid \boldsymbol{\theta})$, and $|\boldsymbol{K}_{\boldsymbol{\theta}}|$ represents the determinant. The gradient of the negative log marginal likelihood [Eq. (17)] is expressed as follows:

$$\frac{\partial \log p(\boldsymbol{g} \mid \boldsymbol{z}, \boldsymbol{\theta})}{\partial \theta} = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}_{\boldsymbol{\theta}}^{-\mathbf{1}}\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \theta}\right) + \left(\boldsymbol{K}_{\boldsymbol{\theta}}^{-\mathbf{1}}\boldsymbol{g}\right)^{\top}\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \theta}\left(\boldsymbol{K}_{\boldsymbol{\theta}}^{-\mathbf{1}}\boldsymbol{g}\right), \tag{18}$$

where $\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \theta}$ denotes the matrix of the same shape as the covariance matrix $\boldsymbol{K}_{\boldsymbol{\theta}}$, and the elements of the matrix are $\frac{\partial}{\partial \theta}k(z, z' \mid \theta)$, which is each element of the covariance matrix $\boldsymbol{K}_{\boldsymbol{\theta}}$ differentiated by the hyper-parameter θ. More accurate modeling and evaluation can be expected by optimizing the hyper-parameters in each training cycle where new input data z^* is given.

To improve the numerical stability of optimization calculations, our system employs multi-start optimization, which starts optimization calculations from multiple initial values by adding values generated by Latin hyper-cube sampling (LHS; McKay et al., 2000) to the hyper-parameters from the one training cycle ago. In addition, we adopted the L-BFGS-B algorithm (Byrd et al., 1995) as the optimization method.

The modeling of response surfaces using the GPR has been described above. Next, we explain evaluation using an acquisition function. The acquisition function is a combination of

301 the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{K}$ obtained by the GPR. First, following Lunderman et al.
302 (2021), we adopted the EI, defined by the following equation, as the acquisition function:

303 $$EI(\mu, \sigma) = (\hat{g} - \mu)\Phi(d) + \sigma\phi(d). \quad (19)$$

304 Here, $\hat{g}$ denotes the provisional optimum solution, i.e., the minimum value of the objective
305 function in the previous training cycle. In addition, σ represents the standard deviation,
306 which is the square root of K. Furthermore, d denotes the difference between the mean and
307 tentative optimal value normalized by the standard deviation and can be written as $d = (\hat{g} -$
308 $\mu)/\sigma$. Here, Φ and ϕ are the normal cumulative distribution function and the normal
309 probability density function, respectively. The EI increases when the GPR mean is small or
310 the GPR standard deviation is large.

311 However, using the EI, the inverse matrix in Eq. (14), (17), and (18) could not be calculated
312 stably due to the redundant exploration of the same input data. Therefore, we then adopted
313 penalized EI. The local penalization method proposed by Gonzalez et al. (2016) is an
314 approach that smoothly decreases the acquisition function value near the input data. This
315 approach assumes that the objective function is a Lipschitz continuous function and
316 prevents the redundant exploration by setting a spherical region centered on the input data
317 and adding a penalty to the acquisition function within that region.

318 In addition, since the algorithm falls into a local solution of the acquisition function, the
319 next input data cannot be obtained appropriately, so we optimized the acquisition function
320 (see also Shahriari et al., 2016). Additionally, multi-start optimization and the L-BFGS-B

algorithm were employed for optimization. To optimize the penalized EI, it is necessary to calculate the penalized EI and its derivative at the input data. The derivative of penalized EI can be described as follows:

$$\nabla \ln \widetilde{EI} = EI^{-1}\nabla EI + \sum_{s=1}^{S} \varphi(z^*, z_s)^{-1}\nabla\varphi(z^*, z_s). \tag{20}$$

Here, $\widetilde{EI}$ indicates the penalized EI. The next input data is explored after calculating the total penalty at all input data. The penalty function takes the following form:

$$\varphi(z^*, z_s) = \frac{1}{2}\mathrm{erfc}(-u), \tag{21}$$

with

$$u = \frac{1}{\sqrt{2\sigma^2}}(L\|z^* - z_s\| - \hat{g} + \mu).$$

Here, erfc is the complementary error function, and L is the Lipschitz constant. In the BO using the penalized EI, changing the ratio of "exploration and exploitation" is possible by adjusting the Lipschitz constant. As a rule of thumb, if L is 0.1 or more and less than 0.5, the setting is exploration-oriented; if L is 0.5 or more and less than 2.0, the setting is general; and if L is 2.0 or more and less than 10.0, the setting is exploitation-oriented. The derivative of the penalty function takes the following form:

$$\nabla\varphi(z^*, z_s) = \frac{e^{-u^2}}{\sqrt{2\pi\sigma^2}}\frac{2L}{\|z^* - z_s\|}(z^* - z_s). \tag{22}$$

The derivative of the EI can be described as follows:

$$\nabla EI = \frac{d\sigma}{dz}\phi(d) - \Phi(d)\frac{d\mu}{dz}. \tag{23}$$

The derivative of the penalized EI is described above. The penalized EI at an input data

340 is written as follows:

341 $$\ln \widetilde{EI} = \ln EI + \sum_{s=1}^{S} \ln \varphi(z^*, z_s)\,. \tag{24}$$

342 The local penalization method calculates the total product of the acquisition function and the

343 penalty at each input data and maximizes it. In Eq. (24), the total sum is calculated by

344 applying a logarithmic characteristic.

345

3. Experimental Setup

a. Lorenz-96 40-variable model

We conducted an observing system simulation experiment (OSSE) using the L96 to investigate whether the BO improves the practicality of the LPF. The L96 is a toy model that simulates atmospheric variables along certain latitudes. The time evolution of the atmospheric variable is expressed as follows:

$$\frac{dx_n}{dt} = (x_{n+1} - x_{n-2})x_{n-1} - x_n + F, \tag{25}$$

where x and t denote the state variables and time, respectively, as described in Section 2a. The subscripts n $(n = 1, \dots, N)$ represent the indices of the grid point. Since the L96 has periodic boundary conditions, the following relationship with respect to state variable at each grid point: $x_{-1} = x_{39}, x_0 = x_{40}$ and $x_{41} = x_1$ are satisfied. Each term on the right side represents the following: the first is advection, the second is diffusion, and the third is forcing F. The shift of the grid point in the advection term causes the nonlinearity of the atmosphere. Here, one variable is simulated at each grid point in 40 grid points. The fourth-order Runge–Kutta scheme is used for time integration, where forecast time step $\Delta t = 0.01$.

b. Data assimilation method

The LETKF can estimate the optimum analysis when the error distribution of observations and forecasts follows a Gaussian distribution. On the other hand, LPF does not require forecast errors to follow a Gaussian distribution. Therefore, when using a nonlinear

observation operator whose the background error distribution does not follow a Gaussian distribution, the LETKF cannot estimate the optimum analysis. In contrast, the LPF can estimate a more accurate analysis by handling observation information more appropriately. Therefore, following Poterjoy (2016), we adopted a nonlinear observation operator that takes the absolute value and logarithm of the state variables,

$$h(\boldsymbol{x}) = \ln(|\boldsymbol{x}|). \tag{26}$$

The root mean square error between the truth and the analysis (RMSE(t vs. a)) and the ensemble spread in the LETKF and the LPF were compared. In addition, the difference of RMSE(t vs. a) was investigated when the inflation factors α, τ, and the localization scale r were changed. In the LETKF, the covariance inflation factor α was varied in increments of 0.001 in the range of 1.01-1.10, and in the LPF, the weight inflation factor τ was varied in increments of 0.01 in the range of 0.1-1.0. In addition, the localization scale r was varied in increments of 0.1 in the range of 1-10 for both the LETKF and the LPF (i.e., 91 * 91 increments). The observations were generated by applying the nonlinear observation operator to the truth, which is a long-term integration of the L96, and adding Gaussian noise $\mathcal{N}(0,1)$ as observation errors. The observations were collected at all grid points at 0.05 time units. Here, 0.05 time units correspond to 6 Earth hours , which is the error-doubling time for synoptic weather. The observation variables are the same as the model variables, and the observation errors are assumed to be uncorrelated. Furthermore, as a gross error check, if the difference between the forecast and the observation exceeds 10 times the observation

error, the observation is rejected.

All observations are assimilated using the LETKF and the LPF for 64 particles over two years. Initial particles are randomly selected from long-term integration using the L96 initialized in a random state.

c. Parameter estimation

To confirm the basic behavior of the BO, we estimate one parameter (weight inflation factor τ) using the 1-dimensional BO, and then perform the 2-dimensional BO experiment to estimate the optimum values of two parameters (τ and localization scale r). In this study, we defined the root mean square error (RMSE(o vs. f)) between the observations and forecasts in the LPF as the objective function, and estimated τ and r that minimize this function using the BO.

$$g(z) = -\frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y_{n,t} - h\left(\bar{x}_{n,t}^{f}(z)\right)\right)^2}. \quad (27)$$

where g and z denote the RMSE(o vs. f) and the input data, respectively, as described in Section 2b; y and h represent the observation and observation operator, respectively, as outlined in Section 2a; $\bar{x}_{n,t}^{f}$ is the mean of the forecast particle at the nth grid point and at the tth time.

The RMSE (o vs. f) was used because the truth cannot be obtained in the real atmosphere. In addition, the analysis may be too close to the observations and is not always appropriate

405 for error estimation. On the other hand, the observations are perturbed around the truth, and
406 the forecast error is expected to be smaller than the observation error in the first guess and
407 larger than the observation error over time (Otsuka and Miyoshi, 2015). Therefore, we
408 evaluated the forecast accuracy by comparing future observations with extended forecasts.
409 This method is equivalent to indirectly assessing the analysis accuracy. Extended forecasts
410 are conducted for all particles. This assumption holds if the optimal analyses are estimated
411 and outliers of observations are rejected. Although this assumption is valid in the
412 experimental settings of this study, it may not always hold in general.

413 Unlike an online system, an offline system executes assimilation cycles and training
414 cycles separately, allowing the use of future observations. In addition, considering that the
415 NWP is executed using the optimum of past parameters, our system is reasonable. The
416 length of the extended forecast was set to 2 Earth days, based on the error doubling time.

417 The offline system of 2-dimensional BO experiment was executed according to the
418 following procedure:

419 1) Execute the OSSE using the weight inflation factor τ and localization scale r generated
420 by the LHS.

421 2) Calculate the RMSE(o vs. f)s and provide them as the initial input data to the BO.

422 3) Estimate the weight inflation factor τ and localization scale r that minimize the RMSE(o
423 vs. f) using the BO.

424 Here, we show the flowchart of the offline system in Fig. 1. The numbers of each process

Fig. 1

425 correspond to the numbers in Fig. 1.

426 We provided the initial input data generated by the LHS to the BO, performed the OSSE
427 with the estimated the weight inflation factor τ and localization scale r, and repeated the
428 training cycle that estimates τ and r, which minimize the RMSE(o vs. f) using the BO. In this
429 experiment, the system was stopped after 20 training cycles, and the weight inflation factor
430 τ and localization scale r with the smallest RMSE(o vs. f) were selected as the estimations
431 by the BO. In our system, we set the number of training cycles to 20 because the GPR
432 prediction distribution hardly changed even when more input data was added. In general,
433 the stopping criterion of the BO is often set based on the amount of computational resources
434 to be invested in advance and the variation of the estimation.

435 To evaluate the estimation accuracy and convergence rate of the BO, this study compared
436 the estimations of the BO and the RS, following Snoek et al. (2012). In addition, when the
437 RMSE(o vs. f) is defined as the objective function, estimation results of the BO depend on
438 the seed of Gaussian noise used to generate the observation errors. Therefore, we
439 conducted 35 experiments using the Gaussian noise with different seeds to investigate the
440 robustness of the BO to changes in the observation set.

441 In addition, we investigated how the estimation accuracy changes when the parameters
442 estimated by the BO are increased from 1 dimension (τ only) to 2 dimensions (τ, r).
443 Moreover, the Lipschitz constant was set to L = 0.5, and the number of initial input data was
444 set to 5 in the 1-dimensional BO. Additionally, in the 2-dimensional BO, the Lipschitz

445 constant was set to L = 2.0, and the number of initial input data was set to 20. We also
446 conducted sensitivity experiments with respect to the Lipschitz constant [Eq. (21)], which
447 determines the ratio of "exploration and exploitation", and the number of initial input data.
448

4. Result and Discussion

a. Comparison of the LETKF and the LPF

First, we investigated the conditions under which the LPF can estimate more accurate analyses than the LETKF. The GS obtained the following localization scale and inflation factor values. Fig. 2a shows the time series of the RMSE(t vs. a) and the ensemble spread in the LETKF. The localization scale r is 6.5, and the covariance inflation factor α is 1.100. The RMSE(t vs. a) fluctuates within the range of 0.5–5.0, showing significant fluctuations, especially during the first half of the experimental period corresponding to the spin-up period. Additionally, the ensemble spread fluctuated within the range of 0.5-1.0. Fig. 2b shows the time series of the RMSE(t vs. a) and the ensemble spread in the LPF. The localization scale r is 1.9, and the inflation factor τ is 0.53. The RMSE(t vs. a) fluctuated within the range of 0.5-2.5. Additionally, the ensemble spread fluctuated within the range of 0.5-1.5. These results show that when the nonlinear observation operator is used, and the RMSE(t vs. a) of the LPF is smaller than that of the LETKF. The LETKF only handles up to the second moment (variance) in the analysis error covariance matrix and posterior distribution updates, so it cannot consider higher-order moments such as the third (skewness) and fourth (kurtosis) moments. On the other hand, the LPF does not have such restrictions, allowing the LPF to adequately assimilate observations using information from the higher-order moments. The nonlinear observation operator in this study produces observation distributions with large skewness, suggesting that the LPF is able to estimate more accurate

Fig. 2

analyses than the LETKF.

