**Authors comments (C442) 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 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.

**Author's response**

- We compare Bayesian optimization (BO) to Covariance Matrix Adaptation Evolution Strategy (CMA-ES) which is an evolutionary algorithm for derivative-free and noisy objective function optimization (see e.g., Hansen and Ostermeier, 1996, 2001; Hansen et al., 2009). We choose this method because first of all it does not require any gradient information and secondly it can handle noise in the objective function evaluations. Both of these properties make CMA-ES suitable for the kind of problems we are solving in the paper. Finally, CMA-ES is also a widely used approach for benchmark optimization problems.
- We test the case of parameter tuning of a chaotic system with "noisy" likelihood evaluations shown in section 5 of the paper using the CMA-ES strategy. In the initialization of the CMA-ES algorithm, we provide similar settings which were used for BO. The CMA-ES implementation we have used is available online at `https://www.lri.fr/~hansen/cmaes_inmatlab.html`. The Fig. 1 shows a comparison between BO and CMA-ES. The lines in the Fig. 1 represent the best value of the log-likelihood obtained upto the current iteration using each method. We ran each algorithm twice from two different initial samples sets. The first sample set was same as in the paper, drawn using the Latin hypercube sampling (LHS) method in the region $5.0 \leq \theta_0 \leq 7.0$ and $0.67 \leq \theta_1 \leq 0.8$. The second sample set was drawn using LHS in the region $5.0 \leq \theta_0 \leq 7.0$ and $0.1 \leq \theta_1 \leq 0.23$. We found that CMA-ES is able to find a maximum as good as the one found using BO. However, in order to achieve this it required a greater number of log-likelihood function evaluations. For BO, we also observed that the initial sample set used to construct the surface (GP approximation) is very critical. This can be done by selecting the initial set of design points several times in order to achieve the best GP approximation. Note that the likelihood function is noisy in this case which means that two runs of the same experiment with the same initialization may result in different paths to the maximum. We may conduct a

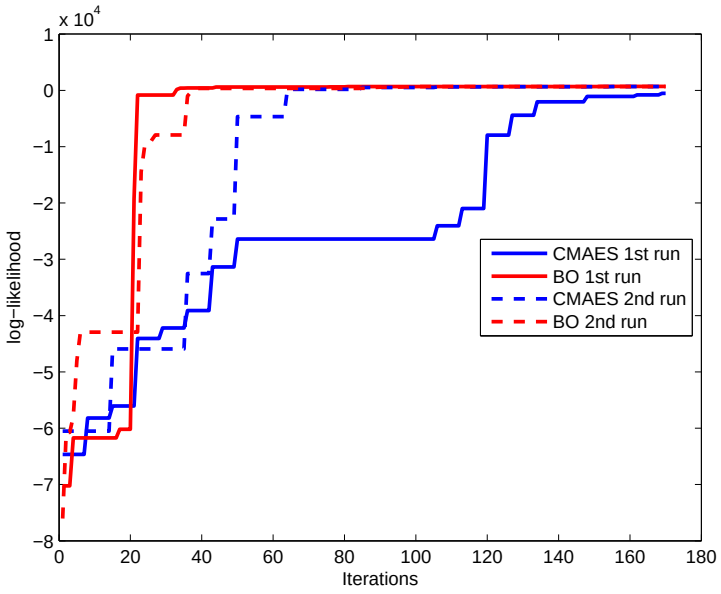



Figure 1: The histories of the best values obtained upto the current iteration using BO and CMAES. Both algorithms were run twice using different initial samples set.

similar test for the parameter tuning of the QG model using CMA-ES. Also, we must emphasize that in the paper, we are focused on presenting a method that is suitable for both 'noisy' and 'noiseless' objective functions. Hence, we hope that this example might be sufficient for demonstrating the suitability of BO for optimizing chaotic systems.

- Because CMA-ES is a evolutionary based algorithm, it uses a specific population size at each iteration of the algorithm. In our test case of parameter tuning of the Lorenz 95 model which has a parameterization of dimensionality 2, CMA-ES used a population size of 7. This implies that at every algorithmic iteration it would evaluate the objective function for each member of the population. In both of the Figures shown above we plot a single evaluation of the objective function at each iteration as shown on the x-axis. This is done only for simplicity and exact comparison of CMA-ES to BO.

