Supplementary materials

Robust Efficient Global Optimization Using Random Forest with Dual Uncertainty Quantification for Microalgae Cultivation Experiments

Mingqi Jiang, [a,b] Yiming Zhao [b] Zhuo Wang [*a,b], Xupeng Cao [c] and Fucheng Pan [a]

a. Shenyang Institute of Automation, Chinese Academy of Sciences, China.

b. University of Chinese Academy of Sciences, China.

c. State Key Laboratory of Catalysis and Division of Solar Energy, Dalian National Laboratory of Clean Energy, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China.

1. Validation of the proposed method with numerical examples

This section evaluates the proposed method using benchmark problems in one and two dimensions. The benchmarking procedure follows these steps: For each optimization in input dimension d, we initialize with a LHS design with size $n_0 = 5 + 5d$ 43. For clarity, we denote algorithms with dual uncertainty and input uncertainty as "_GW" and "_W", respectively. Additionally, a comparison with the RF and GP model is conducted to show their similar functionalities. Computational costs of RF and GP-based EGO algorithms are recorded for both uncertainty measures. The experiments are conducted on two Intel(R) Xeon(R) Platinum 8358P CPU @ 2.60-GHz CPU machine using Python programming language. The implementation on random forest input uncertainty is based on golem 1.0. Each criterion is calculated 10 times with 10 randomly generated inputs, and the average CPU time of the 10 calculations is recorded. The distance metrics is utilized to compare three strategies across methods using the true robust optimum:

$$d\left(x_{m,n}^{r}\right) = \left\|x_{m,n}^{r} - x^{r}\right\| \#(S1)$$

where x^r is the location of the true robust optimum, $x_{m,n}^{r}$ is the recommended value of model m in n^{th} iteration. This metric represents the distance of the input parameter sought by the algorithm from the correct design value in the decision space. In our experiments, lower is better with 0 being floor.

1.1 One-dimensional test problem

For ease of visualize the effect of design and modeling uncertainties, a one-dimensional test function proposed by Forrester et al is applied to an unconstrained minimization problem to test proposed approach44. The mathematical function is shown as:

$$f(x) = (6x-2)^2 \sin(12x-4) + 8x, x \in [0,1] \#(S2)$$

The input variables $z = x + \delta$ follows the Normal distribution $z \sim N(x,\sigma_x)$, $\sigma_x = 0.07$. Fig.S4 shows the shape of the function. It can be seen that a global optimum lies at $(x_1,y_1) = (0.75,0.0067)$ and a local optimum lies at $(x_2,y_2) = (0.12,0.06)$. It's easy to see that the curvature of x_1 is lower then x_2. It means that x_1 is more sensitive to small changes in input variables and applying a certain perturbation to x_2 corresponds to a smaller change in the objective value. Consequently, x_1 demonstrates poor robustness as an optimal design under uncertainty conditions. In contrast, the sub-

optimal design point x_2 shows greater stability against combined uncertainties, with lower sensitivity to external perturbations compared to x_1.

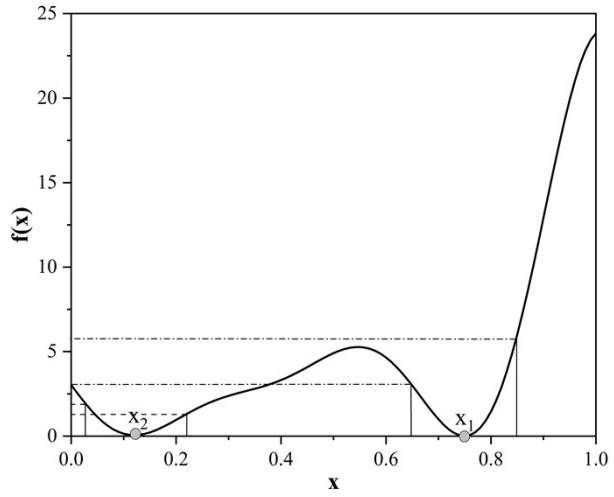

Fig.S1 One-dimensional test problem

First, we compare the performance of prediction and uncertainty quantification based on the GP model and the RF model. As shown in Fig.S2 and Fig.S3, the surrogate models are constructed based on 6 samples at x = [0, 0.22, 0.39, 0.63, 0.86, 1]. Fig.S2(a) and Fig.S3(a) provide the 95% prediction intervals (PI) for the GP model and the RF model considering dual uncertainty, only input uncertainty and model uncertainty.

