Authors comments (C522) on "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

Comments:

Referee specific comments

1. The writing should be revised for clarity. In addition, the descriptions of the experiments are incomplete and do not include sufficient information to reproduce the results. See detailed comments that follow.

2. The manuscript mainly presents very preliminary steps toward understanding how BO might be applied to actual prediction systems and whether it might yield significant or minimal improvements to those systems. Improvements could be made in several aspects:

a) There is little motivation for the experiments chosen and little explanation of what has been learned from the experiments.

b) The manuscript provides no sense of the magnitude of improvements achieved by tuning, since there is no analysis of how predictions are altered by the tuning. See section 4 of Wilks (2005), for examples of diagnostics that might be considered. Without such analysis, the manuscript shows only that BO finds a minimum.

c) There is no exploration (or even much discussion) of how the method might scale to larger problems (both with more parameters and with higher-dimensional states) and what difficulties might be encountered. An especially important issue

is the calculation of the likelihood – see my comment 3.

d) There is no exploration of how the method compares with other possible approaches. For example, would the tuning accomplished here with BO fail with other methods, such as an ensemble Kalman filter with the state augmented by the parameters θ? As another example, how does BO compare with the approach of Wilks (2005) to tune representations of the fast degrees of freedom in the same Lorenz (1995) system?

3. A key issue is how to calculate or approximate the likelihood $p(y_{1:k}|\theta)$, especially when there are many observations, many parameters and the state dimension is large. Typically, we might expect that each individual observation carries little information about the parameters θ, and thus it is crucial to integrate the information in many observations (across space and time) to tune the parameters. Errors in calculating the likelihood may easily swamp that accumulated information Indeed, this may well be an issue even in the experiments presented in the manuscript; the authors use approximate methods but do not demonstrate their accuracy.

4. In general, we would like $p(\theta|y_{1:K})$ rather than a single value of θ that maximizes $p(\theta|y_{1:K})$. Can the method presented here be extended to estimate a pdf for θ?

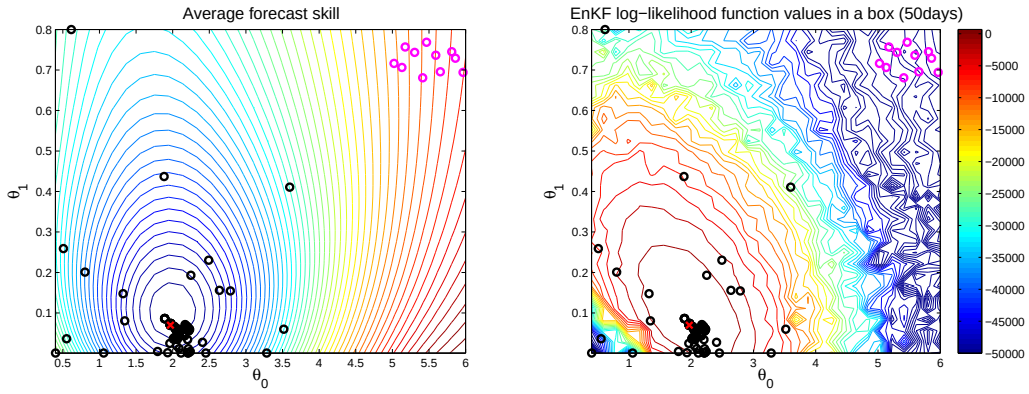
Author's response

1. Please see changes in manuscript.

2.(a,b)-3. For the Lorenz 95 case, we test the goodness of the optimization scheme by computing the forecast accuracy (similar to Hakkarainen J. et al. 2012). We use a 2-dimensional grid of the parameter space and compute the average forecast skill for different parameter values. The average forecast skill was computed using a 6 day forecast starting every 24h for 100 days. The average forecast skill can be written as

$$S(\theta) = \frac{1}{NK\sigma_{clim}^2} \sum_{i=1}^{N} \| M_6(\mathbf{x}_i^{true}, \theta) - \mathbf{x}_{i+6}^{true} \|_2^2 \,, \tag{1}$$

where $N = 100$, $K = 40$ and $\sigma_{clim} = 3.5$. The notation $M_6(\mathbf{x}_i^{true})$ means a 6 day prediction launched from the true state $\mathbf{x}_i^{true}$ with the parameter values θ.

