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."*

Response letter to Anonymous Referee #2

Comment:

This paper presents an application of an efficient parameter optimization technique developed by the authors. As the method is recently published and parts of it have already been presented elsewhere, the paper's contribution lies in the analysis of different surrogate model construction methods with respect to their application to multi-objective optimization of a land surface model. To a certain extent, the paper achieves this aim. There are however a number of improvements that I would expect to see before publication as a final paper.

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:

The most important improvements required are:

1. Acknowledgement and literature review of existing surrogate-based optimisation Techniques.

2. Discussion of the generality of the conclusions and assumptions underlying the results obtained

3. Clarification of some points which are not sufficiently well presented

4. Editing of typos, editing of language and filling in of placeholders

e.g. 'Artificial Neural Network (ANN) (REF) is a time-hornored marchine learning method comparing to the former four' p6736, L3

Response:

Thank you for your helpful comments. The revised manuscript has been improved in the following aspects.

1. A literature review about surrogate-assisted optimization and its application in hydrology has been added in the introduction section.
2. A paragraph about limitations of current work and future research was added to section 5: Discussion and conclusions.
3. Clarification of 'weighting functions', 'uncertainty quantification framework' and 'elbow points' as well as other topics were added to the revised version.
4. The placeholder 'REF' was replaced by [Jain et al., 1996] and typo 'marchine' was also corrected. Other typos were also corrected, as listed in the end of the response letter.

Comment:

While the paper is already of interest, these changes would improve the quality of the manuscript and give the reader a clearer impression of the context and utility of the method proposed by the authors.

More detailed discussion follows:

1. Acknowledgement and literature review of existing surrogate-based optimization

techniques (dating back to at least 2001)
Given that the paper aims to promote the use of the author's new adaptive surrogate model based optimization (ASMO) strategy (p6718 L25), I would expect to see acknowledgement and a brief overview of existing surrogate-based optimization techniques within the main text, even if the authors have mentioned this literature in other papers/journals.

The paper appears to contain only one other reference to existing surrogate-based optimisation techniques: Song, X., Zhan, C., and Xia, J.: Integration of a statistical emulator approach with the SCE-UA method for parameter optimization of a hydrological model, Chinese Sci. Bull., 57, 3397–3403, 2012. p6729, L18 Appendix A1 'MARS method can be used as parameter screening method (Gan et al., 2014; Li et al., 2013; Shahsavani et al., 2010), and also surrogate modeling method (Razavi et al., 2012; Song et al., 2012; Zhan et al., 2013).'

A quick search for the terms "surrogate optimization" brings up at least the following references, dating back to at least 2001 and including several reviews, a book and open-source implementations.

Ong, YewSoon, P B Nair, A J Keane, and K W Wong. 2005. "Surrogate-Assisted Evolutionary Optimization Frameworks for High-Fidelity Engineering Design Problems." In Knowledge Incorporation in Evolutionary Computation SE - 15, edited by Yaochu Jin, 167:307–31. Studies in Fuzziness and Soft Computing. Springer Berlin Heidelberg. doi:10.1007/978-3-540-44511-1_15.

Koziel, Slawomir, and Leifur Leifsson, eds. 2013. Surrogate-Based Modeling and Optimization. New York, NY: Springer New York. doi:10.1007/978-1-4614-7551-4.

Forrester, Alexander I.J., and Andy J. Keane. 2009. "Recent Advances in Surrogate-Based Optimization." Progress in Aerospace Sciences 45 (1-3): 50–79. doi:10.1016/j.paerosci.2008.11.001.

Jin, Yaochu. 2011. "Surrogate-Assisted Evolutionary Computation: Recent Advances and Future Challenges." Swarm and Evolutionary Computation 1 (2): 61–70. doi:10.1016/j.swevo.2011.05.001.

Jones, Donald R. 2001. "A Taxonomy of Global Optimization Methods Based on Response Surfaces." Journal of Global Optimization 21 (4). Kluwer Academic Publishers: 345–83. doi:10.1023/A:1012771025575.

Queipo, Nestor V., Raphael T. Haftka, Wei Shyy, Tushar Goel, Rajkumar Vaidyanathan, and P. Kevin Tucker. 2005. "Surrogate-Based Analysis and Optimization." Progress in Aerospace Sciences 41 (1): 1–28. doi:10.1016/j.paerosci.2005.02.001.

Zhou, Aimin, Bo-Yang Qu, Hui Li, Shi-Zheng Zhao, Ponnuthurai Nagaratnam Suganthan, and

Qingfu Zhang. 2011. “Multiobjective Evolutionary Algorithms: A Survey of the State of the Art.” Swarm and Evolutionary Computation 1 (1): 32–49. doi:10.1016/j.swevo.2011.03.001.

http://www.mathworks.com/matlabcentral/fileexchange/38530-surrogate-model-optimization-toolbox

Optimization methods that aim to provide satisfactory solutions given a limited computational budget should also be mentioned, in light of the comment that "Such parameter set might not be the true global optimum, but it is the “not bad” solution that is cheap enough we can afford." (P6725, L23) e.g. Tolson, B. A., and C. A. Shoemaker (2007), Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, Water Resour. Res., 43, W01413, doi:10.1029/2005WR004723.

Response:

Thank you for your helpful suggestion. We added literature review on surrogate based optimization for LSMs in the introduction section. Please note that this paper is not intended to propose new methods or theories, but to integrate existing techniques to improve the simulation ability of a LSM. See below on the revision:

*“Surrogate based optimization is one of the most commonly used approaches to optimizing large complex dynamic models. Several books and literature reviews have described the advances of surrogate based optimization in recent years [e.g., Jones, 2001; Ong et al., 2005; Jin, 2011; Koziel and Leifsson, 2013; and Wang et al., 2014]. Surrogate based optimization has been applied to economics, robotics, chemistry, physics, civil and environmental engineering, computational fluid dynamics, aerospace designs, et al [Gorissen, 2010]. On the development of surrogate based optimization, Jones et al. [1998] proposed EGO (Effective Global Optimizer) for expensive models using ‘DACE stochastic process model’, namely Kriging interpolation method, as surrogate model. Castelletti et al. [2010] developed a multi-objective optimization method for water quality management using radial basis function, inverse distance weighted and n-dimensional linear interpolator as surrogates. Loshchilov et al. [2010] investigated the use of ranked-based Support Vector Machine and demonstrated that for surrogate based optimization capturing the relative value of the objective functions is more important than reducing the absolute fitting error. Pilát and Neruda [2013] developed a surrogate model selector for multi-objective surrogate-assisted optimization. In hydrology and water resources, Razavi et al. [2012] has summarized recent applications, advantages, and existing problems. Wang et al. [2014] evaluated the influence of initial sampling and adaptive sampling methods for surrogate-assisted optimization of a simple hydrological model, SAC-SMA model. Song et al. [2012] optimized the parameter of a distributed hydrological model-DTVGM model’s parameter with SCE-UA algorithm using MARS method [Friedman, 1991] as surrogate.”*

Comment:

2. Discussion of the generality of the conclusions and assumptions underlying the results obtained The application uses 40 parameters (p6719 L11) of a single land surface model

(CoLM) applied to a single column case study (p6716, L17). It concludes (p6716 L19-22) that: "The result indicated that this framework can achieve optimal parameter set using totally 411 model runs, and worth to be extended to other large complex dynamic models, such as regional land surface models, atmospheric models and climate models."

This seems like a big jump given that doing so might involve:

- scalability of the technique to more parameters
- suitability of the response surface of the "other large complex dynamic models" to be fit by the surrogate technique
- suitability of the runtime of these larger models (411 may still be prohibitively large?)
- case-specific requirements as to how close it is necessary to be to the optimal value
- given that the result is not actually optimal, but rather 'similar' with the one gotten from SCE method using more than 1000 model runs' (p6728, L12)
- availability of software that can be used with those larger models - given that software availability is not discussed at all
- varying impacts of considering only sensitive parameters, where insensitive parameters may have significant interactions with sensitive parameters
- effect of stochasticity in sampling points to build and adapt the surrogate

...

At the very least, it would be useful for the paper to try to explain the factors affecting performance of the method, and its corresponding limitations.

Response:

Thank you for your constructive suggestions.

The last sentence of the abstract was revised as follows.

*"The result indicates that this framework can efficiently archive optimal parameters in a more effective way. Moreover, this result implies the possibility of calibrating other large complex dynamic models, such as regional-scale land surface models, atmospheric models and climate models."*

The following sentence was removed from the conclusions.

*"Consequently this framework is suitable to be applied to more large complex dynamic system models, such as regional land surface models, atmospheric models and even global climate models."*

In the revised section 5, we added a lot of discussion about the factors affecting the performance, the method's limitations, and future works.

*"In the future work, we are going to extend the uncertainty quantification framework to other large complex dynamic models, such as regional-scale land surface models, atmospheric models and climate models. We will look into testing the scalability of the screening, surrogate modeling and optimization techniques on more complex models with more adjustable parameters. We will also investigate the influence of uniformity and stochasticity of initial sampling points, and compare the suitability of different sampling methods. In addition to examining the main and total effects of the parameters, we will also evaluate the interactions among parameters. We will continue to improve the effectiveness, efficiency, flexibility and robustness of Gaussian Processes*

*Regression approach for surrogate modeling, and test with more complex models. Since weighting function based multi-objective optimization methods are simple, intuitive and effective, an inter-comparison of different weighting systems can be an interesting topic worthy of further research. Further, we intend to investigate ways to identify Pareto optimal parameter sets using a surrogate based optimization approach.*

*Discussion and collaborations are warmly welcomed on this and ongoing works. The computer code used in this study is available from the first author, which going to be published as part of the 'UQlab' software package in the future."*

Comment:

3. Clarification of some points which are not sufficiently well presented

- p6716 the paper refers to the framework alternatively as a "uncertainty qualification framework" and "uncertainty quantification framework". In any case it is unclear how uncertainty is addressed at all, given that the result is a single set of optimal parameters corresponding to a single weighted objective function of multiple outputs.

Response:

To be consistent, we use "uncertainty quantification framework" in the revised version.

This "uncertainty quantification framework" includes but not limited to sensitivity analysis, parameter screening, surrogate modeling, single/multi-objective optimization, confidence interval analysis and risk analysis. Parameter specification is one major source of model uncertainty, and parameter uncertainty is the most effective way to reduce the uncertainty. So in this paper, we kept using the term "uncertainty quantification framework".

Comment:

- Discussion of the sufficient number of points should recognize the statistical fact that error commonly continues to decrease as sample size increases and that the sufficient number of points therefore depends on the required error for a particular purpose. It therefore seems misleading to say that: 'error becomes stable' p6721, L28 'elbow points' p6722, L5

It may also be of interest to mention that the absolute error in the surrogate's estimate of the objective function is less important than the ability of the surrogate to capture the relative value of the objective function across parameter space.

Loshchilov, Ilya, Marc Schoenauer, and Michèle Sebag. 2010. "Comparison-Based Optimizers Need Comparison-Based Surrogates", September. Springer-Verlag, 364-73. http://dl.acm.org/citation.cfm?id=1885031.1885071.

Response:

Thank you for your suggestion and the draft was revised as follows:

(1) "The error becomes stable when the sample size is larger than 400. More samples can reduce the error but the benefit of additional samples is marginal." => "The marginal benefits of additional samples becomes less or even negligible if the sample size is larger than 400."

(2) "The elbow points (i.e., the point at which the objective function value changes from rapid decrease to a gradual one) of net radiation, soil temperature and soil moisture are significantly at 200 sample points, while for sensible heat, latent heat and upward long-wave radiation, the elbow points are not clear." => "For net radiation, soil temperature and soil

moisture, the fitting error decreases to nearly zero if the sampling points are more than 200; while for sensible heat, latent heat and upward long-wave radiation, the marginal benefit of adding more points is still significant for more than 200 sample points."