As shown in Fig.S2(a) and Fig.S3(a), the PI incorporating both input parameter and surrogate model uncertainties are significantly wider than those considering only a single uncertainty source within the confidence region. Fig.S2(b) and Fig.S3(b) provides the close look at the difference between the three standard deviation (STD) functions for GP model and RF model. σ represents the prediction STD of surrogate model, σ_W represents the response STD considering only input uncertainty and σ_{WG} represents the compound STD considering dual uncertainty. It can be observed that the STD when considering the dual uncertainties is always larger than the STD when considering only the design parameter uncertainty. Since σ is small at the tail of the curve, other uncertainties with two scenarios are similar. Furthermore, it's obvious that this phenomenon is different from simple linear summation of σ and σ_W.

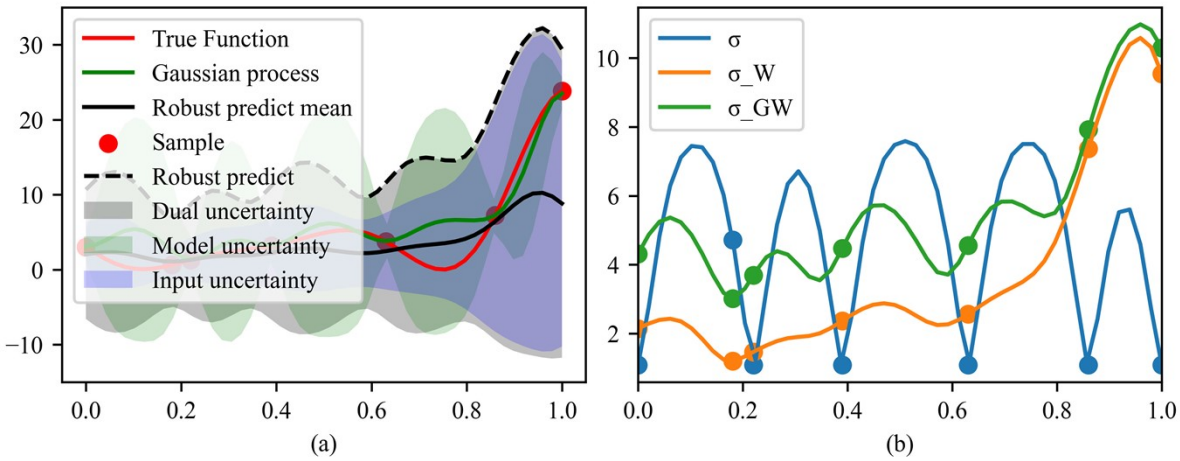

Fig.S2 Robust model prediction based on GP

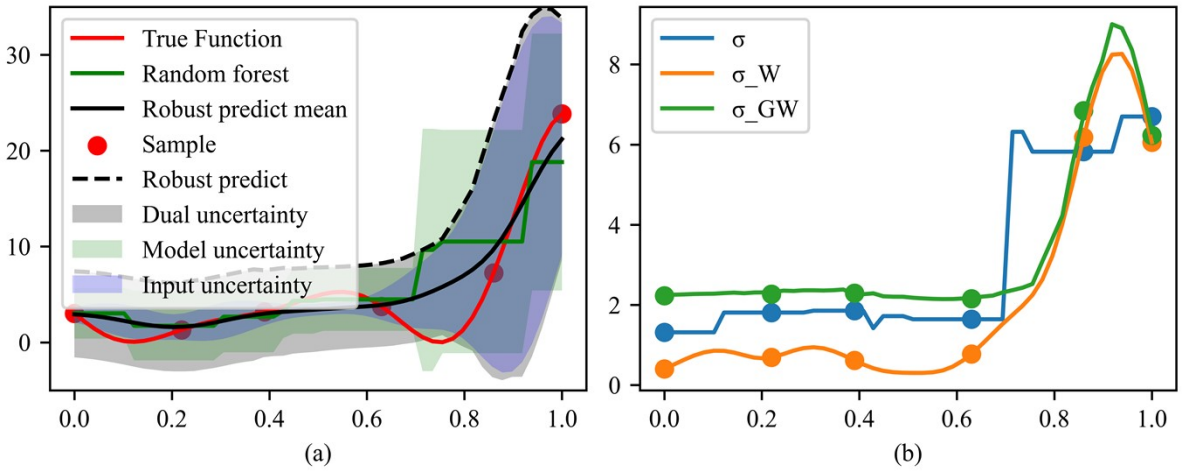

Fig.S3. Robust model prediction based on RF

The optimization trajectories in Fig.S4 demonstrate the convergence behavior of GP and RF under two uncertainty modeling schemes (GW and W). On the left, the mean objective value (solid lines) and variance (shaded regions) reveal that RF-based methods (RF_GW, RF_W) achieve faster convergence and higher stability compared to GP (GP_GW, GP_W). Notably, RF _GW attains the best performance, suggesting that