Next, we investigated how the RMSE(t vs. a) changes when the inflation factors α, τ, and localization scale r are varied. Figure 3a shows the response surface of the RMSE(t vs. a) in the LETKF. In the LETKF, the minimum error of 1.024 was obtained when r = 6.5 and α = 1.100. The region of optimal parameters was distributed on the left shoulder in the range of localization scale r = 2-4 and covariance inflation factor α = 1.04-1.10. In addition, except for the region of r = 1-2 and α = 1.05-1.10, the RMSE(t vs. a) tends to increase as alpha decreases.

Fig. 3

Using the nonlinear observation operator forces the forecast error distribution to become non-Gaussian. Using the multiplicative inflation increases the forecast error covariance, which mitigates this negative impact, and therefore, such a tendency is expected to appear. On the other hand, except for the region where r = 1-2 and α = 1.05-1.10, there was also a tendency for the RMSE(t vs. a) to increase as r increases. This result is due to sampling errors being more pronounced in distant observations, where signals are small and cannot be assimilated effectively. If the inflation factor is too large, the forecast error will be large, and the observation error will be relatively underestimated, causing the analysis to be overly contaminated by observation errors. Additionally, if the localization scale is too small, the observations are not sufficiently assimilated, and the correction to the analysis becomes insufficient. Therefore, the RMSE(t vs. a) becomes large in the region where r = 1-2 and α = 1.05-1.10. The minimum error was not included in the region of optimal parameters, and

489 the boundary of the contour is unclear. This feature was not observed in the response
490 surface when α was changed in increments of 0.01 and r was changed in increments of 1
491 (not shown). In nonlinear dynamical systems, the RMSE(t vs. a) exhibits a nonlinear
492 response to changes in the localization scale and the inflation factor. Therefore, increasing
493 the resolution of the response surface may cause the local optima to appear outside the
494 global optima.

495 Figure 3b shows the response surface of the RMSE(t vs. a) in the LPF. In the LPF, a
496 minimum error of 0.586 was obtained with the localization scale r = 1.9 and the weight
497 inflation factor τ = 0.53. The region of optimal parameters was elliptically distributed within
498 the ranges of the localization scale r = 1-3 and the weight inflation factor τ = 0.4-0.6. Both
499 excessive and insufficient τ resulted in the large RMSE(t vs. a). This result is because
500 excessive τ prevents assimilation of observations, while insufficient τ causes filter instability.
501 Basically, the RMSE(t vs. a) decreases as r decreases. However, in the region where r = 1-
502 2 and τ = 0.1-0.4, the RMSE(t vs. a) increases as r decreases. This result indicates that
503 when the filter with small τ becomes unstable, the observations are not sufficiently
504 assimilated, resulting in a decrease in the analysis accuracy. The fact that the response to
505 changes in τ is more complex than that to changes in r suggests that τ is a more important
506 parameter for stabilizing the LPF.

507 Figure 3c shows the response surface of the RMSE(o vs. f) in observation space in the
508 LPF. The minimum error of 1.282 was obtained with the same r = 1.9 and τ = 0.53 as in the

RMSE(t vs. a) for the RMSE(o vs. f). From this result, it can be seen that the optima of these parameters can be efficiently estimated by estimating τ and r that minimize the RMSE(o vs. f) using the BO. Compared to the response surface of the RMSE(t vs. a), the overall distribution trend was consistent; however, the RMSE(o vs. f) exhibits a larger minimum value and a smaller maximum value. This result is because applying a nonlinear observation operator reduces the variance of the background error distribution in the observation space.

In summary, when using the nonlinear observation operators, the LPF can estimate more accurate analyses than the LETKF, but τ and r must be set to their optima.

b. 1-dimensional BO

Figure 4 is a cross-sectional view of the response surface (Fig. 3b) at the localization scale r = 1.9. In addition, as will be described in detail later, Fig. 6d shows the GPR prediction distribution of the 1-dimensional BO at the 20th training cycle. We compared Fig. 4 and Fig. 6d to verify whether the GPR prediction distribution of the 1-dimensional BO was appropriate. Focusing on the general shape of the GPR mean and GPR standard deviation, the U-shaped distribution was consistent. In addition, the input data were dense around τ = 0.5, which corresponded to the region where the RMSE(t vs. a)s were small in the response surface (Fig. 4). These results show that the 1-dimensional BO can model the response surface with high accuracy.

Fig. 4

The localization scale was fixed at r = 1.9, and only τ was estimated using the BO. The 5

Fig. 5

529 points obtained by LHS are used as initial input data. Fig. 5 shows the time series of the
530 minimum RMSE(o vs. f) from the previous training cycle, estimated by the BO and the RS.
531 Note that for convenience in conducting the OSSE using the BO estimation, the training
532 cycles of τ and the RMSE(o vs. f) are shifted by one cycle. The BO estimation converged at
533 the 8th training cycle, while the RS estimation converged at the 5th training cycle. The
534 estimation accuracy of both methods was equivalent. Although the RS estimation converged
535 in fewer training cycles, the BO was able to estimate a high precision τ from the 1st training
536 cycle. In addition, except for the 6th, 13th, and 19th training cycles, the BO estimation
537 fluctuated within the optimal inflation factor range. In Fig. 5, the parameter with the smallest
538 RMSE(o vs. f) was τ = 0.53 in the 8th training cycle, which was consistent with τ = 0.53, the
539 parameter that minimizes the RMSE(t vs. a) in Fig. 4. This result shows that the 1-
540 dimensional BO can estimate the optima of τ.

541 The estimation results of the 1-dimensional BO were investigated from the viewpoint of Fig. 6
542 the GPR prediction distribution. Fig. 6a-d show the GPR mean (expected value of the
543 RMSE(o vs. f)), the GPR standard deviation (95% confidence interval) (uncertainty of the
544 RMSE(o vs. f)), the EI, the penalty, the penalized EI, and input data variation corresponding
545 to Fig. 5.

546 At the 0th training cycle (Fig. 6a), the RMSE(o vs. f) at τ = 0.5, 0.79, 0.33, 0.14, and 0.97
547 from the LHS were given as the initial input data. The GPR mean showed a U-shaped
548 distribution with a minimum value around τ = 0.5, indicating that the response surface was

549 modeled with high accuracy at this point. The GPR standard deviation was almost constant
550 regardless of τ. Since the EI became large at points where the GPR mean was small, the
551 EI showed a convex distribution with a maximum value around τ = 0.5. The penalty showed
552 a V-shaped distribution with a minimum value around τ = 0.5. Since the penalized EI is
553 calculated as the product of the EI and the penalty, the penalized EI showed a distribution
554 with two connected peaks, with a small value at the point where the penalty was minimum.
555 At the 13th training cycle (Fig. 6b), as the input data accumulated, the GPR mean
556 increased to around τ = 0.8-0.9, and the shape of the GPR prediction distribution became
557 closer to the response surface. The GPR standard deviation increased to around τ = 0.65
558 and 0.85, and the GPR standard deviation became more volatile than in the 0th training
559 cycle. From the 0th to the 13th training cycle, the length scale parameter θ_2 in [Eq. (13)]
560 decreased significantly from 0.292 to 0.071, which is thought to have caused the distribution
561 to become highly volatile. Since the EI became large at points where the GPR standard
562 deviation was large, the EI tended to take the maximum around τ = 0.6. Since the penalty
563 became small at points where the GPR standard deviation was large, the penalty showed a
564 skewed distribution compared to the 0th training cycle. In addition, since the input data were
565 concentrated around τ = 0.5, the minimum penalty decreased significantly from 0.6 to 0.3,
566 and the valley of the V-shaped distribution became deeper. As a result, the penalized EI
567 showed the maximum at τ = 0.61 and a distribution that avoided the dense input data around
568 τ = 0.5.

At the 19th training cycle (Fig. 6c), the GPR mean increased further around τ = 0.8-0.9. During the 13th to 19th training cycle, the length scale parameter θ_2 decreased further, from 0.071 to 0.04. As a result, the GPR standard deviation became more volatile, and the value increased even around τ = 0.2. The EI showed large values around τ = 0.5 and 0.65, and showed a distribution with two connected peaks. The EI was small around τ = 0.2 and 0.85, where the GPR standard deviation was large, because the GPR mean was large at these points. The minimum penalty value increased again from 0.3 to 0.6, and the valley of the V-shaped distribution became shallower. This change is likely because τ = 0.61 was explored at the 13th training cycle, and the GPR standard deviation around this point became small. The penalized EI showed the maximum around τ = 0.5, but τ = 0.67 was explored because the acquisition function was optimized. It should be noted that since the acquisition function is calculated using the GPR mean and the GPR standard deviation, it contains uncertainty; therefore, the point with the maximum acquisition function is not necessarily the optimal exploration point.

At the 20th training cycle (Fig. 6d), the GPR mean decreased slightly around τ = 0.8-0.9. During the 19th to 20th training cycle, the length scale parameter θ_2 increased from 0.04 to 0.071, and the GPR standard deviation showed a smooth distribution overall. Therefore, the EI showed a convex distribution with the maximum around τ = 0.5, where the GPR mean was small. The minimum penalty decreased significantly from 0.6 to 0.15, and the valley of the V-shaped distribution became the deepest in the training cycles so far. This change is

589 likely because a point with τ = 0.67 was explored in the 19th training cycle, the input data
590 were explored evenly, and the GPR standard deviation decreased overall, making the
591 influence of the GPR mean relatively large. As a result, the penalized EI showed a
592 distribution with two peaks around τ = 0.5.

593 To confirm the practicality of the BO, we investigated the robustness of the BO to changes Fig. 7
594 in the observation set. Fig. 7a shows the box plot of τ. In all 5th, 10th, 15th, and 20th training
595 cycles, even when the observation set was changed, the upper and lower limits of the τ box
596 fluctuated by less than 0.1 at most. This fluctuation corresponds to less than 10% of the
597 parameter exploration range, indicating that the BO estimation is robust to changes in the
598 observation set. In addition, the length of the whiskers varied from 0.4, 0.1, 0.5, to 0.9. The
599 median was within the range of the optimal inflation factor in all 5th, 10th, 15th, and 20th
600 training cycles, indicating that the estimation accuracy of the BO is high.

601 Figure 7b shows the box plot of the RMSE(o vs. f). Until the 15th training cycle, the upper
602 and lower limits of the RMSE(o vs. f) box fluctuated by less than 0.1; however, at the 20th
603 training cycle, the RMSE(o vs. f) box fluctuated by approximately 0.3. This change is due to
604 τ becoming more exploration-oriented. The RMSE(o vs. f) whiskers tend to extend in the
605 direction of larger values. On the other hand, the τ whiskers tend to extend in the direction
606 of smaller values. These trends are due to the U-shaped distribution of the GPR mean and
607 the relationship between τ and the RMSE(o vs. f), where smaller τ results in the larger
608 RMSE(o vs. f) (see Fig. 6d). It should be noted that although the BO estimations appear to

609 scatter as the training cycle progresses, in practice, the input data with the smallest RMSE(o
610 vs. f) among the explored input data is adopted (see Section 3c). Focusing on the whiskers
611 of τ and RMSE(o vs. f), it can be inferred that 10 training cycles are sufficient for the 1-
612 dimensional BO, since they are shortest at the 10th training cycle.

613

614 *c. 2-dimensional BO*

615 The weight inflation factor τ and localization scale r were estimated using the BO. The 20 Fig. 8
616 points obtained by LHS are used as initial input data. Figure 8 shows the time series of the
617 minimum RMSE(o vs. f) in the previous training cycle estimated by the BO and the RS, the
618 estimation of τ, and the estimation of r. Note that, for the convenience of conducting the
619 OSSE using the estimation by the BO, the training cycles of τ, r, and the RMSE(o vs. f) are
620 shifted by one cycle. Both the BO and the RS estimation converged at the 17th training cycle.
621 Since the minimum RMSE(o vs. f) by the BO was lower than that by the RS, it can be seen
622 that the BO can estimate τ and r with higher accuracy than the RS. In addition, the
623 estimation of τ fluctuated within the optimal inflation factor range except for the 14th and
624 17th training cycles. The estimation of r fluctuated within the optimal localization scale range
625 except for the 14th training cycle.

626 In Fig. 8, the smallest RMSE(o vs. f) was obtained at the 17th training cycle with τ = 0.48
627 and r = 2.2, which are very close to the parameters that minimize the RMSE(t vs. a) in the
628 response surface (Fig. 3b), τ = 0.53 and r = 1.9. This result demonstrates that the 2-

dimensional BO can accurately estimate the optima of τ and r.

The estimation results of the 2-dimensional BO were investigated from the viewpoint of the GPR prediction distribution. Fig. 9a-e show the GPR mean (expected value of the RMSE(o vs. f)), the GPR standard deviation (95% confidence interval) (uncertainty of RMSE(o vs. f)), the EI, the penalty, the penalized EI, and input data variation corresponding to Fig. 8.

Fig. 9

At the 0th training cycle, the GPR prediction distribution was obtained from 20 initial input data. The GPR mean (Fig. 9a) showed the maximum at τ = 0.75 and r = 9, and the minimum at τ = 0.1 and r = 4.5, indicating a prediction distribution that was a combination of two 2-dimensional normal distributions centered at these points. When compared to the response surface (Fig. 3b), the overall trend was similar, with the RMSE(o vs. f) increasing as τ and r increase and decreasing as τ and r decrease. The GPR standard deviation (Fig. 9b) was large around τ = 0.1, r = 10 and τ = 1.0, r = 1, indicating that uncertainty was large in the regions where the input data were sparse. At this point, the length scale parameter θ_2 in [Eq. (13)] was 0.408 (the minimum: 0.003, the maximum: 2.483) and was small, while the length scale parameter θ_3 was 2.483 (the minimum: 0.003, the maximum: 2.483) and was very large. In this case, varying τ yields a more complex GPR prediction distribution compared to varying r.