(3) The suggested reference was added to the literature review in section 1, as shown in the response to comment 1.

Comment:

- The paper combines multiple objectives using weights. It states: "assign more weight to the objective function output, if that output is simulated more poorly as compared to other outputs" (p6724, L26) Clarification is needed to the effect that the weighting is calculated based on performance of the default parameters, i.e. outputs that need more improvement are emphasized.

Otherwise I would expect that poor performance might be an indicator of error in data or model structure, in which case it is customary to assign weight inversely proportional to the error. The reference cited (Liu et al. 2005) divides each objective function by its performance with the default parameters, which appears to have the opposite effect-of giving less weight to poorly performing outputs?

Response:

The reviewer raised an interesting concern about the weighing system. The reviewer #1 also asked similar question. Actually in [Liu et al., 2005], Wi was proportional to 1/fi(default), which means assign larger weights to smaller error outputs. [Liu et al., 2005]'s explanation is (1) Define objective function fi=RMSEi, and then the normalized objective fi'=NRMSEi, in which the RMSE was normalized by RMSE simulated by default parameters. (2) Assign equal weights to the normalized objectives. Consequently, the weights assigned to RMSEi was actually 1/RMSEi(default). The consideration of [Liu et al., 2005] was to averagely assign weights to each output, considering their magnitude of error, and make sure their weights were approximately the same after normalization.

In this paper, our consideration is as follows: 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.

A paragraph about our thinking about weighing systems, as well as our plan of future works on transforming multi-obj to single-obj, has been added to section 4.2.

> *"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:

4. Editing of typos, editing of language and filling in of placeholders There are sufficiently many improvements necessary that the manuscript should be thoroughly reviewed in full. This includes but is not limited to pages 6716, 6718, 6723-6728, 6730-6733, 6736

Response:

The suggestion is very helpful and the manuscript has been revised thoroughly, the typos are listed below.

Line 23: need => needs; huge=>large

Line 25: quantification

Line 26: aforementioned challenges => aforementioned challenges, which include the following steps

Line 26, 27, 29, 30: use => using

Line 29: promote => improve

Line 32: case study of a => application of a

Line 33: Common Land Model => the Common Land Model

Line 81: multi-objective optimization => multi-objective optimization approach

Line 82: "know how to" deleted

Line 109: "only" deleted

Line 114: in => of

Line 114: models => model

Line 115: on => with

Line 123: on => of

Line 125: will => would

Line 128: conclusion => discussion and conclusions

Line 145: Normalized Mean Squared Error => Normalized Root Mean Squared Error

Line 183: even millions => even up to millions

Line 189: may => can

Line 192: pitfall => pitfalls

Line 199: The sample sizes are set to 50, 100, 200, 400, 800, 1200, and 2000. => add "respectively" to the end of the sentence.

Line 224: The error of surrogate model => The error of a surrogate model

Line 231: good => acceptable

Line 232: speed => speeds
Line 232: is => are
Line 240: “among all methods,” deleted
Lind 240: stable => consistently
Lind 241: “following” deleted
Lind 242: multi-objective optimization analysis. => multi-objective optimization analysis presented later.
Line 257: “in the following text” deleted
Line 260: sample points => model runs
Line 262: converged => optimal
Line 262: from the => given by
Line 263: “runs” deleted
Line 275: “significantly” deleted
Line 277: given => obtained; “the” deleted
Line 282: size => sizes
Line 299: cost => number; “original” deleted; larger than => larger than that
Line 302: efficiency => efficient; SCE => the SCE
Line 305 and 306: for P6 => of P6
Line 307: “measures” deleted
Line 308: “objective” deleted
Line 319: conditions => constrains
Line 321: non-inferior => non-dominated
Line 376:
Such parameter set might not be the true global optimum, but it is the “not bad” solution that cheap enough we can afford. =>
Such parameter set can provide only the approximate global optimum, but this approach is much cheaper than using traditional approaches such as SCE-UA.
Line 387: with => to that by
Line 390: “at” deleted
Line 398: during => for the; validation period => the validation period
Line 399, 406, 420, 468: SCE => SCE-UA
Line 405: parameter => parameters
Line 407: expect => except
Line 407: Even though => Even though soil temperature simulation is degraded
Line 410: of => in; validation period => the validation period
Line 411: is quite similar with => is shown quite similar to; calibration period => the calibration period
Line 424: provide => provides
Line 426: A’rou frozen/thaw station => A’rou station, where frequent freezing and thawing occur.
Line 430: frozen/thaw => freezing/thawing
Line 436: is very different => can be very different
Line 436: for an instance => for instance
Line 438: place => places

Line 440: otherwise => further
Line 442: can’t => cannot
Line 458: many similar works => other studies
Line 465: it is impractical to parameter optimization => parameter optimization is impractical
Lien 468: gotten from => obtained by
Line 470: with only hundreds of model runs => much efficiently
Line 543: test input and output => testing inputs and outputs
Line 545: predict output => predicted outputs
Line 556: positive hyper parameters => are positive hyper parameters
Line 572: Random Forests are => Random Forest is
Line 586, 588, 598: random forests => a random forest
Line 587: output => outputs
Line 599: outstanding performance in => outstanding performance for
Line 601: with each one only provides a little => but each feature provides only a little
Line 604: using => with
Line 642: REF => [Jain et al., 1996]
Line 643: marchine => machine

# Multi-objective Parameter Optimization of Common Land Model Using Adaptive Surrogate Modelling

**Wei Gong[1], Qingyun Duan[1], Jianduo Li[1], Chen Wang[1], Zhenhua Di[1], Yongjiu Dai[1], Aizhong Ye[1], Chiyuan Miao[1]**

[1]College of Global Change and Earth System Science (GCESS), Beijing Normal University, Beijing, 100875, and Joint Center for Global Change Studies, Beijing, 100875, China

*(gongwei2012@bnu.edu.cn, qyduan@bnu.edu.cn, lijianduo@mail.bnu.edu.cn, wangchen@mail.bnu.edu.cn, zhdi@bnu.edu.cn, yongjiudai@bnu.edu.cn, azye@bnu.edu.cn, miaocy@vip.sina.com )*

**Abstract:**

Parameter specification usually has significant influence on the performance of land surface models (LSMs). However, estimating the parameters properly is a challenging task due to the following reasons: (1) LSMs usually have too many adjustable parameters (20 to 100 or even more), leading to the curse of dimensionality in the parameter input space; (2) LSMs usually have many output variables involving water/energy/carbon cycles, so that calibrating LSMs is actually a multi-objective optimization problem; (3) Regional LSMs are expensive to run, while conventional multi-objective optimization methods need~~s~~ a ~~huge~~ large number of model runs (typically $10^5$~$10^6$). It makes parameter optimization computationally prohibitive. An uncertainty ~~qualification~~ quantification framework was developed to meet the aforementioned challenges, which include the following steps: (1) us~~e~~ing parameter screening to reduce the number of adjustable parameters; (2) us~~e~~ing surrogate models to emulate the response~~s~~ of dynamic models to the variation of adjustable parameters; (3) us~~e~~ing an adaptive strategy to ~~promote~~ improve the efficiency of surrogate modeling

based optimization; (4) use~~e~~ing a weighting function to transfer multi-objective optimization to single objective optimization. In this study, we demonstrate the uncertainty quantification framework on a single column application of a ~~case study of a~~ land surface model – the Common Land Model (CoLM) and evaluate the effectiveness and efficiency of the proposed framework. ~~The result indicated that this framework can achieve optimal parameter set using totally 411 model runs, and worth to be extended to other large complex dynamic models, such as regional land surface models, atmospheric models and climate models.~~ The result indicate that this framework can eff~~ec~~icient~~ive~~ly achieve optimal parameters in a more effective way. Moreover, this result implies the possibility of calibrating other large complex dynamic models, such as regional-scale land surface models, atmospheric models and climate models. ~~In the future work we are going to continue the development of the framework and improve its scalability, robustness and flexibility for other expensive models.~~

**Keywords:**
Land surface model; multi-objective optimization; parameter calibration; surrogate modeling; statistical emulator; adaptive sampling;

## 1. Introduction

Land surface models (LSMs), which offer land surface boundary condition for atmospheric models and climate models, are widely used in weather and climate forecasting. They are also a tool for studying the impacts of climate change and human activities on our environment. Parameters of land surface models usually have significant influence on their simulation and forecasting capability. It has been shown that tuning even one or two parameters may significantly enhance the simulation ability of a land surface model (e.g., [*Henderson-Sellers et al.*, 1996; *Liang et al.*, 1998; *Lohmann et al.*, 1998; *Wood et al.*, 1998]). How to specify the parameters in a LSM model properly, however, remains a very challenging task because the LSM parameters are usually not directly measurable at the scale of model applications.

Automatic optimization approaches have been frequently used in calibrating the

parameters of hydrological models. There is a plethora of optimization approaches available, including single-objective optimization methods such as SCE-UA [*Duan et al.*, 1992; *Duan et al.*, 1993; *Duan et al.*, 1994], SCEM-UA [*Vrugt et al.*, 2003], genetic algorithm [*Wang*, 1991], and multi-objective optimization methods such as MOCOM-UA [*Boyle et al.*, 2000; *Boyle*, 2000; *Gupta et al.*, 1998; *Yapo et al.*, 1998] and MOSCEM-UA[*Vrugt et al.*, 2003].

Compared to traditional hydrological models, the parameter calibration approach has not been practiced as much in LSM community, especially for large spatial scale applications. The major obstacles to calibrating land surface models over a large spatial scale are: (1) there are too many parameters to calibrate, (namely, the curse of dimensionality in parameters); (2) dimensionality of the output space is too high (i.e., many processes such as water/energy/carbon cycles are simulated simultaneously); (3) conventional optimization methods, especially multi-objective approach, need a large number (~$10^5$-$10^6$) of model runs; and the large complex dynamic system models such LSMs are usually expensive to run (i.e., costing many CPU hours). There have been numerous attempts to use multi-objective optimization to calibrate the parameters of land surface models and significant improvement in LSM performance measures as a result of optimization have been reported (e.g., [*Bastidas et al.*, 1999; *Gupta et al.*, 1999; *Leplastrier et al.*, 2002; *Xia et al.*, 2002]). However, the optimization efforts in the past were usually limited to cases studies involving only point or limited spatial domain-scale applications of LSMs ([*Liu et al.*, 2003; *Liu et al.*, 2004; 2005]. To take a multi-objective optimization approach to the calibration of LSM parameters for large scale applications, the key is to ~~know how to~~ reduce the number of model runs to an appropriate level that we can afford.

Surrogate based optimization is one of the most commonly used approaches to optimizing large complex dynamic models. Several books and literature reviews have described the advances of surrogate based optimization in recent years [e.g., Jones, 2001; Ong et al., 2005; Jin, 2011; Koziel and Leifsson, 2013; and Wang et al., 2014]. Surrogate based optimization has been applied to economics, robotics, chemistry, physics, civil and environmental engineering, computational fluid dynamics, aerospace

designs, et al [Gorissen, 2010]. On the development of surrogate based optimization, Jones et al. [1998] proposed EGO (Effective Global Optimizer) for expensive models using 'DACE stochastic process model', namely Kriging interpolation method, as surrogate model. Castelletti et al. [2010] developed a multi-objective optimization method for water quality management using radial basis function, inverse distance weighted and n-dimensional linear interpolator as surrogates. Loshchilov et al. [2010] investigated the use of ranked-based Support Vector Machine and demonstrated that for surrogate based optimization capturing the relative value of the objective functions is more important than reducing the absolute fitting error. Pilát and Neruda [2013] developed a surrogate model selector for multi-objective surrogate-assisted optimization. In hydrology and water resources, Razavi et al. [2012] has summarized recent applications, advantages, and existing problems. Wang et al. [2014] evaluated the influence of initial sampling and adaptive sampling methods for surrogate-assisted optimization of a simple hydrological model, SAC-SMA model. Song et al. [2012] optimized the parameter of a distributed hydrological model-DTVGM model's parameter with SCE-UA algorithm using MARS method [Friedman, 1991] as surrogate.