incorporating both model and input uncertainty enhances optimization efficiency. In contrast, GP exhibits slower convergence and higher variance, particularly under W, indicating that since there is no implementation that takes model uncertainty into account, the errors introduced by the proxy model cause the algorithm to converge more slowly, and more iterations are needed to dissipate the effects of such errors. When compared under the same model, it is clear that the GW scheme is superior to the W scheme. The advantage

of the GW scheme is that, although it is more unstable at the beginning of the iteration compared to the W scheme, it is able to quickly reach the performance that can only be achieved by considering the input uncertainty at a later stage as the iteration progresses. This is because by this time, the model uncertainty is already almost negligible.

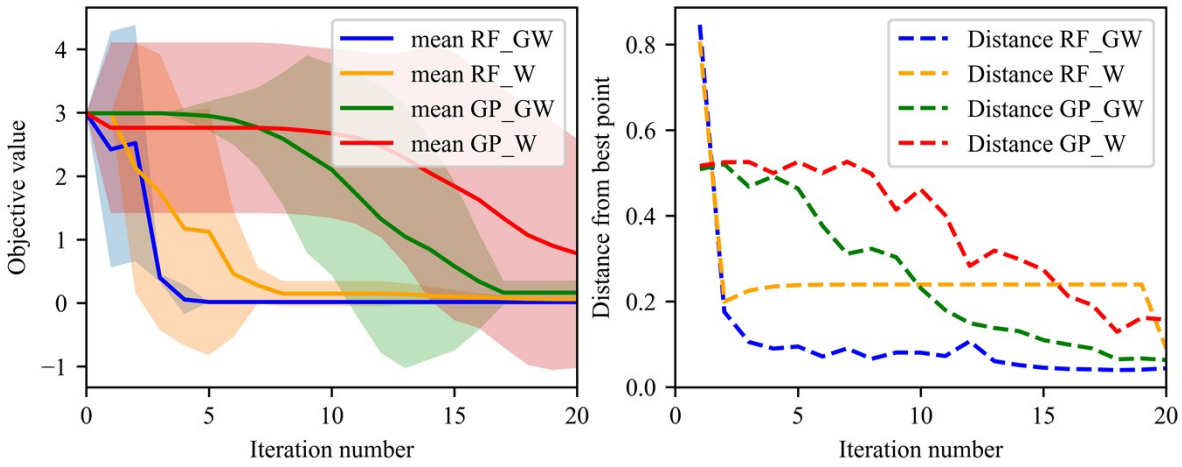

Fig.S4 The convergence of distance indicators under the four uncertainty strategies

The right panel evaluates the distance between optimized inputs and the theoretical optimum. RF_GW consistently maintains the smallest distance, reinforcing its superior convergence. While RF_W and GP_GW show gradual improvement, GP_W lags behind, further highlighting the limitations of RF in uncertainty-aware optimization. These results suggest that RF-based methods, especially when accounting for both uncertainty sources (GW), are more robust and efficient for this class of problems.