Since the GPR standard deviation showed a smooth distribution overall, the EI (Fig. 9c) showed the maximum around τ = 0.1 and r = 4.5, where the GPR mean was small. In

addition, since the penalty decreases as the GPR mean decreases, the penalty (Fig. 9d) showed the minimum around τ = 0.1 and r = 5.0. The distribution was slightly shifted toward r being larger than the EI because the GPR standard deviation was large at τ = 0.1 and r = 10. The penalized EI (Fig. 9e) is calculated as the product of the EI and the penalty, and therefore showed a slightly skewed distribution around τ = 0.1 and r = 5.0. The penalized EI showed the maximum around τ = 0.1 and r = 4.5; however, since the acquisition function is optimized, τ = 0.43 and r = 1.0 were explored.

Following Fig. 9, we investigated how the estimation results of the 2-dimensional BO change as the input data increases, from the perspective of the GPR prediction distribution. Fig. 10a-e shows the GPR mean (expected value of the RMSE(o vs. f)), the GPR standard deviation (95% confidence interval) (uncertainty of the RMSE(o vs. f)), the EI, the penalty, the penalized EI, and input data variation corresponding to Fig. 8.

Fig. 10

At the 20th training cycle, the GPR prediction distribution was obtained from a total of 40 input data. The GPR mean (Fig. 10a) showed the maximum around τ = 0.85, r = 6.5, and τ = 0.2, r = 9, and the minimum around τ = 0.2, r = 5. In addition, the overall prediction distribution was similar to a combination of three 2-dimensional normal distributions centered on these points. Compared to the GPR mean at the 0th training cycle (Fig. 9a), the two-dimensional normal distribution centered on τ = 0.75 and r = 9 was divided into two. Another characteristic is that the two-dimensional normal distribution centered on τ = 0.1 and r = 4.5 was shifted toward larger τ. Compared to the response surface (Fig. 3b), the

GPR mean at the 20th training cycle achieved higher modeling accuracy than that at the 0th training cycle, as the minimum was closer, and the region with the large GPR mean showed an inverted L-shaped distribution. The GPR standard deviation (Fig. 10b) showed a smooth distribution overall due to the increase in input data. At this time, the length scale parameter θ_2 in [Eq. (13)] was 0.226 (the minimum: 0.002, the maximum: 2.198) and was small, and the length scale parameter θ_3 was 0.677 (the minimum: 0.002, the maximum: 2.198) and was also small. In this case, the GPR prediction distribution obtained was considered more volatile than that at the 0th training cycle, as the response became complex for both r and τ.

With sufficient input data, the difference between the provisional optimum solution and the GPR mean decreased, and the GPR standard deviation also decreased, resulting in the EI (Fig. 10c) close to 0 overall. This result indicates that the BO has converged. The GPR standard deviation was smooth overall, and since the GPR mean showed the minimum around τ = 0.2 and r = 5, the penalty (Fig. 10d) also showed the minimum at the same point. Since the EI showed values close to 0 overall, the penalized EI (Fig. 10e) also showed a similar distribution. In the 1-dimensional BO, points with the small GPR means were explored intensively, but this is not the case in the 2-dimensional BO because points with the large GPR standard deviations were explored. The result showed that the input data were dense in the region of optimal parameters (localization scale r = 1-3, inflation factor τ = 0.4-0.6) in the response surface (Fig. 3b), indicating that the 2-dimensional BO can model

689 the response surface with relatively high accuracy.

690 To confirm the practicality of the BO, we investigated the robustness of the BO to changes Fig. 11
691 in the observation set. Fig. 11a shows the box plot of τ. In all 5th, 10th, 15th, and 20th training
692 cycles, even when the observation set was changed, the upper and lower limits of the τ box
693 fluctuated by less than 0.1 at most. From this result, we can see that even when the
694 observation set is changed, the BO estimation converges to τ = 0.1. The observation sets
695 used in Figs. 8, 9, 10, and Table 1 yield estimation close to the optima of τ = 0.48 and r =
696 2.2; however, it can be seen that the estimation may deviate from the optima depending on
697 the observation set. In this case, the estimation accuracy of the BO can be improved by
698 reducing the Lipschitz constant (not shown). In addition, the length of the whiskers varied
699 from 0.9, 0.4, less than 0.1 (including outliers), and 0.4. In all 5th, 10th, 15th, and 20th
700 training cycles, the median was outside the optimal inflation factor range, and the BO
701 estimation accuracy was low.

702 Figure 11b shows the box plot of r. Until the 15th training cycle, the upper and lower limits
703 of the r box fluctuated by 9.0, but at the 20th training cycle, the fluctuation was less than 1.0.
704 This result shows that even when the observation set is changed, the BO estimation tends
705 to shift to the optimal localization scale range. In addition, the length of the whiskers was 9.0
706 in all 5th, 10th, 15th, and 20th training cycles. The median was within the optimal localization
707 scale range in all 5th, 10th, 15th, and 20th training cycles, and the estimation accuracy of
708 the BO was high compared to τ. This difference indicates that r is easier to estimate than τ

because the response of the RMSE(o vs. f) when r is changed is simpler than that of τ.

Figure 11c shows the box plot of the RMSE(o vs. f). In all 5th, 10th, 15th, and 20th training cycles, the upper and lower limits of the RMSE(o vs. f) box fluctuated by less than 0.1. The whiskers of RMSE(o vs. f) fluctuated by 0.5 at the 5th training cycle, but only by 0.3 at the subsequent training cycles. The estimation of τ converged to 0.1, and the estimation of r moved to the optimal range, so the whiskers tended to extend in the direction of the smaller RMSE(o vs. f).

d. Sensitivity experiment for the BO setting changes

Furthermore, we investigated the effects of changes in the response surface dimension, Lipschitz constant, and number of initial input data on the BO estimation. Table 1 summarizes the results of the sensitivity experiment. Although there are multiple cases with the same minimum RMSE(o vs. f), the Lipschitz constant is generally set to L = 0.5-2.0. In addition, the fewer the number of initial input data, the fewer computing resources are required. Therefore, cases with L = 0.5 and 5 initial input data for the 1-dimensional BO and L = 2.0 and 20 initial input data for the 2-dimensional BO are highlighted in bold. In addition, since the ideal number of initial input data is about 10 times the dimension of the response surface (Loeppky et al., 2009), the maximum number of initial input data was 40, which is twice the ideal number. The number of initial input data was changed in increments of 10 for the 2-dimensional BO and 5 for the 1-dimensional BO, resulting in 16 cases for both.

Table. 1

729 In both the 2-dimensional BO and 1-dimensional BO, the minimum RMSE(o vs. f) tended
730 to increase as the Lipschitz constant increased. This result indicates that an excessive
731 Lipschitz constant causes bias in the input data due to an undue emphasis on exploitation,
732 which reduces the estimation accuracy of the BO. In addition, in the 1-dimensional BO, the
733 estimation accuracy became the same regardless of changes in the number of initial input
734 data. As the Lipschitz constant increases, the emphasis shifts to exploitation, and points
735 close to each other continue to be explored regardless of changes in the number of initial
736 input data (not shown). This result was obtained because this tendency was particularly
737 pronounced in the 1-dimensional BO, where there are few exploration points.
738 Furthermore, focusing on cases with each Lipschitz constant, an increase in the number
739 of initial input data did not necessarily improve the estimation accuracy of the BO. When the
740 number of initial input data is large, the GPR prediction distribution approaches the response
741 surface, and the BO estimation does not fluctuate considerably. On the other hand, when
742 the number of initial input data is small, the BO estimation fluctuates considerably, and the
743 optima may be explored by chance (not shown). The "exploration and exploitation" dilemma
744 is usually used in the context of adjusting parameters such as the Lipschitz constant. Still,
745 this expression may also be used for the number of initial input data (for details on the
746 "exploration and exploitation" dilemma, see Russo et al. (2020)).
747 In addition, focusing on the best cases for each dimension of the response surface, the
748 difference in the minimum RMSE(o vs. f) is less than 0.02, indicating that the estimation

accuracy of the BO does not decrease significantly even when the dimension of the response surface increases from 1 to 2. However, it should be noted that in the 2-dimensional BO, the estimations close to the optima may not always be obtained depending on the observation set.

5. Conclusion

The PF is a powerful data assimilation method that does not assume the linearity in the time evolution of errors and Gaussian error distributions. However, the number of required particles increases exponentially with the dimensions of the dynamical system, which is a bottleneck when applying the PF to the NWP systems. The LPF is a method that realizes the PF in high-dimensional systems by the localization. In addition, when using the nonlinear observation operator, the LPF can estimate a more accurate analysis than the LETKF. However, this is limited to cases where the weight inflation factor and the localization scale are set to their optima. In addition, as the resolution of the response surface increases and the number of parameters to be estimated increases (e.g., resampling frequency and Gaussian kernel amplitude), the effort and computational resources required for optimization calculations increase; therefore, efficient parameter estimation methods are needed.

Therefore, we estimated the weight inflation factor and localization scale that minimize the RMSE(o vs. f) using the BO. As a result, in the 1-dimensional case, the BO was able to model the response surface with high accuracy and estimate the inflation factor equivalent to the RS. In addition, this result was robust to changes in the observation set.

In addition, in the 2-dimensional case, the BO was able to model the response surface with relatively high accuracy and estimate the weight inflation factor and localization scale with higher accuracy than the RS. However, depending on the observation set, the BO did not always obtain the estimations close to the optima.

774 Next, we will discuss prospects for the practical application of the LPF. First, as the number
775 of particles decreases, the response surface that enables stable operation of the LPF
776 becomes narrower (not shown), making estimation using the BO difficult. In this case, it
777 would be effective to introduce an approach similar to the annealing method into the BO,
778 where a wide region is explored in the initial stage to identify a promising region, and then
779 the exploration range is narrowed down to the surrounding region.

780 In addition, in models more advanced than the L96, it is expected that the optima of the
781 inflation factor and localization scale will not be uniform throughout the model. In fact, in the
782 LETKF system using the SPEEDY (Simplified Parameterizations, Primitive Equation
783 Dynamics; Molteni, 2003) model, the optimal localization scale is not uniform across the
784 entire domain, and it is desirable to set the larger (smaller) localization scale in regions with
785 the sparse (dense) observations (Kotsuki et al., 2020). Since the optima of covariance
786 inflation factor in the LETKF also depend on the localization scale, it is unlikely to be uniform
787 throughout the entire domain. In this case, it would be appropriate to divide the model
788 domain based on the observation density and perform parameter estimation by the BO.

789 We describe features of the offline optimization. Although not experimented with in this
790 study, the BO uses the extended forecast as an argument for the objective function, enabling
791 parameter estimation that takes into account the model error that develops over time.
792 Additionally, the offline optimization conducts the OSSE multiple times during the same
793 period to estimate parameters that minimize the period-average RMSE(o vs. f). Therefore,

794 extending the experiment period allows for the estimation of parameters that lead to long-
795 term stable operation of the LPF. On the other hand, the online optimization which executes
796 the assimilation cycle and the training cycle in parallel, is also a promising system. Since the
797 optima of parameters in the LPF change moment by moment, the development of this
798 system would enable further stabilization of the LPF and higher accuracy of analysis.

799 Unlike gradient methods, the BO is advantageous in that it can efficiently explore a global
800 optimal solution even when the shape of the response surface of input and output data is
801 unknown or when the response surface is a multi-peaked function that cannot be
802 differentiated. In recent years, libraries such as GPyOpt have become widely available,
803 enabling rapid adoption of such advanced systems. However, to promote the use of the BO
804 within data assimilation frameworks, it is undesirable to treat the BO merely as a tool.
805 Therefore, it is important to conduct analyses focused on the GPR prediction distribution or
806 the Lipschitz constant, as in this study, and to systematically accumulate insights that
807 contribute to a fundamental understanding of the BO.

808 Finally, we summarize the usefulness of the BO in the NWP systems. For example, since
809 there is a nonlinear relationship between radar reflectivity and precipitation intensity, it is
810 necessary to use a nonlinear observation operator for assimilating radar reflectivity to
811 improve the accuracy of heavy rainfall forecasts. In this case, the LPF can estimate the more
812 accurate analyses than the LETKF. However, in order to work the LPF stably, parameters
813 such as the weight inflation factor and the localization scale must be their optima. The BO

814 is a technology that can efficiently explore global optimal solutions for parameters, and its
815 use can be expected to improve the practicality of the LPF. In other words, the BO is a
816 technology that contributes to improving the accuracy of heavy rainfall prediction and will be
817 helpful in atmospheric model experiments for the practical application of the LPF.
818

Data Availability Statement

The source code used in this study is available upon request to the corresponding author.

Acknowledgments

In addition, I would like to express my sincere gratitude to the two reviewers and the editor, Professor Shunji Kozuki (Chiba University), for their honest review. Part of this research was supported by JST SPRING, Grant Number JPMJSP2124. This research was also supported by the Fundamental Technology Research of MRI (M5 and P5), a Grant-in-Aid for Scientific Research (KAKENHI) (Grant Numbers JP23H05494, JP23K17465, and JP21K13995) from the Japan Society for the Promotion of Science, and the Environmental Research and Technology Development Fund (Grant Numbers JPMEERF20245004) of the Environmental Restoration and Conservation Agency of Japan (ERCA).

References

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129**, 420–436, https://doi.org/10.1175/1520-0493(2001)129%3C0420:ASWTET%3E2.0.CO;2.

