Nonlin. Processes Geophys. Discuss., 1, 1283–1312, 2014
www.nonlin-processes-geophys-discuss.net/1/1283/2014/
doi:10.5194/npgd-1-1283-2014
© Author(s) 2014. CC Attribution 3.0 License.

Nonlinear Processes in Geophysics
Discussions
Open Access

This discussion paper is/has been under review for the journal Nonlinear Processes in Geophysics (NPG). Please refer to the corresponding final paper in NPG if available.

Bayesian optimization for tuning chaotic systems

M. Abbas[1], A. Ilin[1], A. Solonen[2], J. Hakkarainen[3], E. Oja[1], and H. Järvinen[4]

[1]Aalto University, School of Science, Espoo, Finland
[2]Lappeenranta University of Technology, Lappeenranta, Finland
[3]Finnish Meterological Institute, Helsinki, Finland
[4]University of Helsinki, Helsinki, Finland

Received: 26 May 2014 – Accepted: 14 July 2014 – Published: 4 August 2014

Correspondence to: M. Abbas (mudassar.abbas@aalto.fi)

Published by Copernicus Publications on behalf of the European Geosciences Union & the American Geophysical Union.

Abstract

In this work, we consider the Bayesian optimization (BO) approach for tuning parameters of complex chaotic systems. Such problems arise, for instance, in tuning the sub-grid scale parameterizations in weather and climate models. For such problems, the tuning procedure is generally based on a performance metric which measures how well the tuned model fits the data. This tuning is often a computationally expensive task. We show that BO, as a tool for finding the extrema of computationally expensive objective functions, is suitable for such tuning tasks. In the experiments, we consider tuning parameters of two systems: a simplified atmospheric model and a low-dimensional chaotic system. We show that BO is able to tune parameters of both the systems with a low number of objective function evaluations and without the need of any gradient information.

1 Introduction

In climate and numerical weather prediction models, accurate simulation and prediction depends upon the selection of optimal tuning parameters. A typical case is the tuning of closure parameters in climate models which describe processes that are imprecisely modeled due to the restrictive grid used for solving the differential equations (Järvinen et al., 2010; Schirber et al., 2013). Similarly, the tuning of parameters which control the stochastic physics components in ensemble prediction systems is a non-trivial task (Leutbecher and Palmer, 2008). Designing efficient procedures for tuning such model parameters is a topic of active research (see, e.g., Annan and Hargreaves, 2007; Solonen et al., 2012; Hauser et al., 2012; Hakkarainen et al., 2013).

The tuning procedure is generally based on a performance metric which measures how well the tuned model fits the data. For example, in numerical weather prediction (NWP), tuning is done by optimizing measures related to forecast skills, while in climate models, tuning is based on optimization criteria which often compare some summary

NPGD
1, 1283–1312, 2014

statistics (spatial and temporal averages) of the model simulation to observed statistics. Evaluating the performance metrics is computationally expensive, since it requires complex model simulations over the observation period.

One of the major difficulties of the tuning process is the high computational cost of the optimization procedure: for every candidate value of the tuning parameters, one has to perform computationally heavy simulations of a complex physical model. Another difficulty is that in many cases the objective function is noisy: two evaluations with the same parameter values lead to different objective function values. Such a situation arises, for instance, when the goal is to tune stochastic systems, such as ensemble prediction systems used to quantify uncertainties in numerical weather predictions (Solonen and Järvinen, 2013). Also, applying stochastic filtering methods for parameter estimation in state space models yields noisy objective functions, as discussed in Hakkarainen et al. (2012).

In this paper, we study the methodology called Bayesian optimization (BO) in the problem of parametric tuning of chaotic systems such as climate models and NWP. In BO, the parameter values where the objective function is evaluated are carefully chosen so that we learn as much as possible about the underlying function. As a result, the optimum can often be found with a small number of function evaluations. We discuss both deterministic and noisy tuning objectives.

We perform two studies: first, we consider tuning of a simplified atmospheric model with a noiseless objective function. The tuned model is a two-layer quasi-geostrophic model with four tuned parameters which define the model error covariance of the corresponding data assimilation system. Second, we consider the problem of tuning a chaotic system with a noisy likelihood function. We use the parameterized Lorenz 95 model as a test model, similarly to previous studies. The goal is to explore the suitability of the BO methodology for tuning full scale climate and weather models.

There are other approaches besides BO to accelerate computations with objective functions that are expensive to evaluate. Various surrogate modeling techniques attempt to describe the parameter-to-output dependence with empirical approximative

NPGD
1, 1283–1312, 2014

models that are cheap to evaluate. Techniques range from polynomial chaos expansions (Marzouk and Xiu, 2009) to GP models (Rasmussen and Williams, 2006), which are also applied in the BO method. In BO, instead of first building a surrogate model and then fixing it for further calculations, the goal is to design the points where the objective function is evaluated on the fly so that the potential of the new point in improving the current best value is maximized. That is, BO is directly built for solving optimization problems efficiently, not to represent the objective function efficiently in a selected region of the parameter space.

The BO method resembles classical response surface techniques for experimental optimization (Box and Draper, 1987), where local quadratic models are used to guide a sequence of experiments to obtain an optimal response. BO uses GP, which is more flexible in describing the behavior of the underlying objective. Also, BO uses a different way for selecting the next point where the objective function is evaluated. We use the Gaussian processes (GP) based BO which has been previously demonstrated as a very efficient and flexible approach in optimization of computationally heavy to compute models in several papers (see, e.g., Brochu et al., 2010a; Lizotte et al., 2012).

The outline of the paper is as follows. In Sect. 2, we present the basic ideas behind the Bayesian optimization. In Sect. 3, we formulate the likelihood for a complex system represented as a state-space model. In Sect. 4, we consider the case of tuning a simplified atmospheric model with a noiseless objective function. In Sect. 5, we demonstrate tuning of a chaotic system with a noisy likelihood. We conclude in Sect. 6.

2 Bayesian optimization

The goal of Bayesian optimization is to find the extrema of black-box functions, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ that are expensive to evaluate (see, e.g., reviews by Brochu et al., 2010a; Snoek et al., 2012). Here, f which is also called the objective function typically does not have a closed form solution. In BO, the objective function is modeled as a random function whose distribution describes our knowledge of the function, given a set of

NPGD
1, 1283–1312, 2014

function evaluations $\{\boldsymbol{\theta}_i, f(\boldsymbol{\theta}_i)\}_{i=1,\ldots,t}$. The posterior distribution over f is handled using the standard Gaussian process methodology which allows for evaluating the mean and variance of objective function values $f(\boldsymbol{\theta})$ in any location $\boldsymbol{\theta}$ at any optimization step t. This information is used to propose a new input location $\boldsymbol{\theta}_{t+1}$ which has the largest potential to improve the current best value of the objective function. In the following, we assume that the objective function is being maximized.