- We would consider the suggestions given for the title and even more suggestions if any which would make the title sound more appropriate. For instance, we

would prefer "Bayesian optimization for parametric tuning of chaotic systems".

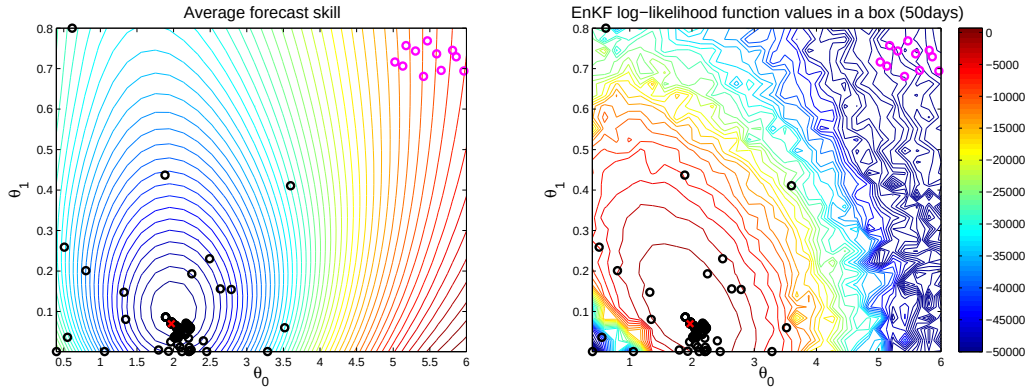
- In the examples considered in this paper, the likelihood formulation is especially relevant for parametric tuning of NWP systems, as it is essentially built around the accuracy of short-term forecasts.

- 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 $\theta$. The contour lines of the Fig. 2(a) show the average forecast skill computed using the method described above. The Fig. 2(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. 2(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. 2(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. 3. 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 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).



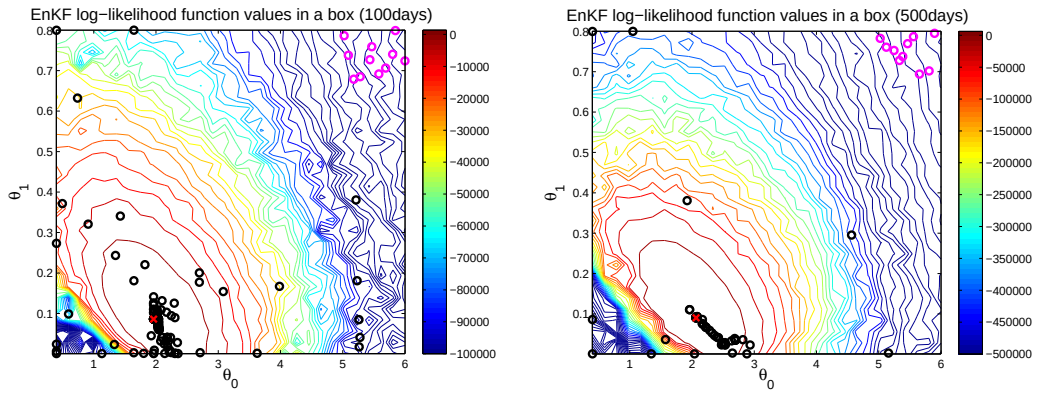

(a) (b)

Figure 2: (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.

- This approach can be used for tuning four parameters of ECHAM5 that are related to clouds and precipitation, as done by (Järvinen et al., 2010).

**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.




(a) (b)

Figure 3: (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.

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 $\theta$ 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 $\sigma$ 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 $\eta$ 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 $\xi$ 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 $\mu^+$ 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 $\theta$ 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)?

**Author's response**

- short paragraph added at the end of section 3.

In this scenario, there are two estimation processes at play. First is the estimation of the model parameters using the filtering likelihood technique. Second is a *meta-estimation* process which optimizes a surrogate model based on the filtering likelihood.

- Noisy: A typical source of noise is approximations made in the likelihood evaluations. 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.

That is why the parameter estimation process should naturally handle possible noise in the likelihood.

- We only use logarithm scale for the parameters when fitting a GP and searching for a new point using the surrogate. The likelihood functions always give log-likelihood values for parameters given to them in the normal scale. The distance in Eq.(2) is measured for $\log(\theta)$.