Byrd, R. H., and P. Lu, J. Nocedal, C. Zhu, 1995: A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.*, **16(5)**, 1190-1208, https://doi.org/10.1137/0916069.

Le Dimet, F. X., and O. Talagrand, 1986: Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*, **38(2)**, 97–110, https://doi.org/10.3402/tellusa.v38i2.11706.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res*. **99**, 10143–10162, https://doi.org/10.1029/94JC00572.

Farchi, A., and M. Bocquet, 2018: Review article: Comparison of local particle filters and new implementations. *Nonlinear Processes Geophys.*, **25**, 765-807, https://doi.org/10.5194/npg-25-765-2018.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723-757, https://doi.org/10.1002/qj.49712555417.

González, J., and Z. Dai, P. Hennig, N. D. Lawrence, 2015: Batch Bayesian Optimization via Local Penalization. arXiv, https://arxiv.org/abs/1505.08052.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith, 1993: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proceedings F (Radar and Signal Processing), 140:2, 107–113, https://doi.org/10.1049/ip-f-2.1993.0015.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Phys. D*, **230**, 112–126, https://doi.org/10.1016/j.physd.2006.11.008.

Kondo, K., and T. Miyoshi, 2019: Non-Gaussian statistics in global atmospheric dynamics: a study with a 10 240-member ensemble Kalman filter using an intermediate atmospheric general circulation model. *Nonlinear Processes Geophys.*, **26**, 211–225, https://doi.org/10.5194/npg-26-211-2019.

Kotsuki, S., T. Miyoshi, K. Kondo, and R. Potthast, 2022: A local particle filter and its Gaussian mixture extension implemented with minor modifications to the LETKF, *Geosci. Model Dev.*, **15**, 8325–8348, https://doi.org/10.5194/gmd-15-8325-2022.

Loeppky, J. L., J. Sacks, and W. J. Welch, 2009: Choosing the Sample Size of a Computer Experiment: A Practical Guide. *Technometrics*, **51(4)**, 366–376, https://doi.org/10.1198/TECH.2009.08040.

Lorenz, E. N., and K. A. Emanuel, 1998: Optimal Sites for Supplementary Weather Observations: Simulation with a Small Model. *J. Atmos. Sci.*, **55**, 399–414,

https://doi.org/10.1175/1520-0469(1998)055<0399:OSFSWO>2.0.CO;2.

Lunderman, S., M. Morzfeld, and D. J. Posselt, 2021: Using global Bayesian optimization in ensemble data assimilation: parameter estimation, tuning localization and inflation, or all of the above. *Tellus A*, **73(1)**, p. 1924952, https://doi.org/10.1080/16000870.2021.1924952.

Mckay, M. D., R. J. Beckman, and W. J. Conover, 2000: A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code. *Technometrics*, **42(1)**, 55–61, https://doi.org/10.1080/00401706.2000.10485979.

Mockus, J., 1989: Mathematics and its Applications: Bayesian Approach to Global Optimization: Theory and Applications. Kluwer Academic Publishers, 270pp. https://doi.org/10.1007/978-94-009-0909-0.

Molteni, F., 2003: Atmospheric simulations using a GCM with simplified physical parametrizations. I: model climatology and variability in multi-decadal experiments. *Climate Dynamics*, **20**, 175–191, https://doi.org/10.1007/s00382-002-0268-2.

Otsuka, S., and T. Miyoshi, 2015: A Bayesian Optimization Approach to Multimodel Ensemble Kalman Filter with a Low-Order Model. *Mon. Wea. Rev.*, **143**, 2001–2012, https://doi.org/10.1175/MWR-D-14-00148.1.

Ott, E., B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. J. Patil, and J. A. Yorke, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56**, 415–428, https://doi.org/10.1111/j.1600-0870.2004.00076.x.

892 Penny, S. G., and T. Miyoshi, 2016: A local particle filter for high-dimensional geophysical
893 systems. *Nonlinear Processes Geophys.*, **23**, 391–405, https://doi.org/10.5194/npg-23-
894 391-2016.
895 Poterjoy, J., 2016: A Localized Particle Filter for High-Dimensional Nonlinear Systems.
896 *Mon. Wea. Rev.*, **144**, 59–76, https://doi.org/10.1175/MWR-D-15-0163.1.
897 Poterjoy, J., and J. L. Anderson, 2016: Efficient Assimilation of Simulated Observations in a
898 High-Dimensional Geophysical System Using a Localized Particle Filter. *Mon. Wea.*
899 *Rev.*, **144**, 2007–2020, https://doi.org/10.1175/MWR-D-15-0322.1.
900 Potthast, R., A. Walter, and A. Rhodin, 2019: A Localized Adaptive Particle Filter within an
901 Operational NWP Framework. *Mon. Wea. Rev.*, **147**, 345–362,
902 https://doi.org/10.1175/MWR-D-18-0028.1.
903 Rasmussen, C. E., and H. Nickisch, 2010: Gaussian Processes for Machine Learning
904 (GPML) Toolbox. *J. Mach. Learn. Res*, **11**, 3011–3015,
905 https://dl.acm.org/doi/abs/10.5555/1756006.1953029.
906 Rasmussen, C. E., and C. K. I. Williams, 2006: Gaussian Processes for Machine Learning.
907 the MIT Press, 266pp.
908 Russo, D., B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen, 2020: A Tutorial on
909 Thompson Sampling. *arXiv*, https://arxiv.org/abs/1707.02038.
910 Sawada, Y., 2020: Machine learning accelerates parameter optimization and uncertainty
911 assessment of a land surface model. *J. Geophys. Res.: Atmos.*, **125**, e2020JD032688,

https://doi.org/10.1029/2020JD032688.

Shahriari, B., and K. Swersky, Z. Wang, R. P. Adams, N. de Freitas, 2016: Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE*, **104(1)**, 148-175, https://doi.org/10.1109/JPROC.2015.2494218.

Snoek, J., H. Larochelle, and R. P. Adams, 2012: Practical Bayesian Optimization of Machine Learning Algorithms, https://doi.org/10.48550/arXiv.1206.2944.

Snyder, C., T. Bengtsson, P. Bickel, and J. Anderson, 2008: Obstacles to High-Dimensional Particle Filtering. *Mon. Wea. Rev.*, **136**, 4629–4640, https://doi.org/10.1175/2008MWR2529.1.

Stordal, A. S., H. A. Karlsen, G. Nævdal, H. J. Skaug, and B. Vallès, 2011: Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter. *Comput. Geosci.*, **15**, 293–305. https://doi.org/10.1007/s10596-010-9207-1.

925 List of Figures

926 Fig. 1. Flowchart of the Bayesian optimization (BO) in the local particle filter (LPF)
927 framework. Since the data assimilation system and the BO are implemented
928 independently, it is possible to replace the LPF with the local ensemble transform
929 Kalman filter. Here, t $(t = 1, \ldots, T)$ is time, g is the objective function, s $(s = 1, \ldots, S)$ is the
930 input data (inflation factor τ, α, and localization scale r), and the number of output data
931 (root mean square error between the observations and the forecasts (RMSE(o vs. f))).
932 In the observing system simulation experiment (OSSE), the observations are
933 assimilated by the LPF every 6 Earth hours, and the RMSE(o vs. f) at the same time is
934 calculated after the 2 Earth days extended forecast. Through this process in the
935 objective function, the input data are converted to the output data. In the BO, the input
936 data that minimizes the objective function is estimated through response surface
937 modeling using Gaussian process regression and evaluation using an acquisition
938 function (penalized expected improvement). The training cycle of conducting the OSSE
939 using the estimated input data is repeated. Note that the BO optimizes τ and r offline.

940

941 Fig. 2. Time series of the root mean square error and ensemble spread between the truth
942 and the analysis (RMSE(t vs. a)) in the local ensemble transform Kalman filter (LETKF)
943 and the local particle filter (LPF) using 64 ensemble members (particles) and the
944 nonlinear observation operators. The vertical axis shows the RMSE(t vs. a) (blue line)

945 and the ensemble spread (red line), while the horizontal axis indicates the assimilation
946 cycle. The localization scale of the LETKF was set to r = 6.5, and the inflation factor was
947 set to α = 1.100 (The optima in Fig. 3a). In addition, the localization scale of the LPF
948 was set to r = 1.9, and the inflation factor was set to τ = 0.53 (The optima in Fig. 3b).
949
950 Fig. 3. Response surface of root mean square error between the truth and the analysis
951 (RMSE(t vs. a)) in the local ensemble transform Kalman filter (LETKF) and the local
952 particle filter (LPF) using 64 ensemble members (particles) and the nonlinear observation
953 operator. The closer the color is to green, the larger the RMSE(t vs. a), and the closer the
954 color is to blue, the smaller the RMSE(t vs. a). The vertical axis shows the localization
955 scale r, and the horizontal axis shows the inflation factor α and τ. The minimum error of
956 1.024 in the LETKF was obtained when r = 6.5 and α = 1.100 (cross mark). In addition,
957 the minimum error of 0.586 in the LPF was obtained when r = 1.9 and τ = 0.53 (cross
958 mark).
959
960 Fig. 4. Response surface of the root mean square error between the truth and the analysis
961 (RMSE(t vs. a)) in the local particle filter using 64 ensemble members (particles) and the
962 nonlinear observation operator. The vertical axis shows the RMSE(t vs. a), and the
963 horizontal axis shows the inflation factor τ. The localization scale was fixed at r = 1.9,
964 and the minimum error of 0.586 was obtained when τ = 0.53 (cross mark). In addition,

965 the light blue shaded area indicates the optimal inflation factor range (τ = 0.34-0.58)
966 where the RMSE(t vs. a) of the LPF is 1.0 or less (the filter operates stably).
967
968 Fig. 5. Time series of the estimation by the 1-dimensional Bayesian optimization (BO). The
969 blue line shows the inflation factor τ, the green line shows the minimum root mean
970 square error between the observations and the forecasts (RMSE(o vs. f)) in the previous
971 training cycle estimated by the BO, and the purple line shows the minimum RMSE(o vs.
972 f) in the previous training cycle estimated by the random sampling. In addition, the light
973 blue shaded area indicates the optimal inflation factor range (τ = 0.34-0.58) for which
974 the root mean square error between the truth and the analysis in the local particle filter is
975 1.0 or less (the filter operates stably). The horizontal axis shows the training cycle, the
976 first vertical axis shows τ, and the second vertical axis shows the minimum RMSE(o vs.
977 f) in the previous training cycle. Note that, due to the convenience of conducting the
978 OSSE using the BO estimation, the training cycles of τ and the RMSE(o vs. f) are
979 shifted by one cycle.
980
981 Fig. 6. Prediction distribution of Gaussian process regression (GPR) using the inflation
982 factor τ and the root mean square error between the observations and the forecasts
983 (RMSE(o vs. f)) in the local particle filter as input and output data. The green line
984 indicates the penalized expected improvement (EI), the purple line indicates the penalty,

985 the yellow line indicates the EI, the red dots indicate the input data that has been
986 explored, and the yellow dots indicate the input data explored in that training cycle. The
987 horizontal axis is τ, the first vertical axis is the RMSE(o vs. f), the second vertical axis is
988 the penalized EI, the third vertical axis is the penalty, and the fourth vertical axis is the
989 EI. (a)-(d) are the prediction distributions for the 0th (i.e., when only the initial input data
990 were given), 13th, 19th, and 20th training cycles, respectively.
991
992 Fig. 7. Variation in the estimation by 1-dimensional Bayesian optimization for different
993 observations. (a) Box plot of inflation factor τ. The blue line is the median, the lower
994 edge of the box is the first quartile, the upper edge of the box is the third quartile, the
995 lower edge of the whiskers is the minimum, and the upper edge of the whiskers is the
996 maximum. In addition, the light blue shaded area indicates the optimum inflation factor
997 range (τ = 0.34-0.58) for which the root mean square error between the observations
998 and forecasts (RMSE(o vs. f)) in the local particle filter is less than 1.0 (the filter
999 operates stably). (b) Box plot of the RMSE(o vs. f). The red line is the median, and the
1000 other plots are the same as in (a). The vertical axes in (a) and (b) indicate τ and the
1001 RMSE(o vs. f), respectively. The horizontal axes indicate the number of training cycles.
1002
1003 Fig. 8. Time series plot of the estimations by the 2-dimensional Bayesian optimization
1004 (BO). The blue line shows the inflation factor τ, the orange line shows the localization

scale r, the green line shows the minimum root mean square error between the observations and the forecasts (RMSE(o vs. f)) in the previous training cycle estimated by the BO, and the purple line shows the minimum RMSE(o vs. f) in the previous training cycle estimated by the random sampling. The light blue shaded area indicates the optimal inflation factor range (τ = 0.32-0.67) for which the root mean square error between the truth and the analysis in the local particle filter is less than 1.0 (the filter operates stably). In addition, the light orange shaded area indicates the optimal localization scale range (r = 1.0-4.2). The light beige shaded area indicates the range that satisfies both the optimal inflation factor and localization scale. The horizontal axis represents the training cycle, the first vertical axis represents τ, the second vertical axis represents r, and the third vertical axis represents the minimum RMSE(o vs. f). Note that due to the convenience of conducting the OSSE using the BO estimations, the training cycles of τ, r, and the RMSE (o vs. f) are shifted by one cycle.

Fig. 9. Prediction distribution of Gaussian process regression (GPR) using the inflation factor τ and the localization scale r in the 0th training cycle (i.e., when only initial input data were given) and the root mean square error between observations and forecasts (RMSE(o vs. f)) in the local particle filter as input and output data. (a) is the GPR mean, (b) is the GPR standard deviation, (c) is the expected improvement (EI), (d) is the penalty, and (e) is the penalized EI prediction distribution. The red dots indicate the input

data that has been explored, and the yellow dots indicate the input data explored in that training cycle. The horizontal axis indicates τ, and the vertical axis indicates r.

