

Notes to describe how Reviewers observations were addressed in the revised manuscript

Paper title: «**Enhancement of machine learning based flash calculations near criticality using a resampling approach**»

By: Eirini Maria Kanakaki, Anna Samnioti and Vassilis Gaganis

Firstly, we would like to thank the Reviewer for their valuable comments on our work. In this note we have tried to respond by providing concrete answers to their comments. Those answers have led to various modifications/additions to the manuscript which are marked in **red color** in the revised version. Detailed responses to the Reviewers observations are given below.

Reviewer 2

Thank you for providing the paper. The title " Enhancement of machine learning based flash calculations near criticality using a resampling approach," suggests a novel resampling technique of the ML models training data population is proposed which aims at the fine-tuning of the training dataset distribution and to the optimal exploitation of the models' learning capacity across various flash conditions. Overall, the paper addresses a valuable topic and has the potential to contribute significantly to the field. However, some minor corrections should be made before publishing.

Comment 1

Consider providing an explanation of the software used to generate the K value. This would enhance the clarity of the methodology and improve reproducibility.

Explanation & Action taken

First of all, we would like to thank the Reviewer for taking the time to carefully review our paper "Enhancement of machine learning based flash calculations near criticality using a resampling approach". We appreciate their positive feedback and suggestions for improvement.

In response to Comment 1, we agree with the Reviewer's recommendation to enhance the clarity of our methodology regarding the software used to generate the **k-values**. To address this, we have developed a dedicated MATLAB code that performs regular flash calculations. This code operates along the reservoir temperature isotherm of the studied fluid, utilizing the feed fluid composition **z**. In detail, the MATLAB code prompts users to input the lower and upper limit pressures, generating 100,000 uniformly selected pressure points within the defined range.

For each pressure point, the code executes regular flash calculations by taking the feed composition **z**, pressure, and temperature as inputs and yields the vapor phase and liquid phase compositions **y** and **x**, respectively. Subsequently, the **k-values** for each

pressure point are determined by calculating the ratio of the vapor phase composition of a given component i to its corresponding liquid phase composition. The algorithm utilized to run the flash calculations is well established in the literature and is further cited in the originally submitted manuscript in line 36.

To clarify this process, we have introduced a detailed discussion in lines 274-280 of the revised manuscript, aiming to ensure that readers can readily comprehend and replicate our methodology. We believe that this addition significantly enhances the transparency and reproducibility of our study.

Comment 2

The paper primarily assigns values to D based on the proximity to the critical point and loss and adjusts the data quantity for each group accordingly. It is suggested to explore the possibility of using Gaussian or other distribution methods to obtain multiple points around the critical point when generating 100,000 data points. This could add robustness to the methodology.

Explanation & Action taken

We thank the Reviewer for their suggestion regarding the distribution methodology employed in our paper. In response to the Reviewer's comment, we would like to clarify our rationale for opting for a uniform distribution. Our primary objective was to illustrate the contrast between uniform and non-uniform pressure- k -value distributions in the base training dataset, as shown in Figure 6. By utilizing a uniform distribution, we aimed to effectively visualize the resulting data imbalance and its consequential impact on the training process.

It is important to emphasize that our final balanced dataset is derived through sub-sampling the source dataset based on the output values (**k -values**), rather than the cardinality of the points. This means that the specific distribution of the source dataset does not exert a substantial influence on the balanced dataset, reinforcing the robustness of our approach.

While we acknowledge the potential benefits of employing Gaussian or other distribution methods, we considered the uniform distribution as the most suitable choice for our study. This decision was driven by the fact that we lack prior knowledge about the expected distribution of **k -values** in real-world flow simulations. The distribution in such scenarios depends on various factors, including the flow model, fluid properties, and production schedule, which are beyond our control.

Our commitment to ensuring the adaptability of our method to diverse real-world scenarios led us to choose the safest option – uniform distribution. This choice allows our methodology to remain robust and applicable across a spectrum of situations where the underlying distribution might be uncertain or variable.

Comment 3

The main variable in the paper is pressure. To strengthen the persuasiveness of the study, it is recommended to simultaneously vary both pressure and temperature.

Additionally, exploring variations in the composition ratios would contribute to a more comprehensive analysis.

Explanation & Action taken

We thank the Reviewer for their valuable feedback on our paper, and we appreciate their suggestion to simultaneously vary both pressure and temperature in evaluating the accuracy of our proposed k-value prediction model. We value this suggestion and would like to offer a comprehensive explanation of our approach.

In our initial investigation, our deliberate emphasis on pressure variations was rooted in the common constancy of temperature observed during the exploitation of gas condensate reservoirs. This strategic choice aligns with the typical operational conditions taking place in these reservoirs, where variations in temperature are not commonly encountered. It is important to note that our decision to focus on pressure dynamics does not negate the significance of the Reviewer's suggestion. Instead, it reflects a targeted exploration of a key variable based on the specific characteristics of the reservoirs under consideration.

Furthermore, we would like to draw attention to the inclusion of both lean and rich gas condensates in the proof-of-concept section (Section 2) of our paper. This inclusion demonstrates the sensitivity of inaccuracies in **k-values** near criticality across a diverse range of fluids. Notably, the deliberate testing of our methodology in rich gas condensate fluid (Section 4) was motivated by the recognition that these fluids pose a more challenging modeling scenario than lean gas condensates. Successfully predicting **k-values** for rich gas condensates serves as a robust demonstration of the efficacy of our methodology in handling complex fluid systems.