Response letter to Z. Hou

Comment:

This is an interesting study applying adaptive surrogates for multi-objective optimization in a land surface model. The surrogate development and optimization approaches are reasonable. I offer the following suggestions:

Response:

First, we would like to thank the editor and all the reviewers for your kind, helpful comments on this manuscript. We have enclosed a revised version and two response letters. Hopefully they can appropriately address the concerns in the review letters.

Comment:

Page 6716, line 20: the number of runs depends on the choice of weighting function, in addition to the choice of the output variables. It could be more reasonable to use a probability-based weighting system, instead of using NRMSE.

Response:

A weighting system that transforms the multi-objective problem to single objective problem is a very interesting research topic in the Multi-Objective Optimization (MOO) community. Both reviewers have raised this issue. In a nutshell, there are many weighting systems including but not limited to probability-based weighting, but the goal of MOO for LSMs is to evenly control the error of different outputs. It seems that, the adopted empirical weighting system has achieved this goal, and furthermore, a comprehensive inter-comparison of weighing systems might be an interesting work to be discussed in the future. In the revised version, we added a paragraph in section 4.2 to elaborate how to assign weights in MOO:

> *"In multi-objective optimization, there have been many methods that can transform multiple objectives to a single objective. Among them, the weighting function based method is the most intuitive and widely used one. In this paper, we assign higher weights to the outputs with larger errors. In the research of Liu et al. [2005], the RMSE of each outputs were normalized by the RMSE of default parameter set, and each normalized RMSE were assigned equal weights. van Griensven and Meixner [2007] developed a weighting system based on Bayesian statistics to define 'high probability regions' that can give 'good' results for multiple outputs. However, both Liu et al. [2005] and van Griensven and Meixner [2007] tended to assign higher weights to the outputs with lower RMSE, and lower weights to the outputs with higher RMSE. This tendency, although reasonable in the probability meaning, conflicts with our intuitive motivations that we want to emphasis on the poorly simulated outputs with large RMSE. [Jackson et al., 2003] assumed Gaussian error in the data and model so that the outputs were in a joint Gaussian distribution, and the multi-objective 'cost function' was defined on the joint Gaussian distribution of multiple outputs. In Gupta et al. [1998], a multiple weighting function method is proposed to fully describe the Pareto frontier, if the frontier is convex and model simulation is cheap enough. If one output is more important than the others, a higher weight should be assigned to it. Marler and Arora [2010] reviewed the applications, conceptual significance and pitfalls of weighting*

function based optimal methods, and gave some suggestions to avoid blind use of it."

Comment:

Page 6720, line 16: NRMSEs were calculated individually for each of the 6 outputs, and used as linear weights in the multi-objective function. Although it is good to look at several output variables at the same time, the outputs in this study are dependent on each other (in fact, sensible heat and latent heat would be strongly corrected). Therefore a linear combination of the misfits is questionable. I wonder if a weighting system based on their covariance matrix or joint pdf would be applicable.

Response:

The author raised a very interesting issue that is worthy of further discussion and experiments. The goal of this paper is to integrate mature and robust techniques to do parameter optimization in order to improve the performance of CoLM. As indicated by the results of both calibration period and validation period, the adopted framework, including the linear weights and NRMSE objects, are seemingly working well.

The outputs, such as sensible and latent heat, might be strongly correlated, but the NRMSEs may not. It is confirmed by the figure 2 that sensible heat requires small P4, but latent heat requires large P4. Both of them prefer large P6 and small P36 as well.

The covariance matrix based weighting system, which assumes Gaussian errors in data and model, is in a way similar to the linear weighting because of the joint Gaussian assumption. The non-Gaussian joint PDF weighting, although might be more flexible, is very rare (to my best knowledge) because high-dimensional non-Gaussian distribution is hard to describe in a simple parametric way, while for Gaussian it's very easy.

For more information, please see the discussion about weighing system in section 4.2 (as shown in the response to comment 1).

Comment:

Page 6720, line 6: the river basin has different land use types, but the study uses data from a single station at the upstream. SO which land use type is used in the study? In addition, is the data from the station representative of the big modeling domain? Page 6720, line 16: "the" should be "then"? What soil properties are linearly interpolated? If the authors meant soil temperature and moisture, how about measurements and interpolation of the other hydraulic properties? Are they vertically and horizontally heterogeneous?

Response:

The land use type of A'rou station is alpine steppe (as shown in Page 6720, line 10 of the original version). The simulated area of the 'single column CoLM' is a 0.05 °×0.05 °square. The land use type and the soil texture in Heihe River basin have variety. We are not using the A'rou station to represent the whole basin, but only use the corresponding land use and soil texture to carry out a single-column simulation and optimize it.

Page 6720, line 16: The simulated soil moisture and soil temperature are interpolated to the measured depth, not the soil hydraulic properties. This sentence has been revised as follows:

"In CoLM, the soil is divided into 10 layers and the simulated soil temperature and soil moisture are linearly interpolated to the measured depth. Currently we have 2 years

observation data."

Comment:

Page 6721, line 29: a solid evaluation of the developed surrogate is to break the dataset into training and testing subsets, and evaluate NRMSE for both. A reasonable surrogate should have low training and testing errors by considering both goodness of fit and avoiding over-fitting.

Response:

Actually, the figure 1 in original draft is the error of testing set. In the revised version, we added an additional figure showing the error of training set, and the following descriptions.