Fig. 10. Prediction distribution of Gaussian process regression (GPR) using the inflation factor τ and the localization scale r at the 20th training cycle, and the root mean square error between the observations and the forecasts (RMSE(o vs. f)) in the local particle filter as input and output data. (a) is the GPR mean, (b) is the GPR standard deviation, (c) is the expected improvement (EI), (d) is the penalty, and (e) is the penalized EI prediction distribution. The training cycle in this figure corresponds to that in Fig. 8. The expected value and uncertainty of the RMSE(o vs. f) are obtained as the mean and standard deviation of the GPR. Red dots indicate input data that has been explored, and yellow dots indicate input data explored in that training cycle. The horizontal axis indicates τ, and the vertical axis indicates r.

Fig. 11. Variation in the estimation by the 2-dimensional Bayesian optimization for the different observation sets. (a) Box plot of inflation factor τ. The blue line is the median, the lower edge of the box is the first quartile, the upper edge of the box is the third quartile, the lower edge of the whiskers is the minimum, the upper edge of the whiskers is the maximum, and the dots indicate outliers. In addition, the light blue shade indicates the optimal inflation factor range (τ = 0.34-0.58) for which the root mean square error

between the truth and the analysis (RMSE(t vs. a)) in the local particle filter is less than 1.0 (the filter operates stably). (b) Box plot of localization scale r. The orange line is the median, and the other plots are the same as in (a). In addition, the light orange shade indicates the optimal localization scale range (r = 1.0-4.2). (c) Box plot of the RMSE(o vs. f). The red line is the median, and the other plots are the same as in (a). The vertical axes in (a), (b), and (c) indicate τ, r, and the RMSE(o vs. f), respectively. The horizontal axes indicate the number of training cycles.

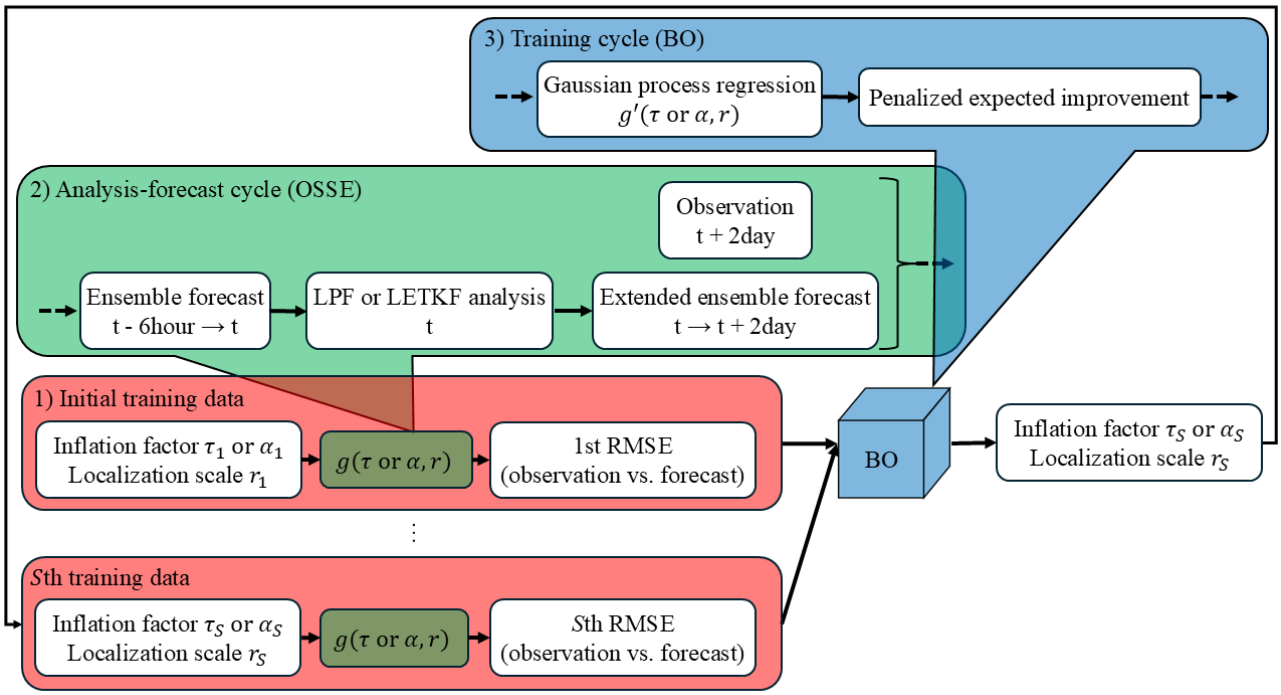


Fig. 1. Flowchart of the Bayesian optimization (BO) in the local particle filter (LPF) framework. Since the data assimilation system and the BO are implemented independently, it is possible to replace the LPF with the local ensemble transform Kalman filter. Here, t $(t = 1, \ldots, T)$ is time, g is the objective function, s $(s = 1, \ldots, S)$ is the input data (inflation factor τ, α, and localization scale r), and the number of output data (root mean square error between the observations and the forecasts (RMSE(o vs. f))). In the observing system simulation experiment (OSSE), the observations are assimilated by the LPF every 6 Earth hours, and the RMSE(o vs. f) at the same time is calculated after the 2 Earth days extended forecast. Through this process in the objective function, the input data are converted to the output data. In the BO, the input data that minimizes the objective function is estimated through response surface modeling using Gaussian process regression and evaluation using an acquisition

function (penalized expected improvement). The training cycle of conducting the OSSE using the estimated input data is repeated. Note that the BO optimizes τ and r offline.

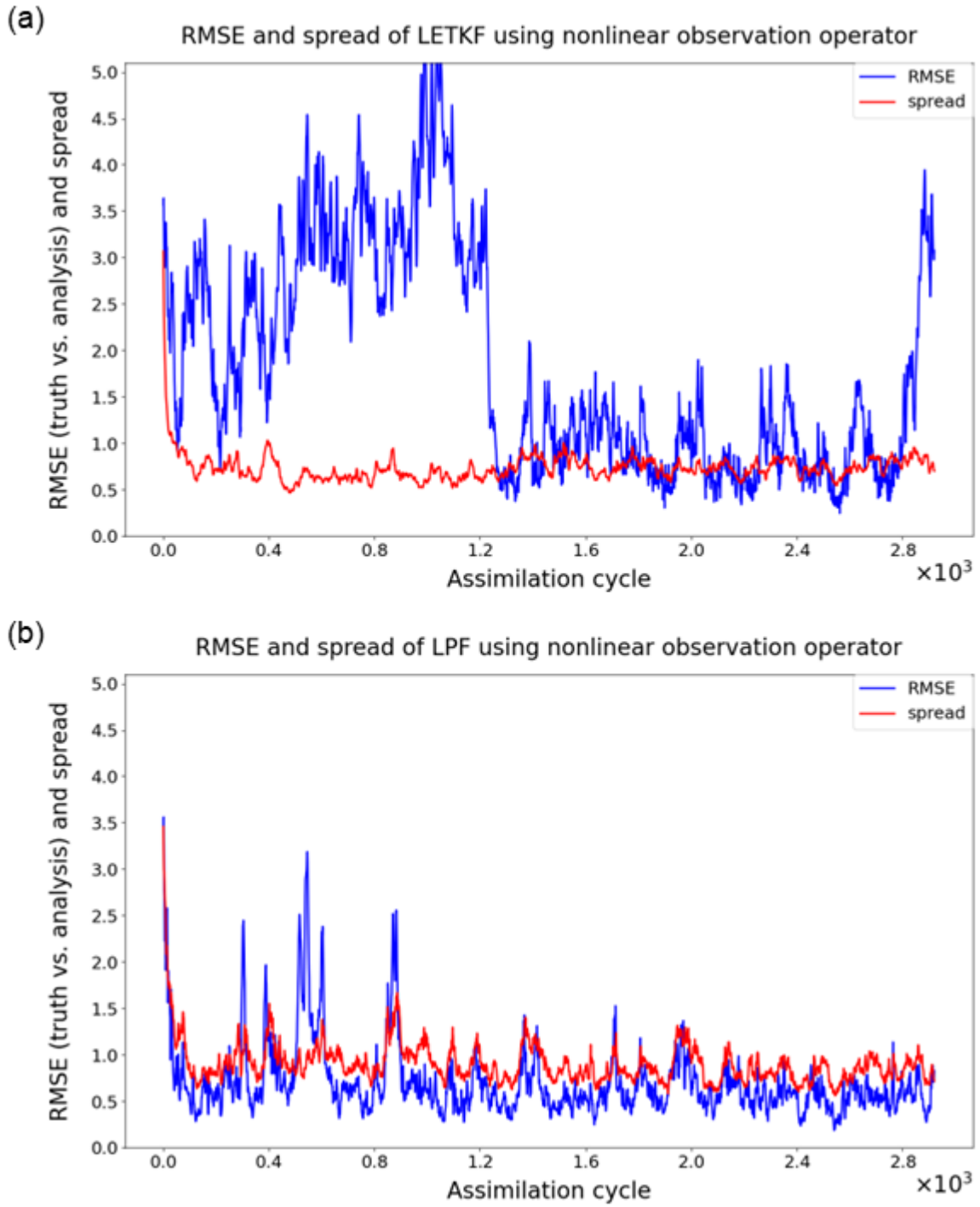
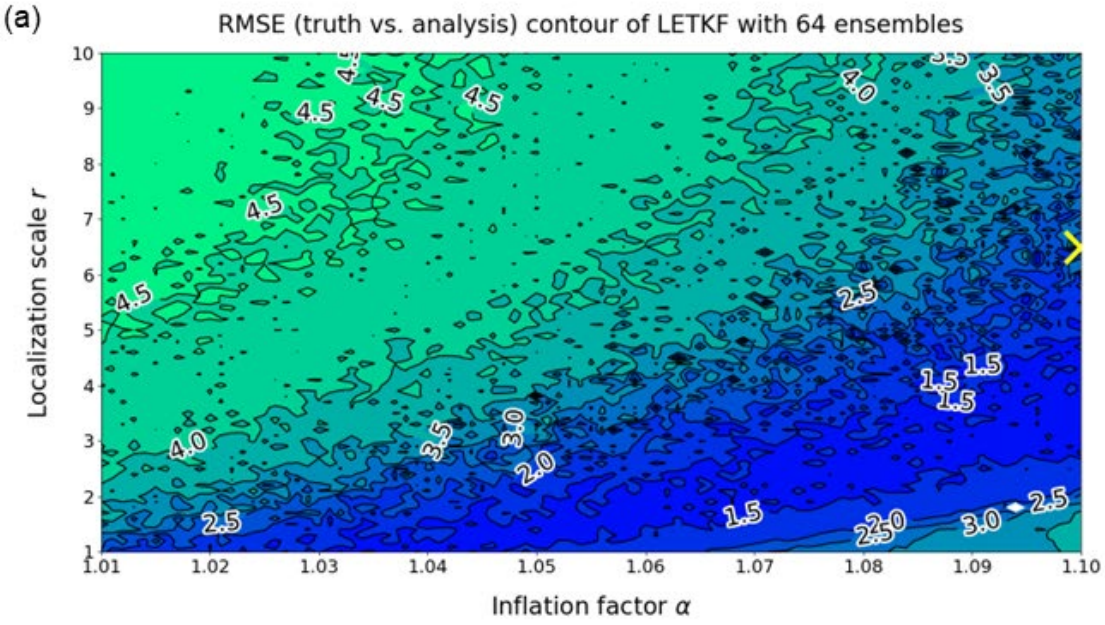
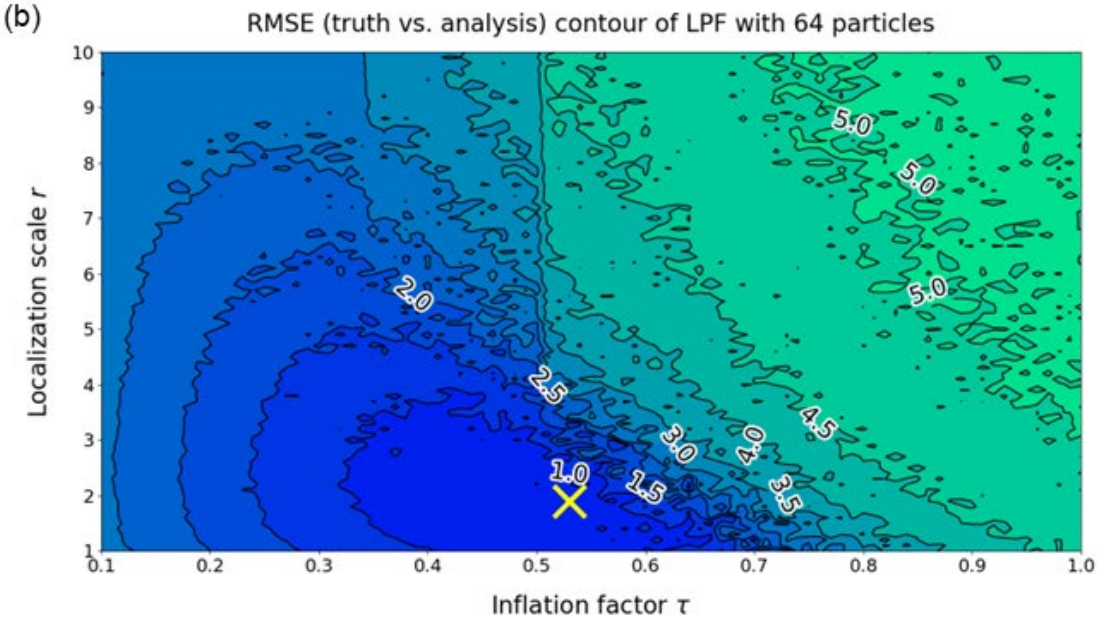


Fig. 2. Time series of the root mean square error and ensemble spread between the truth and the analysis (RMSE(t vs. a)) (a) for the local ensemble transform Kalman filter (LETKF) and (b) for the local particle filter (LPF) using 64 ensemble members (particles) and the nonlinear observation operators. The vertical axis shows the RMSE(t vs. a)

(blue line) and the ensemble spread (red line), while the horizontal axis indicates the assimilation cycle.


