SPE Brazil Flow Assurance Technology Congress
Rio de Janeiro, Brazil, November 06 – 08 2024

Towards Data-Aware Flow Assurance: Enhancing Flash-Calculations using a Machine Learning-based Bayesian Model Calibration

Reza Arefidamghani[1], Gustavo E. O. Celis [2], Daniel O. A. Cruz[2], Hamidreza Anbarlooei[1*], Raiza G. Souza[3,] Jorge Trujillo[3]

[1]Department of Applied Mathematics, Federal University of Rio de Janeiro, Brazil, *hamidreza@im.ufrj.br
[2]Mechanical Engineering Program, Federal University of Rio de Janeiro, Brazil
[3]Upstream-Production Engineering, Galp Energy S.A., Lisbon, Portugal

Abstract

The present work introduces a Machine Learning (ML) technique into the thermodynamics equilibrium of multicomponent mixtures flash-calculations. Particularly, CO_2 enriched mixtures´ systems are analyzed here. Binary interaction parameters play a critical role in flash-calculations as they quantify the molecular interactions within mixtures. However, accurately estimating these parameters poses a significant challenge due to the complex nature of molecular interactions. Fine-tuning, adjusting, and optimization techniques are commonly employed to determine these parameters. However, traditional approaches often require intuitive comprehension and are limited by the availability of data. This study employs sensitivity analysis to identify key parameters for calibration. Then, the machine learning-based Bayesian model calibration (ML-BMC) technique is employed to refine model parameters while constraining associated uncertainties. ML-BMC not only offers comparable calibration to expensive optimization methods but also provides a surrogate model for generate new predictions with constrained uncertainties and cost-effective sensitivity analysis. The developed methods function as a data-driven PVT solver, capable of refining predictions with additional experimental or operational observations.

Keywords

Machine Learning; Calibration; Flash-calculations.

Introduction

The integration of Carbon Capture, Utilization, and Storage (CCUS) stands as a pivotal strategy in steering the transition toward zero-emission energy systems. The effective implementation of CCUS initiatives relies heavily on a thorough comprehension of the thermodynamic principles governing CO_2 mixtures and their accompanying impurities. This deep understanding is essential not only for the effective deployment of such projects but also for optimizing the performance of multi-phase flow modeling techniques tailored to these complex fluid systems.

The flow assurance simulations often necessitate a PVT table derived from flash calculations. Numerous equations of state and computational techniques have been devised for conducting Vapor-Liquid Equilibrium (VLE) and Vapor-Liquid-Liquid Equilibrium (VLLE) calculations. Cubic equations of state (EoS) coupled with van der Waals quadratic mixing rules serve as the basis for numerous commercial and academic software applications.

In this context, Binary Interaction Parameters, k_{ij}, (BIPs) play a pivotal role as they quantify the molecular interactions between different chemical species within a mixture. However, due to the complex nature of these interactions, no formal theory exists to precisely estimate these parameters. While several methods are available for computing them [1 , 2 , 3] evidence suggests that these calculations may lack accuracy and require adjustments. Therefore, fine-tuning, adjusting, and optimization techniques are commonly utilized to determine the appropriate values of these parameters [4 , 5 , 6].

Nevertheless, conventional tuning approaches often demand intuitive understanding and are limited by the availability of data. Also, optimization techniques will become expensive when dealing with numerous parameters (in a mixture with 21 components there are about 210 parameters to be tuned).

In this study, we aim to leverage the advancements in machine learning techniques to discern the most critical parameters and tune them using limited experimental data from various sources. To achieve this objective, we have opted for the Bayesian model calibration (BMC) technique ([7], [8]). Initially, we train a surrogate model using diverse simulations to establish a rapid model. Subsequently, we employ Gaussian Process-based Calibration (GPC) to fine-tune the parameters.

The next critical step involves selecting the parameters for calibration. To address this, we have integrated a sensitivity analysis technique ([9], [10]) to pinpoint parameters that hold potential for optimization.

The subsequent sections of this paper will delve into the implementation of these techniques and present the corresponding results. However, it is essential to outline the case study employed in the present study. Here, we investigate the behavior of a hydrocarbon mixture comprising over 20 components (LiveOil1) in the presence of a high concentration of CO_2 (refer to [4] for detailed information). Specifically, we analyze the vapor-liquid phase envelope of the mixture at fixed temperatures for various concentrations of CO_2 and compare it with experimental observations. Comparing the results of flash calculations with available experimental data reveals discrepancies, particularly when CO_2 is part of the mixture. In this scenario, there are 210 parameters to be fine-tuned, while we have access to only about 20 observations.

