Supporting Information: Computing a Global Degree of Rate Control for Catalytic Systems

Huijie Tian and Srinivas Rangarajan*

Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, PA, 18015

E-mail: srr516@lehigh.edu

Phone: +1 610-758-4219

1 Method: Orthogonal Polynomial Expansions

Orthogonal polynomial expansions, also known as polynomial chaos expansion, is used to calculated the ANOVA-based global sensitivities.[?] Supposing the output variable (y) is a non-linear function of n input variables $(x_1, x_2, ..., x_n)$, such as

$$y = f(x_1, x_2, ..., x_n) \tag{1}$$

Such function can be expanded with an infinite basis set of orthogonal "basis" functions.[S1]

$$y = \sum_{j=0}^{\infty} c_j \Phi_j(x_1, x_2, ..., x_n) \tag{2}$$

where c_j is the coefficient and Φ_j is the multivariate orthogonal polynomial function. The construction of the multivariate polynomial is shown in Sec. 1.2

In practice, the polynomial expansion is truncated to a finite term P,

$$y = \sum_{j=0}^{P} c_j \Phi_j(x_1, x_2, ..., x_n) \tag{3}$$

and the relationship between the number of polynomials P with the highest order of polynomial p and the total number of input variables N is given by

$$P = \frac{(N+p)!}{N!p!} \tag{4}$$

In the following, each multivariate polynomial is represented by a list of integers $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, and α_j is the polynomial order of the jth component.

The computation of the coefficients (c_j) can be calculated by the regression of the output variable with respect to the polynomial basis on the sampling points. Supposing we generate M data points with Monte Carlo sampling, with the outputs $(y_1, y_2, \ldots, y_M)$ and the inputs $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M)$, the system can be expressed as

$$\mathbf{y} = \mathbf{\Phi c} \tag{5}$$

which can be expanded to

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} \Phi_0(\mathbf{X_1}) & \Phi_1(\mathbf{X_1}) & \ldots & \Phi_P(\mathbf{X_1}) \\ \Phi_0(\mathbf{X_2}) & \Phi_1(\mathbf{X_2}) & \ldots & \Phi_P(\mathbf{X_2}) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_0(\mathbf{X_M}) & \Phi_1(\mathbf{X_M}) & \ldots & \Phi_P(\mathbf{X_M}) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_P \end{bmatrix} \tag{6}$$

The coefficient set $(c_0, c_1, \ldots, c_P)$ are calculated via linear regression by minimizing the sum squared error between the expansion and the collected data.

$$\mathbf{c} = \text{Argmin} \left[\sum_{i=1}^{M} (y_i - \sum_{j=0}^{P} c_j \Phi_j(\mathbf{X_i}))^2 \right] \tag{7}$$

To be noted, the regularization such as $L1-$,[S2] $L2-$[S3] can be applied in the linear regression for a sparser representation.

1.1 ANOVA indices

The detail derivation of ANOVA indices based on polynomial chaos expansion (PCE) can be found in.[?] Here we summary it briefly. With the orthogonality of the polynomial basis, the mean ($\bar{y}$) and the variance (D_{PC}) of the responsive variable can be easily calculated, such as

$$\bar{y} = E[f(\mathbf{x})] = c_0 \tag{8}$$

$$D_{PC} = E[f^2(\mathbf{x})] - (E[f(\mathbf{x})])^2 = \sum_{j=1}^{P-1} c_j^2 E[\Phi_j^2(\mathbf{x})] \tag{9}$$

It is straightforward to derive the Sobol' indices with the current setting. The Sobol indices (SI) of given a parameter set $i_1, \ldots, i_s$ is expressed as

$$SI_{i_1,\ldots,i_s} = \sum_{\boldsymbol{\alpha} \in \Xi(i_1,\ldots i_s)} c_{\boldsymbol{\alpha}}^2 E[\Phi_{\boldsymbol{\alpha}}^2]/D_{PC} \tag{10}$$

where $\Xi(i_1, ...i_s)$ is a subset of the full polynomial set in which the polynomial order of parameter $i_1, ..., i_s$ are non-zero, so that the $SI_{i_1,\ldots,i_s}$ represents the dependence of the responsive variable on the input parameter set $i_1, \ldots, i_s$ only.

The total Sobol' indices (SI^T) of a given parameter set $j_1, \ldots, j_t$ is calculated as

$$SI^T_{j_1,\ldots,j_t} = \sum_{(i_1,\ldots,i_s) \in \Xi_{j_1,\ldots,j_t}} SI_{i_1,\ldots,i_s} \tag{11}$$

where the the subset Ξ is defined as

$$\Xi_{j_1,\ldots,j_t} = \{(i_1, \ldots, i_s),\ (j_1, \ldots, j_t) \subset (i_1, \ldots, i_s)\} \tag{12}$$

To be noted, the major computational cost is the construction of polynomial basis function and the calculation of linear coefficient with linear regression.

