XLIII CILAMCE
21–25, November, 2022
FOZ DO IGUAÇU – PR – BRAZIL

A MULTI-FIDELITY REDUCED-ORDER MODEL APPLIED TO RESERVOIR ENGINEERING

Matheus S. Gonçalves[1], Yuri N. Saraiva[1], Mathieu Ducros[1]

[1]*Kognitus, Rio de Janeiro, Brazil*
matheus@kognitus.com.br, yuri@kognitus.com.br, mathieu@kognitus.com.br

Abstract. Although the rapid evolution of computational power, some applications (e.g., optimization, uncertainty propagation), combined with the increase of simulation complexity, may still require computational times that are unsuitable for practical purposes. Therefore, in the present work, a multi-fidelity surrogate strategy for expensive reservoir engineering numerical models is presented. The main idea of the approach is to combine information from few high-fidelity samples (of long computational duration) and accurate simulations with faster evaluations from a low-fidelity model with lower accuracy. This strategy provides enough information for creating a surrogate model, however with a lower computational cost than what would be obtained by evaluating just high-fidelity samples (for the same level of quality). Furthermore, conversely to the majority of multi-fidelity models, the present strategy includes a dimensionality reduction technique to deal with multi-dimensional outputs. The presented approach is employed in an example of a reservoir engineering problem for predicting both the net present value and the saturation of oil in the reservoir. The results obtained are suitable when evaluating both time and accuracy, indicating that the presented approach is promising for practical applications. Moreover, it also suggests that the same framework may be used in other computationally demanding applications.

Keywords: Multi-fidelity modeling, reservoir engineering, surrogate models

1 Introduction

Although the computational power available for simulations has presented a continuous increase in the last decades, many numerical models still present a highly expensive computational cost. Reservoir simulations, computational fluid dynamics, finite element analysis are some examples. In this regard, applications involving outer loops (e.g., optimization, uncertainty propagation and inference) may require a prohibitive execution time [1]. A technique frequently employed to overcome the computational burden necessary for outer loop analyses is replacing the time expensive numerical model by some approximation function which may properly predict the outputs from the expensive model at a negligible cost.To reach this goal, such methods uses samples from the expensive model in order to construct the approximation. These techniques are named as meta-models, surrogates, proxy models or emulators [2]. Although the successful application of such techniques for a wide range of numerical models, some specific problems may exhibit simulation times that prevent the calculation of a number of samples large enough to construct suitable surrogate models.

In order to reduce the number of necessary samples from the expensive solver, some current proxy model approaches apply multi-fidelity techniques. The main idea behind multi-fidelity approaches rely on the use of several low cost simulations together with a few expensive simulations to create suitable surrogate models. The fidelity is associated with the accuracy of the model, in which high-fidelity models provide the desired level of accuracy while being expensive to simulate, and low-fidelity ones are cheaper but not as accurate as the high-fidelity models. For the proper application of multi-fidelity approaches, it is necessary that the response from high- and low-fidelity models present significant correlation [3, 4]. Among the simplifications that a low-fidelity model may present to reduce the computational cost are the reduction of mesh size [2, 5–8], the use of simplified physical models [6, 9, 10], and the application of broader convergence criteria [3].

Multi-fidelity models that address single output problems are abundant in literature. When the objective is to estimate complete field solutions (e.g. the complete pressure field around an airfoil), however, the number of approaches is considerably lower [11]. A common strategy to address field outputs is to combine a regression model and a dimensionality reduction technique. In this regard, both steps are studied almost independently. First, it is searched for a reduced representation of the solution field, in which variations of non-intrusive Proper Orthog-

onal Decomposition (POD), also known as Principal Component Analysis (PCA), are usually employed. Then, a regression method is used for estimating the coefficients of this representation for unseen inputs. When employing this simplified representation, the number of coefficients that must be estimated by the regression strategy are substantially reduced. It occurs since, instead of creating a surrogate for each cell in the field, just the coefficients for the most important components are necessary. This approach turns the problem computationally tractable, however it also introduces errors in the process due to the simplified representation of the field of results.