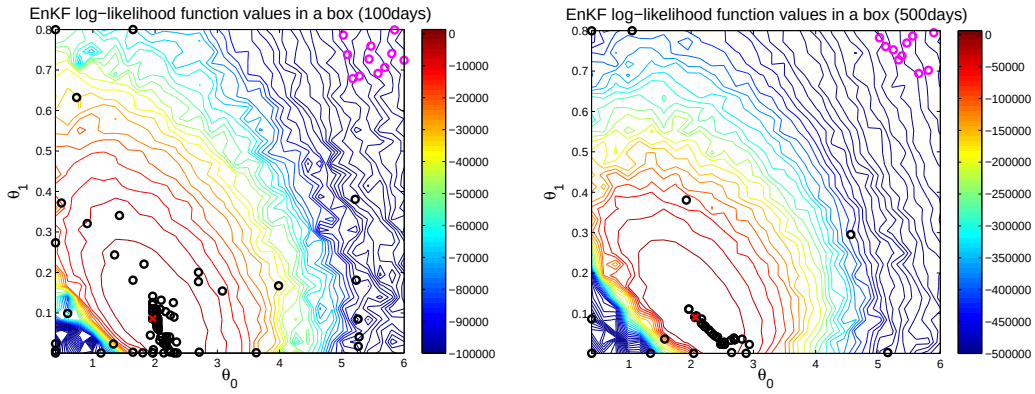


(a) (b)

Figure 1: (a) An Illustration of the average forecast skill. The magenta circles show the initial samples used for BO. The black circles show the samples obtained from BO. The red cross indicates the maximum obtained using BO. Blue contour colors indicate high forecast skill. (b) An illustration of the result of a 50 day run using EnKF based likelihood function. The stochastic filtering method produces a noisy likelihood function. The circles and cross mark represent the same things as in (a). It must be noted that the simulation length of the EnKF likelihood function used for optimization with BO was also 50 days.

The contour lines of the Fig. 1(a) show the average forecast skill computed using the method described above. The Fig. 1(a) also shows the same result of tuning the Lorenz 95 model using BO which was illustrated first in the paper. Note that the simulation length of the EnKF likelihood function we used for optimization with BO was 50 days and we used the parameters in the log-scale.

In the Fig. 1(b), the contour lines show the EnKF log-likelihood function values. Again we use a 2-dimensional grid of the parameter space and compute the EnKF likelihood function for different parameter values. Again the Fig. 1(b) also shows the result of tuning the Lorenz 95 model using BO method which was illustrated first in the paper. We also run experiments with longer simulation lengths of the EnKF likelihood function. The results from 100 days and 500 days runs are shown in the Fig. 2. Note the the contour lines for the EnKF likelihood function plots are not smooth. This is because the stochastic filtering method produces noisy likelihood function. With these results we observe that there is a good agreement between the tuned parameters obtained with BO using the likelihood



(a) (b)

Figure 2: (a) An illustration of the result of a 100 day run using EnKF likelihood function. The magenta circles show the initial samples used for BO. The black circles show the samples obtained from BO. The red cross indicates the maximum obtained using BO. The simulation length of the EnKF likelihood function used for optimization with BO was also 100 days. (b) Similar plot as (a) but with 500 days simulation length of the EnKF likelihood function.

approach and the forecast skill when the simulation length is sufficient. The performance and analysis of the likelihood approach for tuning the model error covariance matrix in case of the QG model (we have used) is shown by (Solonen et al., 2014).

2.(c)-3. For BO, some recent papers have shown that it works on real-world problems for the following dimensions: 12 in (Brochu et al., 2010a), 9 in (Snoek et al., 2012), or 9 in (Brochu et al., 2010b). More recent work on BO has generated improvements to the basic BO technique so that it could handle larger dimensionality. However, this is only for special cases where there is some type of projection space of the parameters being utilized. However, in general this problem still remains and active research area, especially, in machine learning.