1.2 Construction of multivariate polynomials

Supposing we have a set of univariate polynomial basis ($U_k(x)$, $k \in \mathbb{N}$), the multivariate polynomial Φ of $x_1, \ldots, x_n$ is constructed by products of univairate polynomials, such as

$$\Phi_{\boldsymbol{\alpha}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} U_{\alpha_i}(x_i) \tag{13}$$

The order of $\Phi_{\boldsymbol{\alpha}}$ is less or equal to the summation of $\{\alpha_i,\ i = 1, \ldots, n\}$

1.3 Choice of polynomials

In general, if the inputs $\mathbf{x}$ are uncorrelated, a good choice for the orthogonal polynomial basis set is the Legendre polynomial; in this work, the expansions are based on Legendre polynomials. If the natural distribution of $\mathbf{x}$ is known to be Gaussian, on the other hand, a Hermite polynomial can be used; Beta distributions can use Jacobi polynomials and Gamma distributions the Laguerre polynomials. If the distribution is neither of these, arbitrary polynomial chaos expansions can be used.

1.4 Example

Here, we use an example to illustrate the construction of a multivariate polynomial, the construction of a polynomial chaos expansion model, and the calculation of Sobol' indices.

Supposing we have the responsive variable (y) which is a function of three input variables (x_1, x_2, x_3). Each of them are independent uniformly distributed over [-1, 1]. The natural choice of polynomial is Legendre polynomial. The maximum PCE order is presumed as three

for simplicity. The first three Legendre polynomials are

$$L_1(x) = x \tag{14}$$

$$L_2(x) = \frac{1}{2}(3x^2 - 1) \tag{15}$$

$$L_3(x) = \frac{1}{2}(5x^3 - x) \tag{16}$$

They are orthogonal within the uniform probability, and the expectation of the product between the two polynomial is given by:

$$E[L_m(x)L_n(x)] = \frac{1}{2n+1}\delta_{mn} \tag{17}$$

The multivariate polynomial with the set of order $(\alpha_1, \alpha_2, \alpha_3)$ is expressed as

$$\Phi_{\alpha_1,\alpha_2,\alpha_3} = L_{\alpha_1}(x_1)L_{\alpha_2}(x_2)L_{\alpha_3}(x_3) \tag{18}$$

For example,

$$\begin{aligned} \Phi_{2,1,3} &= L_2(x_1)L_1(x_2)L_3(x_3) \\ &= \left[\frac{1}{2}(3x_1^2 - 1)\right][x_2]\left[\frac{1}{2}(5x_3^3 - x_3)\right] \end{aligned} \tag{19}$$

The PCE with maximum order three is constructed as

$$
\begin{aligned}
PCE(x_1, x_2, x_3) = c_0 + \\
& c_{0,0,1}\Phi_{0,0,1} + c_{0,1,0}\Phi_{0,1,0} + c_{1,0,0}\Phi_{1,0,0} + \\
& c_{0,0,2}\Phi_{0,0,2} + c_{0,2,0}\Phi_{0,2,0} + c_{2,0,0}\Phi_{2,0,0} + \\
& c_{1,1,0}\Phi_{1,1,0} + c_{1,0,1}\Phi_{1,0,1} + c_{0,1,1}\Phi_{0,1,1} + \\
& c_{0,0,3}\Phi_{0,0,3} + c_{0,3,0}\Phi_{0,3,0} + c_{3,0,0}\Phi_{3,0,0} + \\
& c_{2,1,0}\Phi_{2,1,0} + c_{1,2,0}\Phi_{1,2,0} + c_{0,1,2}\Phi_{0,1,2} + c_{0,2,1}\Phi_{0,2,1} + c_{1,0,2}\Phi_{1,0,2} + c_{2,0,1}\Phi_{2,0,1} + \\
& c_{1,1,1}\Phi_{1,1,1}
\end{aligned}
\tag{20}
$$

Once the linear coefficients $\mathbf{c}$ is calculated, the variance of the responsive variable (from Eqn. 9) is estimated as

$$
D_{PC} = \sum_{\boldsymbol{\alpha}} c^2_{\alpha_1,\alpha_2,\alpha_3} \frac{1}{(2\alpha_1+1)(2\alpha_2+1)(2\alpha_3+1)} \tag{21}
$$