Sensitivity Analysis

Sensitivity analysis examines how uncertainty in the results of a mathematical model or system is linked to various sources of uncertainty in its input parameters. In theory, such an analysis enables us to systematically identify the most crucial parameters for fine-tuning, without the need for prior knowledge or physical intuition.

In this study, Sobol Sensitivity Analysis [9] implemented in SALib [10] has been primarily utilized. Two different criteria have been used for comparing predictions with experimental observations: Mean Squared Error (MSE) and the Integral Error (Int-Err) as the area between the experimental and simulation phase envelopes. Different forms of sensitivity indices (SI) include first-order indices, which assess the contribution to the variability in the model's output attributed to a single input parameter in isolation. Second-order indices gauge the contribution to output variability arising from the interaction between two specific input parameters (and so on).

In our specific benchmark problem, we are confronted with 210 parameters to analyze. However, even with the GPU implementation of our PVT solver (which is approximately 50 times faster than the CPU implementation), it is impractical to scrutinize the sensitivity of all 210 parameters. Hence, we are narrowing our focus to a subset of three binary interactions to compare their sensitivity indices: $k_{(CO_2,CH_4)}$ as suggested in [4], $k_{(CH_4,Squ)}$ due to highest molar weight of Squalane (Squ) among the components [4] and $k_{(CO_2,Squ)}$, based on our observations and same reasoning as before. We will discuss the solution to the issue of computational hunger of sensitivity analysis method in the next section.

Table 1 shows the results of first order sensitivity analysis (SAlib [10]) of these parameters. Our investigation suggests that due to the very small and negligible second sensitivity indices, there are no higher-order interactions observed among these parameters.

Table 1. Comparative Sensitivity Analysis Index of Binary Interaction Parameters.

Parameters	SAlib	AS-Surr
$k_{(CO_2,CH_4)}$	0.5090	0.2035
$k_{(CO_2,Squ)}$	0.0802	0.0810
$k_{(CH_4,Squ)}$	0.0052	0.0016

As can be seen, $k_{(CO_2,CH_4)}$ displays the highest SI, which is consistent with previous observations [4], indicating that optimizing this parameter significantly improves the accuracy of model predictions. Also, it is evident that the SI of $k_{(CO_2,Squ)}$ exceeds that of $k_{(CH_4,Squ)}$. This suggests that the former has a more pronounced effect on the output results, particularly in aligning the phase envelope with experimental data. Interestingly, this finding contradicts the physical rationale for prioritizing $k_{(CH_4,Squ)}$ in [4]. Consequently, relying solely on physical reasoning or intuition may not be adequate for selecting the most sensitive interaction parameter. Instead, the presented sensitivity analysis offers a systematic approach to identify parameters for calibration or optimization.

To obtain these results, approximately 8,000 simulations were conducted by sensitivity analysis algorithm, each generating phase envelopes based on different binary interaction parameter settings. Despite considering just three parameters to analysis, the computational cost remains significant, even with the utilization of resources such as GPUs. This underscores the necessity for faster solvers when analyzing all BIPs. The next section tries to tackle this issue.

It is important to note that the objective of this analysis is to identify parameters requiring optimization, while the primary task remains unaddressed. To tackle these challenges, we adopt a calibration approach employing Gaussian processes.

Gaussian Process-Based Calibration

Science-based simulations are commonly used to predict the behavior of complex physical systems. Physical observations, on the other hand, are commonly used to solve the inverse problem, that is, to learn about the values of parameters within the model. This is referred to as calibration. In what follows, we will use Kennedy and O'Hagan's [7] terminology and notation.

A simulator generally comprises two types of inputs: control variables and calibration parameters. Control variables, such as pressure, temperature and mixture composition, clarify the specific characteristics of the physical system to be predicted. In contrast, calibration parameters are parameters of the physical models, such as gravitational constant and critical temperature or pressure of a pure component. These parameters can introduce an element of uncertainty due to their imprecise values, for example, binary interaction parameters BIPs.