The search of the new point where the objective function has to be evaluated is done by optimizing a complementary function called *acquisition function*, that measures the potential to improve the current best point. The two statistics often utilized in designing acquisition functions are the predictive mean and the predictive variance of f at possible location $\boldsymbol{\theta}$. In designing new points where the function is evaluated, one typically has to choose between two extremes: sampling from locations of high predicted mean value (*exploitation* strategy) and locations of high uncertainty value (*exploration* strategy), which is illustrated in Fig. 1. BO provides a tool that is able to automatically trade off between exploration and exploitation, which often yields a reduced number of objective function evaluations needed (Lizotte et al., 2012). It can also prove useful for objective functions with multiple local optima, and noise in the objective function can be handled in a straight-forward manner.

Even though BO was introduced in the seventies (Mockus et al., 1978), the methodology has been under active development in the recent years due to its successful application to a number of machine learning problems (Boyle, 2007; Frean and Boyle, 2008; Lizotte, 2008; Osborne et al., 2009; Brochu et al., 2010a; Snoek et al., 2012). To our knowledge, BO has not been studied in connection with parameter tuning in complex dynamic models.

BO is a methodology that suits very well to the problem of complex system tuning. First, evaluation of the objective function in this task requires computationally expensive model simulations, typically several days to complete. Second, the sampling region of the parameters is often unknown and it is manually selected using expert knowledge. Third, the gradient information is unavailable and direct optimization is infeasible.

NPGD
1, 1283–1312, 2014

2.1 Gaussian processes

The Gaussian processes (GP) methodology is the key element of BO, as it is an elegant tool for describing distributions over unknown functions (Rasmussen and Williams, 2006). In this methodology, the prior distribution over f is chosen such that the function values $\boldsymbol{f} = [f(\boldsymbol{\theta}_1), \ldots, f(\boldsymbol{\theta}_t)]$ are assumed to be normally distributed:

$$\boldsymbol{f} \mid \eta \sim \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \mathbf{K}_f) \tag{1}$$

where the mean is typically taken to be zero and the covariance matrix $\mathbf{K}_f$ is constructed such that its ijth element is computed using covariance function $k(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j|\eta)$ for the corresponding inputs $\boldsymbol{\theta}_i$, $\boldsymbol{\theta}_j$ and hyperparameters η. The covariance function k is the key element of the GP modeling: it encodes our assumptions about the behavior of function f, such as its smoothness properties.

In our experiments, we use the squared exponential covariance function which is one of the most common covariance functions:

$$k(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j|\eta) = \sigma^2 \prod_k \exp\left(\frac{-(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)^2}{2l_k^2}\right), \tag{2}$$

where σ^2 is the scaling parameter which specifies the magnitudes of the function values and l_k is the parameter defining the smoothness of the function. Both belong to hyperparameters η.

At every iteration of the BO algorithm, the properties of the unknown function f are learned by adapting the hyperparameters η to fit well the observed data $\{\boldsymbol{\theta}_i, f(\boldsymbol{\theta}_i)\}_{i=1,\ldots,t}$. This is typically done by maximizing the marginal likelihood of the hyperparameters η. Then, GP is used to evaluate the predictive distribution over the function values $f(\boldsymbol{\theta}_{\text{new}})$ at any new location $\boldsymbol{\theta}_{\text{new}}$. Assuming that the observed values of the objective function are noisy and the noise is Gaussian, the predictive distribution is normal:

$$p(f(\boldsymbol{\theta}_{\text{new}})|\boldsymbol{f}, \eta) \sim \mathcal{N}\left(\mu(\boldsymbol{\theta}_{\text{new}}), \sigma^2(\boldsymbol{\theta}_{\text{new}})\right) \tag{3}$$

NPGD
1, 1283–1312, 2014

with the mean and variance given by

$$\mu(\boldsymbol{\theta}_{\text{new}}) = \boldsymbol{k}_{\text{new}}^T (\mathbf{K}_{\boldsymbol{f}} + \boldsymbol{\Sigma})^{-1} \boldsymbol{f} \tag{4}$$

$$\sigma^2(\boldsymbol{\theta}_{\text{new}}) = k(\boldsymbol{\theta}_{\text{new}}, \boldsymbol{\theta}_{\text{new}}) - \boldsymbol{k}_{\text{new}}^T (\mathbf{K}_{\boldsymbol{f}} + \boldsymbol{\Sigma})^{-1} \boldsymbol{k}_{\text{new}}, \tag{5}$$

where $\boldsymbol{k}_{\text{new}} = [k(\boldsymbol{\theta}_{\text{new}}, \boldsymbol{\theta}_1), \ldots, k(\boldsymbol{\theta}_{\text{new}}, \boldsymbol{\theta}_t)]^T$. $\boldsymbol{\Sigma}$ is the covariance matrix of the noise in the objective function, which is often parameterized as $\sigma^2\mathbf{I}$ and estimated in the optimization procedure. For more details on training GP, see, for example, the book by Rasmussen and Williams (2006).

2.2 Acquisition functions

The acquisition functions are used to search for a new location $\boldsymbol{\theta}_{\text{new}}$ which has the highest potential to improve the best value of the objective function obtained so far, denoted by

$$\mu^+ = \max_t \mu(\boldsymbol{\theta}_t).$$

At each BO iteration, the new sample is chosen to maximize the value of the acquisition function:

$$\boldsymbol{\theta}_{\text{new}} = \arg\max_{\boldsymbol{\theta}} g(\boldsymbol{\theta}),$$

where

$$g(\boldsymbol{\theta}) = g(\mu^+, \mu(\boldsymbol{\theta}), \sigma(\boldsymbol{\theta})).$$

High values of the acquisition function correspond to regions where the expected value $\mu(\boldsymbol{\theta})$ of the objective function value is high or where the prediction uncertainty $\sigma(\boldsymbol{\theta})$ is high or both. Deciding which areas have the largest potential is known as the exploration vs exploitation trade off (see, e.g., Jones, 2001).

NPGD
1, 1283–1312, 2014

The choice of possible acquisition criteria is quite large with developments still taking place (see, e.g., Lizotte et al., 2012; Brochu et al., 2010b). Here, we illustrate two of the most popular acquisition functions called *probability of improvement* (Kushner, 1964) and *expected improvement* (Mockus, 1989). The probability of improvement (PI) is formulated as

$$g_{\mathrm{PI}}(\boldsymbol{\theta}) = \Phi\left(d/\sigma(\boldsymbol{\theta})\right) \tag{6}$$

$$d = \mu(\boldsymbol{\theta}) - \mu^{+} - \xi \tag{7}$$

where $\Phi(\cdot)$ is the normal cumulative distribution function. When $\xi = 0$, $g_{\mathrm{PI}}(\boldsymbol{\theta})$ is simply the probability of improving the best value μ^{+} by taking a sample at location $\boldsymbol{\theta}$. The problem with using $\xi = 0$ is that PI favors locations that have even a slight improvement over the current best μ^{+}. This means that in this setting PI has a higher tendency to exploit rather than explore and it practically always gets stuck at a local optimum (Lizotte et al., 2012). The tuning parameter $\xi > 0$ allows for tuning PI in order to reduce this problem. However, the choice of ξ is always subjective, although it has a great impact on the performance. For a detailed study of the effect of ξ, we recommend the work of Lizotte et al. (2012).