The individual and total Sobol' sensitivity indices can be calculated straightforward (See Eqn. 10 and 11). For example,

$$
\begin{aligned}
SI(x_1, x_2) &= \frac{1}{D_{PC}} \sum_{\boldsymbol{\alpha}:\{\alpha_1\neq 0, \alpha_2 \neq 0, \alpha_3 = 0\}} c^2_{\boldsymbol{\alpha}} E[\Phi^2_{\boldsymbol{\alpha}}] \\
&= \frac{1}{D_{PC}} \left[\frac{c^2_{1,1,0}}{9} + \frac{c^2_{2,1,0}}{15} + \frac{c^2_{1,2,0}}{15} \right]
\end{aligned}
\tag{22}
$$

$$
\begin{aligned}
SI^T(x_1) &= \frac{1}{D_{PC}} \sum_{\boldsymbol{\alpha}:\{\alpha_1\neq 0\}} c^2_{\boldsymbol{\alpha}} E[\Phi^2_{\boldsymbol{\alpha}}] \\
&= \frac{1}{D_{PC}} \left[\frac{c^2_{1,0,0}}{3} + \frac{c^2_{2,0,0}}{5} + \frac{c^2_{3,0,0}}{7} + \frac{c^2_{1,1,0}}{9} + \frac{c^2_{1,2,0}}{15} + \frac{c^2_{2,1,0}}{15} + \frac{c^2_{1,1,1}}{27} \right]
\end{aligned}
\tag{23}
$$

1.5 Sampling and Convergence

The accuracy of the PCE the efficiency and accuracy of the sensitivity indices calculation, especially when extrapolating. It is also advised to split the dataset, and check the model performance outside the training data. Uncertainty quantification (UQ) on PCE can, in principle, identify the data when doing extrapolation. Since PCE is linear with respect to the basis function, uncertainty quantification (UQ) on linear regression can be easily applied to PCE. In addition, sparsifying the basis set with regularized regression, such as LASSO, can also be used to improve the model accuracy and reduce the overfitting with less training data.

For high dimensional problem, convergence of the model can be an issue. It is has been shown that quasi-Monte Carlo sampling, such as Sobol sampling, has better convergence for training the model with higher dimensional space. In addition, model-based design of experiments using linear methods such as D-optimal sampling can be used to further improved the convergence.

1.6 Additional Example about Global Sensitivity Analysis

Assuming we have two functions $f_1(x_1, x_2) = x_1 x_2$, $f_2(x_1, x_2) = x_1^2 + x_2^2$, and x_1 and x_2 are two uniform random variable between [0, 1], the global sensitivity indices can be calculated analytically:

$$f_1 : S_1 = 0,\ S_2 = 0,\ S_{12} = 1 \tag{24}$$

$$f_2 : S_1 = 0.5,\ S_2 = 0.5,\ S_{12} = 0 \tag{25}$$

In both cases, the derivative of f with respect to parameters (x_1 and x_2) is not constant across the space (i.e. the model is not linear), however, their combined sensitivity (S_{ij}) are in stark contrast. This particularly shows that S_{ij} represents the extent to which the derivative

of f with respect to x_i depends on the value of x_j or the degree to which the parameter contributions cannot be fully separated.

For another case where $f = x_1^2 + x_2^2 + ax_1x_2$; if a is small and we pick a parameter space for x_i such that $ax_1x_2 \sim 0$, then $f \sim x_1^2 + x_2^2$ in that space, we will see that $S_1 \sim 0.5$, $S_2 \sim 0.5$, $S_{12} \sim 0$. However, as we increase the parameter space of x_i, we may encounter regions of the subspace where $ax_1x_2 \neq 0$ (i.e it is non-negligible), and as a result $S_{12} > 0$. This is what is happening to our microkinetic model as we increase our parameter space.

2 Water gas shift reaction

2.1 Validation of microkinetic model

Figure S1 shows the parity between the kinetic experimental data[S4] and microkinetic model predictions after Bayesian calibration. The good agreement shows the calibrated microkinetic model describes the kinetic experiments and underlying mechanism very well.