It is important to remark that the POD decomposition of outputs from the training set is not a straightforward task in the multi-fidelity scenario. The reason is that low- and high-fidelity models can have distinct behaviors and the coefficients of independent POD bases for each dataset may not be directly comparable (e.g., cell outputs for models with different mesh sizes). After the model reduction step, it is necessary to fit a regression model which takes into account the distinction between high- and low-fidelity model coefficients. This step can be performed following several approaches. For instance, there are methods that work with multiplicative and/or additive corrections, whose goal is to adjust a low-fidelity response to the high-fidelity data employing a surrogate for the difference between them [12]. In this regard, virtually any machine learning model would be used. A common approach in the multi-fidelity literature, with many successful application examples, is co-kriging [11]. Co-kriging is the extension of kriging for multivariate functions [13]. The reader is referred to the works of Peherstorfer et al. [1] and Fernández-Godino et al. [12] for more information related to multi-fidelity models.

In the present work the methodology proposed by Perron [11] is used to construct multi-fidelity surrogates based on data coming from simulations with different mesh sizes. In this approach, manifold alignment is used to fuse data from distinct fidelities by projecting them into a common latent space. This approach avoids the necessity of matching dimensions of outputs from high- and low-fidelity models, which prevents imprecisions that may occur in the process (e.g., when interpolating results from distinct meshes). Furthermore, the co-kriging algorithm is employed as the regression strategy for predicting the coefficients of the components for new inputs. This algorithm is the natural option due to its constant application in the multi-fidelity literature. The results are evaluated for a synthetic example of reservoir model, in which both scalar and high-dimensional outputs are predicted.

2 Multi-fidelity approach

In this work, just two levels of accuracy are considered even though more levels can be accounted for. The applied high-fidelity model is a reservoir model with a mesh size considered suitable for practical purposes. The low-fidelity model, on the other hand, is an upscaled version of this model, where the resolution of the mesh is highly reduced to provide faster estimates. It is worth noting that it is not intended to construct an accurate low-fidelity model since high-fidelity data is also used for constructing the multi-fidelity surrogate and the goal of the low-fidelity model is to provide cheap function evaluations.

For a given vector of inputs $\boldsymbol{x}$, the objective of the multi-fidelity model is to estimate the response of the high-fidelity model based on the combination of both high- and low-fidelity information:

$$y_h(\boldsymbol{x}) = \rho(\boldsymbol{x}) y_l(\boldsymbol{x}) + \delta(\boldsymbol{x}), \tag{1}$$

where y_h and y_l represent the high- and low-fidelity scalar model responses, $\rho(\boldsymbol{x})$ is a multiplicative correction and $\delta(\boldsymbol{x})$ is a discrepancy function. The main idea behind eq. (1) is to correct a low-fidelity surrogate (y_l) to match the desired high-fidelity response (y_h) by means of the use of both a multiplicative correction and a discrepancy function.

In the present work a co-kriging approach is employed for learning the mapping from the input vector $\boldsymbol{x}$ to the high-fidelity model response. The training phase of the method uses samples for both high- and low-fidelity model responses, in which the number of low-fidelity samples is significantly higher. For the method used in the present work, it is necessary that a low-fidelity sample be drawn for the same input parameters of every high-fidelity model sample. Or, in other words, the training set for the high-fidelity model is a subset of the low-fidelity model. A nested space filling design proposed by Le Gratiet [13] is employed to adapt low-fidelity experimental designs to high-fidelity ones, both of them initially drawn from Latin Hypercube Sampling (LHS).

The approach previously described is able to predict scalar responses. However, in the present study some outputs are high-dimensional. For instance, the saturation of oil is estimated for each cell of the high-fidelity mesh, which easily reaches thousands of cells. For modeling high-dimensional outputs, the use of Proper Orthogonal Decomposition (POD) to reduce the dimensionality of the model is a common strategy. In this approach, the problem of predicting the output for every cell of the model is replaced by the estimation of the coefficients of

the principal components of the dataset. In this regard, the high-dimensional output is decomposed in orthogonal components and just the most significant ones are included in the model:

$$\boldsymbol{y}_h(\boldsymbol{x}) \approx \sum_{k=1}^{K} \alpha_k(\boldsymbol{x})\boldsymbol{\phi}_k, \tag{2}$$

in which $\boldsymbol{y}_h(\boldsymbol{x})$ is the high-dimensional model output, $\boldsymbol{\phi}_k$ is the vector representing the k POD basis and α_k is the coefficient related to this basis function. Both $\boldsymbol{\phi}_k$ and α_k are calculated from the POD algorithm. To enable this calculation, the high-dimensional output of every training sample is transformed into a vector and plugged as a column of the same matrix, called snapshot matrix, and the POD algorithm is applied to this matrix. Moreover, K is the number of components used in the approximation. To simplify the process, usually just the first components are selected since they are able to accurately represent the system response. In this work the selection corresponds to the components representing 99.9% of the variance of the training set. Notice that this procedure is an approximation, since a truncation of the full POD is performed.