In our recent works, we proposed a framework that can potentially reduce the number of model runs needed for parameter calibration of large complex system models [*Wang et al.*, 2014]. This framework involves the following steps: (1) a parameter screening step using global sensitivity analysis to identify ~~only~~ the most sensitive parameters to be included in the optimization; (2) surrogate modelling that can emulate the response surface of the dynamic system model to the change in parameter values; (3) an adaptive sampling strategy to improve the efficiency of the surrogate model construction; (4) a multi-objective optimization step to optimize the most sensitive parameters of~~in~~ the dynamic system model~~s~~. In this paper, we will illustrate this parametric uncertainty quantification framework with~~on~~ the Common Land Model (CoLM), a widely used, physically based land surface model developed by Yongjiu Dai and colleagures [*Dai et al.*, 2003; *Dai et al.*, 2004; *Ji and Dai*, 2010]. The work related to parameter screening and surrogate modeling based optimization (ASMO) method for single objective has already been published [*Li et al.*, 2013; *Wang et al.*, 2014]. This

paper will emphasize on the analysis of different surrogate model construction methods and multi-objective optimization method and results.

This paper contains the following parts: section 2 introduces the basic information ~~on~~of CoLM, the study area and dataset, the adjustable parameters and the output variables to be analyzed; section 3 presents an inter-comparison of 5 surrogate modeling methods, and discusses how many model runs ~~will~~ would be sufficient to build a surrogate model for optimization; section 4 carries out single and multiple objective optimization using an adaptive surrogate model based optimization strategy (ASMO); section 5 provides the discussion and conclusions.

## 2. Experiment setup

### Model and Parameters

Common Land Model (CoLM) proposed by Yongjiu Dai and colleagues [*Dai et al.*, 2003; *Dai et al.*, 2004; *Ji and Dai*, 2010] is one of the most widely used land surface model in the world. It combines the advantages of Land Surface Model (LSM) [*Bonan*, 1996], Biosphere-atmosphere transfer scheme (BATS) [*Dickinson et al.*, 1993] and Institute of Atmospheric Physics land-surface model (IAP94) [*Dai and Zeng*, 1997; *Dai et al.*, 1998]. CoLM considers physical processes of energy and water transmission in soil vegetation, snow cover and atmosphere. It also implements glacier, lake, wetland and dynamic vegetation processes. Similar to previous research in presented in [*Li et al.*, 2013], we select 40 adjustable parameters from CoLM. The parameter names, physical meanings and value ranges are shown in **Table 1**.

**[Table 1]**

This study considers six output variables simulated by CoLM: sensible heat, latent heat, upward longwave radiation, net radiation, soil temperature and soil moisture. The Normalized Root Mean Squared Error is used as the objective function in our analysis:

$$NRMSE_i = \frac{\sqrt{\sum_{j=1}^{N}\left(y_{i,j}^{sim} - y_{i,j}^{obs}\right)^2}}{\sum_{j=1}^{N} y_{i,j}^{obs}} \quad (1)$$

where $i$ is the index of output variables, $j$ is the index of time step, $N$ is the total number

of observations, $y_{i,j}^{sim}$ and $y_{i,j}^{obs}$ are the simulated and observed values, respectively. Objective functions represent the performance of model simulation and a smaller objective function means better performance.

**Study area and datasets**

The study area and associated datasets are from the Heihe river basin, the same as in [*Li et al.*, 2013]. The Heihe river basin, which is located between 96 °42'-102 °00'E and 37 °41'-42 °42'N, is an inland river basin in the arid region of northwest China. The basin area is approximately 130,000 km$^2$ and its altitude varies from sea level to 5500m. The Heihe river basin has a variety of land using types including forest, grassland, farmland, and glacier, among others, making it an ideal research region for LSM simulation. In this research we use the data from A'rou observation station located at the upstream region of the Heihe river basin. Its geographic coordinate is 100 °28'E, 38~~0~~ °038~~8~~'N, altitude is 3032.8m above sea level and the land cover type is alpine steppe.

The forcing data used include downward shortwave and longwave radiation, precipitation, air temperature, relative humidity, air pressure and wind speed [*Hu et al.*, 2003]; and the observation data used to validate the simulation of CoLM include: sensible heat, latent heat, upward longwave radiation, net radiation, soil temperature and soil moisture. The soil temperature and moisture were measured at depth 10cm, 20cm, 40cm and 80cm. 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. The data from year 2008 was used for spin up and that of 2009 was used for parameter screening, surrogate modeling and optimization. The simulation time step is set to 30 minutes and the simulation outputs are averaged to 3 hours in order to compare with the observation data.

## 3. Comparison of Surrogate models

After the sensitive parameters are identified using global sensitivity methods (see [Li et al., 2013]), the next step is to calibrate the sensitive parameters through multi-

objective optimization. Since the calibration of CoLM in real world applications can be very expensive, we aim to establish a surrogate model to represent the response surface of the dynamic CoLM. Surrogate model, also called response surface, meta-model, statistical emulator, is a statistical model that describes the response of output variable to the variation of input variables. Because the surrogate model only considers the statistical relationship between input and output, it is usually much cheaper to run than the original large complex dynamic model ("original model" for short). Parameter optimization usually needs thousands, or even up to millions times of model runs. It will be impossible to calibrate large complex dynamic models if running the original dynamic model is too time consuming. If we can do parameter optimization with surrogate model instead of original model, the time of running original model will be dramatically reduced, making it possible to calibrate the large complex dynamic models, such as land surface models, atmospheric models, or even global climate models. However, optimization based on surrogate models ~~may~~ can be a challenging task because the response surface might be very bumpy and has many local optima. *Razavi et al.* [2012] gave a comprehensive review of the surrogate modeling methods and applications in water resources, and discussed the pitfalls of surrogate modeling as well.

In this research, we first compared 5 different surrogate models: Multivariate Adaptive Regression Spline (MARS), Gaussian Process Regression (GPR), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN). A brief introduction of these methods is provided in the **Appendix**. To build a surrogate, we need to choose a sampling method first. The sampling method used in this study is Latin Hypercube Sampling (LH) [*McKay et al.*, 1979]. The sample sizes are set to 50, 100, 200, 400, 800, 1200, and 2000, respectively. The inter-comparison results are shown in **Figure 1** and **Figure 2**, in which the x-axis is the sample size, and y-axis is the NRMSE (i.e., the ratio of the root mean square error (RMSE) of the simulation model and the surrogate model). **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. For each output variable, we only construct surrogate models for the most sensitive parameters based on the screening results ~~of several global sensitivity methods performed~~obtained by *Li* [2012] and *Li et al.* [2013] (see ~~. The list of screened sensitive parameters is~~are ~~shown in~~ **Table 2**).

**[Table 2]**

As shown in **Figure 1**, for some cases, such as upward longwave radiation, the fitting ability of the training set does no~~t~~t ~~significantly~~ change significantly with sample size, but for soil moisture, larger sample size leads to better fitted surrogate models. Such phenomenon indicated that the ~~variousness~~specific features of the response surfaces ha~~s~~ve significant influence ~~t~~on the fitting ability, and good surrogate models must ha~~s~~ve the ability to adapt~~ively identify and mitigate the various properties~~ to those features. As shown in **Figure 1**, GPR has the best fitting ability for almost every case~~s~~ except soil temperature. ~~Because a~~As described in Appendix 2, the hyper-parameters used by GPR can be adaptively ~~be adaptively~~ determined~~s~~ 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 ~~can get~~note more remarkable findings ~~as follows~~:~~**Figure 1** indicated that:~~ (1) The error of a surrogate model decreases as the sample size increases. The marginal benefits of using additional samples become~~s~~ less or even negligible if the sample size is larger than 400.~~;~~; ~~The error becomes stable when the sample size is larger than 400. More samples can reduce the error but the benefit of additional samples is marginal.~~ (2) Among the 5 different surrogate models, GPR has the best performance, while ANN ranks the second. RF and MARS have lower accuracy. For some output variables (e.g., sensible and latent heat), the performance of SVM seems ~~good~~acceptable, while for other variables (e.g., soil temperature), SVM's performance is not satisfactory~~;~~. (3) The convergence speeds for the 6 output variables ~~is~~ are different. For net radiation, soil temperature and soil moisture, the fitting error decreases to nearly zero if the sampling points are more than 200; while for sensible heat, latent heat and upward long-wave radiation, the marginal benefit of adding more points is still

significant ~~for~~beyond more than 200 sample points. ~~The elbow points (i.e., the point at which the objective function value changes from rapid decrease to a gradual one) of net radiation, soil temperature and soil moisture are significantly at 200 sample points, while for sensible heat, latent heat and upward long-wave radiation, the elbow points are not clear.~~ Since ~~among all methods,~~ the GPR method can ~~stably~~ consistently give the best performance for all the 6 output variables, we choose GPR in the ~~following~~ multi-objective optimization analysis presented later.

**[Figure 1]**

**[Figure 2]**

## 4. Optimization

### 4.1 Single-objective optimization

Before we conduct multi-objective optimization, we first carried out single-objective optimization for each output variable using the GPR surrogate model. The Shuffled Complex Evolution (SCE) method [*Duan et al.*, 1992; *Duan et al.*, 1993; *Duan et al.*, 1994] is used to find the optima of the surrogate models. In order to figure out how many sample points are sufficient to construct a surrogate model for optimization, different sample sizes (i.e., 50, 100, 200, 400, 800, 1200, and 2000) are experimented. To evaluate the optimization results based on the surrogate model, we also set up two control cases: (1) No optimization using the default parameters as specified in CoLM. (2) Optimization using the original CoLM (i.e., no surrogate model is used). The second case is referred as "direct optimization" ~~in the following text~~. The control cases are used to confirm the following hypotheses: (1) Parameter optimization can indeed enhance the performance of CoLM. (2) Optimization using the surrogate model can achieve similar optimization result as using the original model, but with fewer ~~sample points~~model runs.

The ~~converged~~ optimal parameters ~~from the~~given by single-objective optimization ~~runs~~ are shown in **Figure 3~~2~~**. In each subfigure the optimal parameter values are normalized to [0, 1]. The bold black line is the optimal parameter set obtained by direct optimization using the original CoLM, and other lines are optimal parameters given by

surrogate models created with different sample sizes. **Table 3** summarizes the optimized NRMSE values of all surrogate model based optimization runs with different sample sizes, as well as the control cases. The numbers of original model runs that SCE takes are also listed in the brackets.