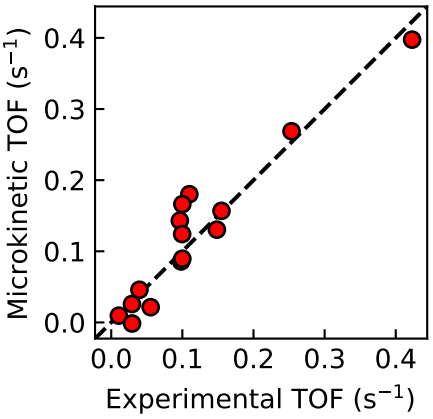

Figure S1: Parity plot of the microkinetic modeling prediction (kinetic parameters are in Table 2 in the manuscript) vs. experimental data of water gas shift on Cu(111) at P_{H_2O}=0.69 - 2.94 kPa, P_{CO}=2.01 - 12.6 kPa, and 560 - 680 K.[S4]

2.2 Validation of PCE model

Figure S2 shows the parity of the polynomial expansion surrogate model with the full microkinetic model for different parameter ranges. Clearly, the surrogate model is close enough to the original model, which indicates the reliability of the coefficients of the expansion.

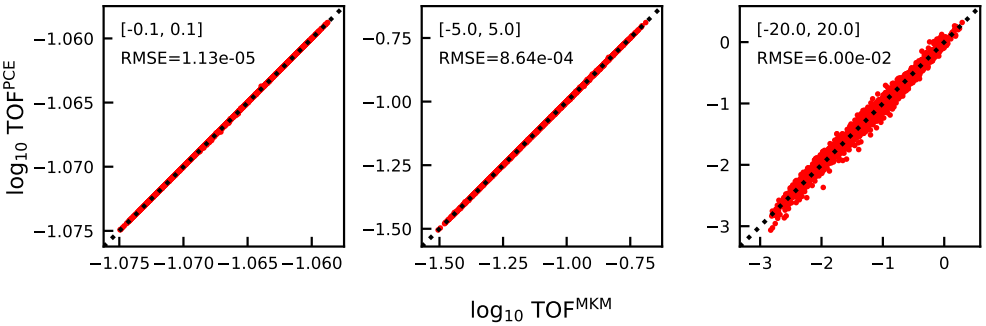

Figure S2: The fitted PCE model vs. the microkinetic model predictions on the test dataset with 5000 sampling.

2.3 Global sensitivity analysis on kinetic parameters only

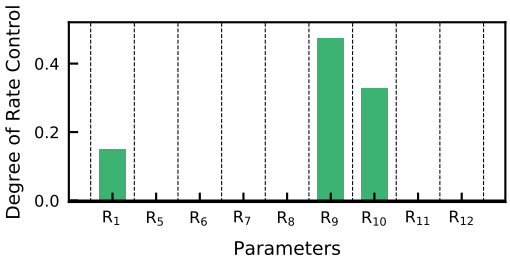

Figure S3: The degree of rate control with respect to the kinetic parameters of the water gas shift reaction studied in the main manuscript (kinetic parameters are in Table 2 in the manuscript) at 610 K, P_{H_2O}=1.3 kPa and P_{CO}=3.5 kPa.

Figures S3 and S4 gives the degree of rate control of only the reaction steps. Clearly, when we compare the global sensitivities at small parameter ranges [-0.1,0.1], we can note that the values are similar.

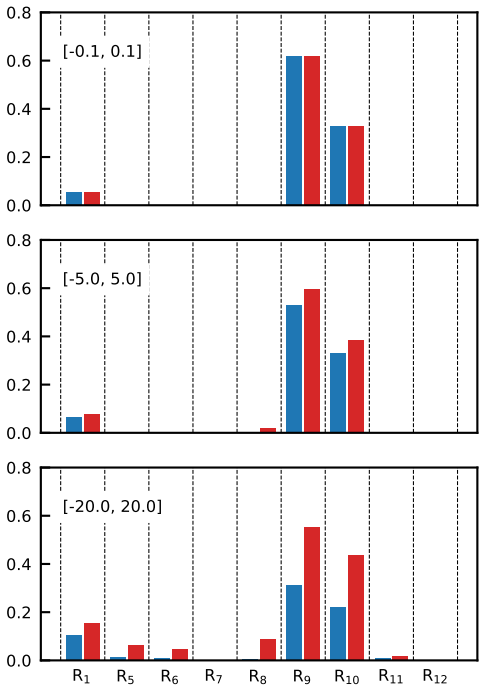

Figure S4: Global sensitivity values with respect to the kinetic parameters of the water gas shift reaction studied in the main manuscript (kinetic parameters are in Table 2 in the manuscript) at 610 K, P_{H_2O}=1.3 kPa, and P_{CO}=3.5 kPa.