When including data from a second level of fidelity, an extra procedure should be used for allowing the transference of information from the low- to the high-fidelity models. This is due to the difference on length of the $\boldsymbol{\phi}_k$ vector extracted from different meshes. In order to avoid the use of interpolation steps, a manifold alignment strategy proposed by Perron [11] in the context of multi-fidelity analysis is employed in the present work. The idea of the manifold alignment is to enable the transference of information from low- to high-fidelity models by uncovering a shared latent space. The method uses Procrustes analysis to align the intrinsic latent spaces, which are obtained from independent POD for each level of fidelity. Then, to model the high-dimensional output, the coefficients of each high-fidelity principal component, after the manifold alignment, are modeled as in eq. (1). Therefore, for every k component in the model, a co-kriging is trained. The response of the model is based on the known high-fidelity POD basis, extracted from the training data, and the prediction of their respective coefficients:

$$\boldsymbol{y}_h(\boldsymbol{x}) = \sum_{k=1}^{K} \hat{\alpha}_k^h(\boldsymbol{x})\boldsymbol{\phi}_k^h, \tag{3}$$

in which $\hat{\alpha}_k^h(\boldsymbol{x})$ is the prediction of the k high-fidelity POD coefficient and $\boldsymbol{\phi}_k^h$ is the k POD basis for the high-fidelity model. The multi-fidelity approach can be summarized as :

- Training phase:
  1. Use the POD algorithm for both high- and low-fidelity training samples and select the most influential components for both datasets
  2. Perform the manifold alignment to match the components of both datasets in a latent space
  3. Train a co-kriging algorithm for each component, considering both datasets in the latent space
- Prediction phase:
  1. Calculate each coefficient of the high-fidelity POD basis using the co-kriging surrogate previously trained for each new input vector $\boldsymbol{x}^*$
  2. Perform the projection of the solution to the original space employing the high-fidelity principal components and the predicted coefficients, recovering the original dimensionality

3 Case study

The reservoir model for the Brugge case is used in all following analyses. This benchmark project was developed as part of the SPE Applied Technology Workshop (ATW) held in Brugge in June 2008.The information used in this section comprises one geological realization of permeability, porosity, and net-to-gross (NTG) thickness ratio for an upscaled grid (60,000 grid cells, with 139 x 48 x 9 elements in each direction) from the 104 stochastic geological realization provided with the case. The reservoir contains 30 wells, 20 producers and 10 injectors. For more information related to this benchmark case, the reader is referred to Peters et al. [14]. The example analyzed uses the geological realization n°96 and 9 uncertain parameters as defined in Table 1. The NTG, porosity multiplier and all permeability multipliers are constant throughout the reservoir. The OpenSim Technology reservoir simulator, considering a Black-Oil model that solves the generalized Darcy's equations, was used [15].

The goal of the surrogate models is to predict the total oil production (FOPT) and the saturation of oil in all cells (SOIL) for the final production step. Notice that, although the whole production curve is predicted, the FOPT

Table 1. Parameters for the Brugge model

Parameter	Description	Lower bound	Upper bound	Distribution
NTG	Net-to-gross ratio	0.5	1	Uniform
krw1	End point relative permeability for water in the first range of porosities of the Corey model	0.55	0.6	Uniform
krw5	End point relative permeability for water in the fifth range of porosities of the Corey model	0.55	0.6	Uniform
kro1	End point relative permeability for oil in the first range of porosities of the Corey model	0.35	0.4	Uniform
kro5	End point relative permeability for oil in the fifth range of porosities of the Corey model	0.35	0.4	Uniform
por	Porosity multiplier	0.7	1.3	Uniform
permx	Multiplier of permeability in x direction	0.5	2	Uniform
permy	Multiplier of permeability in y direction	0.5	2	Uniform
permz	Multiplier of permeability in z direction	0.5	2	Uniform

is analyzed as a scalar quantity. Conversely, SOIL is a high-dimensional output (about 60,000 values, one value per active cell). A total of 50 independent simulations are performed for testing the methods, using two distinct metrics: explained variance (Q2) and root mean squared error (RMSE). This sample is constructed using Latin Hypercube Sampling (LHS).