**[Figure 3~~2~~]**

**[Table 3]**

The optimization results reveal that: (1) Parameter optimization can significantly improve the simulation ability of CoLM ~~significantly~~ for all output variables~~.~~; (2) For sensible heat, upward longwave radiation, net radiation, soil moisture, the optimal parameters ~~given~~ obtained by ~~the~~ surrogate model optimization runs are very similar to those obtained by direct optimization. The optimal parameters obtained for different sample sizes are also close to each other. For latent heat and soil temperature, however, the optimal parameters given by surrogate model optimization and direct optimization are significantly different. The discrepancy between the results with different sample sizes is also significant, comparing to the previous 4 outputs~~.~~; (3) 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. As shown in **table 3**, even a surrogate model constructed with 50 samples is similar to the one constructed with 2000 samples and with the direct optimization. For soil temperature, 200 samples are sufficient, and for latent heat, more than 400 samples are enough. 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; (4) The number of original model runs that SCE takes before convergence is also listed in **Table 3**. The result indicated that SCE can get better, or similar optimal NRMSE, but the ~~cost~~ number of ~~original~~ model runs is larger than that using surrogate model. If the original dynamic model costs too much CPU time to run, surrogate based optimization can be more efficien~~ct~~y than the SCE~~;~~. (5) Different output variables require different optimal parameters, indicating the necessity of multi-objective optimization. For example, P6, the Clapp and Hornberger "b" parameter, is sensitive to many outputs. For sensible heat, latent heat and soil moisture, the optimal value ~~for~~ of P6 is high, while for upward longwave radiation, net radiation and soil temperature, the optimal value ~~for~~ of P6 is low. In order to balance the performance ~~measures~~ of all output variables, it is necessary to choose a compromised value for P6. Multi-objective optimization is an ~~objective~~ approach that can provide such a compromised optimal parameter that balances the requirements of many output variables.

### 4.2 Multi-objective optimization

The results of single-objective optimization from previous section have highlighted the necessity for multi-objective optimization. Many multi-objective optimization methods have been proposed and validated in numerous studies (e.g., [*Boyle et al.*, 2000; *Boyle*, 2000; *Gupta et al.*, 1998; *Yapo et al.*, 1998; *Vrugt et al.*, 2003; *Bastidas et al.*, 1999; *Gupta et al.*, 1999; *Leplastrier et al.*, 2002; *Pollacco et al.*, 2013; *Xia et al.*, 2002])**,** but in the context of this research, we need a method that can satisfy the following ~~conditions~~constrains: (1) the method should be compatible with surrogate model optimization; (2) for practical reasons, it should provide a single best parameter set instead of a full Pareto optimum set with many non-~~inferior~~ dominated parameter sets. The Pareto optimal set usually contains hundreds of points, ~~with hundreds, or more parameter sets have appeals in that it can provide an ensemble of model outputs, which in turn can be used to assess parametric uncertainty. For~~ but for large complex dynamic models such as regional or global land surface models~~CoLM~~, it is generally

impractical, and also unnecessary to run the model in an ensemble mode with hundreds of model runs. For regional or global land surface models coupled with atmospheric models, providing only one parameter set that has good simulation ability for most outputs is a more economical and convenient choice.

In multi-objective optimization, there have been many methods that can transform multiple objectives to 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 of *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 outputs is more important than 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.

In this study, we use a weighting function method to convert the multi-objective optimization into a single objective optimization. ~~The weight assigned to each objective function is based on [Liu et al., 2005].~~ The general idea is that we assign more weight to the objective function of an output, if that output is simulated more poorly as compared to the other outputs. **Table 4** shows the RMSE and NRMSE of CoLM using default parameterization scheme, and the weight of each output is proportional to the

NRMSE.

**[Table 4]**

After the weights are determined, the weighted objective function is as follows:

$$F = \sum_{i=1}^{n} w_i RMSE_i \quad (2)$$

in which the $RMSE_i$ is the Root Mean Squared Error of each output variable that defined as $RMSE_i = \frac{1}{N}\sqrt{\sum_{j=1}^{N}\left(y_{i,j}^{sim} - y_{i,j}^{obs}\right)^2}$, $w_i$ is the weight of each output, and $\sum_{i=1}^{n} w_i = 1$.

In order to use the information offered by surrogate model more effectively, we developed an adaptive surrogate modeling based optimization method called ASMO [*Wang et al.*, 2014]. The major steps of ASMO are as follows: (1) Construct a surrogate model with initial samples, and find the optimal parameter of the surrogate model. (2) Run the original model with this optimal parameter and get a new sample. (3) Add the new sample to the sample set and construct a new surrogate model, and go back to the 1st step. The effectiveness and efficiency of ASMO have been validated in [*Wang et al.*, 2014] using 6D Hartman function and a simple hydrologic model SAC-SMA. As shown in the comparison between ASMO and SCE-UA, ASMO is more efficient that can ~~get archive convergence~~ converge with less model runs, while SCE-UA is more effective that can get closer to the true global optimal parameter. So making a choice between ASMO and SCE-UA is a "cost-benefit" trade-off: if the model is very cheap to run, SCE-UA is preferred because it is more effective to find the global optimum; while if the model is very expensive to run, ASMO is preferred because it can find a fairly good parameter within a limited time of model runs. Such parameter set ~~might not be~~can provide only the ~~true~~approximate global optimum, but ~~it~~his approach is ~~the "not bad" solution that is~~much cheaper ~~enough we can afford~~than using traditional approaches such as SCE-UA.

We carried out multi-objective optimization with ASMO using weighting function defined in **equation (2)** and the optimization results are shown in **figure ~~3~~4 and ~~4~~5**. To compare, we also carried out the direct optimization using SCE-UA. **Figure ~~3~~4** presents

the default parameter, the optimal parameter given by ASMO and that given by SCE-UA. **Figure 45** shows the improvements given by ASMO and SCE-UA comparing to the default parameters. From **Figure 45** we can find that all of the outputs are improved nearly 10% except soil temperature, and the improvements made by ASMO is similar ~~with~~ to that by SCE-UA. The results indicated that multi-objective optimization can indeed enhance the performance of CoLM using either ASMO or SCE-UA method. The ASMO method get converged after 11 iterations, namely, the total number of model runs is 411, while the number of model runs for SCE-UA is ~~at~~ 1000, which is the maximum model runs set for SCE-UA. Obviously ASMO is a more efficient method compared to SCE-UA in this case.

**[Figure ~~3~~4]**

**[Figure 45]**

We also used the Taylor diagram [*Taylor*, 2001] to compare the simulation results ~~during~~ for the calibration period and the validation period (see **figure ~~5~~6 and ~~6~~7)**. The optimization results given by SCE-UA and ASMO are compared against the performance of default parameterization scheme. Since only 2 years observation data of the 6 output variables are available, we use the first year (2008) data as the warm-up period, use the second year (2009) data as calibration period, and then use the previous 2008 year data as the validation period. The missing records have been removed from the comparison.

As indicated in **figure ~~5~~6,** the performance of optimized parameters given by both SCE-UA and ASMO (Case C and D in the Taylor diagram) are better than default parameterization scheme (Case B) exc~~p~~e~~c~~pt soil temperature. Even though soil temperature simulation is degraded, the correlation coefficients given by all the three cases are higher than 0.9, indicating that this imperfection will not cause significant inconsistency ~~of~~in the land surface modelling. In **figure ~~6~~7**, the performance of the validation period is shown quite similar ~~with~~to that in the calibration period, indicating that the optimal parameters are well identified and the over-fitting problem is avoided.

**[Figure ~~5~~6]**

**[Figure ~~6~~7]**

The four energy fluxes (sensible/latent heat, upward long-wave radiation, net radiation) and soil surface temperature have very good performance. However, the performance of soil moisture seems not satisfactory. The correlation coefficient of soil moisture of Case B(default parameter) is less than 0, while with the help of SCE-UA and ASMO optimization the correlation coefficient is only slightly larger than 0. The possible reasons might be as follows: (1) The default soil parameters of CoLM is derived from the soil texture in the 17-category FAO-STATSGO soil dataset [*Ji and Dai*, 2010], which provides top-layer (30cm) and bottom-layer (30-100cm) global soil textures and has a 30 seconds resolution. The resolution and accuracy of this dataset may be not accurate enough in A'rou ~~frozen/thaw~~ station, where frequent freezing and thawing occur. A finer soil database, such as 'The Soil Database of China for Land Surface Modeling' [*Shangguan et al.*, 2013], or an in-situ field survey for soil texture, should be used to improve the quality of default parameterization scheme~~.~~: (2) Simulating fr~~ozen~~eezing/thawing processes is a challenging task in land surface modeling, and we are still lack of knowledge about the details of the physical processes. Parameter optimization can improve the model performance if the model physics are correct, but ~~optimization~~ is helpless if the model structure is inconsistent with the true underlying physical processes. Although CoLM's performance of simulating frozen soil and snow cover has been evaluated in the experiment in Valdai, Russia [*Dai et al.*, 2003], the situation of Heihe in China ~~is~~can be very different. For ~~an~~ instance, in CoLM the soil depth is set to 2.86m globally, but actually the soil depth varies in different places. Fundamentally such a simplification may not introduce significant error to the simulation of energy flux, but it definitely influence the performance of hydrological processes such as soil moisture. ~~Otherwise~~Further, the altitude of Heihe is much higher than Valdai, and the influence of human activities is also significantly different. All these reasons can potentially influence the performance of CoLM and can~~'~~not be

mitigated by parameter optimization., so we should revise the model physics before parameter optimization.

In the optimization results, five of the outputs were improved but only soil temperature became worse. In multi-objective optimization, a 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 might 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. 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.

## 5. Discussion and Conclusions

We have carried out multi-objective parameter optimization for a land surface model, CoLM, at the Heihe river basin. Although there have been many similar worksother studies, such as multi-objective calibration of hydrological models [*Gupta et al.*, 1998; *Vrugt et al.*, 2003], land surface models [*Gupta et al.*, 1999], single column land-atmosphere coupled model [*Liu et al.*, 2005], and SVAT model [*Pollacco et al.*, 2013], the novel contribution of this research lies in the significant reduction of model runs. In previous researches, a typical multi-objective optimization needs $10^5$~$10^6$ or even more model runs. For large complex dynamic models which are very expensive to run, parameter optimization it is impractical to parameter optimization because of lack of computational resources. In this research, we managed to achieve a multi-objective optimal parameter set with only 411 model runs. The performance of the optimal parameter set is similar with the one gotten obtainedfrom by SCE-UA method using more than 1000 model runs. Such a result indicates that the proposed framework in this paper is able to provide optimal parameters with only hundreds of model runsmuch efficiently. Consequently this framework is suitable to be applied to more large complex dynamic system models, such as regional land surface models,

~~atmospheric models and even global climate models.~~

In the future work, we are going to extend the uncertainty quantification framework to other large complex dynamic models, such as regional-scale land surface models, atmospheric models and climate models. We will look into testing the scalability of the screening, surrogate modelling and optimization techniques on more complex models with more adjustable parameters. We will also investigate the influence of uniformity and stochasticity of initial sampling points, and compare the suitability of different sampling methods. In addition to examining the main and total effects of the parameters, we will also evaluate the interactions among parameters. We will continue to improve the effectiveness, efficiency, flexibility and robustness of Gaussian Processes Regression approach for surrogate modelling, and test with more complex models. Since weighting function based multi-objective optimization methods are simple, intuitive and effective, an inter-comparison of different weighting systems can be an interesting topic worthy of further research. Further, we intend to investigate ways to identify P~~p~~areto optimal parameter sets using a surrogate based optimization approach.

Discussion and collaborations are warmly welcomed on this and ongoing works. The computer code used in this study is available from the first author, which going to be published as part of the 'UQlab' software package in the future.

**Acknowledgements**

This research is supported by Natural Science Foundation of China (Grant No.41075075, No.41375139 and No.51309011), Chinese Ministry of Science and Technology 973 Research Program (No. 2010CB428402) and the Fundamental Research Funds for the Central Universities - Beijing Normal University Research Fund (No.2013YB47). Special thanks are due to "Environmental & Ecological Science Data Center for West China, National Natural Science Foundation of China" (http://westdc.westgis.ac.cn) for providing the meteorological forcing data, to the group of Prof. Shaomin Liu at State Key Laboratory of Remote Sensing Science, School of Geography and Remote Sensing Science of Beijing Normal University for providing the surface flux validation data.

批注 [g1]: Change the funding resource.