2.(d) Unlike some other optimization procedures, BO is a universal technique that is applicable to any kind of likelihood formulations. For example, in the original paper by Wilks (2005), the parameterization of the effect of the fast variables of the Lorenz 95 system was tuned using the fact that the tuned component is an additive term of the differential equation governing the system dynamics. Then,

the parameters were tuned using a simple metrics computed from observed variables. In practice, there can be many difficulties that break that simple scenario: missing observations, noise in the observations and complex effect of parameters on the observed variables. Not also that using such a simple fitting approach may in practice yield badly tuned models (see, e.g., a discussion in Hakkarainen et al, 2013). /* A dilemma of the uniqueness of weather and climate model closure parameters Authors Janne Hakkarainen, Antti Solonen, Alexander Ilin, Jouni Susiluoto, Marko Laine, Heikki Haario, Heikki Järvinen */ Other popular optimization methods may fail in some situations too. For example, the popular approach of state-augmentation can work well for parametric tuning of chaotic systems (see, e.g., Hakkarainen et al, 2012). However, it fails in optimization of variance parameters such as the problem considered in the QG example. A simple illustration can be found in (DelSole and Yang, 2010).

4. Yes, it is possible to extend BO so that we can get a pdf for θ. One way to do this is by substituting the expected improvement (EI) function with a sampling based technique. Hybrid Monte Carlo (HMC) based sampling has already been done by [Rasmussen, C.E. (2003)].

Author's response on detailed comments and typos

section 1: Neelin et al. (2010 PNAS) deserve a citation both in the introductory paragraph and where response-surface techniques are discussed.

Reference included

p 1284, l 25: forecast skills -> forecast skill

changed

p 1286, l 15-16: awkward phrase

please see manuscript changes

p 1288, l 9-10: Change "is computed using covariance function $k(\theta_i, \theta_j|\eta)$ for the corresponding inputs θ_i, θ_j " to "is computed using a covariance function $k(\theta_i, \theta_j|\eta)$ that depends on θ_i, θ_j ". If that changes the meaning you intend, then I don't understand how you're specifying the covariance.

please see manuscript changes

p 1289, l 6: i) Is this σ the same as that used in (2)? If so, why? ii) σ is used again, with a different meaning (I think), in (20).

σ are different, notation changed now

p 1290, the EI acquisition function: I'm confused what EI is doing and how (8-9) describe that. Equation (8) (and the explanatory text) says that the acquisition function $g(\theta)$ is equal to the expectation of $f(\theta)$ minus μ^+, the current maximum value found for the mean of $f(\theta)$ at some θ_i . This seems to me to imply $g(\theta) = E(f(\theta)) - \mu^+ = \mu(\theta) - \mu^+$, which is not what (9) says. Please clarify.

please see manuscript changes

p 1292, following (12): Usually, the EnKF is described as sampling correctly from the predictive distribution [i.e. the r.h.s. of (12)], but approximating the update (13).

p 1293, l 11-12: I suggest saying " ... C_{k-1}^{est} is the estimated covariance of $p(s_{k-1}|y_{1:k-1})$."

changed

p 1293, l 23-25: I can't parse this sentence.

please see manuscript changes

p 1294, l 4-5: "... which results in noiseless likelihood evaluations." The approximate likelihood (15) is a deterministic function of θ, but it is only an approximation and therefore contains error. Shouldn't the likelihood be consider "noisy" here too, in the sense that we shouldn't require the GP estimate of $f(\theta)$ to match the likelihood exactly? Is your assumption that (15) has errors that are too small to matter?

Yes, the errors due to approximation are too small to matter

p 1294, l 7-8: "fluid motion dynamics on a rotating cylinder" This is a confusing way to think of the QG model (and it leads the authors to compute distances in a odd way later, p 1296, l 12-13). The periodic-channel geometry of the model is a computational artifice and not a physically realizable property – note that the planetary rotation is normal to the horizontal plane of the model. Calling this a "periodic channel" or a "zonally periodic channel" will make more sense to readers familiar with idealized QG simulations and will not lead astray those who are not.

Explanation corrected. It was not correct to think 'fluid motion dynamics on a rotating cylinder'

p 1295, l 1: physic –> physical

changed

section 4.1-2: What are boundary conditions at the channel walls? Does the model include any dissipation?

