Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 55-7 (2022) 520–525

When is gray-box modeling advantageous for virtual flow metering?

M. Hotvedt * **B. Grimstad** *,** **D. Ljungquist** *** **L. Imsland** *

* *Engineering Cybernetics Department, NTNU, Trondheim, Norway (e-mail: {mathilde.hotvedt, lars.imsland}@ntnu.no)*

** *Solution Seeker (e-mail: bjarne.grimstad@solutionseeker.no)*

*** *TechnipFMC (e-mail: Dag.Ljunquist@technipfmc.no)*

Abstract: Integration of physics and machine learning in virtual flow metering applications is known as gray-box modeling. The combination is believed to enhance multiphase flow rate predictions. However, the superiority of gray-box models is yet to be demonstrated in the literature. This article examines scenarios where a gray-box model is expected to outperform physics-based and data-driven models. The experiments are conducted with synthetic data where properties of the underlying data generating process are controlled. The results show that a gray-box model yields increased prediction accuracy over a physics-based model in the presence of process-model mismatch, and improvements over a data-driven model when the amount of available data is small. On the other hand, gray-box and data-driven models are similarly influenced by noisy measurements. Lastly, the results indicate that a gray-box approach may be advantageous in nonstationary process conditions. Unfortunately, model selection prior to training is challenging, and overhead on gray-box model development and testing is unavoidable.

Keywords: Gray-box, hybrid model, virtual flow metering, neural networks

1. INTRODUCTION

Gray-box modeling is a methodology that integrates physics-based modeling with machine learning techniques in process model development (Willard et al., 2020). The gray-box models are placed on a gray-scale dependent on the degree of integration, ranging from physics-based to data-driven models. A common perception is that physics-based models require little data in development and are more robust to noisy measurement than data-driven models. This perception arguably stems from the high extrapolation capabilities demonstrated by many physics-based models (Oerter, 2006). Nevertheless, complex physical phenomena can be challenging to model in detail using first principles, and simplifications are generally necessary for suitability in real-time control and optimization applications (Roscher et al., 2020). Simplifications reduce the model capacity and thereby the ability to capture complex physical behavior. Therefore, physics-based models often have a bias, or process-model mismatch (Hastie et al., 2009).

In contrast, many data-driven models have a large capacity, typically reducing model bias. Furthermore, some data-driven models are computationally cheap to evaluate and are therefore suitable for real-time applications. Moreover, they commonly have lower development and maintenance costs compared to physics-based models (Solle et al., 2016). On the other side, due to the inherent bias-variance trade-off (Hastie et al., 2009), a large capacity often results in high variance. High variance causes data-driven models to struggle with extrapolation to future process conditions and to yield low performance in the small data regime (Roscher et al., 2020). Gray-box modeling is expected to leverage the complementary and advantageous properties of physics and data to minimize both bias and variance. In other words, create a model that achieves high performance in the presence of process-model mismatch, little or noisy data, which extrapolates well to previously unseen process conditions and is computationally efficient. Gray-box modeling is similar to introducing strong priors in a data-driven model. In image classification using convolutions neural networks, strong priors in terms of parameter sharing resulted in state-of-the-art performance (Hastie et al., 2009).

One application where accurate process models are of high importance is in virtual flow meters (VFMs): a soft-sensor able to predict the multiphase flow rate in real-time at convenient locations in a petroleum asset (Toskey, 2012). The standard practice in the industry today is physics-based models, and several commercial simulators exist (Amin, 2015). In later years, data-driven VFM models have demonstrated high performance (AL-Qutami et al., 2017a,b,c, 2018; Bikmukhametov and Jäschke, 2019; Grimstad et al., 2021). On the other hand, due to the inherently complex multiphase flow rate characteristics and that the available data typically resides in the small data regime (Grimstad et al., 2021), gray-box VFMs have gained increasing attention, see (Bikmukhametov and Jäschke, 2020; Hotvedt et al., 2020, 2021, 2022) and references therein. However, superior performance over physics-based or data-driven models has yet to be demonstrated. This article contributes in this direction by investigating four scenarios where a gray-box approach is believed to excel over non-gray-box alternatives. These are formulated as four hypotheses:

10.1016/j.ifacol.2022.07.496

Hypothesis 1 Under mismatch between a physics-based VFM and the process, a gray-box VFM developed from the physics-based VFM achieves higher performance.

Hypothesis 2 With little available data, a gray-box VFM obtains higher performance than a data-driven VFM.

Hypothesis 3 Increasing the noise level on the data, a gray-box VFM is less influenced than a data-driven VFM.