## Appendix A. Surrogate modelling approaches

### A.1 Multivariate Adaptive Regression Splines (MARS)

The Multivariate Adaptive Regression Splines (MARS) model is a kind of flexible regression model of high dimensional data [*Friedman*, 1991]. It automatically divide the high-dimensional input space into different partitions with several knots and carry out linear or nonlinear regression in each partition. It takes the form of an expansion in product spline basis functions as follows:

$$y = f(\mathbf{x}) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} [s_{km}(x_{v(k,m)} - t_{km})]_+ \quad \text{(A.1)}$$

where $y$ is the output variable and $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is the n-dimensional input vector; $a_0$ is a constant, $a_m$ are weightings of each basis functions, $m$ is the index of basis functions and $M$ is the total number of basis functions; in each basis function $B_m(\boldsymbol{x}) = \prod_{k=1}^{K_m} [s_{km}(x_{v(k,m)} - t_{km})]_+$, $k$ is the index of knots and $K_m$ is the total number of knots; $s_{km}$ take on value $\pm 1$ and indicate the right/left sense of associated step function, $v(k,m)$ is the index of the input variable in vector $\mathbf{x}$, and $t_{km}$ indicates the knot location of the k-th knot in the m-th basis function.

MARS model is built in two stages: the forward pass and the backward pass. The forward pass builds an over-fitting model includes all input variables, while the backward pass removes the insensitive input variables one at a time. According to statistical learning theory, such a build-prune strategy can extract information from training data and meanwhile avoid the influence of noise [*Hastie et al.*, 2009]. Because of its pruning and fitting ability, MARS method can be used as parameter screening method[*Gan et al.*, 2014; *Li et al.*, 2013; *Shahsavani et al.*, 2010], and also surrogate modeling method[*Razavi et al.*, 2012; *Song et al.*, 2012; *Zhan et al.*, 2013].

### A.2 Gaussian Processes Regression (GPR)

Gaussian Processes Regression (GPR) [*Rasmussen and Williams*, 2006] is a new machine learning method based on statistical learning theory and Bayesian theory. It is

suitable for high-dimensional, small-sample nonlinear regression problems. In the function-space view, a Gaussian process can be completely specified by its mean function and covariance function:

$$\begin{cases} m(\mathbf{x}) = \mathrm{E}[f(\mathbf{x})] \\ k(\mathbf{x},\mathbf{x}') = \mathrm{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] \end{cases} \tag{A.2}$$

where $f(\mathbf{x})$ is the Gaussian process with n-dimensional input vector $\mathbf{x} = (x_1, x_2, \dots, x_n)$, $m(\mathbf{x})$ is its mean function and $k(\mathbf{x},\mathbf{x}')$ is its covariance function between two input vectors $\mathbf{x}$ and $\mathbf{x}'$. For short this Gaussian process can be written as $f(\mathbf{x}) = GP(m(\mathbf{x}), k(\mathbf{x},\mathbf{x}'))$.

Suppose a nonlinear regression model

$$y = f(\mathbf{x}) + \varepsilon \tag{A.3}$$

where $\mathbf{x}$ is the input vector, $y$ is the output variable, and $\varepsilon$ is the independent identically distributed Gaussian noise term with zero mean and variance $\sigma_n^2$. Suppose $\mathbf{y}$ is the training outputs, $X$ is the training input matrix in which each column is an input vector, $\mathbf{f}_*$ is the test outputs, $X_*$ is the test input matrix, $K(X,X)$, $K(X,X_*)$ and $K(X_*,X_*)$ denote covariance matrixes of all pairs of training and test inputs. We can easily write the joint distribution of training and testing inputs and outputs as a joint Gaussian distribution:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) \tag{A.4}$$

We can derive the mean and variance of predicted outputs from Bayesian theory. The predictive equations are presented as follows:

$$\mathrm{E}(\mathbf{f}_*) = K(X_*,X)[K(X,X) + \sigma_n^2 I]^{-1}\mathbf{y} \tag{A.5}$$

$$\mathrm{cov}(\mathbf{f}_*) = K(X_*,X_*) - K(X_*,X)[K(X,X) + \sigma_n^2 I]^{-1}K(X,X_*) \tag{A.6}$$

In this example, the outputs $\boldsymbol{y}$ is centered to zero so that the mean function is $m(\mathbf{x}) = 0$, while each element of covariance matrixes equals to the covariance function $k(\mathbf{x},\mathbf{x}')$ of input pairs.

The covariance function is the crucial ingredient of Gaussian Processes Regression, as it encodes the prior knowledge about the input-output relationship. There are many kinds of covariance functions to choose and users can construct special type of cov-function depending on their prior knowledge. In this paper, we choose Martérn

covariance function:

$$k(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{l}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}r}{l}\right) \quad \text{(A.7)}$$

where $r = |\mathbf{x} - \mathbf{x}'|$ is the Euclidian distance between input pair $\mathbf{x}$ and $\mathbf{x}'$, $K_{\nu}(.)$ is a modified Bessel function, $\nu$ and $l$ are positive hyper parameters, $\nu$ is the shape factor and $l$ is the scale factor (or characteristic length). The Matérn covariance function is an isotopic cov-function that the covariance only depends on the distance between $\mathbf{x}$ and $\mathbf{x}'$. The shape scale $\nu$ controls the shape of cov-function: larger $\nu$ leads to a smoother process while small $\nu$ leads to a rougher one. If the shape scale $\nu \to \infty$ we obtain squared exponential covariance function $k(r) = \exp(-r^2/2l^2)$, which is also called radial basis function (RBF). The Matérn covariance function becomes a product of a polynomial and an exponential when $\nu$ is half-integer: $\nu = p + 1/2$. The most widely used cases are $\nu = 3/2$ and $\nu = 5/2$, as follows:

$$k_{\nu=3/2}(r) = \left(1 + \frac{\sqrt{3}r}{l}\right)\exp\left(-\frac{\sqrt{3}r}{l}\right) \quad \text{(A.8)}$$

$$k_{\nu=5/2}(r) = \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right)\exp\left(-\frac{\sqrt{5}r}{l}\right) \quad \text{(A.9)}$$

批注 [g2]: A typo, v=3/2 shuld be v=5/2

In this paper, a value of $\nu = 5/2$ was used.

To adaptively determine the values of hyper parameters $l$ and $\sigma_n$, we use maximum marginal likelihood method. Because of the properties of Gaussian distribution, the log-marginal likelihood can be easily obtained as follows:

$$\log[p(\mathbf{y}|X)] = -\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{n}{2}\log 2\pi \quad \text{(A.10)}$$

where $K = K(X, X)$. In the training process of GPR, we used SCE-UA optimization method [*Duan et al.*, 1993] to find the best $l$ and $\sigma_n$.

### A.3 Random Forests (RF)

Random Forest~~s~~ (RF) [*Breiman*, 2001] ~~are~~ is a combination of Classification and Regression Trees (CART) [*Breiman et al.*, 1984]. Generally speaking, Tree-based methods split the feature space into a set of rectangles and fit the samples in each rectangle with a class label (for classification problems) or a constant value (for

regression problems). In this paper only regression tree was discussed. Suppose $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is the n-dimensional input feature vector and $y$ is the output response, the regression tree can be expressed as follows:

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} c_m I(\mathbf{x} \in R_m) \tag{A.11}$$

$$I(\mathbf{x} \in R_m) = \begin{cases} 1, & \mathbf{x} \in R_m \\ 0, & \mathbf{x} \notin R_m \end{cases} \tag{A.12}$$

where $M$ is the total number of rectangles, $m$ is the index of rectangle, $R_m$ is its corresponding region, $c_m$ is a constant value equals to the mean value of $y$ in region $R_m$. To effectively and efficiently find the best binary partition, a greedy algorithm is used to determine the feature to split and the location of split point. This greedy algorithm can be very fast especially for large dataset.

Because of the major disadvantages of a single tree, such as over-fitting, lack of smoothness and high variance, many improved methods have been proposed, such as MARS and random forests. A ~~R~~random forest~~s~~ construct many trees using randomly selected outputs and features, and synthetic the output~~s~~ of all the trees to give the prediction result. A ~~R~~random forest~~s~~ only have two parameters: the total number of trees $t$, and the selected feature number $\hat{m}$. Constructing random forests needs following steps:

1) Bootstrap aggregating (Bagging): From total $N$ samples $(\mathbf{x}_i, y_i), i = 1,2, \dots, N$, randomly select one point at one time with replacement, and replicate $N$ times to get a resample set containing $N$ points. This set is called a bootstrap replication. We need $t$ bootstrap replications for each tree.
2) Tree construction: For each splitting of each tree, randomly select $\hat{m}$ features from the total $M$, and select the best fitting feature among the $\hat{m}$ to split. The $\hat{m}$ selected features should be replaced in every splitting step.
3) The prediction result of a random forest~~s~~ is given by averaging the output of $t$ trees.

$$\hat{f}_{rf}(\mathbf{x}) = \sum_{j=1}^{t} \hat{f}_j(\mathbf{x}) \tag{A.13}$$

Random forests have outstanding performance for~~in~~ very high dimensional

problems, such as medical diagnosis and document retrieval. Such problems usually have hundreds or thousands of input variables (features), ~~with~~ but each ~~one~~ feature ~~only~~ provides only a little information. A single classification or regression model usually has very poor skill that only slightly better than random prediction. However, by combining many trees trained ~~using~~ with random features, a random forest can give improved accuracy. For big-data problems that have more than 100 input features and more than one million training samples, random forests become the only choice because of its outstanding efficiency and effectiveness.

**A.4 Support Vector Machine (SVM)**

Support Vector Machine (SVM) is an appealing machine learning method for classification and regression problems depending on the statistical learning theory [*Vapnik*, 1998; 2002]. The SVM method can avoid over-fitting problem because it employs the structural risk minimization principle. It is also efficient for big-data because of its scarcity. A brief introduction to support vector regression is presented below.

The aim of SVM is to find a function $f(\mathbf{x})$ that can fit the output $y$ with minimum risk given a $N$ point training set $(\mathbf{x}_i, y_i), i = 1,2,\dots,N$. Take a simple linear regression model for example, the function $f(\mathbf{x})$ can be:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{A.14}$$

where $\mathbf{w}$ is the weighting vector and $\mathbf{x}$ is the n-dimensional input feature vector. This function is actually determined by a small subset of training samples called support vectors (SVs).

Nonlinear problems can be transferred to linear problems by applying a nonlinear mapping from low-dimensional input space to some high-dimensional feature space:

$$f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b \tag{A.15}$$

where $\phi(\mathbf{x})$ is the mapping function. The inner product of mapping function is called Kernel Function: $K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T\phi(\mathbf{x}')$ and this method is called Kernel method. The commonly used kernel functions are: linear kernel function, polynomial, sigmoid and radial basis function (RBF). In this paper we use RBF kernel:

$$K(\mathbf{x},\mathbf{x}') = \exp(-\gamma|\mathbf{x}-\mathbf{x}'|^2) \tag{A.16}$$

where $|\mathbf{x}-\mathbf{x}'|$ is the Euclidian distance between $\mathbf{x}$ and $\mathbf{x}'$, $\gamma$ is a user defined parameter that controls the smoothness of $f(\mathbf{x})$.