(a)
RMSE (truth vs. analysis) contour of LETKF with 64 ensembles
Localization scale r
Inflation factor α


(b)
RMSE (truth vs. analysis) contour of LPF with 64 particles
Localization scale r
Inflation factor τ

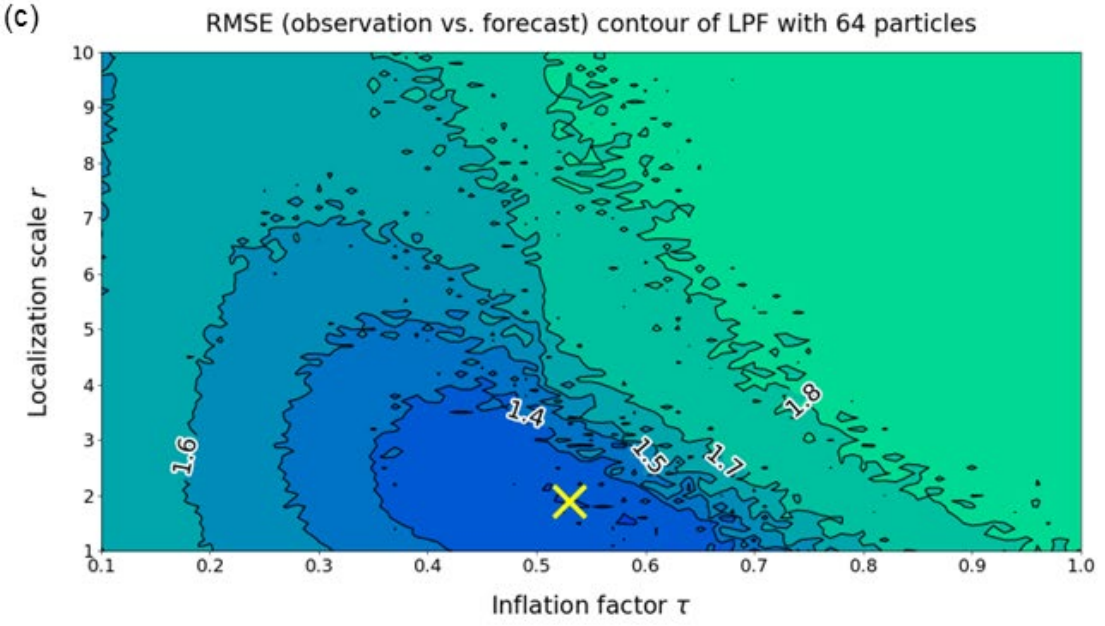

(c)
RMSE (observation vs. forecast) contour of LPF with 64 particles
Localization scale r
Inflation factor τ

Fig. 3. Response surface of root mean square error between the truth and the analysis (RMSE(t vs. a)) in the local ensemble transform Kalman filter (LETKF) and the local particle filter (LPF) using 64 ensemble members (particles) and the nonlinear observation operator. The closer the color is to green, the larger the RMSE(t vs. a), and the closer the color is to blue, the smaller the RMSE(t vs. a). The vertical axis shows the localization scale r, and the horizontal axis shows the inflation factor α and τ. The minimum error of 1.024 in the LETKF was obtained when r = 6.5 and α = 1.100 (cross mark). In addition, the minimum error of 0.586 in the LPF was obtained when r = 1.9 and τ = 0.53 (cross mark).

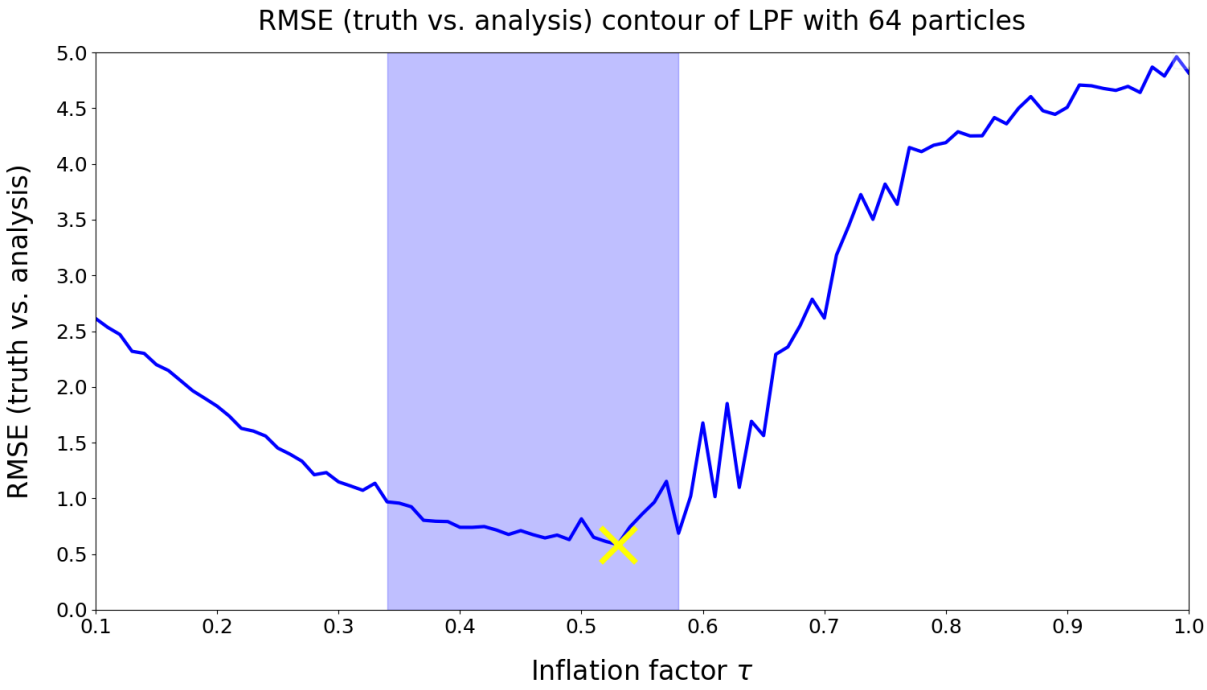


Fig. 4. Response surface of the root mean square error between the truth and the analysis (RMSE(t vs. a)) in the local particle filter using 64 ensemble members (particles) and the nonlinear observation operator. The vertical axis shows the RMSE(t vs. a), and the horizontal axis shows the inflation factor τ. The localization scale was fixed at r = 1.9, and the minimum error of 0.586 was obtained when τ = 0.53 (cross mark). In addition, the light blue shaded area indicates the optimal inflation factor range (τ = 0.34-0.58) where the RMSE(t vs. a) of the LPF is 1.0 or less (the filter operates stably).

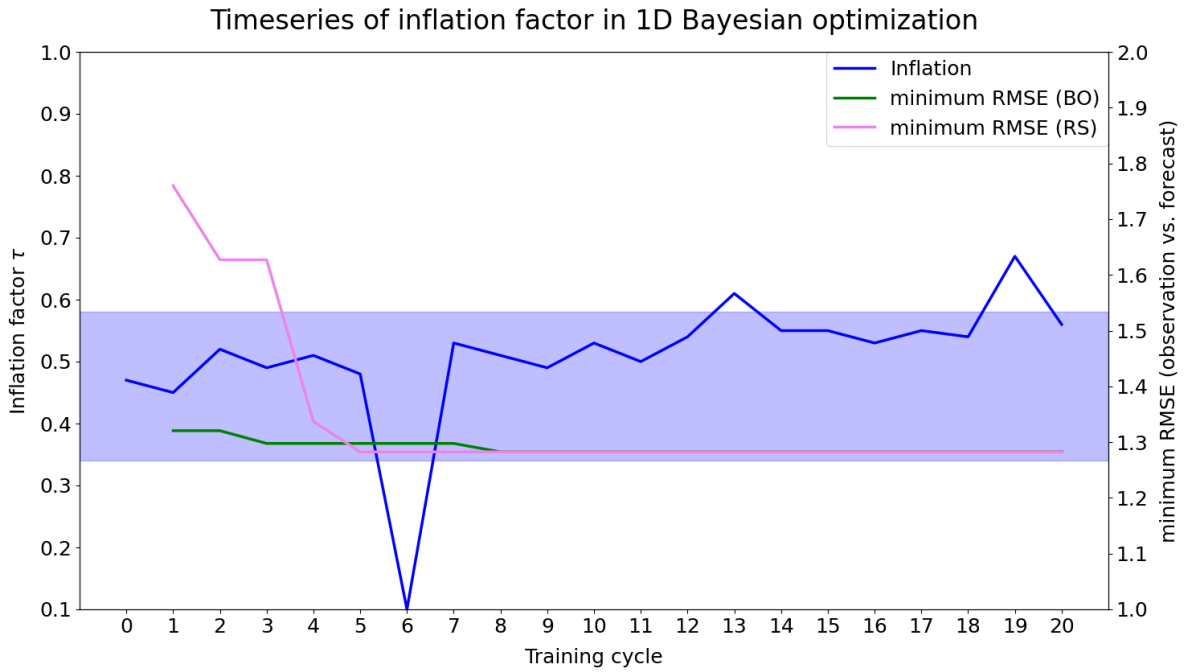


Fig. 5. Time series of the estimation by the 1-dimensional Bayesian optimization (BO). The blue line shows the inflation factor τ, the green line shows the minimum root mean square error between the observations and the forecasts (RMSE(o vs. f)) in the previous training cycle estimated by the BO, and the purple line shows the minimum RMSE(o vs. f) in the previous training cycle estimated by the random sampling. In addition, the light blue shaded area indicates the optimal inflation factor range (τ = 0.34-0.58) for which the root mean square error between the truth and the analysis in the local particle filter is 1.0 or less (the filter operates stably). The horizontal axis shows the training cycle, the first vertical axis shows τ, and the second vertical axis shows the minimum RMSE(o vs. f) in the previous training cycle. Note that, due to the convenience of conducting the OSSE using the BO estimation, the training cycles of τ and the RMSE(o vs. f) are shifted by one cycle.

1110

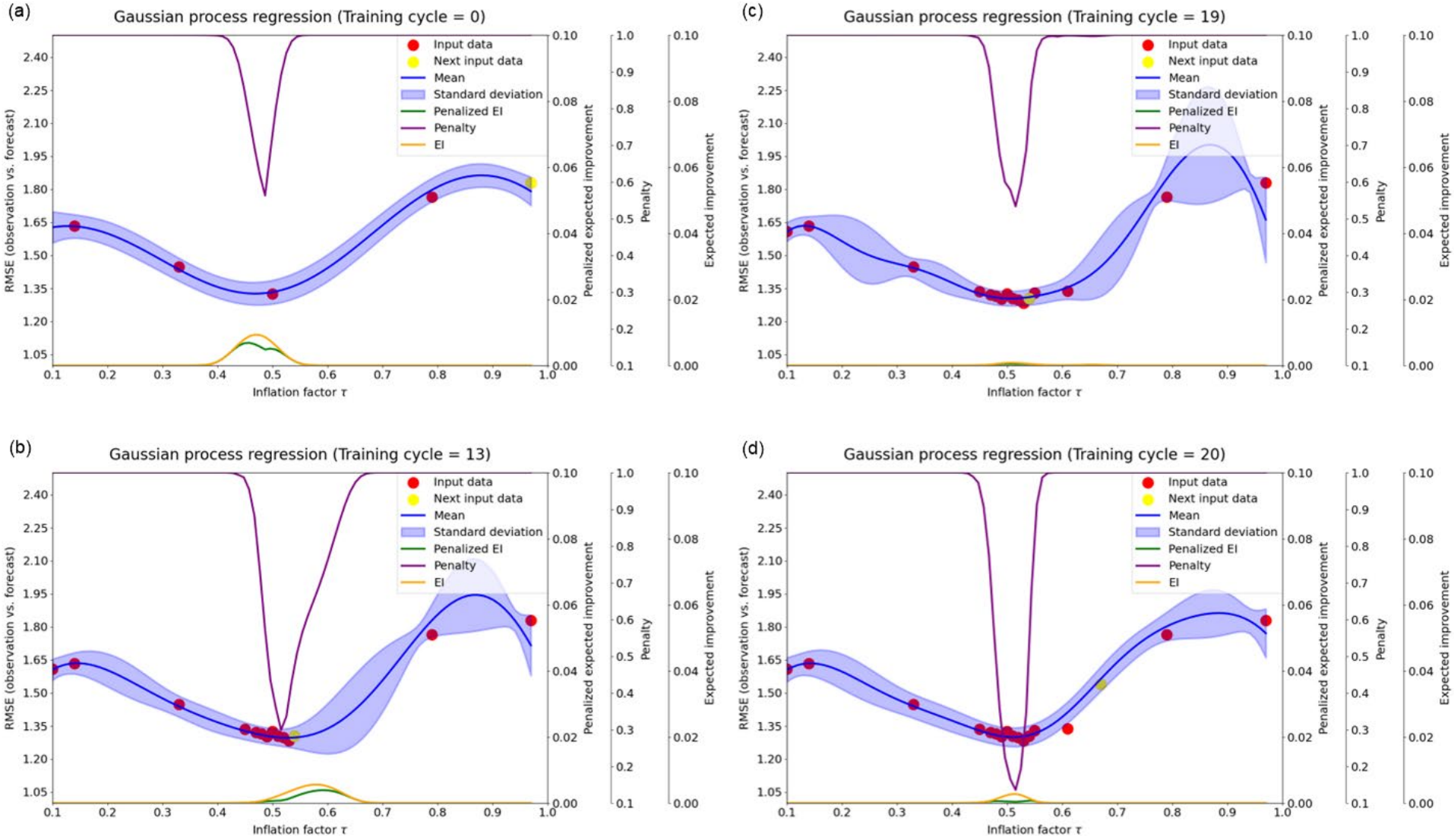

(a)
Gaussian process regression (Training cycle = 0)
(b)
Gaussian process regression (Training cycle = 13)
(c)
Gaussian process regression (Training cycle = 19)
(d)
Gaussian process regression (Training cycle = 20)
Input data
Next input data
Mean
Standard deviation
Penalized EI
Penalty
EI
RMSE (observation vs. forecast)
Penalized expected improvement
Penalty
Expected improvement
Inflation factor τ

Fig. 6. Prediction distribution of Gaussian process regression (GPR) using the inflation factor τ and the root mean square error between the observations and the forecasts (RMSE(o vs. f)) in the local particle filter as input and output data. The green line indicates the penalized expected improvement (EI), the purple line indicates the penalty, the yellow line indicates the EI, the red dots indicate the input data that has been explored, and the yellow dots indicate the input data explored in that training cycle. The horizontal axis is τ, the first vertical axis is the RMSE(o vs. f), the second vertical axis is the penalized EI, the third vertical axis is the penalty, and the fourth vertical axis is the EI. (a)-(d) are the prediction distributions for the 0th (i.e., when only the initial input data were given), 13th, 19th, and 20th training cycles, respectively.