The expected improvement (EI) is formulated as

$$g_{\mathrm{EI}}(\boldsymbol{\theta}) = \langle f(\boldsymbol{\theta}) - \mu^{+} \rangle \tag{8}$$

$$= d\Phi\left(d/\sigma(\boldsymbol{\theta})\right) + \sigma(\boldsymbol{\theta})\phi\left(d/\sigma(\boldsymbol{\theta})\right) \tag{9}$$

where $\langle\cdot\rangle$ denotes expectation, d is defined in Eq. (7) and $\phi(\cdot)$ is the normal probability density function. The EI criterion is derived as the expected difference between the function value in a new location $f(\boldsymbol{\theta})$ and the current best μ^{+}. Thus, EI aims at maximizing f by the biggest margin and yields optimization which is less prone to getting stuck in a local optimum. Nevertheless, using a tuning parameter $\xi > 0$ allows for control of the exploration vs exploitation trade-off (Lizotte et al., 2012).

Figure 4 illustrates a one-dimensional maximization procedure using BO. We start with three function evaluations and show the sampled points, the GP fits (with the

NPGD
1, 1283–1312, 2014

red line) and the posterior uncertainty (with the pink filled areas). The acquisition functions are shown at the subplots below. The new location (marked with a magenta vertical line) is chosen so that it maximizes the acquisition function. As the optimization proceeds, we collect more samples and finally find the maximum of the objective function. One can notice that, compared to PI, EI favors exploration as it samples from regions with higher uncertainty. The objective function approximation obtained with EI improves much faster compared to PI. EI is also able to find the global maximum earlier. In this case, one could argue that using a larger value of ξ in PI could result in more exploration and faster optimization. However, choosing the right ξ value for PI is generally difficult. We use the EI acquisition function with BO for tuning parameters in our experimental study.

As stated earlier, the acquisition function is maximized at every step of BO in order to find a sample with the best potential. The acquisition function typically has multiple local optima (see Fig. 4) and the ability to find the *global* optimum of the acquisition function is extremely important for the efficiency of BO. Thus, global optimization procedures are typically used, which can be computationally demanding. Nevertheless, this procedure is usually far cheaper computationally because the global optimization only evaluates the GP and does not touch the objective function, which is computationally the most expensive part. Any global optimization method can be used in this task. In this work we have used the *DIRECT* method by Jones et al. (1993).

3 Filtering methods for likelihood evaluation

In tuning chaotic systems, we use the approach where the likelihood is computed using filtering techniques (Hakkarainen et al., 2012). The tuned system is represented as a state-space model

$$\boldsymbol{s}_k = \mathcal{M}(\boldsymbol{s}_{k-1}) + \boldsymbol{E}_k \tag{10}$$

$$\boldsymbol{y}_k = \mathcal{K}(\boldsymbol{s}_k) + \boldsymbol{e}_k, \tag{11}$$

NPGD
1, 1283–1312, 2014

where $\boldsymbol{s}_k$ is the state of the model, $\mathcal{M}$ is the forward model which can be implemented by a solver of partial differential equations and $\mathcal{K}$ is the observation operator. $\boldsymbol{E}_k$ and $\boldsymbol{e}_k$ are noise terms which account for model imperfection and observation noise. In climate science applications, model parameters $\boldsymbol{\theta}$ usually appear in the formulation of $\mathcal{M}$ or/and they can govern the distribution of the model error term $\boldsymbol{E}_k$.

Filtering methods evaluate the likelihood by sequentially estimating the dynamically changing model state $\boldsymbol{s}_k$ for a given observation sequence $\boldsymbol{y}_{1:k} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$. Filters work by iterating two steps: prediction and update. In the prediction step, the current distribution of the state is evolved with the dynamical model to the next time step. The predictive distribution is given by the integral

$$p(\boldsymbol{s}_k|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta}) = \int p(\boldsymbol{s}_k|\boldsymbol{s}_{k-1}, \boldsymbol{\theta}) p(\boldsymbol{s}_{k-1}|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{s}_{k-1}\,. \tag{12}$$

As this integral generally does not have a closed form solution, it is usually approximated in one way or another. This yields different filtering techniques such as extended Kalman filter, ensemble Kalman filter, particle filter and so on.

The state distribution is updated using a new observation $\boldsymbol{y}_k$ using the Bayes rule:

$$p(\boldsymbol{s}_k|\boldsymbol{y}_{1:k}, \boldsymbol{\theta}) \propto p(\boldsymbol{y}_k|\boldsymbol{s}_k, \boldsymbol{\theta}) p(\boldsymbol{s}_k|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta}). \tag{13}$$

This posterior is used inside the integral Eq. (12) to obtain the prior for the next time step.

The likelihood $p(\boldsymbol{y}_{1:n}|\boldsymbol{\theta})$ of the model parameters can be computed from the quantities evaluated in the filtering procedure:

$$p(\boldsymbol{y}_{1:K}|\boldsymbol{\theta}) = p(\boldsymbol{y}_1|\boldsymbol{\theta}) \prod_{k=2}^{K} p(\boldsymbol{y}_k|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta}), \tag{14}$$

where $p(\boldsymbol{y}_k|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta})$ is calculated based on the marginal posterior of the states:

$$p(\boldsymbol{y}_k|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta}) = \int p(\boldsymbol{y}_k|\boldsymbol{s}_k, \boldsymbol{\theta}) p(\boldsymbol{s}_k|\boldsymbol{y}_{1:k-1}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{s}_k\,.$$

NPGD
1, 1283–1312, 2014

Our goal is to search for parameters that maximize this likelihood $f(\boldsymbol{\theta}) = p(\boldsymbol{y}_{1:K}|\boldsymbol{\theta})$ in Eq. (14).

Extended Kalman filter (EKF) is a filtering technique in which the integrals are approximated by linearization of the forward model $\mathcal{M}$ and the observation operator $\mathcal{K}$ around the current state estimate. Assuming that the observation error is normally distributed with zero mean and covariance matrix $\mathbf{R}_k$, the linearization yields:

$$p(\boldsymbol{y}_{1:n}|\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\sum_{k=1}^{n} \boldsymbol{r}_k^T \left(\mathbf{C}_k^y(\boldsymbol{\theta})\right)^{-1} \boldsymbol{r}_k + \log|\mathbf{C}_k^y(\boldsymbol{\theta})|\right) \tag{15}$$

$$\mathbf{C}_k^y(\boldsymbol{\theta}) = \mathbf{K}_k\left(\mathbf{M}_k \mathbf{C}_{k-1}^{\text{est}} \mathbf{M}_k^T + \mathbf{Q}_k(\boldsymbol{\theta})\right)\mathbf{K}_k^T + \mathbf{R}_k \tag{16}$$

where $\boldsymbol{r}_k = \boldsymbol{y}_k - \mathcal{K}(\boldsymbol{s}_k^p)$ are the prediction residuals, $\mathbf{M}_k$ and $\mathbf{K}_k$ are the linearization of $\mathcal{M}$ and $\mathcal{K}$ operators, respectively, $\mathbf{C}_{k-1}^{\text{est}}$ is the covariance matrix of the state distribution Eq. (13) at time $k-1$ and $|\cdot|$ denotes the matrix determinant.