To qualify the 'risk' of function $f(\mathbf{x})$, a loss function is defined as follows:

$$|y-f(\mathbf{x})|_\varepsilon = \begin{cases} 0, & if\ |y-f(\mathbf{x})| \le \varepsilon \\ |y-f(\mathbf{x})|-\varepsilon, & otherwise \end{cases} \tag{A.17}$$

The loss function means regression errors less than tolerance $\varepsilon$ are not penalized. The penalty-free zone is also called $\varepsilon$-tube or $\varepsilon$-boundary. As explained in statistical learning theory[*Vapnik*, 1998], the innovative loss function is the key point that SVM can balance empirical risk (risk of large error in the training set) and structure risk (risk of an over-complex model, or over-fitting). The problem of simultaneously minimizing both empirical risk (represented by regression error) and structure risk (represented by the width of $\varepsilon$-tube) can be written as a quadratic optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\xi,\xi^*} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i + C\sum_{i=1}^{n}\xi_i^* \\ \text{subject to} \quad & \mathbf{w}^T\phi(\mathbf{x}_i) + b - y_i \le \varepsilon + \xi_i \\ & y_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b \le \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \ge 0, i = 1,2,\dots,n \end{aligned} \tag{A.18}$$

The problem can be transferred to the dual problem:

$$\begin{aligned} \min_{\mathbf{w},b,\xi,\xi^*} \quad & \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T\boldsymbol{K}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) + \varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*) \\ & + \sum_{i=1}^{n} y_i(\alpha_i-\alpha_i^*) \\ \text{subject to} \quad & \boldsymbol{e}^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) = 0 \\ & y_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b \le \varepsilon + \xi_i^* \\ & 0 \le \alpha_i, \alpha_i^* \le C, i = 1,2,\dots,n \end{aligned} \tag{A.19}$$

where $\boldsymbol{K}$ is the kernel function matrix with $K_{ij} = K(\mathbf{x}_i,\mathbf{x}_j)$. Solving the dual problem and we can get the predictive function:

$$f(\mathbf{x}) = \sum_{i=1}^{n}(-\alpha_i+\alpha_i^*)\,K(\mathbf{x}_i,\mathbf{x}) + b \tag{A.20}$$

where the vectors $(\boldsymbol{\alpha}^*-\boldsymbol{\alpha})$ are the support vectors (SVs).

### A.5 Artificial Neural Network (ANN)

Artificial Neural Network (ANN) ~~REF~~[*Jain et al.*, 1996] is time-hornored

批注 [g3]:

machine learning method comparing to the former four. It is a data-driven process that can solve complex nonlinear relationships between input and outpur data. A nerual network is constructed by many interconnected neurons. Each neuron can be mathematically described as a linear weighing function and a nonlinear activation function:

$$I_i = \sum_{j=1}^{n} w_{ij} x_j \tag{A.21}$$

$$f_i(I) = \frac{1}{1 + \exp(-I_i)} \tag{A.22}$$

where $x_j$ is the j-th input variable, $w_{ij}$ is the weight and $I_i$ is the weighted sum of the i-th neuron. The output of the i-th neuron $f_i(I)$ is given by the nonlinear activation function of the weighted sum input. Here we use Sigmoid function.

[*Minsky and Papert*, 1969] shows that single layer neural network can only solve linear problem. [*Cybenko*, 1989] extended ANN to multiple layer and demostrated that multi-layer ANN can infinitely approximate any nonlinear function (the universal approximation theorem). The training procedure of ANN is optimizing the value of weights. There are many training methods for ANN and we use the Levenberg-Marquardt (LM) [*Marquardt*, 1963] algorithm, a modification of the classic Newton algorithm provided in matlab ANN toolbox.

# References:

Bastidas, L. A., H. V. Gupta, S. Sorooshian, W. J. Shuttleworth, and Z. L. Yang (1999), Sensitivity analysis of a land surface scheme using multicriteria methods, *Journal of Geophysical Research: Atmospheres*, *104*(D16), 19481-19490.

Bonan, G. B. (1996), A Land Surface Model (LSM Version 1.0) for Ecological, Hydrological, and Atmospheric Studies: Technical Description and User's Guide, NCAR, Boulder, CO, USA.

Boyle, D. P., H. V. Gupta, and S. Sorooshian (2000), Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resour Res*, *36*(12), 3663-3674.

Boyle, D. P. (2000), Multicriteria calibration of hydrological models, Ph.D. dissertation thesis, University of Arizona, Tucson, USA.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984), *Classification and Regression Trees*, Chapman and Hall/CRC, Bota Raton, Florida, USA.

Breiman, L. (2001), Random forests, *Mach Learn*, *45*(1), 5-32.

673 Castelletti, A., F. Pianosi, R. Soncini-Sessa, and J. P. Antenucci (2010), A multiobjective response
674 surface approach for improved water quality planning in lakes and reservoirs, *Water Resour Res*, *46*(6),
675 W6502.
676 Cybenko, G. (1989), Approximation by superpositions of a sigmoidal function, *Mathematics of Control,*
677 *Signals and Systems*, *2*(4), 303-314.
678 Dai, Y., and Q. Zeng (1997), A Land Surface Model (IAP94) for Climate Studies Part I:Formulation and
679 Validation in Off-Line Experiments, *Adv Atmos Sci*, *14*(4), 433-460.
680 Dai, Y., F. Xue, and Q. Zeng (1998), A Land Surface Model (IAP94) for Climate Studies Part
681 II:Implementation and Preliminary Results of Coupled Model with IAP GCM, *Adv Atmos Sci*, *15*(4), 47-
682 62.
683 Dai, Y. J., X. B. Zeng, R. E. Dickinson, I. Baker, G. B. Bonan, M. G. Bosilovich, A. S. Denning, P. A.
684 Dirmeyer, P. R. Houser, G. Y. Niu, K. W. Oleson, C. A. Schlosser, and Z. L. Yang (2003), The Common
685 Land Model, *B Am Meteorol Soc*, *84*(8), 1013-1023.
686 Dai, Y. J., R. E. Dickinson, and Y. P. Wang (2004), A two-big-leaf model for canopy temperature,
687 photosynthesis, and stomatal conductance, *J Climate*, *17*(12), 2281-2299.
688 Dickinson, R. E., A. Henderson-Sellers, and P. J. Kennedy (1993), Biosphere-atmosphere Transfer
689 Scheme (BATS) Version 1e as Coupled to the NCAR Community Climate Model, NCAR, Boulder, CO.
690 USA.
691 Duan, Q. Y., S. Sorooshian, and V. K. Gupta (1992), Effective and Efficient Global Optimization for
692 Conceptual Rainfall-Runoff Models, *Water Resour Res*, *28*(4), 1015-1031.
693 Duan, Q. Y., V. K. Gupta, and S. Sorooshian (1993), Shuffled Complex Evolution Approach for
694 Effective and Efficient Global Minimization, *J Optimiz Theory App*, *76*(3), 501-521.
695 Duan, Q. Y., S. Sorooshian, and V. K. Gupta (1994), Optimal Use of the SCE-UA Global Optimization
696 Method for Calibrating Watershed Models, *J Hydrol*, *158*(3-4), 265-284.
697 Friedman, J. H. (1991), Multivariate Adaptive Regression Splines, *Ann Stat*, *19*(1), 1-14.
698 Gan, Y., Q. Duan, W. Gong, C. Tong, Y. Sun, W. Chu, A. Ye, C. Miao, and Z. Di (2014), A
699 comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological
700 model, *Environ Modell Softw*, *51*(0), 269-285.
701 Gorissen, D. (2010), Grid-enabled Adaptive Surrogate Modeling for Computer Aided Engineering, Ph.D.
702 thesis, Ghent University, Ghent, Flanders, Belgium.
703 Gupta, H. V., S. Sorooshian, and P. O. Yapo (1998), Toward improved calibration of hydrologic models:
704 Multiple and noncommensurable measures of information, *Water Resour Res*, *34*(4), 751-763.
705 Gupta, H. V., L. A. Bastidas, S. Sorooshian, W. J. Shuttleworth, and Z. L. Yang (1999), Parameter
706 estimation of a land surface scheme using multicriteria methods, *J Geophys Res*, *104*(D16), 19491-19503.
707 Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning 2nd*, Springer,
708 New York, USA.
709 Henderson-Sellers, A., K. McGuffie, and A. J. Pitman (1996), The Project for Intercomparison of Land-
710 surface Parametrization Schemes (PILPS): 1992 to 1995, *Clim Dynam*, *12*(12), 849-859.
711 Hu, Z. Y., M. G. Ma, R. Jin, W. Z. Wang, G. H. Huang, Z. H. Zhang, and J. L. Tan (2003), WATER:
712 dataset of automatic meteorological observations at the A'rou freeze/thaw observation station, Cold and
713 Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences,
714 Lanzhou, China.
715 Jackson, C., Y. Xia, M. K. Sen, and P. L. Stoffa (2003), Optimal parameter and uncertainty estimation
716 of a land surface model: A case study using data from Cabauw, Netherlands, *Journal of Geophysical*

*Research: Atmospheres*, *108*(D18), 4583.
Jain, A. K., M. Jianchang, and K. M. Mohiuddin (1996), Artificial neural networks: a tutorial, *Computer*, *29*(3), 31-44.
Ji, D., and Y. Dai (2010), The Common Land Model (CoLM) technical guide, GCESS, Beijing Normal University, Beijing, China.
Jin, Y. (2011), Surrogate-assisted evolutionary computation: Recent advances and future challenges, *Swarm and Evolutionary Computation*, *1*(2), 61-70.
Jones, D. R., M. Schonlau, and W. J. Welch (1998), Efficient global optimization of expensive black-box functions, *J Global Optim*, *13*(4), 455-492.
Jones, D. R. (2001), A Taxonomy of Global Optimization Methods Based on Response Surfaces, *J Global Optim*, *21*(4), 345-383.
Koziel, S., and L. Leifsson (2013), *Surrogate-Based Modeling and Optimization*, Springer New York, New York, NY.
Leplastrier, M., A. J. Pitman, H. Gupta, and Y. Xia (2002), Exploring the relationship between complexity and performance in a land surface model using the multicriteria method, *J Geophys Res*, *107*(4443D20).
Li, J. (2012), Screening and Analysis of Parameters Most Sensitive to CoLM, Master thesis, BNU, Beijing, China.
Li, J., Q. Y. Duan, W. Gong, A. Ye, Y. Dai, C. Miao, Z. Di, C. Tong, and Y. Sun (2013), Assessing parameter importance of the Common Land Model based on qualitative and quantitative sensitivity analysis, *Hydrol. Earth Syst. Sci.*, *17*(8), 3279-3293.
Liang, X., E. F. Wood, D. P. Lettenmaier, D. Lohmann, A. Boone, S. Chang, F. Chen, Y. Dai, C. Desborough, R. E. Dickinson, Q. Duan, M. Ek, Y. M. Gusev, F. Habets, P. Irannejad, R. Koster, K. E. Mitchell, O. N. Nasonova, J. Noilhan, J. Schaake, A. Schlosser, Y. Shao, A. B. Shmakin, D. Verseghy, K. Warrach, P. Wetzel, Y. Xue, Z. Yang, and Q. Zeng (1998), The Project for Intercomparison of Land-surface Parameterization Schemes (PILPS) phase 2(c) Red-Arkansas River basin experiment:: 2. Spatial and temporal analysis of energy fluxes, *Global Planet Change*, *19*(1 – 4), 137-159.
Liu, Y. Q., L. A. Bastidas, H. V. Gupta, and S. Sorooshian (2003), Impacts of a parameterization deficiency on offline and coupled land surface model simulations, *J Hydrometeorol*, *4*(5), 901-914.
Liu, Y. Q., H. V. Gupta, S. Sorooshian, L. A. Bastidas, and W. J. Shuttleworth (2004), Exploring parameter sensitivities of the land surface using a locally coupled land-atmosphere model, *J Geophys Res*, *109*(D21101D21), 13.
Liu, Y. Q., H. V. Gupta, S. Sorooshian, L. A. Bastidas, and W. J. Shuttleworth (2005), Constraining land surface and atmospheric parameters of a locally coupled model using observational data, *J Hydrometeorol*, *6*(2), 156-172.
Lohmann, D., D. P. Lettenmaier, X. Liang, E. F. Wood, A. Boone, S. Chang, F. Chen, Y. Dai, C. Desborough, R. E. Dickinson, Q. Duan, M. Ek, Y. M. Gusev, F. Habets, P. Irannejad, R. Koster, K. E. Mitchell, O. N. Nasonova, J. Noilhan, J. Schaake, A. Schlosser, Y. Shao, A. B. Shmakin, D. Verseghy, K. Warrach, P. Wetzel, Y. Xue, Z. Yang, and Q. Zeng (1998), The Project for Intercomparison of Land-surface Parameterization Schemes (PILPS) phase 2(c) Red – Arkansas River basin experiment:: 3. Spatial and temporal analysis of water fluxes, *Global Planet Change*, *19*(1 – 4), 161-179.
Loshchilov, I., M. Schoenauer, and M. E. L. Sebag (2010), Comparison-based Optimizers Need Comparison-based Surrogates, in Proceedings of the 11th International Conference on Parallel Problem Solving from Nature: Part I, pp. 364-373, Berlin, Heidelberg.