Let us denote our flash simulator as $\eta(u, K)$, where u signifies the control variables, and K embodies the calibration parameters. Also, suppose that one has *n* experimental observations (points on phase envelope in our case), labeled as $z_1, z_2, \ldots, z_n$. Each instance z_i corresponds to a set of control inputs u_i. The *i*-th observation z_i can be expressed as:

$$z_i = f(u_i) + \epsilon_i, \tag{1}$$

where $f(u_i)$ represents the true state of the physical system at control variable setting u_i, and ϵ_i accounts for independent observational errors. Given numerous factors like imperfect physical modeling or uncertain model parameters, there naturally exists a mismatch between the real system state and our simulator predictions. This discrepancy, modeled as:

$$f(u) = \eta(u, K) + \delta(u) \tag{2}$$

where $\delta(u)$ characterizes the deviation between physical reality and our simulations. By merging Eqs. (1) and (2), one has

$$z_i = \eta(u_i, K) + \delta(u_i) + \epsilon_i, \tag{3}$$

which relates the predictions, discrepancies and error in observations to the observed values.

To model the discrepancy term, $\delta(u)$ is constructed using a basis representation, placing GP models on the basis weights [7]. The observation error term is assumed to follow normal distribution as $N(0, \varepsilon_i)$. Ultimately, we construct a surrogate model to emulate our flash simulator. To accomplish this, we utilize a constrained set of phase envelopes derived from various control and calibration parameters to train a Gaussian process model. This surrogate model provides a cost-effective approximation of the simulator. The surrogate model uses principle component analysis to limit the dimensionality of the outputs and represents the simulator as:

$$\eta(u_i, \alpha) = B_i \, W(u_i, K), \tag{4}$$

where B_i are the basis from Principal Component Analysis (PCA) and $W(u_i, K)$ are weights defined as GP. The goal will be to maximize the probability of the observations, conditioned on the inputs, discrepancy and basis. The Bayesian model calibration (BMC) method, as outlined by Higdon et al. [7], has been put into practice. Further details can be found in the same reference [7]. One advantage of this model is that the surrogate model can be utilized for sensitivity analysis due to its cost-effectiveness. In summary, to train the surrogate model, an ensemble of test runs is generated by sampling model parameters from plausible distributions, such as a normal distribution centered around BIPs provided by group theory [2]. Subsequently, the trained emulator is integrated into a Bayesian framework to derive a posterior distribution (tuned) for the model parameters (BIPs). This step aims to refine the emulator predictions, aligning them more closely with experimental observations.

Results

In the present work, we have chosen to use the SEPIA package [11]. SEPIA (Simulation-Enabled Prediction, Inference, and Analysis) implements Bayesian emulation and calibration with the ability to handle multivariate outputs.

In the Calibration process, if each parameter's posterior marginal distribution has a single, well-defined peak, then the model parameters corresponding to those peaks can be considered calibrated model parameters. Otherwise (if sharp peaks are not created), further experimental data is likely required to fully investigate the model.

Our results indicate that the mean of the posterior distribution for each parameter obtained by GPC, closely aligns with the values obtained from expensive optimization methods like grid search and AdaGrad (Adaptive Gradient Method); see [5]. Furthermore, the small standard deviation indicates a concentrated distribution around the mean, underscoring the effectiveness of our calibration approach (see Fig. 1).

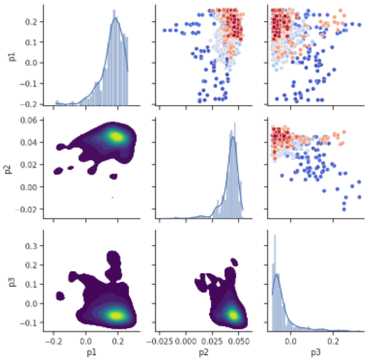

Figure 1. Posterior distribution of three selected parameters after calibration.

This model can also quantify uncertainties in the predictions. Figure 2 shows the uncertainty in the predictions at fixed temperature of 373.15 K. As can be seen, regular flash calculation underestimates the phase envelope. While, our Bayesian model response shows complete agreement with the actual observed data. This graphical representation serves as a valuable tool for evaluating the reliability of the model and offers a confidence measure regarding its performance. It is important to note that this narrow uncertainty range depicted in the plot indicates an impressive performance from our calibration model. A slim uncertainty range suggests that the model's predictions consistently closely match the observed data points. This level of precision and reliability in the model's predictions is an encouraging indication, indicating that our calibration model adeptly captures the underlying relationships within the data and produces dependable results.