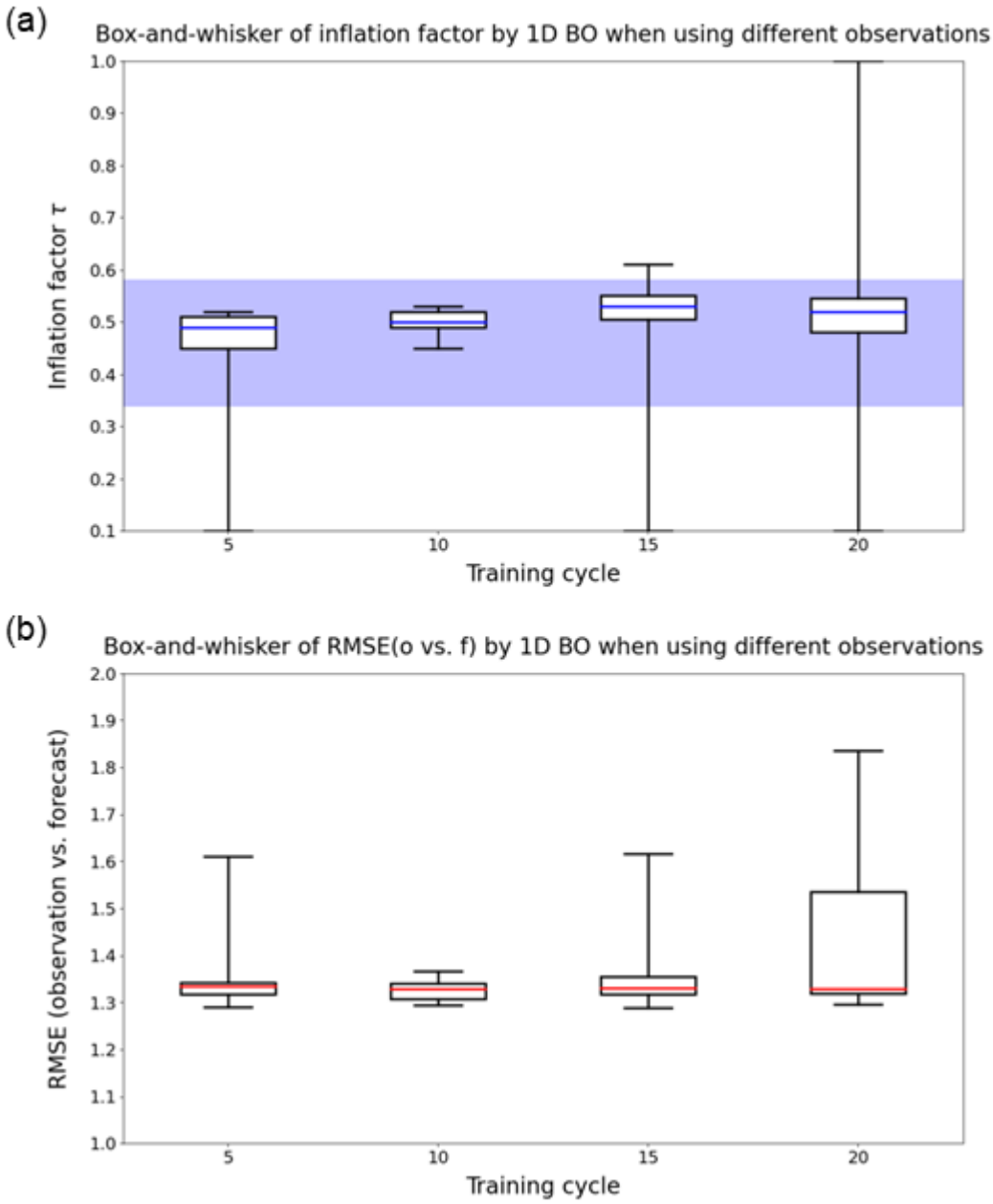


Fig. 7. Variation in the estimation by 1-dimensional Bayesian optimization for different observations. (a) Box plot of inflation factor τ. The blue line is the median, the lower edge of the box is the first quartile, the upper edge of the box is the third quartile, the lower edge of the whiskers is the minimum, and the upper edge of the whiskers is the maximum. In addition, the light blue shaded area indicates the optimum inflation factor range (τ = 0.34-0.58) for which the root mean square error between the observations

and forecasts (RMSE(o vs. f)) in the local particle filter is less than 1.0 (the filter operates stably). (b) Box plot of the RMSE(o vs. f). The red line is the median, and the other plots are the same as in (a). The vertical axes in (a) and (b) indicate τ and the RMSE(o vs. f), respectively. The horizontal axes indicate the number of training cycles.

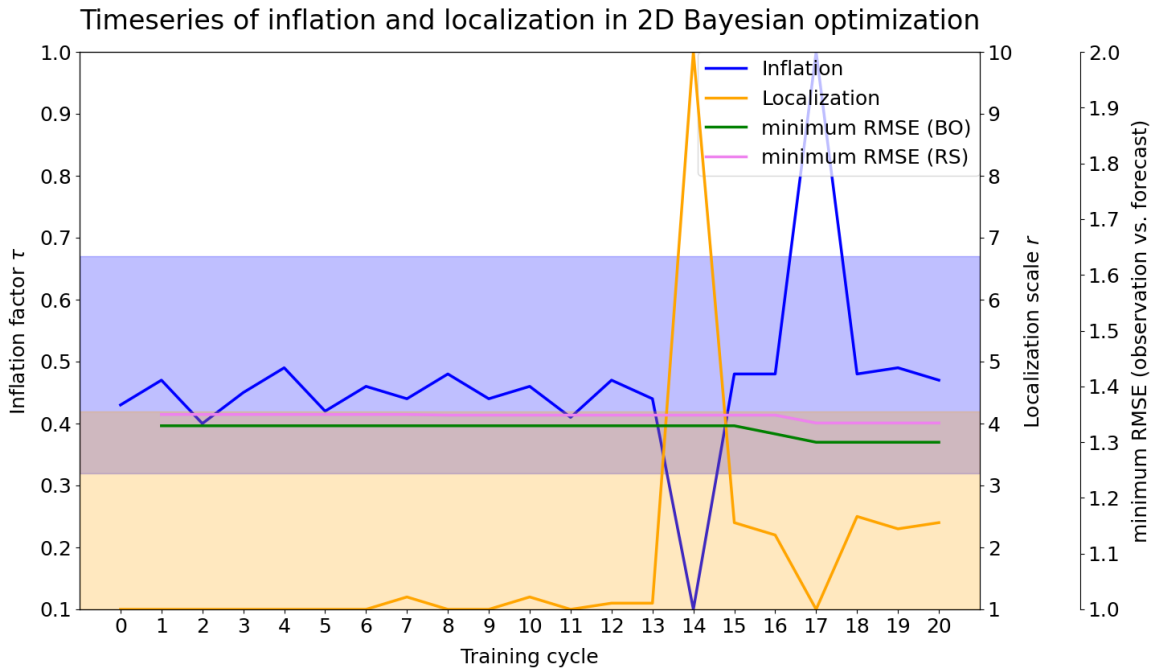


Fig. 8. Time series plot of the estimations by the 2-dimensional Bayesian optimization (BO). The blue line shows the inflation factor τ, the orange line shows the localization scale r, the green line shows the minimum root mean square error between the observations and the forecasts (RMSE(o vs. f)) in the previous training cycle estimated by the BO, and the purple line shows the minimum RMSE(o vs. f) in the previous training cycle estimated by the random sampling. The light blue shaded area indicates the optimal inflation factor range (τ = 0.32-0.67) for which the root mean square error between the truth and the analysis in the local particle filter is less than 1.0 (the filter operates stably). In addition, the light orange shaded area indicates the optimal localization scale range (r = 1.0-4.2). The light beige shaded area indicates the range that satisfies both the optimal inflation factor and localization scale. The horizontal axis represents the training cycle, the first vertical axis represents τ, the second vertical axis represents r, and the third vertical axis

represents the minimum RMSE(o vs. f). Note that due to the convenience of conducting the OSSE using the BO estimations, the training cycles of τ, r, and the RMSE (o vs. f) are shifted by one cycle.

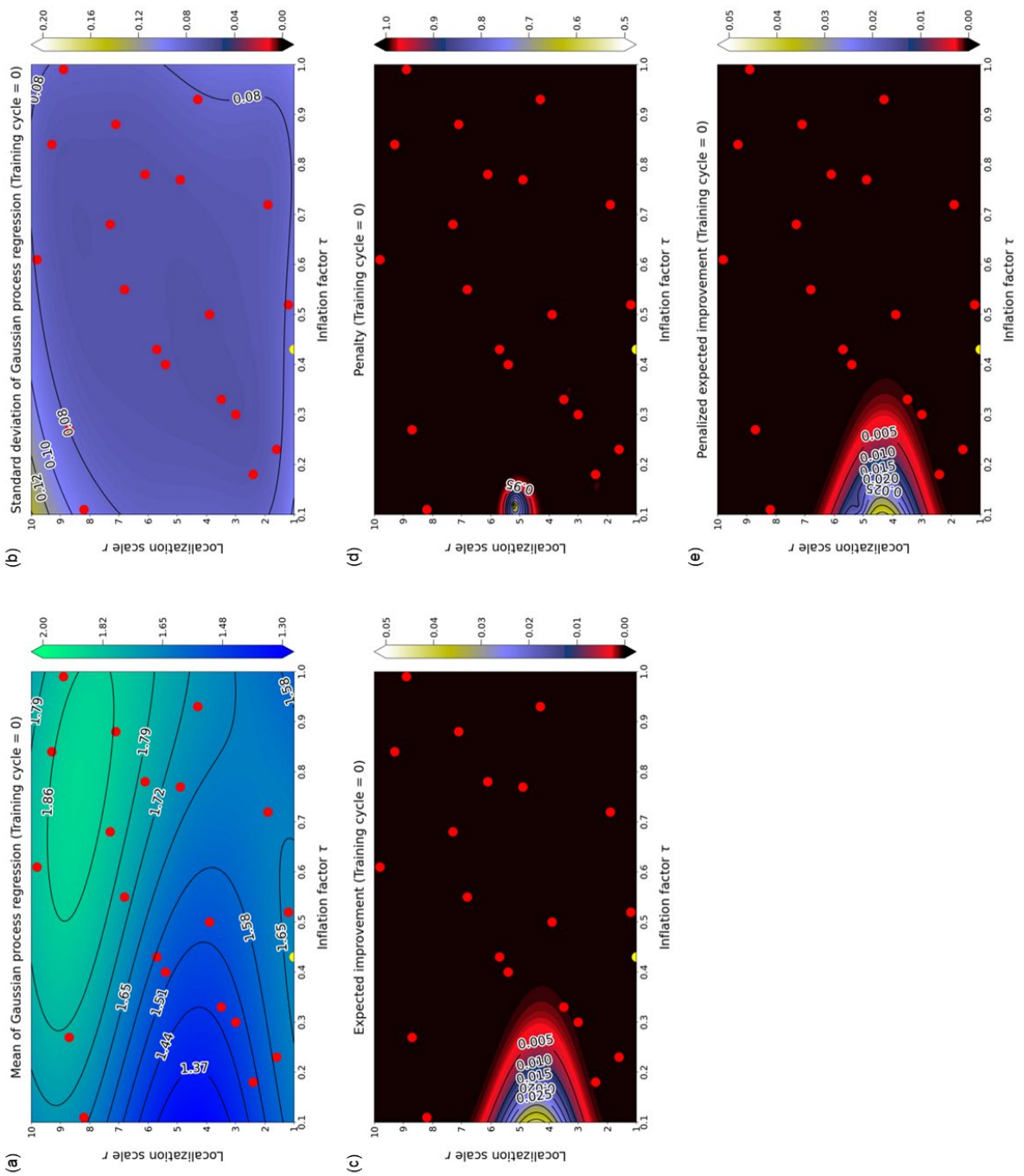


Fig. 9. Prediction distribution of Gaussian process regression (GPR) using the inflation factor τ and the localization scale r in the 0th training cycle (i.e., when only initial input data were given) and the root mean square error between observations and forecasts (RMSE(o vs. f)) in the local particle filter as input and output data. (a) is the GPR mean, (b) is the GPR standard deviation, (c) is the expected improvement (EI), (d) is the