761 Marler, R. T., and J. Arora (2010), The weighted sum method for multi-objective optimization: new
762 insights, *Struct Multidiscip O*, *41*(6), 853-862.
763 Marquardt, D. W. (1963), An Algorithm for Least-Squares Estimation of Nonlinear Parameters, *Journal*
764 *of the Society for Industrial and Applied Mathematics*, *11*(2), 431-441.
765 McKay, M. D., R. J. Beckman, and W. J. Conover (1979), A Comparison of Three Methods for Selecting
766 Values of Input Variables in the Analysis of Output from a Computer Code, *Technometrics*, *21*(2), 239-
767 245.
768 Minsky, M., and S. A. Papert (1969), *Perceptrons: An Introduction to Computational Geometry*, MIT
769 Press, Cambridge, Massachusetts, USA.
770 Ong, Y., P. B. Nair, A. J. Keane, and K. W. Wong (2005), Surrogate-Assisted Evolutionary Optimization
771 Frameworks for High-Fidelity Engineering Design Problems, In 'Studies in Fuzziness and Soft
772 Computing', edited by Y. Jin, pp. 307-331, Springer Berlin Heidelberg.
773 Pil á, M., and R. Neruda (2013), Surrogate model selection for evolutionary multiobjective optimization,
774 paper presented at Evolutionary Computation (CEC), 2013 IEEE Congress on, 2013-01-01.
775 Pollacco, J. A. P., B. P. Mohanty, and A. Efstratiadis (2013), Weighted objective function selector
776 algorithm for parameter estimation of SVAT models with remote sensing data, *Water Resour Res*, *49*(10),
777 6959-6978.
778 Rasmussen, C. E., and C. K. I. Williams (2006), *Gaussian Processes for Machine Learning*, MIT Press,
779 Massachusetts, USA.
780 Razavi, S., B. A. Tolson, and D. H. Burn (2012), Review of surrogate modeling in water resources, *Water*
781 *Resour Res*, *48*(7), W7401.
782 Shahsavani, D., S. Tarantola, and M. Ratto (2010), Evaluation of MARS modeling technique for
783 sensitivity analysis of model output, *Procedia - Social and Behavioral Sciences*, *2*(6), 7737-7738.
784 Shangguan, W., Y. Dai, B. Liu, A. Zhu, Q. Duan, L. Wu, D. Ji, A. Ye, H. Yuan, Q. Zhang, D. Chen, M.
785 Chen, J. Chu, Y. Dou, J. Guo, H. Li, J. Li, L. Liang, X. Liang, H. Liu, S. Liu, C. Miao, and Y. Zhang
786 (2013), A China data set of soil properties for land surface modeling, *Journal of Advances in Modeling*
787 *Earth Systems*, 5(2), 212-224.
788 Song, X., C. Zhan, and J. Xia (2012), Integration of a statistical emulator approach with the SCE-UA
789 method for parameter optimization of a hydrological model, *Chinese Sci Bull*, *57*(26), 3397-3403.
790 Taylor, K. E. (2001), Summarizing multiple aspects of model performance in a single diagram, *Journal*
791 *of Geophysical Research: Atmospheres*, *106*(D7), 7183-7192.
792 van Griensven, A., and T. Meixner (2007), A global and efficient multi-objective auto-calibration and
793 uncertainty estimation method for water quality catchment models, *JOURNAL OF*
794 *HYDROINFORMATICS*, *9*(4), 277-291.
795 Vapnik, V. N. (1998), *Statistical Learning Theory*, John Wiley & Sons, NewYork, USA.
796 Vapnik, V. N. (2002), *The Nature of Statistical Learning Theory*, 2nd ed., Springer, New York.
797 Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian (2003), A Shuffled Complex Evolution
798 Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters,
799 *Water Resour Res*, *39*(12018), 1201.
800 Vrugt, J. A., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian (2003), Effective and efficient
801 algorithm for multiobjective optimization of hydrologic models, *Water Resour Res*, *39*(12148), 1214.
802 Wang, C., Q. Duan, W. Gong, A. Ye, Z. Di, and C. Miao (2014), An evaluation of adaptive surrogate
803 modeling based optimization with two benchmark problems, *Environ Modell Softw*, *60*(0), 167-179.
804 Wang, Q. J. (1991), The Genetic Algorithm and Its Application to Calibrating Conceptual Rainfall-

Runoff Models, *Water Resour Res*, *27*(9), 2467-2471.
Wood, E. F., D. P. Lettenmaier, X. Liang, D. Lohmann, A. Boone, S. Chang, F. Chen, Y. Dai, R. E. Dickinson, Q. Duan, M. Ek, Y. M. Gusev, F. Habets, P. Irannejad, R. Koster, K. E. Mitchel, O. N. Nasonova, J. Noilhan, J. Schaake, A. Schlosser, Y. Shao, A. B. Shmakin, D. Verseghy, K. Warrach, P. Wetzel, Y. Xue, Z. Yang, and Q. Zeng (1998), The Project for Intercomparison of Land-surface Parameterization Schemes (PILPS) Phase 2(c) Red – Arkansas River basin experiment:: 1. Experiment description and summary intercomparisons, *Global Planet Change*, *19*(1 – 4), 115-135.
Xia, Y., A. J. Pittman, H. V. Gupta, M. Leplastrier, A. Henderson-Sellers, and L. A. Bastidas (2002), Calibrating a land surface model of varying complexity using multicriteria methods and the Cabauw dataset, *J Hydrometeorol*, *3*(2), 181-194.
Yapo, P. O., H. V. Gupta, and S. Sorooshian (1998), Multi-objective global optimization for hydrologic models, *J Hydrol*, *204*(1-4), 83-97.
Zhan, C., X. Song, J. Xia, and C. Tong (2013), An efficient integrated approach for global sensitivity analysis of hydrological model parameters, *Environ Modell Softw*, *41*(0), 39-52.

**Figure list:**

837 **Table 1:** Adjustable parameters and their categories, meanings and ranges.

| Num | Para | Units | Category | Physical meaning | Feasible range |
|---|---|---|---|---|---|
| P1 | dewmx | | canopy | maximum dew ponding of leaf area | [0.05, 0.15] |
| P2 | hksati | mm/s | soil | maximum hydraulic conductivity | [0.001, 1] |
| P3 | porsl | - | soil | porosity | [0.25, 0.75] |
| P4 | phi0 | mm | soil | minimum soil suction | [50, 500] |
| P5 | wtfact | | soil | fraction of shallow groundwater area | [0.15, 0.45] |
| P6 | bsw | -- | soil | Clapp and Hornberger "b" parameter | [2.5, 7.5] |
| P7 | wimp | | soil | water impermeable if porosity less than wimp | [0.01, 0.1] |
| P8 | zlnd | m | soil | roughness length for soil surface | [0.005, 0.015] |
| P9 | pondmx | mm | soil | maximum ponding depth for soil surface | [5, 15] |
| P10 | csoilc | -- | soil | drag coefficient for soil under canopy | [0.002, 0.006] |
| P11 | zsno | m | snow | roughness length for snow | [0.0012, 0.0036] |
| P12 | capr | | soil | tuning factor of soil surface temperature | [0.17, 0.51] |
| P13 | cnfac | | canopy | Crank Nicholson factor | [0.25, 0.5] |
| P14 | slti | | canopy | slope of low temperature inhibition function | [0.1, 0.3] |
| P15 | hlti | | canopy | 1/2 point of low temperature inhibition function | [278, 288] |
| P16 | shti | | canopy | slope of high temperature inhibition function | [0.15, 0.45] |
| P17 | sqrtdi | $m^{-0.5}$ | canopy | the inverse of square root of leaf dimension | [2.5, 7.5] |
| P18 | effcon | mol $CO^2$/ mol quanta | canopy | quantum efficiency of vegetation photosynthesis | [0.035, 0.35] |
| P19 | vmax25 | mol $CO^2$/ $m^2$s | canopy | maximum carboxylation rate at 25℃ | [$10^{-6}$, $200^{-6}$] |
| P20 | hhti | | canopy | 1/2 point of high temperature inhibition function | [305, 315] |
| P21 | trda | | canopy | temperature coefficient of conductance-photosynthesis model | [0.65,1.95] |
| P22 | trdm | | canopy | temperature coefficient of conductance-photosynthesis model | [300, 350] |
| P23 | trop | | canopy | temperature coefficient of conductance-photosynthesis model | [250, 300] |
| P24 | gradm | | canopy | slope of conductance-photosynthesis model | [4, 9] |
| P25 | binter | | canopy | intercept of conductance-photosynthesis model | [0.01, 0.04] |
| P26 | extkn | | canopy | coefficient of leaf nitrogen allocation | [0.5, 0.75] |
| P27 | chil | | canopy | leaf angle distribution factor | [-0.3, 0.1] |
| P28 | ref(1,1) | | canopy | VIS reflectance of living leaf | [0.07, 0.105] |
| P29 | ref(1,2) | | canopy | VIS reflectance of dead leaf | [0.16, 0.36] |
| P30 | ref(2,1) | | canopy | NIR reflectance of living leaf | [0.35, 0.58] |
| P31 | ref(2,2) | | canopy | NIR reflectance of dead leaf | [0.39, 0.58] |
| P32 | tran(1,1) | | canopy | VIS transmittance of living leaf | [0.04, 0.08] |
| P33 | tran(1,2) | | canopy | VIS transmittance of dead leaf | [0.1, 0.3] |
| P34 | tran(2,1) | | canopy | NIR transmittance of living leaf | [0.1, 0.3] |

| | | | | | |
|---|---|---|---|---|---|
| P35 | tran(2,2) | | canopy | NIR transmittance of dead leaf | [0.3, 0.5] |
| P36 | z0m | m | canopy | aerodynamic roughness length | [0.05, 0.3] |
| P37 | ssi | | snow | irreducible water saturation of snow | [0.03, 0.04] |
| P38 | smpmax | mm | soil | wilting point potential | [-2.e5, -1.e5] |
| P39 | smpmin | mm | soil | restriction for min of soil potential | [-1.e8, -9.e7] |
| P40 | trsmx0 | mm/s | canopy | maximum transpiration for vegetation | [1.e-4, 1. e-2] |

838

839

840 **Table 2:** Screened parameters of CoLM in A'rou Station [Li et.al., 2013]

| Output variables (fluxes) | Screened parameters |
|---|---|
| Sensible Heat | P2, P4, P6, P30, P34, P36 |
| Latent Heat | ~~P2, P3, P4, P6, P18, P30, P36, P38~~ P3, P4, P6, P18, P19, P23, P25, P36 |
| Upward Longwave Radiation | P6, P17, P36 |
| Net radiation | P6, P17, P30, P34, P36 |
| Soil Temperature | P3, P6, P36 |
| Soil Moisture | P3, P6 |

841

842

批注 [g4]: These screened parameters were suggested by [Li, J. (2012)]. It is corrected in order to be consistent with [Li, J. (2012)] and other parts of this paper.