"Figure 1 shows the error of the training set, namely the NRMSE between the outputs predicted by the surrogate model and the outputs of the training samples, and figure 2 shows the NRMSE of the testing set. Since every sample set of each size was independently generated, we use the 2000 points set to test 50, 100, 200, 400, 800 and 1200 points set, and use the 1200 one to test the 2000 one."

Following discussion about the goodness of fit and over-fitting was also added.

"As shown in Figure 1, for some cases, such as upward longwave radiation, the fitting ability of the training set does not change significantly with sample size, but for soil moisture, larger sample size leads to better fitted surrogate model. Such phenomenon indicated that the specific features of the response surfaces have significant influence on the fitting ability, and good surrogate models must have the ability to adapt to those features. As shown in Figure 1, GPR has the best fitting ability for almost every case except soil temperature. As described in Appendix 2, the hyper-parameters used by GPR can be adaptively determined using the maximum marginal likelihood method.

Figure 2 shows the NRMSE of the testing sets, indicating the risk of over-fitting. In Figure 2 we can note more remarkable findings:"

Comment:

Page 6723, line 19: "sample size does not . . ." not true for the latent heat and soil moisture data.

Response:

May be the expression is misleading. Please see the line above. "…For some variables (sensible heat, upward longwave radiation, net radiation, soil moisture),…" Latent heat and soil moisture were not in the list.

In the revised version this sentence has been replaced by the following one.

"Surprisingly, for four of the outputs, namely some variables (e.g., sensible heat, upward longwave radiation, net radiation, and soil moisture), sample size does not have significant influence on the optimization results."

Comment:

Page 6723, line 23: "200 sample points might be sufficient. . ." the number of samples needed should vary for different observational data (e.g., sensible heat vs soil temperature)

Response:

It is true that the number of samples needs varies for different outputs. As shown in table 3,

200 samples might be sufficient for soil temperature, 400 samples are enough for latent heat. For others, surprisingly, only 50 samples may be enough. Interestingly, the LH50's NRMSE of sensible heat is even smaller than that of LH2000. It might be because LH sampling is a random sampling, and in the LH50 there is a sample point which happened to be very close to the global optimum, while for LH2000 the best sample point may not be as close. We inserted discussion on this point:

"Interestingly, the LH50's optimization result for sensible heat is even smaller than that of LH2000. This is because LH sampling is random and the LH 50 sampling may have produced a sample point very close to the global optimum, while the best sample point of LH2000 sampling may be further away from the global optimum. Consequently, the number of samples required for surrogate based optimization varies for different outputs because of the randomness of sampling designs, and the complexity of response surfaces. A more complex surface needs more sample points to build an effective surrogate model, compared to simple surface. Even so, this result is very encouraging that with the help of surrogate models we can possibly reduce the number of model runs required by optimization down to hundreds of times."

Comment:

Page 6724, line 19: I agree that for practical reasons, we want to have a single best parameter set, but people have preferences assigning the weights to data. It is fine to assign higher weights to better-simulated outputs (i.e., smaller NRMSE). However, Table 4 shows that the authors assigned higher weights to outputs with larger NRMSE, that is, more poorly-simulated ones.

Moreover, a probability-based weight (i.e., Wi~exp(-NRMSEiˆ2)) could be easier to interpret than NRMSE itself.

Response:

Let me explain why we assign higher weights to the outputs with larger NRMSE. Consider two outputs A and B, if we want to optimize A without considering B, we assign Wa = 1 and Wb = 0. Similarly, if we want to consider A twice as important as B, assign Wa = 2/3 and Wb = 1/3. In this case study, every output is important but we want to improve the worst ones, so a larger weight was assigned to outputs with larger NRMSE.

We are aware that someone may prefer the Bayesian based weighting, as is the case by van Griensven, and Meixner [2007]. But in our opinion, if one assigns lower weights to large error outputs and higher weights to small error outputs, the optimization would emphasize the small error outputs, and the large error outputs would have less improvement. The reviewer suggested "probability-based weight (i.e., Wi~exp(-NRMSEiˆ2))" also assign large weight to small error outputs, which would have the opposite effect to the ones we employed. So in this manuscript, we didn't use the "probability based" and stayed with our original approach.

van Griensven, A. and T. Meixner, A global and efficient multi-objective auto-calibration and uncertainty estimation method for water quality catchment models. JOURNAL OF HYDROINFORMATICS, 2007. 9(4): p. 277-291.

Comment:

Page 6726, line 6: Figures 4/5: the performance for soil temperature is worse, due to the low weight assigned to the temperature data. It is useful to expand the discussion, including the mathematical form and shape of the surrogates.

Response:

We have added some discussion in the end of section 4 as suggested.

"In the optimization results, five outputs were improved but only soil temperature became worse. In multi-objective optimization, compromise is necessary. In this case study, soil temperature requires small P6 and large 36, which conflict with all other outputs. Consequently, improving every output is impossible and some output must be sacrificed. If the cost is affordable and the gain is big enough, such compromise might be worthwhile. In this case study, the smallest weight was assigned to soil temperature so that its priority is the lowest. In the optimal solution, the RMSE of soil temperature increases from 2.66 degree to 2.90 degree (only 0.24 degrees larger), but other outputs RMSE can all be improved by about 10%. We think the sacrifice of soil temperature is worthwhile because a negligible degradation of one output can lead to significant improvement of all other outputs."