When the dimensionality of the tuned model is too large, the extended Kalman filter suffers from memory issues. Another problem is that linearization is often too cumbersome for highly complex models. In such scenarios, more sophisticated techniques like stochastic ensemble Kalman filters (EnKF) are often used for filtering.

The basic idea of EnKF is that the posterior distribution of the states is approximated using sample statistics, which are computed using a relatively small number of ensembles propagated by the model at every assimilation step. Stochastic filters involve random perturbations of the model states and observations, which introduces randomness in the likelihood evaluation. More details on EnKF can be found, for example, in (Evensen, 2007).

From the explanation above, we can see that such likelihood evaluation techniques using filtering methods is the example of a target that is very heavy to evaluate, and thus, efficient optimization techniques are needed.

NPGD
1, 1283–1312, 2014

4 Parameter tuning of an atmospheric model with "noiseless" likelihood evaluations

In the following experiment, we tune a model of synoptic-scale chaotic dynamics. The likelihood is evaluated using the extended Kalman filter, which results in noiseless likelihood evaluations.

4.1 A two-layer quasi-geostrophic model

The quasi-geostrophic (QG) model simulates fluid motion dynamics on a rotating cylinder (see, e.g., Fisher et al., 2011). The chaotic nature of the dynamics generated by the QG model is shown in Vannitsem and Nicolis (1997). In our case, the system includes two atmospheric layers (see Fig. 5) indicated as the top layer and the bottom layer. The orography in our model is such that there is a "hill" formation which affects the flow in the bottom layer. The system is simulated on a uniform grid for each layer so that the values at the top and bottom of the cylinder are set to pre-defined values (as boundary conditions).

The model dynamics are governed by the potential vorticity equations

$$\frac{\mathcal{D}_1}{\mathcal{D}t}\left(\nabla^2\psi_1 - F_1(\psi_1 - \psi_2) + \beta y\right) = 0, \tag{17}$$

$$\frac{\mathcal{D}_2}{\mathcal{D}t}\left(\nabla^2\psi_2 - F_2(\psi_2 - \psi_1) + \beta y + R_s\right) = 0, \tag{18}$$

where ψ_i denotes the model state vector called stream function and index i specifies the top atmospheric layer ($i = 1$) and the bottom layer ($i = 2$). $\mathcal{D}_i$ denotes the substantial derivatives for latitudinal wind u_i and longitudinal wind v_i:

$$\frac{\mathcal{D}_i \cdot}{\mathcal{D}t} = \frac{\partial \cdot}{\partial t} + u_i \frac{\partial \cdot}{\partial x} + v_i \frac{\partial \cdot}{\partial y}.$$

Parameters R_s and β denote dimensionless orography component and the northward gradient of the Coriolis parameter f_0.

NPGD
1, 1283–1312, 2014

The relationship between the model physic attributes and parameters F_1 and F_2 in Eqs. (17) and (18) is defined by

$$F_1 = \frac{f_0^2 L^2}{\acute{g} D_1},\ F_2 = \frac{f_0^2 L^2}{\acute{g} D_2},\ \acute{g} = g\frac{\Delta\theta}{\bar{\theta}},$$
$$R_s = \frac{S(x,y)}{\eta D_2},\ \beta = \beta_0 \frac{L}{U},$$

where D_1 and D_2 are the depths of the two layers, $\Delta\theta$ defines the potential temperature change on the layer interface, $\bar{\theta}$ is the mean potential temperature, g is acceleration of gravity, $\eta = \frac{U}{f_0 L}$ is the Rossby number associated with the defined system, and $S(x,y)$ and β_0 are dimensional representations of $R_s(x,y)$ and β, respectively. We used the implementation of the two-layer QG-model developed by Bibov (2011).

4.2 Experimental setup

The described QG-model is used to formulate a synthetic problem of chaotic system tuning. The data used in the tuning process are generated by a QG resolved on a dense 120 × 62 grid with the following parameters:

- layer depths are $D_1 = 6000$ units and $D_2 = 4000$ units
- distance between the grid points is 100 000 units.

This is our "true" system which is only used for generating the data.

The tuned system is a model which is governed by the same equations but it is resolved on a sparser grid 40 × 20 with the distance between grid points to be 300 000 units. This truncation of the grid size is a common practice in actual climate model testing. Thus, bias is being added to our tuning model because the fast processes affecting the observations on the finer scale will remain unmodeled.

The tuned system is represented as a state-space model Eqs. (10) and (11) where the forward model $\mathcal{M}$ is implemented by a solver of Eqs. (17) and (18). The state is

NPGD
1, 1283–1312, 2014

the 1600-dimensional vector of the stream function values in every point on the grid ($2 \times 40 \times 20$). In our scenario, the tuned quantity is the covariance matrix $\mathbf{Q}_k$ of the model error term E_k which is assumed to be normally distributed:

$$E_k(\boldsymbol{\theta}) \sim N(0, \mathbf{Q}(\boldsymbol{\theta})). \tag{19}$$

The model error covariance matrix $\mathbf{Q}(\boldsymbol{\theta})$ is parameterized such that its ijth element is given by

$$\tau^2 \delta_{i=j} + \sigma^2 \rho \exp\left(-\frac{x_{ij}^2}{2\alpha^2}\right), \tag{20}$$

where $\delta_{i=j}$ equals 1 when $i = j$ and 0 otherwise. Component $\exp\left(-\frac{x_{ij}^2}{2\alpha^2}\right)$ represents the covariance function which models the dependency of the correlation on the distance between points: x_{ij} is the distance between points i and j projected on the same layer and α is a tuning parameter. Parameter ρ defines correlations between the two layers such that $0 \leq \rho \leq 1$. To guarantee that the parameter stays within the bounds, we use parameterization

$$\rho = \exp\left(-\frac{h_{ij}^2}{2\gamma^2}\right),$$

where h_{ij} is the distance between the layers if i and j are in the same layer and $h_{ij} = 0$ otherwise. Thus, the actual tuning parameter is γ. Parameter σ^2 is the scaling parameter and τ is the nugget term often used to assure numerical stability.

Thus, there are four tuning parameters in total. The parameterization in Eq. (20) assures the positive-definiteness of $\mathbf{Q}$ for any combinations of the tuned parameters, which is important for the stability of BO. This corresponds to describing the model error

NPGD
1, 1283–1312, 2014

as a Gaussian process with a covariance function separable in x and h domains. The surface direction can be imagined horizontal to the cylinder shown in Fig. 5 and height is in the vertical direction. In order to use a valid covariance function, we computed the distance x_{ij} in the three-dimensional space, not on the cylinder surface.