Hypothesis 4 In nonstationary conditions, a gray-box VFM yields higher performance than a data-driven VFM.

In Hypothesis 1, the increased capacity of the gray-box compared to the physics-based model is believed to be significant. In Hypothesis 1-3, the decreased capacity of the gray-box compared to the data-driven model is believed to be decisive. In real life, available process data can have several uncontrolled characteristics, for instance, faulty sensor measurements. Such characteristics make it challenging to examine and conclude on the hypotheses as it is difficult to deduce whether a poor model performance results from the modeling technique or the available data. This has been experienced in previous work with gray-box VFMs (Hotvedt et al., 2022). Therefore, in this work, synthetic data designed to explore the hypotheses are generated by a simulator of a petroleum production choke. In several idealized experiments, the properties of gray-box production choke models are compared to physics-based and data-driven models. Hopefully, the results obtained can act as a guide to when gray-box modeling is likely to be advantageous, also in practical applications.

2. THE SIMULATOR

The simulator is a physics-based petroleum production choke valve model. A typical production choke along with available measurements is illustrated in Fig. 1. The

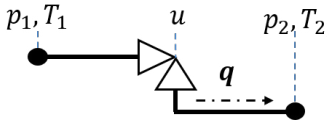

Fig. 1. Illustration of the production choke valve and typically available measurements.

multiphase mass flow rate (a mixture of oil, gas, and water) $\dot{m}$ through the choke restriction is calculated using an advanced version of the Sachdeva model (Sachdeva et al., 1986), where slip effects, allowing the gas and liquid phases to move with unequal velocity, are included in the model. The slip model is taken from (Alsafran and Kelkar, 2009). The model requires measurements of the pressure upstream (p_1) and downstream (p_2) of the choke valve, the upstream temperature (T_1), the choke opening (u), and the mass fractions of the phasic fluids $\boldsymbol{\eta} = (\eta_{\text{oil}}, \eta_{\text{gas}}, \eta_{\text{wat}})$. The mass fractions are assumed to sum to one. The volumetric multiphase flow rate $q = q_{\text{oil}} + q_{\text{gas}} + q_{\text{wat}}$ can be obtained from the $\dot{m}$ using the $\boldsymbol{\eta}$ and fluid densities ρ at standard conditions (SC) (ISO, 1996):

$$q_i = \frac{\eta_i \dot{m}}{\rho_{i,SC}}, \quad i \in \{\text{oil}, \text{gas}, \text{wat}\}. \tag{1}$$

In the simulator, an area function relates the choke opening to the effective flow area through the choke $A(u)$. This function will mimic an equal percentage valve, where an equal increment in u results in an equal percentage changed area. The simulator, or process, is referred to as $\mathcal{P}$ and defined by the notation:

$$y = f(\boldsymbol{x}; \boldsymbol{\phi}) + \varepsilon \in \mathbb{R}, \tag{2}$$

where the model output is $y = q$, f is the first principle equations, the input measurements are $\boldsymbol{x} = (p_1, p_2, T_1, u, \eta_{\text{oil}}, \eta_{\text{wat}}) \in \mathbb{R}^6$, and $\boldsymbol{\phi}$ are constant model parameters. Noise is added to q by sampling ε from a probability distribution, for instance, a Gaussian distribution.

3. DATASET GENERATION

Process $\mathcal{P}$ in Section 2 is used to generate three different datasets $\mathcal{D}_k = \{(\boldsymbol{x}_t, y_t)\}_{t=1}^{N_k}$, $k = \{1, 2, 3\}$. The index t reflects time. The datasets are designed to investigate the hypotheses in Section 1. The sequence of observations in each dataset is sampled from the joint probability distribution of $\mathcal{P}$: $p_t(\boldsymbol{x}, y) = p_t(y \,|\, \boldsymbol{x}) p_t(\boldsymbol{x})$, where $p_t(\boldsymbol{x})$ is the marginal distribution of the inputs and the output y_t follow the conditional distribution $p_t(y \,|\, \boldsymbol{x})$ expressed with (2). Notice, $\mathcal{P}$ is allowed to be nonstationary resulting in $p_{t_1}(\boldsymbol{x}, y) \neq p_{t_2}(\boldsymbol{x}, y)$ for $t_1 \neq t_2$.