3.1 High- and low-fidelity models

The high-fidelity model is described by the properties previously cited, whose mesh has 139 x 48 x 9 cells in each direction. The low-fidelity model is an upscaled version of the high-fidelity one. The mesh size for the low-fidelity model, also called hereafter as coarse mesh, is nearly 36 times lower (35 x 16 x 3). The geological parameters are upscaled using a simple arithmetical mean. The coarse mesh was generated based on the PyGRDECL_Upsc_3D python library [16, 17]. It is worth noting that it is not intended to construct a highly precise low-fidelity model. Conversely, it is expected that the low-fidelity model is not able to precisely predict all the field properties. The usefulness of the low-fidelity model is to be able to explore the domain defined by the uncertain parameters, providing information for the shape of the surrogate in regions where there are not high-fidelity samples. Even if some bias is present in results, the high-fidelity samples can be used to correct it and improve predictions. Figure 1 presents the comparison of the permeability in x direction for high- and low-fidelity models. It is noticed that the coarse mesh contains a more homogeneous pattern of permeabilities due to the reduction in the mesh resolution. In addition, the process of upscaling the mesh was not able to precisely model the sealing fault present in the fine model. As already discussed, it is not necessarily a problem since high-fidelity samples are used for correcting the predictions.

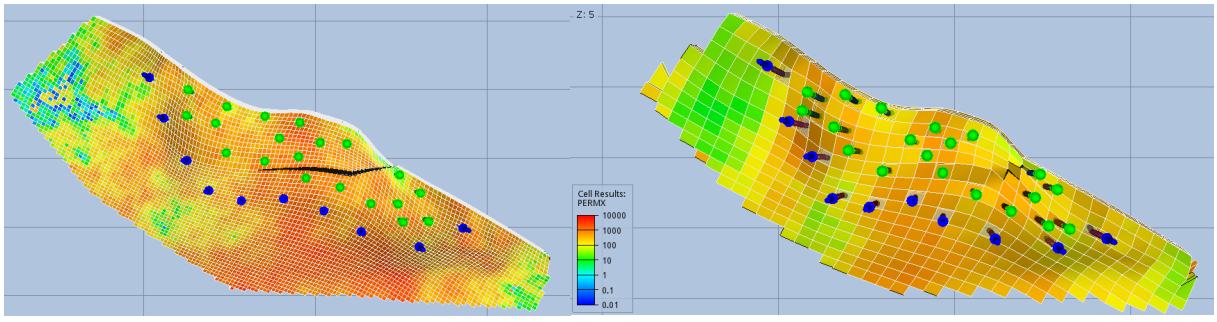

Figure 1. Comparison of the permeability in x for the top layer of the fine (up) and coarse (bottom) meshes

4 Results

In order to assess the suitability of the multi-fidelity model, some baseline performance should be available. The natural candidate is the single-fidelity surrogate using just high-fidelity model evaluations for training, since it is the standard approach. This approach is called hereafter a high-fidelity surrogate. As many machine learning strategies are available to construct this surrogate, it is opted to train three different machine learning models: a Random forest, a Gaussian process with linear regression term and a simple least squares considering polynomial degree of 1 and without interactions. Such methods differ considerably in their representation capacity, which allows a more complete analysis. For the multi-fidelity approach, a co-kriging with linear regression term is used, named here as MFK since it is based on the MFK function present in the python library SMT [18]. The implementation of this function relies on the work of Le Gratiet [13]. The implementation of the scikit-learn [19] (Random forest and least squares) and SMT [18] (Gaussian process and co-kriging) python libraries are used. Both Gaussian process and co-kriging uses the same configuration: Matérn 5/2 correlation function and TNC solver for the optimization of the hyperparameters. The remaining parameters for all methods apply the default configuration. Then, the results for 3 high-fidelity surrogates and 1 multi-fidelity surrogate are evaluated in what follows. The quality metrics are calculated for different numbers of samples for all analyzed surrogates. However, it is difficult to directly compare results from high- and multi-fidelity surrogates since the former employs just high-fidelity (HF) simulations whereas the later includes also low-fidelity (LF) simulations. Then, to allow a fair comparison, the computational cost associated to each surrogate must be converted to a same time scale, expressed in number of equivalent computational time of high-fidelity simulations. Therefore, the computational time of the LF simulations need to be converted into equivalent number of high-fidelity simulations. In this work, it is assumed that the simulation time depends linearly on the number of cells in the model. Thus, running a HF simulation has a same theoretical computational cost as running 36 LF simulations (e.g., a multi-fidelity model with 9 HF samples and 36 LF samples result in 10 equivalent HF samples).