There is no explicit dissipation. We used the QG model as described by Fisher et al. (2011). Please see the report by Fisher et al. (2011) in the supplementary materials of this manuscript on NPGD

section 4.2: Judging by the definitions of the Froude numbers F_1 and F_2 and the Rossby number R_s , it seems that the parameters D_1 , D_2 should be dimensional. My guess is the "units" are in fact meters, so that the layer depths are 6 km and 4 km, and the grid spacing is 100 km.

This is correct, the units are in meters

section 4.2: Please specify $S(x, y)$.

$S(x, y)$ is dimensional orography

p 1296, l 17: if i and j are in the same layer –> if i and j are in different layers

Although the explanation was correct, we have made changes for simplicity

p 1296, l 17: It is not obvious what the distance between layers should be (i.e. numerical value is assigned to h_{ij} when points are in different layers).

please see manuscript changes

p 1296, l 19: The "nugget" term and its motivation are mysterious.

We have performed test experiments without the use of the nugget term. We observed that the log-likelihood function values are higher in this case. However, the likelihood surface seems to contain more approximation errors. Therefore, we include the four parameter results in the manuscript

p 1297, l 4: It seems that a covariance function based on the 2D (Euclidean, allowing for periodicity in the zonal direction) distance would be fine. The model definitely does not represent fluid on the surface of a cylinder.

please see manuscript changes

section 4.3: Plese give other BO parameters, in addition to $\zeta = 0$.

Other BO parameters are computed from the marginal likelihood using the Maximum Likelihood at each iteration of BO

p 1298, l 5: ... iterations to 200. While –> ... iterations to 200, though

changed

p 1298, l 13: bad samples –> good samples (I think??)

changed

section 5: An obvious and very helpful step would be to explain how these experiments relate to those of Wilks (2005) with the same model.

please see answer 2.(d) above

p 1298, l 19-20: I don't see that a "noisy" likelihood (i.e. one that is stochastic function of the observations and θ) is a crucial distinction Even if you used a de-

terministic EnKF, the likelihood would only be approximate, even though it was a deterministic function of the inputs $(y_{1:k}, \theta)$. This error is equally important and is present in your first example (the QG model) as well.

We agree that there is a approximation error but as commented above that this error is very small to matter in the QG model case. While in the Lorenz 95 case we want to show how BO behaves in a noisy case. For example, when the goal is to optimize ensemble prediction systems with a stochastic mechanism of ensemble member selection, two evaluations of the likelihood function with the same parameter values generally leads to distinct function values. Another possible source of noise is the chaoticity of the tuned model. Small perturbations of the model parameters can result in significantly different simulation trajectories and therefore significant differences in the computed likelihood

p 1300, l 12: Eq.(23) –> $g(x_k, \theta)$

changed

p 1300, l 13-16: Please give details of how the likelihood is calculated.

We use the same implementation as described in [Section 3.2 of Hakkarainen et al. (2012)], reference is now added to the text

p 1300, l 19: LHS not defined

changed

section 5.3: Again, please give more details of set-up for BO. How were the parameters η selected and adjusted? What was the prior on θ?

η is found using Maximum Likelihood at each step of BO. In the initialization, the prior samples are used to compute the standard deviation in each dimension. More details on the marginal likelihood added to the manuscript

Manuscript changes

Please see text file attached (Manuscriptchanges.txt)

Best regards,
Authors

References

Brochu, E., Brochu, T., and de Freitas, N. (2010a). A bayesian interactive optimization approach to procedural animation design. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 103–112. Eurographics Association.

Brochu, E., Hoffman, M. W., and de Freitas, N. (2010b). Portfolio allocation for bayesian optimization. *arXiv preprint arXiv:1009.5419*.

Fisher, M., Tremolet, Y., Auvinen, H., Tan, D., and Poli, P. (2011). Weak-constraint and long window 4DVAR. Technical Report 655, European Centre for Medium-Range Weather Forecasts.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959.

Solonen, A., Hakkarainen, J., Ilin, A., Abbas, M., and Bibov, A. (2014). Estimating model error covariance matrix parameters in extended kalman filtering. *Nonlinear Processes in Geophysics*, 21(5):919–927.