Concluding our study, we compare the sensitivity analysis results obtained from the surrogate model

(SA-Surr) with those from SAlib. Remarkably, both methods identify the most critical parameters for optimization in the same order, albeit with slight variations in the indexes. These results are summarized in Tab. 1. However, a crucial observation lies in the significant disparity between the computational demands of the two approaches (240 times faster than SAlib).

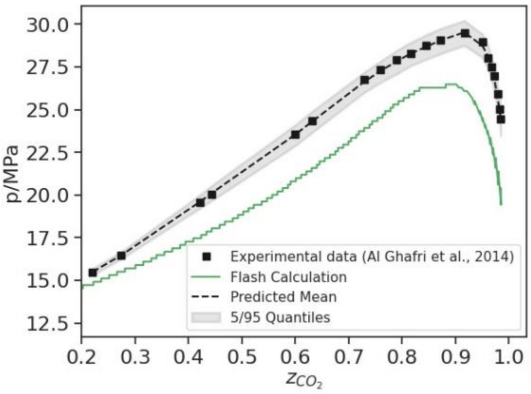

Figure 1. Uncertainty Analysis for Fixed Temperature (373.15 K)

Conclusion

While traditional fine-tuning techniques can adjust the predictions of the PVT solver with experimental observations, we opted to explore well-known machine learning techniques for this purpose. Our study demonstrates that these techniques, without the need for prior physical knowledge or experience, can effectively identify parameters suitable for optimization. Moreover, these techniques can provide vital data, such as uncertainty in the predictions. Leveraging this valuable information, it becomes feasible to explore various outcomes of flow assurance studies or plan future experiments. Finally, integrating traditional PVT solvers with these techniques transforms them into data-aware solutions.

Acknowledgments

The authors acknowledge the technical and financial support of Petrogal Brasil S.A. (Joint Venture Galp | Sinopec) and the promotion of Research, Development and Innovation (R,D&I) in Brazil by the National Agency of Petroleum, Natural Gas and Biofuels (ANP) for the execution of this project.

Responsibility Notice

The authors are the only responsible for the paper content.

References

[1] J. A. Coutinho, G. M. Kontogeorgis and E. H. Stenby, *Fluid phase equilibria,* pp. 31-60, 1994.

[2] J.-N. Jaubert and F. Mutelet, *Fluid Phase Equilibria,* vol. 224, no. 2, pp. 285-304, 2004.

[3] A. Kordas, K. Magoulas, S. Stamataki and D. Tassios, *Fluid Phase Equilibria,* vol. 112, no. 1, pp. 33-44, 1995.

[4] S. Z. Al Ghafri, G. C. Maitland and M. J. Trusler, *Fluid Phase Equilibria,* vol. 365, no. 1, pp. 20-40, 2014.

[5] G. E. O. Celis, R. Arefidamghani, H. Anbarlooei and D. O. A. Cruz, *arXiv preprint arXiv:2306.16327,* 2023.

[6] R. Arefidamghani, G. E. O. Celis, J. Trujillo, R. G. d. Souza, H. Althoff, D. O. d. A. Cruz and H. Anbarlooei, in *Proceedings of the ROG.e*, Rio de Janeiro, RJ, 2024.

[7] M. C. Kennedy and A. O'Hagan, *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 63, no. 3, pp. 425-464, 2001.

[8] D. Higdon, M. Kennedy, J. C. Cavendish, J. A. Cafeo and R. D. Ryne, *SIAM Journal on Scientific Computing,* vol. 26, no. 2, pp. 448-466, 2004.

[9] I. M. Sobol Prime, *Mathematics and Computers in Simulation,* vol. 55, no. 1-3, pp. 271-280, 2001.

[10] W. Usher, J. Herman, C. Whealton, D. Hadka, Xantares, F. Rios, Bernardoct, C. Mutel and J. Van Engelen, *SALib/SALib: Launch!,* Zenodo, 2016.

[11] J. Gattiker, N. Klein, G. Hutchings and E. Lawrence, *lanl/SEPIA: v1.1,* Zenodo, 2020.