Figure 2 shows the FOPT prediction provided by each surrogate model. The predictions from multi-fidelity and high-fidelity surrogates are similar for the initial range of values but the best high-fidelity surrogate outperforms the multi-fidelity one for the larger sample size evaluated. However, the results provided by both models achieved a high Q2 and a small RMSE (around 2×10^6 compared to the FOPT range from 0.8 to 1.2×10^8). Indeed, even the Least Squares results may be argued as suitable for this case. Then, the high-fidelity surrogate seems a better choice in this context. However, results from the multi-fidelity surrogate may also be argued as suitable for practical purposes, mainly when using small sample sizes. Another relevant aspect to discuss is the proportion of computational budget allocated to HF and LF models. In this work, it is opted to spend most of the time running HF models since LF models are quite cheap and just a few HF model evaluations correspond to dozens of LF model evaluations. Then, it is possible to cover a large range of the domain, however keeping almost the same number of HF samples.

Figure 3 presents the results for SOIL. The observed trend is different to that of the first case. The multi-fidelity surrogate was able to present better results for all sample sizes. The cost, measured in equivalent high-fidelity samples, to achieve the RMSE results of the multi-fidelity approach was roughly 40% lower than the cost necessary by any high-fidelity surrogate to reach the same accuracy. When analyzing the obtained RMSE for each cost, the multi-fidelity results were between 15% and 30% lower than any high-fidelity surrogate. Then, for SOIL predictions, improvements were similar across different sample sizes and the multi-fidelity surrogate was able to outperform the high-fidelity surrogates.

It is worth pointing out that the configuration applied in the present work for MFK and Gaussian Process have a higher computational cost, in the order of a few minutes to run the training for each POD coefficient. It has a significant cost when compared to the time to run the present test problem. However, the approach is derived for highly expensive reservoir models, for which each training simulation may take several hours (even considering that more powerful computers are available). Then, when compared to the practical scenario, this higher training cost does not seem significant.

5 Conclusions

The results of the multi-fidelity surrogate are promising. For FOPT predictions, the results are comparable to the high-fidelity ones, however the best high-fidelity surrogate achieved better results for the higher sample size. On the other hand, both RMSE and Q2 of the multi-fidelity approach are better than those obtained by any high-fidelity surrogate for SOIL predictions in all equivalent sample sizes. The main advantages observed for the multi-fidelity approach is for low computational budget, since even the models trained with the minimum available sample sizes were often able to achieve performance that would be judged acceptable for practical purposes.