843

844 **Table 3:** The NRMSE between simulated and observed outputs after single objective optimization

| | Sensible heat | Latent heat | Upward longwave radiation | Net radiation | Soil Temperature | Soil Moisture |
|---|---|---|---|---|---|---|
| Default | 0.8586 | 0.5833 | 0.0590 | 0.2357 | 0.0096 | 0.4527 |
| SCE Optimized | 0.7450 (1492) | 0.4921 (1354) | 0.0380 (458) | 0.1963 (982) | 0.0073 (473) | 0.3900 (210) |
| LH50 | 0.7672 | 0.5255 | 0.0377 | 0.1913 | 0.0080 | 0.4222 |
| LH100 | 0.7841 | 0.5785 | 0.0372 | 0.1908 | 0.0077 | 0.4130 |
| LH200 | 0.7821 | 0.5885 | 0.0374 | 0.1928 | 0.0069 | 0.3947 |
| LH400 | 0.7798 | 0.5627 | 0.0374 | 0.1928 | 0.0070 | 0.3971 |
| LH800 | 0.7683 | 0.5024 | 0.0377 | 0.1909 | 0.0068 | 0.3956 |
| LH1200 | 0.7760 | 0.5150 | 0.0374 | 0.1919 | 0.0068 | 0.3962 |
| LH2000 | 0.7705 | 0.5048 | 0.0375 | 0.1912 | 0.0070 | 0.3946 |

845

846

847

848 **Table 4:** Weights assigned to each output variables.

| Flux name | Label | Unit | RMSE | NRMSE | Weights |
|---|---|---|---|---|---|
| Sensible heat | fsena | $W/m^2$ | 49.14 | 0.8586 | 0.3905 |
| Latent heat | lfevpa | $W/m^2$ | 43.59 | 0.5833 | 0.2653 |
| Upward longwave radiation | orlg | $W/m^2$ | 19.43 | 0.0590 | 0.0268 |
| Net radiation | sabvg | $W/m^2$ | 42.78 | 0.2357 | 0.1072 |
| Soil temperature | tss | K | 2.66 | 0.0096 | 0.0044 |
| Soil moisture | wliq | $kg/m^2$ | 21.14 | 0.4527 | 0.2059 |

849

850

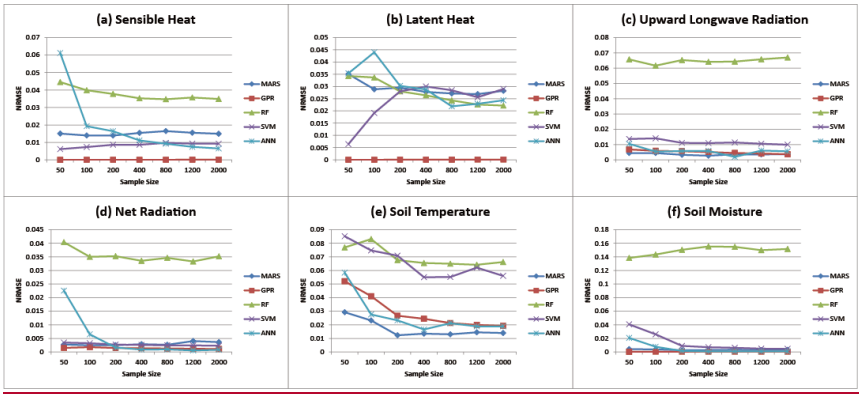



**Figure 1:** Inter-comparison of 5 surrogate modelling methods, error of training set.

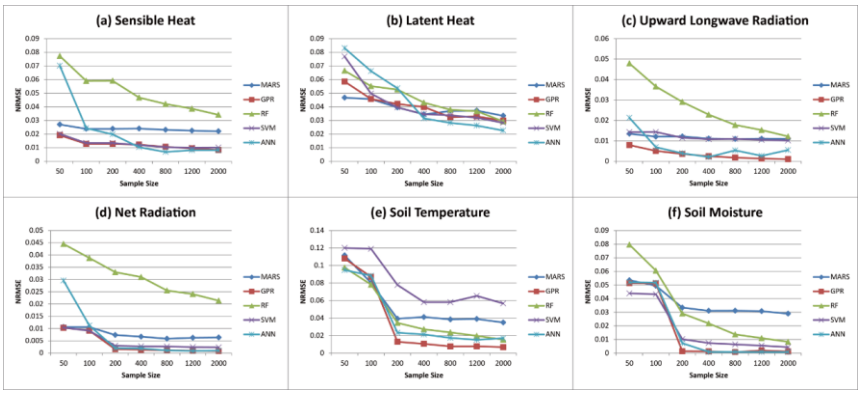



**Figure 21:** Inter-comparison of 5 surrogate modelling methods, error of testing set.

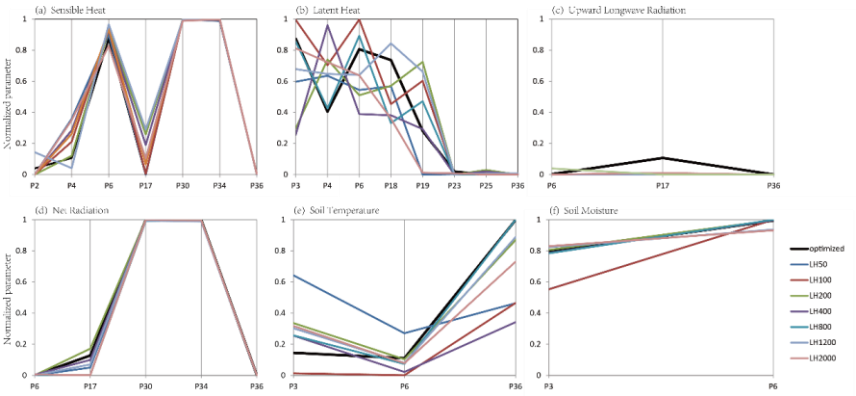
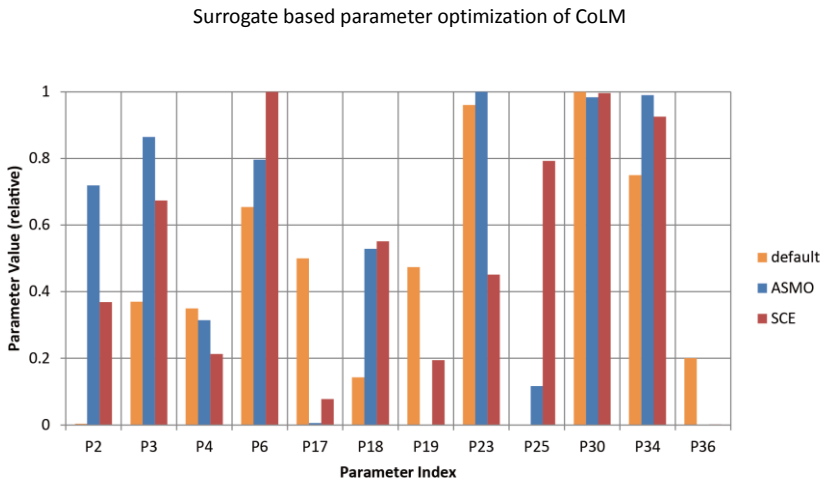



**Figure 2:** Single-objective optimization result: optimal parameters.




862

863 **Figure 3:** Optimal value of CoLM given by multi-objective optimization (comparing default

864 parameter, optimal parameter given by ASMO and SCE-UA)

865

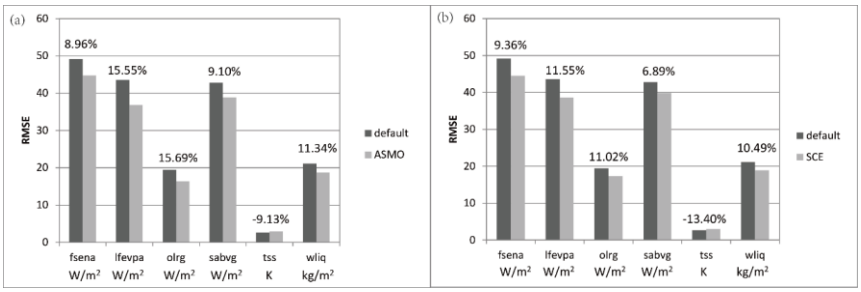



**Figure 4:** Comparing the improvements given by ASMO and SCE.

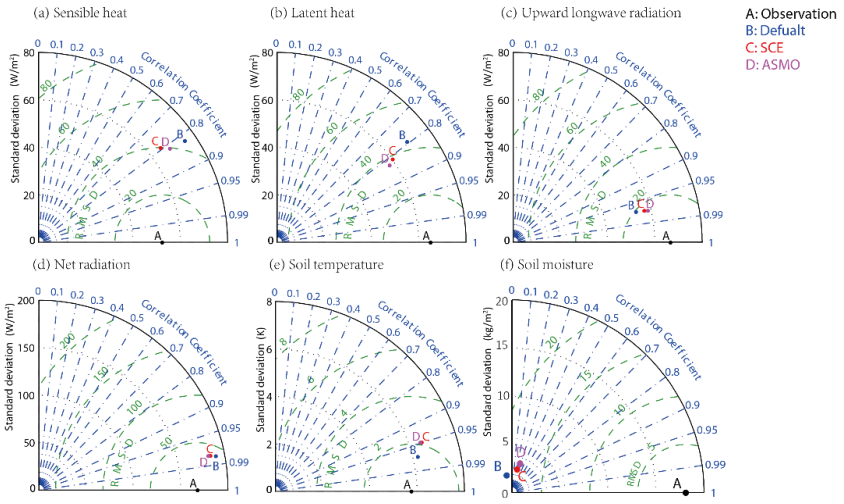



**Figure 5:** Taylor diagram of simulated fluxes during calibration period (Jan-1-2009 to Dec-31-2009).

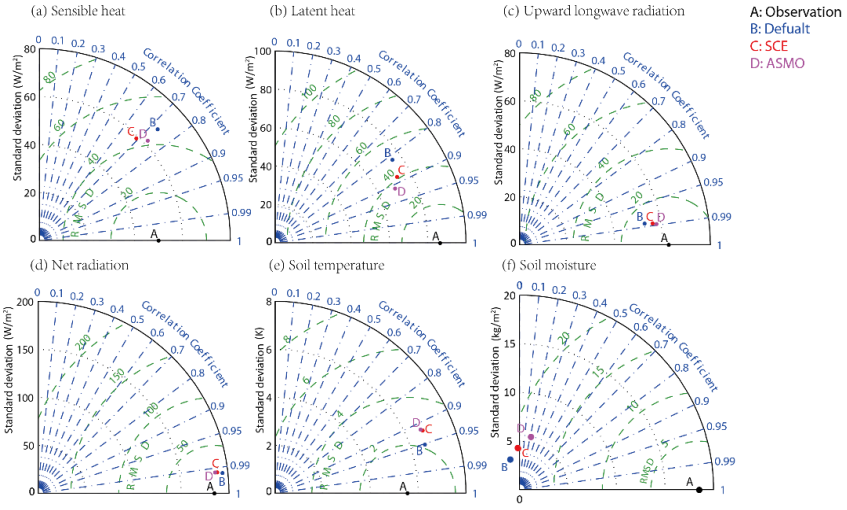



**Figure 6:** Taylor diagram of simulated fluxes during validation period (Here we use warm-up period as validation period, Jan-1-2008 to Dec-31-2008).