1.2 Two -dimensional test problem

In this section, proposed method is tested on a 2d test problem from Bertsimas et al, a common test case in robust optimization, which is defined as:

$$f(x_1,x_2) = -2x_1^6 + 12.2x_1^5 - 21.2x_1^4 + 6.4x_1^3 + 4.7x_1^2 - 6.2x_1 - x_2^6 + 11x_2^5 - 43.3x_2^4 + 74.8x_2^3 - 56.9x_2^2$$

$$+ 10x_2 + 4.1x_1x_2 + 0.1x_1^2x_2^2 - 0.4x_1x_2^2 - 0.4x_1^2x_2 \#(S3)$$

We negated this function compared to Bertsimas et al., who were interested in maximization. The input variables $x = [x_1,x_2]$, $x_i \in [0,1]$ and follow the uniform distribution $X \sim U(x - 0.15, x + 0.15)$. Fig.S5 shows the true response of 2-

dimontion problem. As shown by the star symbol in Fig.S5, the global minima of the function are at x^* = (0.908, 0.919). Intuitively, x^* is the desired optimal solution, but it is clearly that the perturbation resistance is stronger at the suboptimal solution (In this case, design at [0.198, 0.085] and [0.135, 0.111] is more reasonable as a robust optimal solution than [0.435, 0.915]). Both the GP and RF models were established..

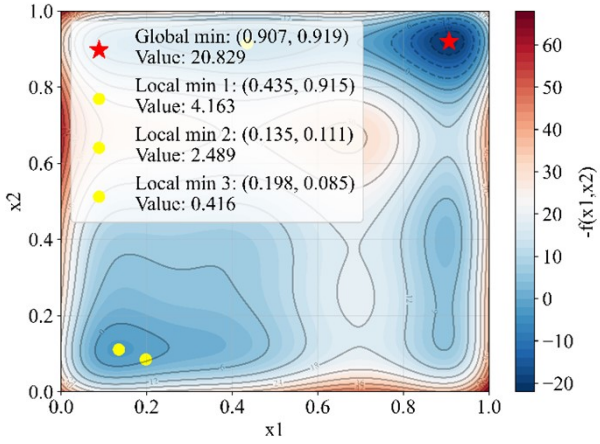

Fig.S5 Truly two-dimensional test problem

Fig. S6 to Fig. S7 show the robust prediction results of the GP model and the RF model considering the dual uncertainty after 50 iterations. They represent the dual-robust mean, dual-robust variance, and comprehensive robust model, from left to right. In Fig.S6, the GP model finds the robust optimal value located at [0.175,0.085] and [0.193,0.091] and marked by red stars. These two results represent robust solutions for the expectation-based robustness model $\hat{g}_{EBR,(\delta,S_f)}$and the composite robustness model $\hat{g}_{CR,(\delta,S_f)}$ under dual uncertainty. Because of the characteristics of RF segmented prediction, the predicted values given on a two-dimensional problem usually behave as if they were equal in a certain region. Therefore, we take the center of the region as the robust design point. As shown in Fig.S7, the RF model finds the robust optimal value located at [0.162,0.093]. As can be seen from Fig.S7, the region given by the robust prediction model covers the robust design points shown in Fig. S8, and the width of the region does not exceed the uniform distribution parameter of our design. This means that the results given by the GW method satisfy the robust design requirements for both the GP model and the RF model. The benchmarks carried out suggest that proposed can be able to efficiently guide optimization campaigns towards robust solutions.