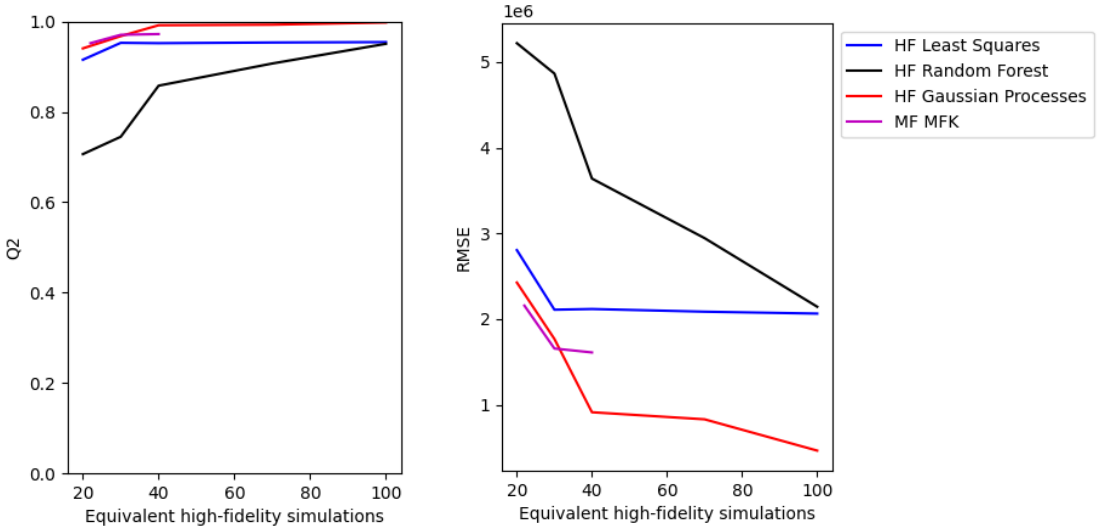

Figure 2. Explained variance (left) and RMSE (right) for FOPT predictions computed with the 50 testing simulations for all evaluated surrogates as a function of the number of equivalent high-fidelity samples. Black, blue and red curves show results of surrogates using only HF simulations (20, 30, 40, 70 and 100 simulations). The multi-fidelity (MF) surrogates were obtained using 20 HF + 72 LF, 27 HF + 108 LF and 36 HF + 144 LF simulations, respectively.

Acknowledgements. We thank the Rio de Janeiro Research Foundation for the research grants conceded to Yuri Nunes Saraiva and Matheus Silva Gonçalves. Grants "Inserção de Pesquisadores em Empresas", Processo SEI-260003/001639/2021, Rio de Janeiro Research Foundation (FAPERJ) and the OpenSim Technology for providing the licenses of the software used for the numerical simulations of reservoir models.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] B. Peherstorfer, K. Willcox, and M. Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *Siam Review*, vol. 60, n. 3, pp. 550–591, 2018.
[2] A. Thenon, V. Gervais, and M. L. Ravalec. Multi-fidelity meta-modeling for reservoir engineering-application to history matching. *Computational Geosciences*, vol. 20, n. 6, pp. 1231–1250, 2016.
[3] M. J. Mifsud, D. G. MacManus, and S. T. Shaw. A variable-fidelity aerodynamic model using proper orthogonal decomposition. *International Journal for Numerical Methods in Fluids*, vol. 82, n. 10, pp. 646–663, 2016.
[4] D. J. J. Toal. On the potential of a multi-fidelity g-pod based approach for optimization and uncertainty quantification. In *Proceedings of the ASME Turbo Expo 2014: Turbine Technical Conference and Exposition*, volume 2B: Turbomachinery, pp. V02BT45A002. American Society of Mechanical Engineers, 2014.
[5] S. Bunnell, S. Gorrell, and J. Salmon. Multi-fidelity surrogates from shared principal components. *Structural and Multidisciplinary Optimization*, vol. 63, n. 5, pp. 2177–2190, 2021.
[6] M. Kast, M. Guo, and J. S. Hesthaven. A non-intrusive multifidelity method for the reduced order modeling of nonlinear problems. *Computer Methods in Applied Mechanics and Engineering*, vol. 364, pp. 112947, 2020.
[7] T. Benamara, P. Breitkopf, I. Lepot, and C. Sainvitu. Multi-fidelity extension to non-intrusive proper orthogonal decomposition based surrogates. In *Proceedings of VII European Congress on Computational Methods in Applied Sciences and Engineering*, pp. 4129–4145. ECCOMAS, 2016.
[8] T. Benamara, P. Breitkopf, I. Lepot, C. Sainvitu, and P. Villon. Multi-fidelity POD surrogate-assisted optimization: Concept and aero-design study. *Structural and Multidisciplinary Optimization*, vol. 56, n. 6, pp. 1387–1412, 2017.