2.4 Application of global sensitivity analysis for uncertainty quantification

In this section, we applied both global and local sensitivity analysis in the application of uncertainty quantification on microkinetic model predictions. We compared the performance of the different methods (that we use to compute global and local sensitivities) for predicting the distribution of the turnover frequency under the assumption that each kinetic and thermodynamic parameter has an error of ± 20 kJ/mol (in terms of energies). We compared (i) the global orthogonal polynomial model (also called the polynomial chaos expansion or the PCE model that we use for computing global sensitivity analysis), (ii) linear extrapolation model (based on the local derivatives that we use to compute the degree of rate control), (iii)

the full-fledged microkinetic model (with no approximations), and (iv) a "reduced" kinetic model where the unimportant parameters identified using global sensitivity analysis were fixed to their nominal values.

The specific procedure is as follows. First, we ran the full microkinetic model of water gas shift on Cu(111) at 50000 randomly sampled data points within a parameter range of [-20, 20] (for all the parameters) and collected the distribution data of the turnover frequency predicted by the model. This represents the true uncertainty of the model predictions under the assumption that the parameters all have an uncertainty range of [-20, 20]. Second, we used the previously trained PCE model (used in our global sensitivity analysis, see SI 2.2 and 2.5) to predict the TOF distribution. Third, we used the results of our global sensitivity analysis (see Table 3 in the manuscript) to fix energies corresponding to R_7, R_{11}, R_{12}, I_1, I_2, I_6, I_7, I_8 at their nominal points in the microkinetic model and used the reduced set of parameters to obtain the prediction distribution. Finally, we use the degree of rate control values at the nominal point to make a simple linear extrapolation to predict the TOFs by using the truncated Taylor expansion: $(f(\mathbf{x}) = f(\mathbf{x_0}) + \nabla f(\mathbf{x})|_{\mathbf{x}=\mathbf{x_0}}(\mathbf{x} - \mathbf{x_0}))$

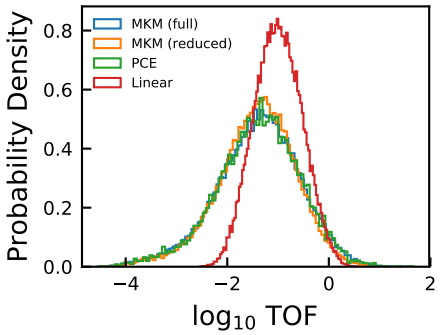

Figure S5: Predicted distribution of turnover frequencies of water gas shfit on Cu(111) at 610 K, P_{H_2O}=1.3 kPa, and P_{CO}=3.5 kPa for the parameter range [-20,20] by different models. MKM (full): full-fledged microkinetic model with all seventeen parameters; MKM (reduced): "reduced" microkinetic model with nine parameters (R_7, R_{11}, R_{12}, I_1, I_2, I_6, I_7, I_8 are fixed to their nominal points); PCE: polynomial chaos expansion; Linear: linear extrapolation with DRC values based on truncated Taylor expansion

Clearly, we can see that the PCE model is very close to the full microkinetic model

(MKM) predictions because PCE is trained on the data sampled from the same parameters space [-20, 20]. The prediction distribution of the "reduced" model is also close to these two cases, validating the results of our global sensitivity analysis. We expect that if we used the reduced set of parameters along with the PCE model, the resulting distribution would match the previous two distributions. The distribution obtained from the linear model is shifted and narrower compared to the other three. In addition, since the linear model neglects the nonlinear effect in the system, the distribution of linear model is symmetric with respect to the nominal point. This comparison clearly reveals the discrepancy of using a local method to perform uncertainty quantification when the parameter space is wide.

It should be noted that the computational cost of PCE model (45000 microkinetic runs) is much higher than local linear model (1 microkinetic run + 17 local derivative calculations). Clearly therefore there is an accuracy-cost trade-off here and the local sensitivity analysis is preferable when under tight computational budget.

2.5 Implementation detail

The maximum order of PCE with seventeen parameters (9 reactions + 8 intermediates) is set to four. So the total number of polynomials term is $\left(\frac{(17+4)!}{17!4!}\right) = 5985$. The training data are sampled from a uniform distribution within certain range, and the data not satisfying the thermodynamic consistency are discarded. For each uncertainty range, 50000 data points are generated by simulated the CSTR model. 45000 of them are randomly selected to train the PCE, and the rest are used to validate the model (shown in the parity plot).

References

(S1) Szeg, G. *Orthogonal polynomials*; American Mathematical Soc., 1939; Vol. 23.

(S2) Tikhonov, A. N.; Goncharsky, A.; Stepanov, V.; Yagola, A. G. *Numerical methods for the solution of ill-posed problems*; Springer Science & Business Media, 2013; Vol. 328.

(S3) Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **1996**, *58*, 267–288.

(S4) Campbell, C. T.; Daube, K. A surface science investigation of the water-gas shift reaction on Cu(111). *J. Catal.* **1987**, *104*, 109 – 119.