Nonlin. Processes Geophys. Discuss., 1, C442–C449, 2014
www.nonlin-processes-geophys-discuss.net/1/C442/2014/

Nonlinear Processes
in Geophysics
Discussions
Open Access

Interactive comment on "Bayesian optimization for tuning chaotic systems" *by* M. Abbas et al.

Anonymous Referee #1

Received and published: 8 September 2014

1 General comments

This work presents an application of the Bayesian Optimization (BO) technique to the currently relevant topic of tuning parameterizations of chaotic systems. The technique aims to find the global optimum of a very expensive objective function which may have multiple local optima. This is achieved by first approximating the objective function by a computationally cheaper function. This function can be seen as a smoothed version of the objective function, much like the interpolation obtained by Kriging. BO is Bayesian in the sense that there is a mechanism for belief updates: the approximation is not static, but it is updated with every optimization step. At each optimization step, a new point in parameter space is chosen by maximizing an acquisition function. Two possible acquisition functions with a tunable parameter are proposed, which differ in the way they balance exploration and exploitation. Note that the convexity properties

of the acquisition function are also unknown, so the optimization of this function is not very different from the initial problem; however, the acquisition function is designed to be cheap to calculate, so optimizing it is not computationally costly.

In the first application, BO is used to find an optimal parameterization of the error covariance matrix in a data assimilation scheme (using extended Kalman filter) for a 2-layer QG model. The second example features a noisy objective function, as it makes use of the Ensemble Kalman filter (EnKF). In this example, BO is applied to the Lorenz 95 model, where it is used to find an optimal (linear) parameterization for unresolved, fast processes. In both cases the technique is shown to find the optimum with very few evaluations of the objective function, however without comparing the performance to a similar method.

This work aims to address a pertinent question in geosciences, and seems to be technically correct, although it appears to lack some necessary information to verify this. The innovation of the work is adequate. However, there are several issues with this article in its current form, both with regard to the scientific value and clarity. This paper discusses a very complex process, and more work is needed to explain it clearly. Both the structure and the notation can be improved in order to guide a non-specialist reader through the steps.

2 Scientific comments

A first major concern is that the two benchmarks discussed in this paper cannot objectively be considered as benchmarks, as there is no comparison with a baseline method. The authors should compare BO with a commonly used, advanced optimization method, using an objective metric, in order to convince the reader of the added value of BO. Although this would require a lot of extra work, it would greatly increase the scientific value of this paper.

It should also be noted that the concept of "tuning" is interpreted quite liberally in this work. In its current form, the title would lead the reader to believe that the technique is used to tune the parameters of chaotic systems in order to obtain some desired behavior. Perhaps it should be reformulated to reflect the fact that the work presents the optimization of the parameterization of an error covariance matrix and that of the parameterization of a chaotic model, respectively. A title such as "Bayesian optimization for parameterizing chaotic systems" or "[...] for parameterizations in chaotic systems" would be more accurate.

In the introduction, the distinction is made between the optimization metrics for weather models and for climate models. Please clarify which of the two are chosen here for the two models under consideration. It is not clear whether parameter estimation of the covariance matrix falls in either one of these categories. Please put into context.

In Section 5, how is the parameter optimization scheme sensitive to the lead time up to which the likelihood is evaluated? I guess, at least for the Lorenz 95 system, this can make a huge difference as the model and observation diverge from each other as time evolves.

As an outlook, the authors mention that they want to apply the BO technique to large-scale models like ECHAM5, without specifying what they want to do. Please specify this.

3 Suggestions to improve clarity

I realize that the clarity of this work suffers greatly from an unfortunate coinciding of nomenclature (e.g. this work features various likelihoods, noise terms, error covariances,... all unrelated); care should be taken so that the reader does not confuse these quantities. For example, while reading the article a second time, I found Eq. (1) very confusing, and it took me quite some time to understand its meaning again. One

of the main sources of this confusion is that you call it the "prior distribution over f". Later in the text, f is revealed to be a likelihood. Having a prior distribution over a likelihood did not make any sense to me at first. The meaning of Eq. (3), being a pdf of a variable which is a likelihood itself, also took a while to sink in.

It should be mentioned very clearly that there are two estimation processes at play: one is for estimating the optimal parameters of the parameterization, and the other "meta-estimation" which estimates the likelihood function of the first estimation process. The fact that the nomenclature for both processes overlap can be very confusing to the reader. In fact, the first application adds yet another layer of complexity by not estimating the parameters of the system itself, but the parameters of the error covariance matrix. I realize that there's no quick fix for this issue, but it would help if the authors add some caveats or highlight this distinction somehow, e.g. in the beginning of Section 3.

The introduction does not clearly explain what is meant by "noisy".

How does the chaotic nature of the model relate to the parameter estimation (process)? Please clarify or add a reference.

According to Eqs. (1) and (2), the prior for the set of t observed values for f is a Gaussian distribution in the t-dimensional space of possible parameter values, with a covariance matrix that depends on the distance between these parameter values. Later on, however, the authors mention that in practice, they use the logarithm of the parameters, as well as log-likelihoods. In the two examples, does f represent the log-likelihood or the likelihood? Furthermore, is the distance in Eq. (2) measured for θ or for $\log(\theta)$? Please specify.