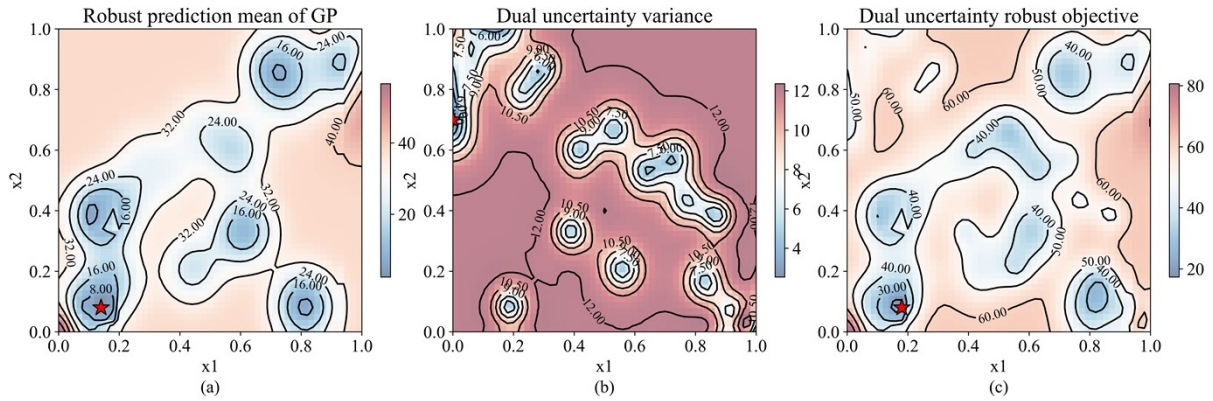

Fig.S6 Robust design use GP for minimize objective function of dual uncertainty scenarios

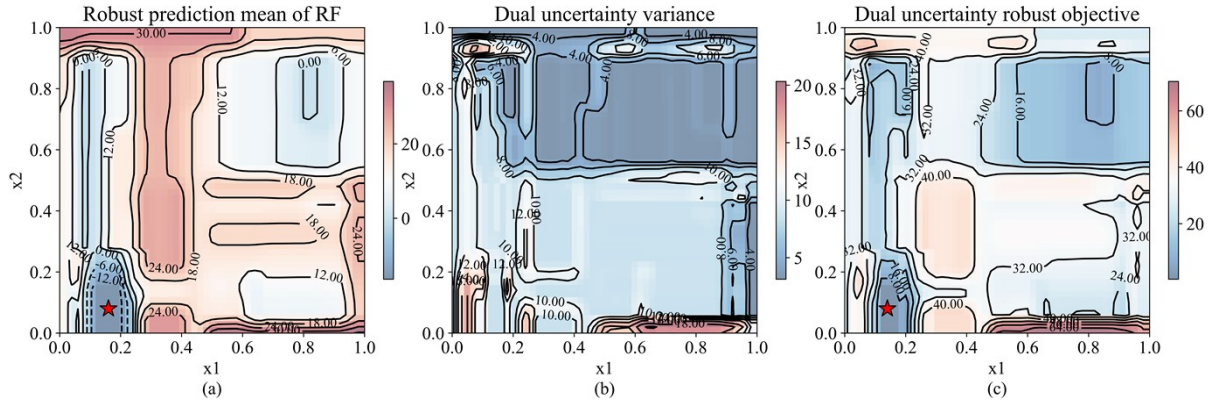

Fig.S7 Robust design use RF for minimize objective function of dual uncertainty scenarios

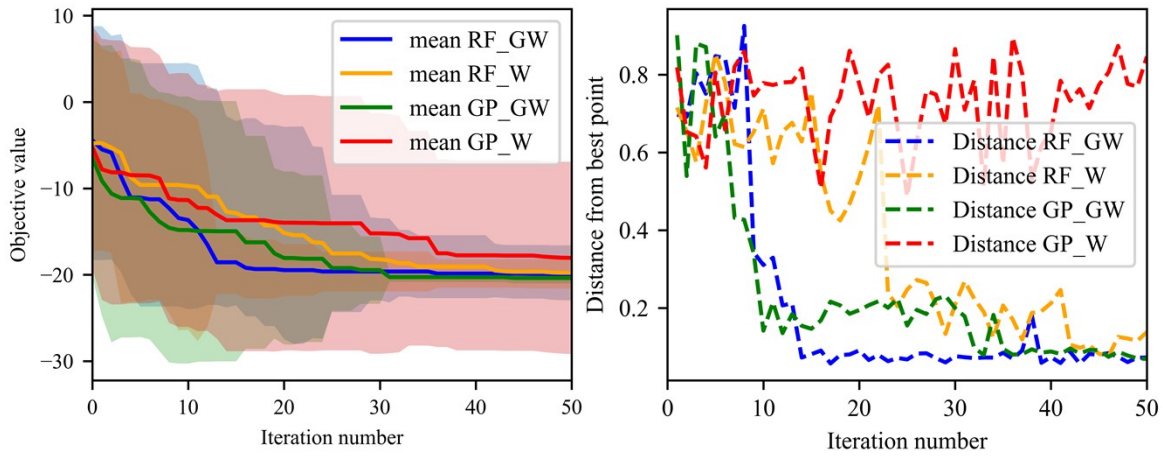

Fig.S8 The convergence of distance indicators under the four uncertainty strategies on the 2D test problem