In the experiments, we assume that noisy observations of the simulated stream function are available at 50 randomly selected grid points with the interval of six hours. Thus, the observation operator $\mathcal{K}$ in Eq. (11) simply selects some of the elements of the state vector as being observed. The standard deviation of the iid Gaussian noise added to the simulated ψ values is $\sigma_y = 0.1$. The same value was used to form the covariance matrix of the observation noise in the tuned system in Eq. (11). The observation sequence contained 400 time instances.

We evaluate the likelihood using the extended Kalman filter, as described in Sect. 3.

4.3 Experimental results

We used BO with the EI acquisition function for tuning the parameters of the model error covariance matrix. Initially, we draw 40 samples using the Latin hypercube sampling (LHS, see, e.g., Lizotte et al., 2012) in the region $\alpha \in [10 \quad 500]$, $\sigma^2 \in [0.01 \quad 0.81]$, $\rho \in [0.61 \quad 0.97]$ and $\tau^2 \in [0.25 \quad 0.81]$. In practice, we worked with the logarithm of the tuned parameters.

Figure 6 presents the results of BO using EI with $\xi = 0$. Here, we plot the approximation of the objective function using the mean of GP fitted after 200 iterations over all the data (initial 40 samples and BO 200 samples). Since there are four tuning parameters, each subplot presents the surface of the objective function when parameters α and $\log(\sigma)$ are varied and the other two parameters ρ and τ are fixed (see the corresponding fixed values above each subplot). The circles represent the samples used at the stage of initial sampling (the magenta circles) and the ones gathered during the optimization procedure (the white circles). Note that the locations of the samples are approximate: they are projected to the nearest plane corresponding to one of the subplots.

NPGD
1, 1283–1312, 2014

Note that the initial samples were gathered from an area with relatively low likelihoods. Figure 2 shows the log-likelihood function values (log-scale) and the maximum of the initial data and the ones obtained after every BO iteration t. We observe that the number of likelihood function values required by the BO method to find the best point was 141. Here, we fixed the total number of iterations to 200. While other optimization performance criteria can be used as well (see, e.g., Huang et al., 2006, p. 457).

This experiment shows that BO with the EI acquisition function is able to find the maximum of the posterior computed with GP while overcoming the local optima corresponding to small values of α. These values can be seen more clearly in the top four subplots (first row) of the Fig. 6. Note that we initialize the experiment with samples taken from an arbitrary region with quite bad log-likelihood values. This implies that the possibility of finding bad samples compared to the initial samples is highly likely due to the exploration property of the method. Therefore, we can see from Fig. 2 that the maximum found with the method gradually keeps improving over long number of iterations.

5 Parameter tuning of a chaotic system with "noisy" likelihood evaluations

In the following example, we simulate a scenario of tuning a large-scale chaotic system in which the evaluation of likelihood Eq. (14) using the extended Kalman filter is infeasible and therefore a stochastic EnKF is used. This results in noisy evaluations of the likelihood. As a tuned system we use a parameterized Lorenz 95 model.

NPGD
1, 1283–1312, 2014

5.1 Parameterized Lorenz 95 model

The model generating the data is the classical Lorenz 95 model (Lorenz, 1995; Wilks, 2005) whose dynamics is given by

$$\frac{\mathrm{d}x_k}{\mathrm{d}t} = -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F - \frac{hc}{b}\sum_{j=J(k-1)+1}^{Jk} z_j, \tag{21}$$

$$\frac{\mathrm{d}z_j}{\mathrm{d}t} = -cby_{j+1}(z_{j+2} - z_{j-1}) - cz_j + \frac{c}{b}F_z + \frac{hc}{b}x_{1+\lfloor\frac{j-1}{J}\rfloor} \tag{22}$$

where $k = 1,\ldots,K$ and $j = 1,\ldots,JK$. We use values $K = 40$, $J = 8$, $F = F_z = 10$, $h = 1$ and $c = b = 10$. In this model, the evolution of the slowly changing state variables x_i is affected by fast variables z_j and vice versa.

The tuned model is designed such that only the evolution of the slow variables is modeled and the net effect of the fast variables is represented with a deterministic component, such that

$$\frac{\mathrm{d}x_k}{\mathrm{d}t} = -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F - g(x_k, \boldsymbol{\theta}), \tag{23}$$

and $g(x_k, \boldsymbol{\theta})$ is selected to be a polynomial $g(x_k, \boldsymbol{\theta}) = \sum_{i=0}^{d} \theta_i x_k^i$, similarly to (Hakkarainen et al., 2012). In our experiments, we use the polynomial of order $d = 1$ which corresponds to slope θ_1 and intercept θ_0. The forcing term remains unchanged, that is $F = F_z = 10$.

5.2 Experimental setup and likelihood formulation

The training data were generated using the Lorenz 95 model Eqs. (21) and (22) with the discretization interval $\Delta t = 0.0025$. The state of the system is represented by a 40-dimensional vector of the slow variables x_k. We assume that noisy observations

NPGD
1, 1283–1312, 2014

of the slow variables are available at 24 locations each day. The last three state variables from every set of five states are picked and thus we observe the states 3, 5, 8, 9, 10, . . . , 38, 39, 40. The standard deviation of the iid Gaussian noise added to the simulated x_j values is $(0.1\sigma_{\text{clim}})^2$, where $\sigma_{\text{clim}} = 3.5$ corresponds to a climatological standard.

The tuned model is the parameterized Lorenz 95 model Eq. (23) simulated with the discretization interval $\Delta t = 0.025$ using 50 days of observations (one day corresponds to 0.2 time units). The tuned system is formulated as a state-space model Eqs. (10) and (11) with a diagonal covariance matrix $\mathbf{Q} = \sigma^2\mathbf{I}$ with σ^2 fixed to 0.0065, the value found to be optimal in our previous studies (Hakkarainen et al., 2012). Thus, the two tuned parameters are slope θ_0 and intercept θ_1 of the polynomial parameterization Eq. (23).

The objective function is the likelihood Eq. (14) computed via stochastic EnKF with 100 ensemble members. Since we use a version of EnKF that involves random perturbations, the likelihood evaluations are noisy. Noisy likelihood introduces difficulties in standard methods that try to explore or optimize the likelihood surface.

5.3 Experimental results

We present the results of a single run of BO with the EI acquisition function. The initial 20 samples were drawn using the LHS method in the region $5.0 \leq \theta_0 \leq 7.0$ and $0.67 \leq \theta_1 \leq 0.8$. These samples are represented with the magenta circles in Fig. 7b. Figure 7b shows the GP mean approximation over all the data (initial 20 samples and BO 150 samples). The samples obtained using BO are shown with the white circles and black cross indicates the found optimum. The found optimal values of the parameters are $\theta_0 = 1.97$ and $\theta_1 = 0.07$. Note that the initial samples come from a rather bad region with respect to the parameter likelihood. Figure 7a shows a scatter of uniformly random sampling of 1000 likelihood evaluations.

In Fig. 8b, we show the region closer to the optimum. The standard deviation value of the estimated noise in the total collected data was 647, which is rather large compared

NPGD
1, 1283–1312, 2014

to the variability of the systematic component of the objective function. We also perform uniformly random sampling in this smaller region and the scatter of 1000 samples is shown in Fig. 8a. At the found optimum the standard deviation of the noise is around 200.