1158 penalty, and (e) is the penalized EI prediction distribution. The red dots are input data
1159 obtained in advance using the LHS, and the yellow dots indicate the input data explored
1160 in that training cycle. The horizontal axis indicates τ, and the vertical axis indicates r.
1161

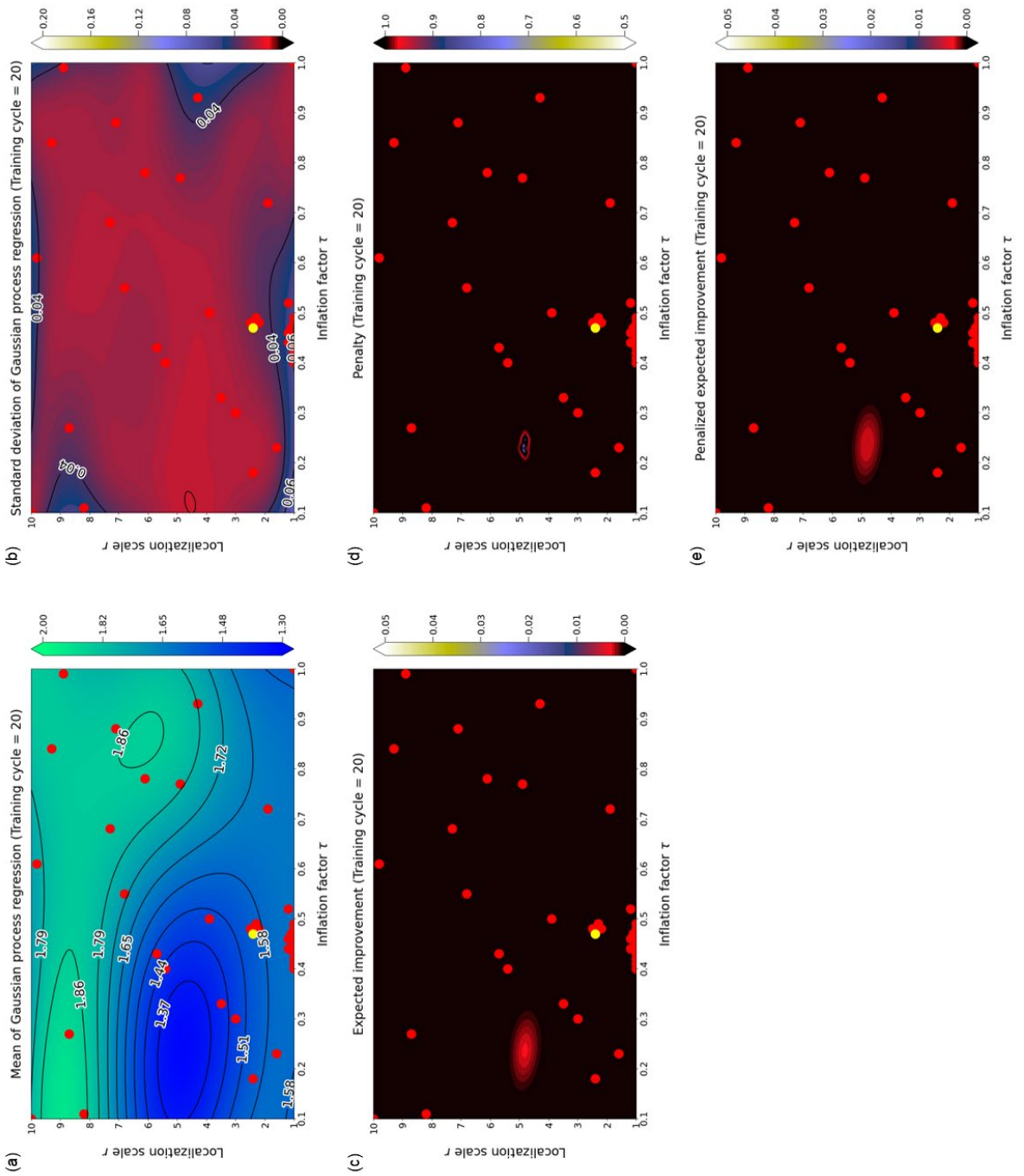
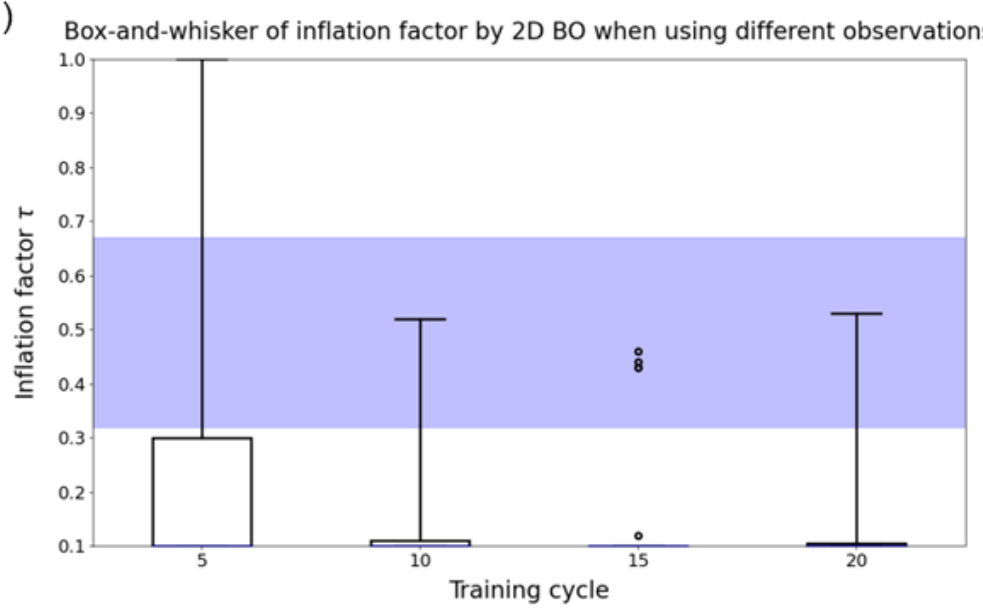


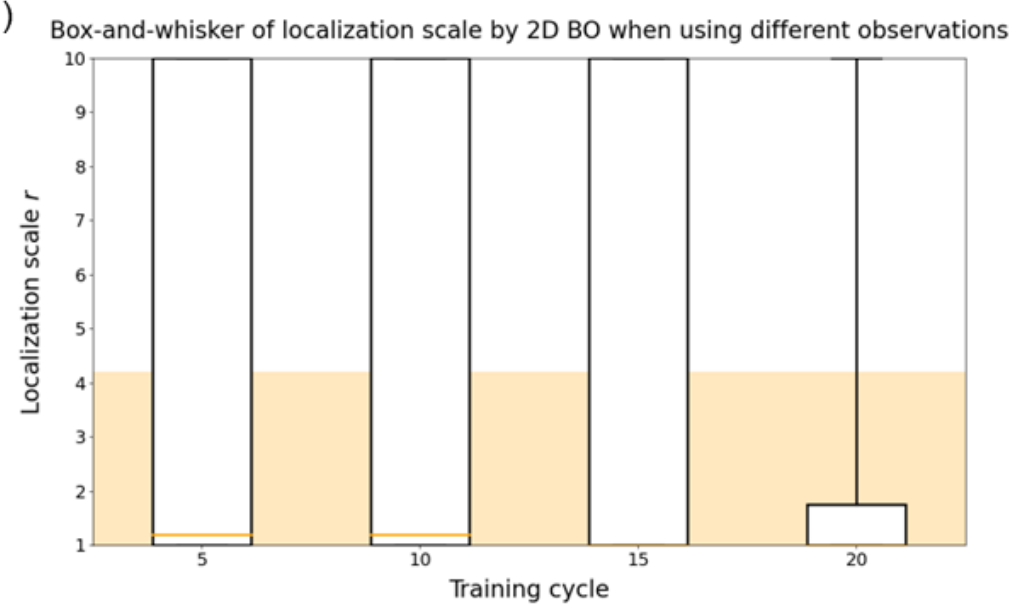
Fig. 10. Prediction distribution of Gaussian process regression (GPR) using the inflation factor τ and the localization scale r at the 20th training cycle, and the root mean square error between the observations and the forecasts (RMSE(o vs. f)) in the local particle filter as input and output data. (a) is the GPR mean, (b) is the GPR standard deviation, (c) is the expected improvement (EI), (d) is the penalty, and (e) is the penalized EI

prediction distribution. The training cycle in this figure corresponds to that in Fig. 8. The expected value and uncertainty of the RMSE(o vs. f) are obtained as the mean and standard deviation of the GPR. Red dots indicate input data that has been explored, and yellow dots indicate input data explored in that training cycle. The horizontal axis indicates τ, and the vertical axis indicates r.

(a)


Box-and-whisker of inflation factor by 2D BO when using different observations
Inflation factor τ
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
5
10
15
20
Training cycle

(b)


Box-and-whisker of localization scale by 2D BO when using different observations
Localization scale r
10
9
8
7
6
5
4
3
2
1
5
10
15
20
Training cycle

(c)

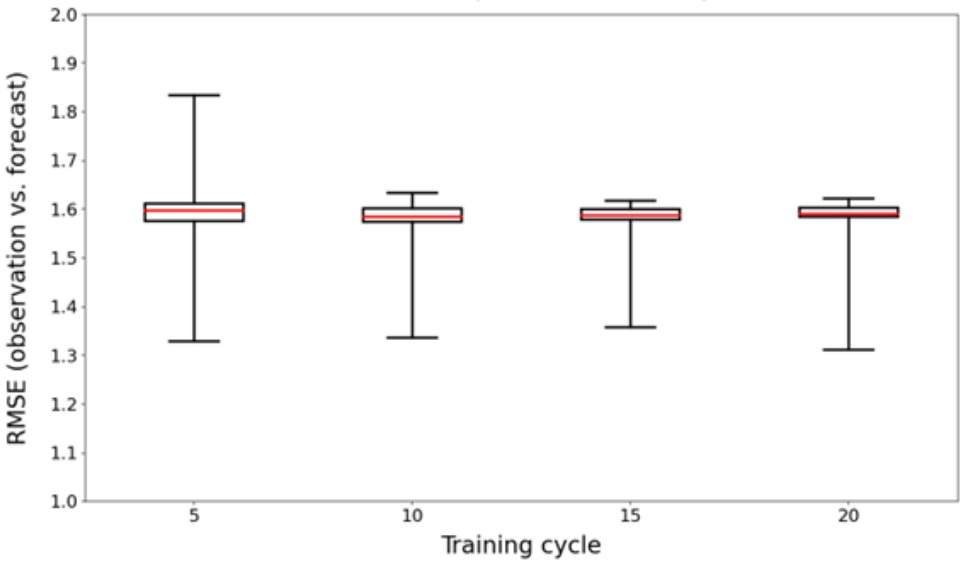

Box-and-whisker of RMSE(o vs. f) by 2D BO when using different observations
RMSE (observation vs. forecast)
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
5
10
15
20
Training cycle

Fig. 11. Variation in the estimation by the 2-dimensional Bayesian optimization for the different observation sets. (a) Box plot of inflation factor τ. The blue line is the median, the lower edge of the box is the first quartile, the upper edge of the box is the third quartile, the lower edge of the whiskers is the minimum, the upper edge of the whiskers is the maximum, and the dots indicate outliers. In addition, the light blue shade indicates the optimal inflation factor range (τ = 0.34-0.58) for which the root mean square error between the truth and the analysis (RMSE(t vs. a)) in the local particle filter is less than 1.0 (the filter operates stably). (b) Box plot of localization scale r. The orange line is the median, and the other plots are the same as in (a). In addition, the light orange shade indicates the optimal localization scale range (r = 1.0-4.2). (c) Box plot of the RMSE(o vs. f). The red line is the median, and the other plots are the same as in (a). The vertical axes in (a), (b), and (c) indicate τ, r, and the RMSE(o vs. f), respectively. The horizontal axes indicate the number of training cycles.

List of Tables

Table 1 Variation in the estimation by the Bayesian optimization with respect to the dimension of the response surface, the Lipschitz constant, and the number of initial input data. The rightmost column shows the minimum root mean square error between the observations and the forecasts for 20 training cycles, with the best cases for each dimension of the response surface highlighted in bold.

Table 1 Variation in the estimation by the Bayesian optimization with respect to the dimension of the response surface, the Lipschitz constant, and the number of initial input data. The rightmost column shows the minimum root mean square error between the observations and the forecasts for 20 training cycles, with the best cases for each dimension of the response surface highlighted in bold.

2-dimension (91 * 91)

Lipschitz constant	Number of initial input data	minimum RMSE
0.1	10	1.319
	20	1.302
	30	1.312
	40	1.319
0.5	10	1.319
	20	1.304
	30	1.316
	40	1.315
2.0	10	1.319
	20	**1.300**
	30	1.323
	40	1.332
10.0	10	1.319
	20	1.329
	30	1.354
	40	1.352

1-dimension (91)

Lipschitz constant	Number of initial input data	minimum RMSE
0.1	5	1.282
	10	1.282
	15	1.304
	20	1.282
0.5	5	**1.282**
	10	1.301
	15	1.298
	20	1.282

2.0	5	1.298
	10	1.301
	15	1.301
	20	1.282
10.0	5	1.301
	10	1.301
	15	1.301
	20	1.282

1201