- Notation changes made to manuscript.

- Marginal likelihood added to manuscript. There exists an analytical expression.

- $\xi = 0$ for both noiseless and noisy case examples shown in the paper.

- changes will be added to Figures

- changes made to manuscript

- line 15-16 is rephrased: It describes that the improvement is gradually achieved. The maximum was found in 141 iterations from total 200 iterations. The global optimum is not known in this case. The optimum might improve very little if the iterations are doubled.

- The model is assumed to have cyclic boundary conditions.

- The standard deviation of the log-likelihood for all the collected points in the parameter space.
- In Figure. 2 each point on the black line is computed as

$$l_t = f(\theta_t) - \mu^*$$

where $\mu^*$ is the GP mean value corresponding to the maximum (log-likelihood function value) found using BO. In the figure, $\log(l_t)$ is plotted. The red line shows $\log(\max(l_{1:t}))$: maximum upto the current iteration $t$.

In Figure. 3 each point on the black like is computed as

$$l_t = f(\theta_t) - f^*$$

where $f^*$ is the maximum (log-likelihood function value) found using BO. In the figure, $\log(l_t)$ is plotted. The red line shows $\log(\max(l_{1:t}))$: maximum upto the current iteration $t$.

This is done to show better scaling of the lines in the plots so that we are able to highlight smaller changes in the log-likelihood function values during the experiments.

**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?

**Author's response**

- BO and kriging are closely related. The goal in kriging is interpolation of a random field by using a linear predictor. However, the errors on this model are typically assumed to be not independent. Kriging combines a linear regression model and a stochastic model that is fitted to the residual errors of the linear model. The residual is modelled with a zero-mean Gaussian process. The regression model then depends on the type of kriging, for example, ordinary kriging, universal kriging, or stochastic kriging. - In practice, there are several differences between BO and kriging. For example, in BO the model is usually fit using

the maximum likelihood while kriging uses variogram (a measure of the average dissimilarity between samples versus their separation distance), and in BO it is normal to use all the data in order to learn a global model while kriging normally restricts the prediction model to use only a small number of neighbors making it fit locally. - Bayesian optimization and kriging are more closely related in the experimental design literature. DACE (Design and Analysis of Computer Experiments) is a well known technique and several toolboxes are freely available which implement this method. The efficient global optimization (EGO) is an algorithm which combines the DACE model with expected improvement (EI). However, experimental design is typically non-adaptive and the whole experiment is designed before the data is collected. (brief comparison description from /*A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning by Eric Brochu, Vlad M. Cora and Nando de Freitas (2010)*/ )
- Another set of methods which use the kriging framework in a sequential manner are called sequential parameter optimization (SPO) [Hutter F. et al. (2009)] and sequential kriging optimization (SKO) [Huang D. et al. (2006)]. These methods are able to model noise in the objective function and are shown to be efficient for black-box optimization. We consider our approach of Bayesian optimization (BO) more closely related to these methods. We have not compared our method to any other surrogate based or response surface methods because the technique that will work in our problems is a sequential one (because of expensive model computations) and we find these later methods too similar to our approach. Hence, such an analysis might not be very useful.

- Some recent papers have shown that BO 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.

**Technical corrections**

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 $\mu$ and $\sigma$ 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"?

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 $\theta_1 \dots \theta_3$ . 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)

**Author's response**

Please see manuscript changes. Any remaining suggested corrections in the text and the figures will be made to the revised 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*.

Hansen, N., Niederberger, A. S., Guzzella, L., and Koumoutsakos, P. (2009). A method for handling uncertainty in evolutionary optimization with an application to feedback control of combustion. *Evolutionary Computation, IEEE Transactions on*, 13(1):180–197.

Hansen, N. and Ostermeier, A. (1996). Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Evolutionary Computation, 1996., Proceedings of IEEE International Conference on*, pages 312–317. IEEE.

Hansen, N. and Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195.

Järvinen, H., Räisänen, P., Laine, M., Tamminen, J., Ilin, A., Oja, E., Solonen, A., and Haario, H. (2010). Estimation of ECHAM5 climate model closure parameters with adaptive MCMC. *Atmos. Chem. Phys.*, 10:9993–10002.

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.