Figure 3 shows the log-likelihood function values (log-scale) and the maximum of the initial data and the ones obtained after every BO iteration t. We observe that the number of likelihood function values required by the BO method to find the best point was 132. Although, we initialize the method with quite bad likelihood function values, we see that BO is able to find a good solution, even though there was significant noise in the likelihood function.

6 Conclusions

In this paper, we considered Bayesian optimization as a tool for tuning chaotic systems. We used two benchmark systems for testing the BO procedure: a simplified atmospheric model and a low-dimensional chaotic system. In the two-layer QG-model, the tuning parameters were four variables that constructed a model error covariance matrix used in the filtering with EKF. In the Lorenz 95 model, the tuning parameters were two variables that were used in the construction of a polynomial parameterization. For both experiments, the learning was based on the filtering likelihood.

The experiments showed that by using BO we were able to find the optimal solutions. The expensive models were tuned without the need of any gradient information. BO required only a small number of likelihood function evaluations as compared to random sampling.

The tested technique can be a practical tool in many tuning problems in climate sciences. Possible applications include tuning of large-scale climate models (Schirber et al., 2013) and ensemble prediction systems (Solonen and Järvinen, 2013). However, there are some known issues when using the BO method for large systems: first, with the increase in the number of tuning parameters of the model the required number

NPGD
1, 1283–1312, 2014

of samples to explore the domain would automatically increase. Second, with limited availability of training data due to expensive computational cost of the model, little is known about the objective function which means the design of the prior becomes more critical. Third, with the exploration vs exploitation property of acquisition functions, one has to decide on how to handle such parameters. We see the above mentioned problems as a very interesting direction for researchers, especially, in weather and climate applications. Applying the BO technique to large-scale models like ECHAM5 will be the future direction of our research.

The Supplement related to this article is available online at doi:10.5194/npgd-1-1283-2014-supplement.

Acknowledgements. This work has been supported by the Academy of Finland (project number 134935) and the Finnish Centre of Excellence in Computational Inference Research (COIN).

References

Annan, J. and Hargreaves, J.: Efficient estimation and ensemble generation in climate modelling, Philos. T. R. Soc. A, 365, 2077–2088, 2007. 1284

Bibov, A.: Variational Kalman filter data assimilation for two-layer Quasi-Geostrophic model, M.S. thesis, Lappeenranta University of Technology, Finland, 2011. 1295

Box, G. E. P. and Draper, N. R.: Empirical model-building and response surfaces, John Wiley & Sons, Oxford, England, 1987. 1286

Boyle, P.: Gaussian Processes for Regression and Optimisation, Ph. D. thesis, Victoria University of Wellington, New Zealand, 2007. 1287

Brochu, E., Cora, V. M., and Freitas, N. D.: A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning, Tech. rep., arXiv:1012.2599, arXiv.org, Cornell University Library, USA, 2010a. 1286, 1287

Brochu, E., Hoffman, M., and de Freitas, N.: Portfolio Allocation for Bayesian Optimization, Tech. rep., arXiv:1009.5419, arXiv.org, Cornell University Library, USA, 2010b. 1290

NPGD
1, 1283–1312, 2014

Evensen, G.: Data Assimilation: The Ensemble Kalman Filter, Springer, New York, USA, 2007. 1293
Fisher, M., Tremolet, Y., Auvinen, H., Tan, D., and Poli, P.: Weak-constraint and long window 4DVAR, Tech. Rep. 655, European Centre for Medium-Range Weather Forecasts, Reading, UK, 2011. 1294
Frean, M. and Boyle, P.: Using Gaussian processes to optimize expensive functions, Lect. Notes Artif. Int., 5360, 258–267, 2008. 1287
Hakkarainen, J., Ilin, A., Solonen, A., Laine, M., Haario, H., Tamminen, J., Oja, E., and Järvinen, H.: On closure parameter estimation in chaotic systems, Nonlin. Processes Geophys., 19, 127–143, doi:10.5194/npg-19-127-2012, 2012. 1285, 1291, 1299, 1300
Hakkarainen, J., Solonen, A., Ilin, A., Susiluoto, J., Laine, M., Haario, H., and Järvinen, H.: A dilemma of the uniqueness of weather and climate model closure parameters, Tellus A, 65, 20147, doi:10.3402/tellusa.v65i0.20147, 2013. 1284
Hauser, T., Keats, A., and Tarasov, L.: Artificial neural network assisted Bayesian calibration of climate models, Clim. Dynam., 39, 137–154, 2012. 1284
Huang, D., Allen, T. T., Notz, W. I., and Zeng, N.: Global optimization of stochastic black-box systems via sequential kriging meta-models, J. Global Optim., 34, 441–466, 2006. 1298
Järvinen, H., Räisänen, P., Laine, M., Tamminen, J., Ilin, A., Oja, E., Solonen, A., and Haario, H.: Estimation of ECHAM5 climate model closure parameters with adaptive MCMC, Atmos. Chem. Phys., 10, 9993–10002, doi:10.5194/acp-10-9993-2010, 2010. 1284
Jones, D. R.: A taxonomy of global optimization methods based on response surfaces, J. Global Optim., 21, 345–383, 2001. 1289
Jones, D. R., Perttunen, C. D., and Stuckman, B. E.: Lipschitzian optimization without the Lipschitz constant, J. Optimiz. Theory App., 79, 157–181, 1993. 1291
Kushner, H. J.: A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise, J. Basic Eng.-T. ASME, 86, 97–106, 1964. 1290
Leutbecher, M. and Palmer, T.: Ensemble forecasting, J. Comput. Phys., 227, 3515–3539, 2008. 1284
Lizotte, D. J.: Practical bayesian optimization, Ph. D. thesis, University of Alberta, Edmonton, Canada, 2008. 1287
Lizotte, D. J., Greiner, R., and Schuurmans, D.: An experimental methodology for response surface optimization methods, J. Global Optim., 53, 699–736, 2012. 1286, 1287, 1290, 1297

NPGD
1, 1283–1312, 2014

Lorenz, E.: Predictability: a problem partly solved, in: Proceedings of the Seminar on Predicability, 4-9 September 1995, European Center on Medium Range Weather Forecasting, 1, Reading, UK, 1–18, 1995. 1299

Marzouk, Y. and Xiu, D.: A stochastic collocation approach to Bayesian inference in inverse problems, Commun. Comput. Phys., 6, 826–847, 2009. 1286

Mockus, J.: Bayesian Approach to Global Optimization: Theory and Applications, Kluwer Academic, Dordrecht, 1989. 1290

Mockus, J., Tiesis, V., and Zilinskas, A.: Bayesian methods for seeking the extremum, in: Toward Global Optimization, Elsevier, Amsterdam, North Holland, 2, 117–128, 1978. 1287

Osborne, M. A., Roman, G., and Roberts, S. J.: Gaussian processes for global optimization, in: 3rd International Conference on Learning and Intelligent Optimization (LION3), 14–18 January 2009, Trento, Italy, 1–15, 2009. 1287

Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning, The MIT Press, Cambridge MA, USA, 2006. 1286, 1288, 1289

Schirber, S., Klocke, D., Pincus, R., Quaas, J., and Anderson, J. L.: Parameter estimation using data assimilation in an atmospheric general circulation model: from a perfect toward the real world, Journal of Advances in Modeling Earth Systems, 5, 58–70, 2013. 1284, 1301

Snoek, J., Larochelle, H., and Adams, R. P.: Practical Bayesian optimization of machine learning algorithms, arXiv preprint arXiv:1206.2944, arXiv.org, Cornell University Library, USA, 2012. 1286, 1287

Solonen, A. and Järvinen, H.: An approach for tuning ensemble prediction systems, Tellus A, 65, 20594, doi:10.3402/tellusa.v65i0.20594, 2013. 1285, 1301

Solonen, A., Ollinaho, P., Laine, M., Haario, H., Tamminen, J., and Järvinen, H.: Efficient mcmc for climate model parameter estimation: parallel adaptive chains and early rejection, Bayesian Analysis, 7, 715–736, 2012. 1284

Vannitsem, S. and Nicolis, C.: Lyapunov vectors and error growth patterns in a T21L3 quasigeostrophic model, J. Atmos. Sci., 54, 347–361, 1997. 1294

Wilks, D. S.: Effects of stochastic parametrizations in the Lorenz '96 system, Q. J. Roy. Meteor. Soc., 131, 389–407, 2005. 1299

NPGD
1, 1283–1312, 2014

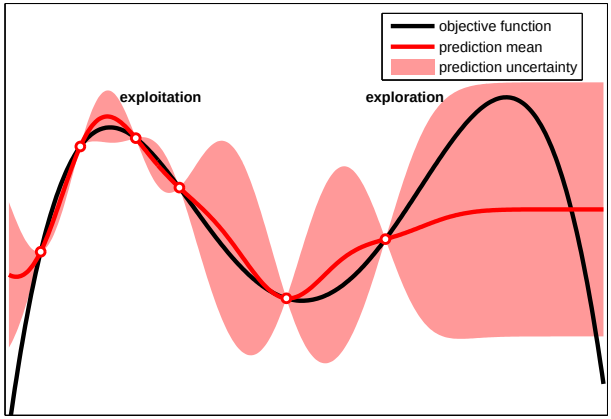


Figure 1. One-dimensional example of a Gaussian process (GP) approximation over an objective function. The GP is based on the points shown with red circles. The pink fill shows uncertainty of ±2 standard deviation from the prediction mean. The typical dilemma is between sampling in locations of high prediction mean (exploitation) and locations of high prediction uncertainty (exploration).

NPGD
1, 1283–1312, 2014

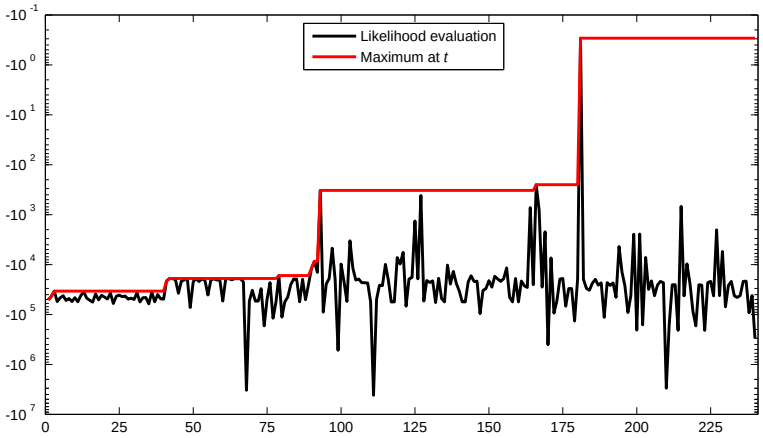


Figure 2. Log-likelihood function values and the maximum values for the initial data (40 points), and the ones obtained after every BO iteration t (200 points). Note the log-scale on the y axis. For plotting, we select the final best value between the maxima obtained from the likelihood evaluations and the GP mean predictions, and subtract this best value from the likelihood evaluations. In this case, the best value is the maximum of the GP mean predictions, therefore, the best value can be directly used for the subtraction.

NPGD
1, 1283–1312, 2014

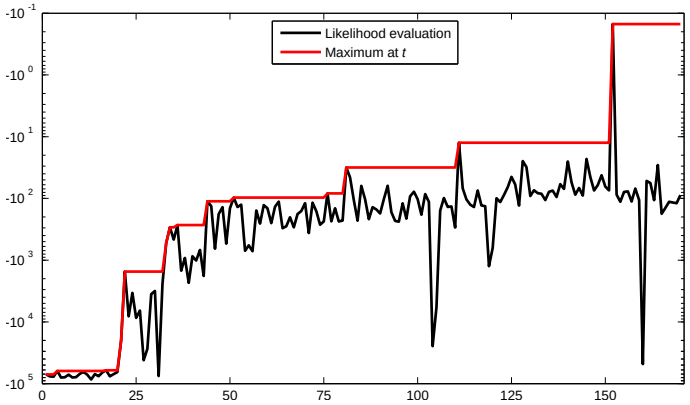


Figure 3. Log-likelihood function value and the maximum of the initial data (20 points), and the ones obtained after every BO iteration t (150 points). Note the log-scale on the y axis. For plotting, we select the final best value between the maxima obtained from the likelihood evaluations and the GP mean predictions, and subtract this best value from the likelihood evaluations. In this case, the best value is the maximum of the likelihood evaluations, therefore, a small nugget term (0.15) is added to the best value before the subtraction.

NPGD
1, 1283–1312, 2014

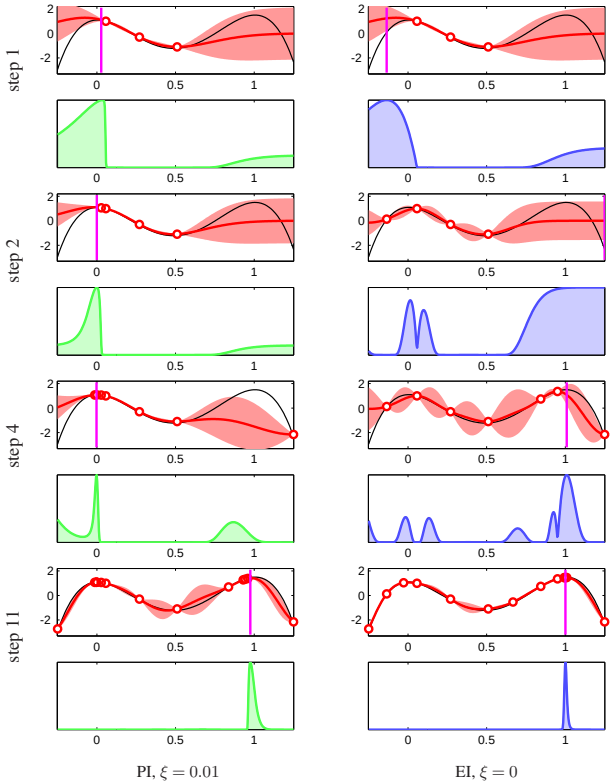


Figure 4. One-dimensional demonstration of Bayesian optimization. The objective function is shown with the black line. The circles represent sampled values of the objective function. The red line is the prediction mean and pink fill color is the uncertainty (±2 standard deviation). The vertical magenta lines show the new sample locations proposed by BO so as to maximize the acquisition function. The two columns correspond to two acquisition functions: PI (left) and EI (right).