There appear to be at least five different meanings for the symbol "k". In Section 2.1, the authors introduce two k functions, which is fine. However, in Section 3 another, unrelated, K function is introduced. This might be confusing. Also, in Eq. (2) an index k is used for iteration. In Eqs. (10) and (11) of Section 3 the k seems to denote a time step and K the total amount of time steps. Please change. Also, clarify that the σ of Eq. (2) is different from that in Eq. (5).

Please clarify the product over k in Eq. (2). What values does k assume? Is k related to i, j?

Please explain the η hyperparameters that appear on page 1288 and their meaning in the rest of the article. When you mention that they are determined by maximizing the marginal likelihood, please provide the expression for this likelihood and for the marginal likelihood (which involves an integration if I am not mistaken) – is there an analytical expression, or is it integrated numerically?

What value of ξ was chosen for the acquisition function in the two examples?

Fig. 1 concerns the difference between exploitation and exploration. This is not immediately clear to me. In fact, in my opinion, this figure is not necessary as more or less the same figure is reproduced in Fig. 4, and since the contrast between exploitation and exploration is well explained in the introduction. Preferably, Fig. 4 is clarified a bit more (see suggestions below) while Fig. 1 is replaced with a schedule of the different steps required to optimize theta. Please indicate μ^+ in the figure as well.

Section 3 starts with the sentence "In tuning chaotic systems, we use the approach where the likelihood is computed using filtering techniques". The relation with the previous sections is unclear. Only later in this section do you mention that the likelihood concerns the parameters θ but, most importantly, that the likelihood is actually the function f which is being maximized. It should be stressed much earlier, preferably in the introduction when f is introduced, that you take f equal to the likelihood. Also it should be explained that the "Bayesian" part of BO has nothing to do with this likelihood but rather points to the exploration of the parameter space.

Section 4, p. 1297 lines 1-4: This part appears to be a bit out of place; perhaps it should be mentioned earlier. The description of surface and height directions on the

cylinder are not clear. According to this explanation, the surface direction is not along the surface of the cylinder, and the height is not normal to the cylinder surface. Is this correct? Improve this description and if possible, improve the figure. Also, please clarify why the distance measured on the cylinder surface does not yield a valid covariance function.

p. 1298 lines 4-5 and lines 15-16: These two sentences appear to contradict one another. First, it is said that the number of likelihood function evaluations required to find the optimum was only 141. The other sentence seems to imply that we need a large number of iterations to reach the maximum. How far is the BO result from the global optimum (if this is known)? Or equivalently, how much does the optimum improve if the number of iterations is, say, doubled?

Does the Lorenz 95 model in Section 5 have periodic boundary conditions?

On p. 1300, last line, it is mentioned that the standard deviation of the noise is 647. Is this the standard deviation of the log-likelihood at a single point in parameter space?

The procedure for plotting that is mentioned in the caption of Figures 2 and 3 is not at all clear. What is the purpose of this small nugget term (0.15)?

3.1 Other questions

There are some other questions which arose while reading the paper; I would appreciate it if the authors could briefly address these in their reply. * Is there a relation between the smoothing function used in BO and the one obtained by Kriging? At first sight these appear to be quite similar. * The authors mention briefly that there is a problem when applying BO to high-dimensional problems. What number of dimensions would you regard as "high"? Could you discuss how the method scales with the problem's dimensionality? You already mentioned the increased number of samples required. Are there other issues, e.g., could it become tricky to optimize the acquisition function for high-dimensional problems?

4 Technical corrections

4.1 Text

p. 1286 line 2: "GP models" are introduced and only in the paragraph below the abbreviation GP is explained.
p. 1286 line 14-15: please rephrase "demonstrated as a very efficient and flexible approach in optimization of computationally heavy to compute models in several papers".
p. 1286 line 26-27: random function: this has a very specific mathematical meaning. I suppose this is not the meaning that you intended.
p. 1290 Eq. (7): mention that μ and σ are as defined in Eqs. (4) and (5).
p. 1293 line 18-19: "small number of ensembles": did you mean "small number of ensemble members"?
p. 1295 line 1: "physic" should be "physical".
p. 1296 line 17: "is the distance between the layers if i and j are in the same layer"? I would say that this distance is always zero.
p. 1297 line 6: "with the interval of six hours" should be "every six hours".
p. 1298 line 15: please rephrase "[...] the maximum found with the method gradually keeps improving over long number of iterations."
p. 1299 Eq. (22): "y" should be "z".
p. 1300 line 1: day is used before it is defined in this context.
p. 1300 line 3: 4 is missing from the list.
p. 1300 line 4-5: "a climatological standard": did you mean "the climatological standard deviation"?

4.2 Figures

The figures appear to be ordered haphazardly. Please put the figures in the order in which they are referred to in the text.

Most of the plots lack labels on x- and y-axis. Please add labels where necessary, and note that an explanation in the caption is not sufficient.

Figures 2 and 3 have several issues.

- First of all, it is unclear from the caption for which system the results are shown.
- According to the captions, the log-likelihood is shown; however, the legend indicates that the black curve is "likelihood evaluation".
- The legend is ambiguous. Please correct this, so that it is clear that the black curve represents the log-likelihood evaluation at optimization step t, while the red curve is the maximum of the log-likelihoods up to step t.

  The meaning of the subplots of Fig. 4 is not mentioned in the caption. Please relate the symbols as introduced in 2.2 to the curves and symbols in this figure. For example, the open circles denote θ_1, ..., θ_3. The red curve is $\mu(\theta)$ of Eq. (4) and the shaded areas denote Eq. (5) while the acquisition functions are given by Eq. (6) and Eq. (8)