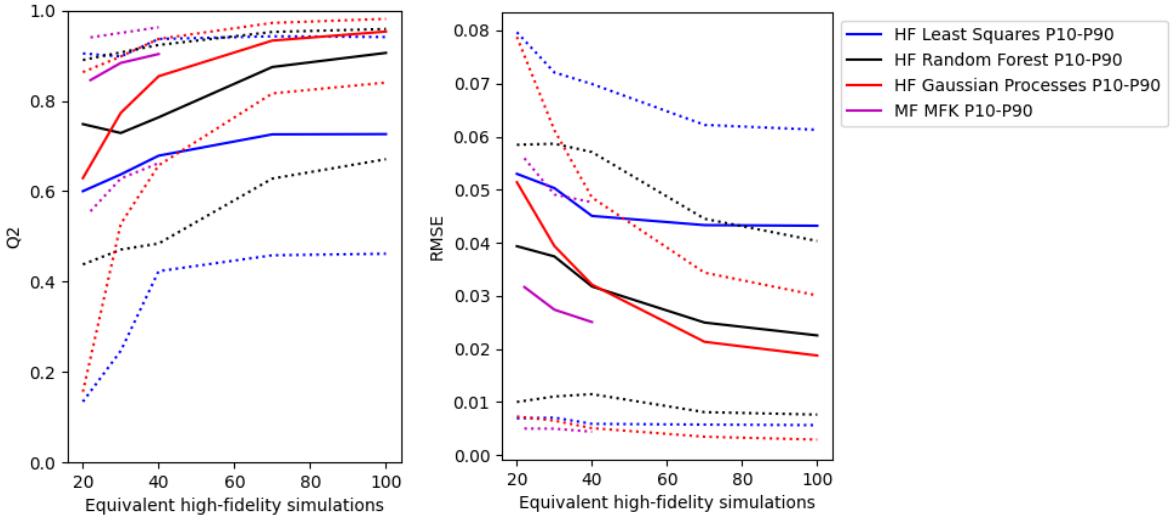

Figure 3. Explained variance (left) and RMSE (right) for SOIL predictions computed with the 50 testing simulations for all evaluated surrogates as a function of the number of equivalent high-fidelity samples. Solid lines refer to the median whereas dotted lines indicate the P10-P90 interval. Black, blue and red curves show results of surrogates using only HF simulations (20, 30, 40, 70 and 100 simulations). The multi-fidelity surrogate models were obtained using 20 HF + 72 LF, 27 HF + 108 LF and 36 HF + 144 LF simulations, respectively.

[9] A. Bertram, C. Othmer, and R. Zimmermann. Towards real-time vehicle aerodynamic design via multi-fidelity data-driven reduced order modeling. In *2018 AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, 2018.

[10] X. Wang, J. Kou, and W. Zhang. Multi-fidelity surrogate reduced-order modeling of steady flow estimation. *International Journal for Numerical Methods in Fluids*, vol. 92, n. 12, pp. 1826–1844, 2020.

[11] C. Perron. *Multi-fidelity reduced-order modeling applied to fields with inconsistent representations*. PhD thesis, Georgia Institute of Technology, 2020.

[12] M. G. Fernández-Godino, C. Park, N. H. Kim, and R. T. Haftka. Review of multi-fidelity models. *arXiv preprint arXiv:1609.07196*, vol. , 2016.

[13] L. Le Gratiet. *Multi-fidelity Gaussian process regression for computer experiments*. PhD thesis, Université Paris-Diderot-Paris VII, 2013.

[14] E. Peters, R. J. Arts, G. K. Brouwer, C. R. Geel, S. Cullick, R. J. Lorentzen, Y. Chen, K. N. B. Dunlop, F. C. Vossepoel, R. Xu, P. Sarma, A. H. Alhutali, and A. C. Reynolds. Results of the brugge benchmark study for flooding optimization and history matching. *SPE Reservoir Evaluation & Engineering*, vol. 13, n. 03, pp. 391–405, 2010.

[15] A. A. Rodriguez, J. E. Monteagudo, N. E. Nieto, and A. Primera. Coupled flow-geomechanics simulation of fracturing-to-production in unconventional reservoirs. In *SPE Latin American and Caribbean Petroleum Engineering Conference*. OnePetro, 2020.

[16] M. Zakari. PyUpsc_3D a python-based corner point grid upscaling library. `https://github.com/mzakari31/PyGRDECL`, 2021.

[17] B. Wang. Pygrdecl a python-based grdecl visualization library. `https://github.com/BinWang0213/PyGRDECL`, 2018.

[18] M. A. Bouhlel, J. T. Hwang, N. Bartoli, R. Lafage, J. Morlier, and J. R. R. A. Martins. A python surrogate modeling framework with derivatives. *Advances in Engineering Software*, vol. , pp. 102662, 2019.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.