NPGD
1, 1283–1312, 2014

Bayesian optimization for tuning chaotic systems

M. Abbas et al.

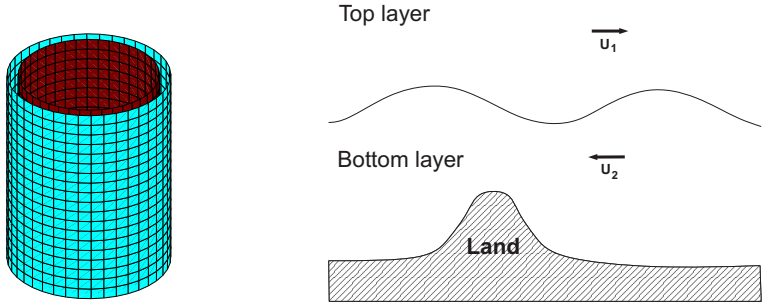


Figure 5. Geometrical layout of the two-layer quasi-geostrophic model. Left: schematic representation of the two model layers on a rotating cylinder. Right: the model layout at one of the longitudes.

NPGD
1, 1283–1312, 2014

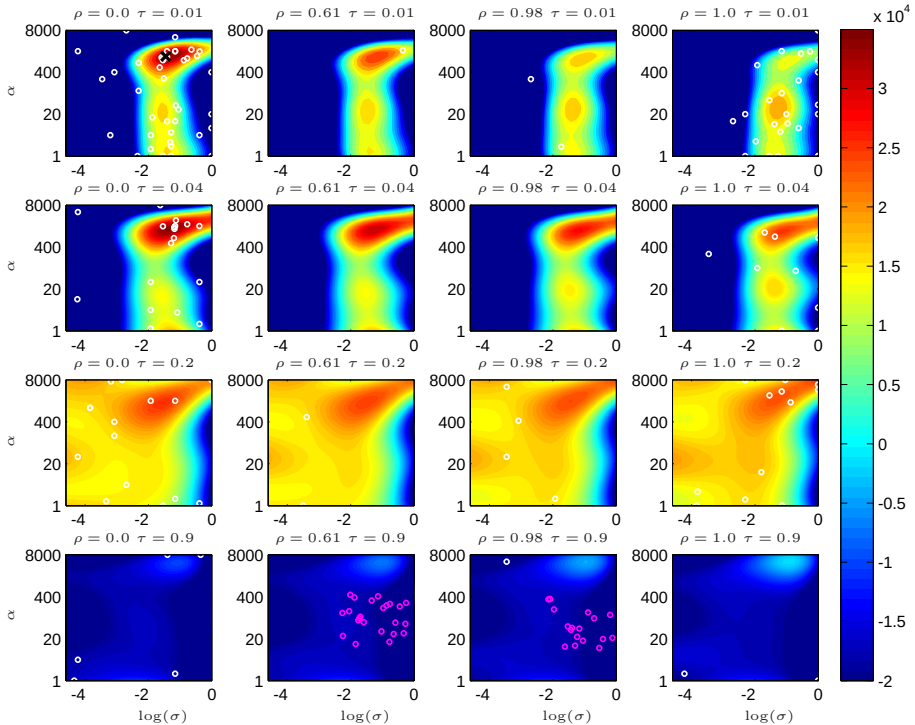
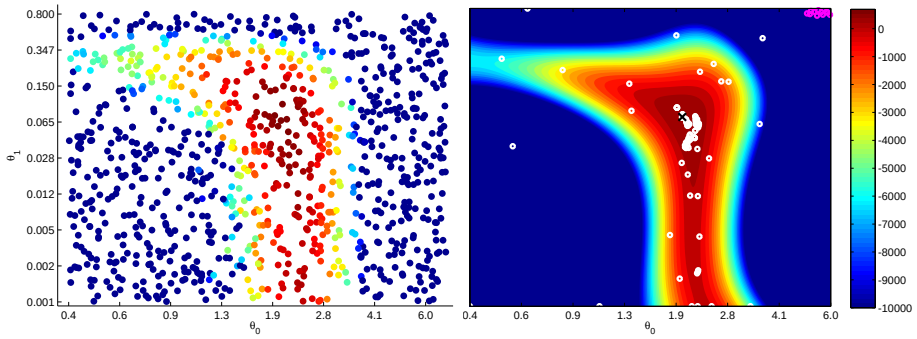


Figure 6. The illustration of the surface of the objective function estimated during BO for tuning the QG model. Each subplot corresponds to a surface obtained by varying two parameters and keeping the other two parameters fixed. The circles represent the sampled locations: initial (magenta) and obtained during optimization (white). The black cross indicates the maximum of the likelihood function. Note the logarithmic scale for parameter σ. See text for a more thorough explanation.

NPGD
1, 1283–1312, 2014

Bayesian optimization for tuning chaotic systems

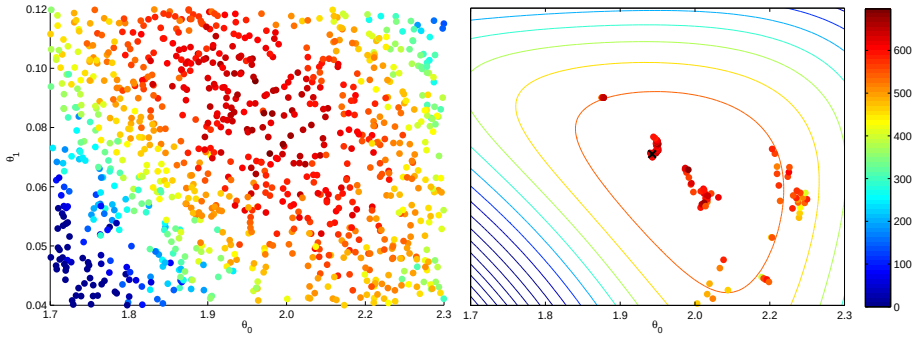
M. Abbas et al.



(a) (b)

Figure 7. (a) The scatter plot of the objective function values calculated on uniform random samples. **(b)** Single result of BOEI for tuning the Lorenz 95 model parameters using the EnKF likelihood. The magenta circles indicate the initial set of samples. The white circles indicate the BO samples and the maximum found is shown with the black cross. The contours represent the objective function approximation at the last iteration of BO.

NPGD
1, 1283–1312, 2014



(a) (b)

Figure 8. The same result as in Fig. 7 but in a smaller region near the optimum. The BO samples in **(b)** are shown with circles whose colors represent the evaluated objective function values. The black cross indicates the maximum.

NPGD
1, 1283–1312, 2014