Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Viktor Duplyakov

PhD Program: Petroleum Engineering

Title of Thesis: Machine learning on field data for hydraulic fracturing design optimization

Supervisor: Professor Andrei Osiptsov
Co-supervisor: Professor Evgeny Burnaev

Name of the Reviewer: Professor Dmitry Garagash

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 19-09-2023**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense. *If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.*
Reviewer's Report
Reviewers report should contain the following items: • Brief evaluation of the thesis quality and overall structure of the dissertation. • The relevance of the topic of dissertation work to its actual content • The relevance of the methods used in the dissertation • The scientific significance of the results obtained and their compliance with the international level and current state of the art • The relevance of the obtained results to applications (if applicable) • The quality of publications The summary of issues to be addressed before/during the thesis defense

- Brief evaluation of the thesis quality and overall structure of the dissertation.

  I found the thesis very well written, logically structured, and of very good written quality. The overarching subject of the thesis (application of the machine learning in hydraulic fracturing design) is very well introduced, followed by systematic exposition of the main contributions of this work.

- The relevance of the topic of dissertation work to its actual content

  The main topic of the thesis is the evaluation and application of the machine learning (ML) methods to a large database of hydraulic fracturing treatments to identify key treatment parameters and optimize treatment design. The actual content of the thesis is fully relevant to its announced topic.

- The relevance of the methods used in the dissertation

  The methods used in the thesis, as relate to a large database building and the ML methods, are well described and appear to be relevant (this reviewer is not an ML practicioner).

- The scientific significance of the results obtained and their compliance with the international level and current state of the art
  The result are significant and comply with international level and state of the art – see the comments below.

- The relevance of the obtained results to applications (if applicable)
  The results of this work are very important to oil and gas field applications – indeed assimilating existing large sets of reservoir characterization, stimulation and production data for the use in ML forward models is a paramount for optimization of reservoir treatment and subsequent production.

- The quality of publications
  The research in this thesis has been published as two journal articles and one conference (Society of Petroleum Engineers - SPE) Proceedings. The former journal publications are on the database building and ML inverse problem, respectively, are published in the international journals that are very well reputed in the Oil and Gas science and engineering field. The conference paper, which addresses practical aspects of HF optimization using a field test (21 wells), provides an important link to oil field practitioners (SPE publications are the single most important outlet for petroleum industry researchers and practitioners).

- **Comments and questions to the Candidate**

  1) Thesis researches two sets of problems with regard to design of ML approach to reservoir forecasting: a) what constitutes the optimum set of data (for a given well) and number of wells in given groups in the database for the method traning purposes; b) what constitutes a reliable ML methodology and corresponding optimal set of parameters to design a successful HF job.

  The preferred trained ML model in this thesis, when tested on hold-out sample of the database, yielded R^2=0.64. Is this a good result for an ML forward model?

Related to this, the author writes:
p.71 "Thus, an accurately formed, verified and validated field database on stimulation treatments may lead to the results that are not "ideal" (in terms of the predictive power), because of its inherent heterogeneities. Poor results may also be explained by lack of data due to the nature of the problem: we have very limited knowledge about the reservoir structure and its properties."

Does "heterogeneity" refers to fundamentally different reservoir and well/fracture classes? (for example: vertical, multi-layer well vs horizontal presumably single layer contained well. Or different type of reservoir: conventional vs unconventional?)
Would an ML-method, given sufficiently representative training database, is expected to learn the structure of heterogeneity and adapt and recognize it accordingly, allowing for accurate forward ML forecasting fore distinct input setups?

The fact that the accuracy is "not ideal" - does this mean that the database perhaps is still under-represented for some subclasses of input conditions?

Was there an attempt to select a more 'homogeneous' sub-set of the database and provide forward ML forecasting for this subset - perhaps in some ways similar to previous studies which looked at much smaller data samples? If so, would that allow for a more accurate ML prediction? And if so, should one then suggest to have individually trained ML forecasters for defined subclasses of data ? In other words, it is not clear to this reviewer - is there an advantage in ML trained on very large but heterogeneous dataset vs. a set of MLs trained for subsets of the data?

2)
p. 64 of the thesis "Generally, the accuracy of the prediction model ($MAPE$ = 37.28%, $wMAPE$ = 27.46%) in field tests is close to the hold-out set accuracy check ($MAPE$ = 37.78%, $wMAPE$ = 29.95%). The distribution of well types (vertical, horizontal) for field tests is close to that presented in the hold-out set."

What is the distinction between 'hold-out' and 'field test' wells?
From the description, it follows that 21 field test wells belong to the same field (and ML trained on the subset of the database is used) - is the correct? While hold-out wells are a wider set (which would include the 'field test' ones) from the full-database.
The fact that similar prediction error is observed for sub-set (field test) and full population of hold-out wells - would that be an indication that algorithm works equally well (or not) irrespective of how heterogeneous are the field conditions. And the prediction error is rather dominated by poor prediction/modeling of multi-lateral vertical wells (again irrespective of a field).

What improvements to the ML model could be suggested to improve handling of multilaterals? How would the performance of ML model would be like if multi-laterals were excluded from the database?

3)
On other, physics based forward modeling and possible future work.

How would physics based forward modeling using the same inputs compare to the ML-approach of the thesis? Given the knowledge base of the Skoltech group (Prof Osiptsov and co-workers) in forward physics-based HF computer simulation, it is somewhat surprising that this has not be pursued. This would also make for a logical 'future work' item!

Provisional Recommendation

☒ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report*

☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*