Fig.S8 shows the convergence behavior of four strategies on 2d test problem. As with the conclusion of the 1d test problem, the schemes that consider the double uncertainty (RF_GW and GP_GW) converge faster. The distance metrics displayed on the right side of Fig.S8 show that the algorithm performs intense exploratory behavior early in the iteration due to multiple robust solutions to the 2d test problem. Comparison of the left and right plots reveals significant instability in the GP_W-based results. This instability is quantified by the right-panel distance metric, where GP_W-derived values show substantial deviation from the true robust design point.

We measured the computational time of the EGO algorithm using both GP and RF models under two scenarios across benchmark test problems. Each iteration's computational time comprises with surrogate model training, robust counterpart construction, maximizing robust EI criterion, and infill sample evaluation. Fig.S9 compares the computational costs of different robust counterpart models in 1D and 2D test problems. Reported values are averaged over 10 independent runs in seconds. The time complexity of the GP model scales cubically with the training data size. Consequently, the computational cost of the GP-based EGO algorithm grows with iterations, while the EGO algorithm based on the RF model maintains consistent computational requirements across iterations. Because the RF-based EGO with dual uncertainty avoids complex integration, its total computational time is significantly shorter than that of the GP-based EGO. For RF-based EGO (both input and dual uncertainty variants), the input uncertainty calculation step is shared. Additionally, the dual-uncertainty RF-EGO requires extra computations to fuse model and input uncertainties. Therefore, the computation time of EGO-RF_GW is approximately three times that of input-uncertainty RF-EGO. Nevertheless, it remains substantially faster than GP-based EGO.

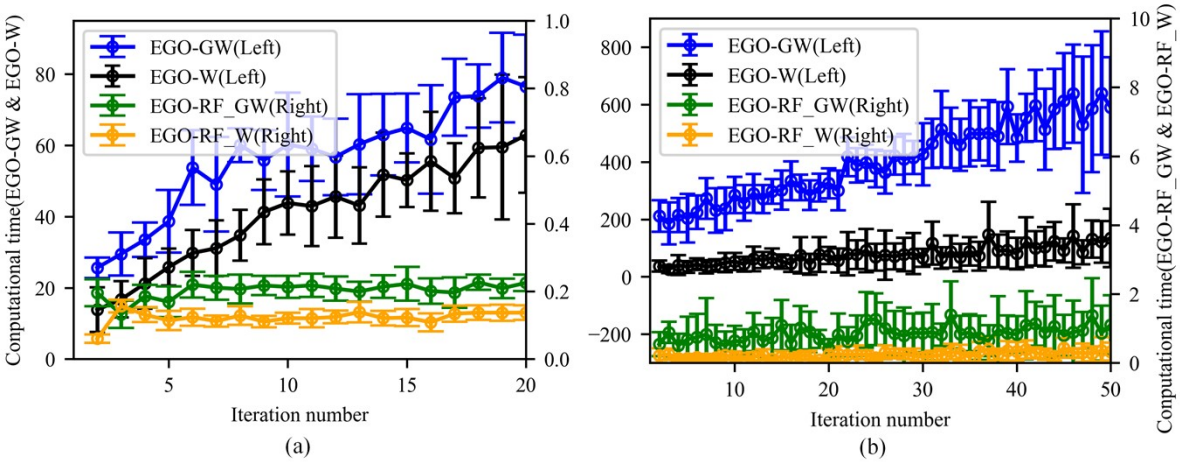

Fig.S9 Computational time of the EGO algorithm with GP and RF model with two scenarios