Dataset $\mathcal{D}_1$ is generated as a best-case scenario to fairly examine Hypothesis 1-3 in Section 1. Firstly, the process is assumed stationary: $p_{t_1}(\boldsymbol{x}, y) = p_{t_2}(\boldsymbol{x}, y) \forall t$. Secondly, the $\boldsymbol{x}$ are independently drawn. This is idealized as measurements in real data are often strongly correlated (Hotvedt et al., 2022). Thirdly, a large range of common process conditions through the lifetime of a petroleum well is covered by sampling the inputs from:

$$\begin{aligned} p_1 &\sim \mathcal{U}(30, 70)\ bar, \\ p_2 &\sim \mathcal{N}(22, 0.5)\ bar, \\ T_1 &\sim \mathcal{N}(50, 2)\ ^\circ C \\ u &\sim \mathcal{U}(0, 100)\ \%, \\ \eta_{\text{oil}} &\sim \mathcal{U}(0, 80)\ \%, \\ \eta_{\text{wat}} &\sim \mathcal{U}(0, 20)\ \%. \end{aligned} \tag{3}$$

for any t. The p_1, u, η_{oil}, and η_{wat} are sampled from wide uniform distributions as they commonly vary much, whereas p_2 and T_1 vary little, which is mimicked by drawing from narrow normal distributions. To ensure a sufficient dataset size $N_1 = 10000$ observations are sampled. Lastly, only normally distributed noise $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is considered. The included noise levels are $\sigma_\varepsilon \in \{1, 2, 3, 4, 5, 10\}$, yielding a coefficient of variation of $\sigma_\varepsilon / \mu \in \{0.02, 0.05, 0.07, 0.1, 0.12, 0.24\}$, where μ is the mean of the noise-free flow rate measurements. Normally distributed noise is an idealized case as measurement sensors may comprise different noise types. However, it is interesting to investigate how the models are influenced by increasing level of idealized noise before introducing noise of higher complexity. The dataset is randomly separated into a training and a test dataset with $N_{1,\text{test}} = 2000$. From the training dataset, 20% are randomly extracted as a validation dataset.

The $\mathcal{D}_2$ and $\mathcal{D}_3$ mimics two typical real case scenarios where the process is nonstationary. In this study, only virtual drift is simulated, meaning that nonstationarity is caused by the marginal distribution $p_t(\boldsymbol{x})$ shifting in time while the conditional distribution $p_t(y \mid \boldsymbol{x})$ stays

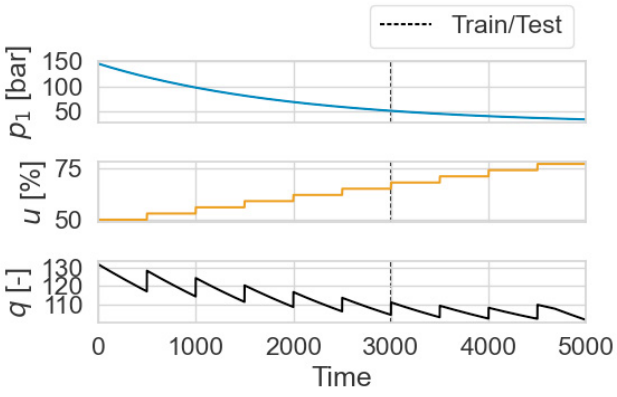

Fig. 2. Illustration of the dataset mimicking typical behavior when the reservoir is depleted with time.

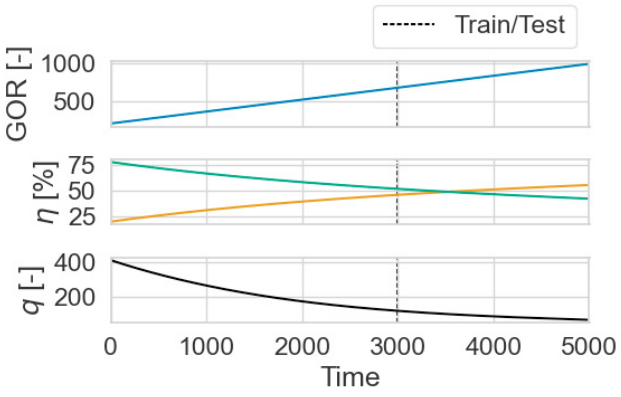

Fig. 3. Illustration of the dataset mimicking typical behavior when the gas-to-oil ratio increases. The mass fractions of oil and gas are the green and orange curve, respectively.

constant (Ditzler et al., 2015). Virtual drift is commonly seen for a petroleum asset. For instance, in time with the reservoir being depleted, the pressure in the reservoir and the upstream part of the choke decreases. If the petroleum asset is producing on plateau, process engineers increase the choke opening to maintain high production rates (Jahn et al., 2008). Real drift, which is the opposite of virtual drift, is typically a consequence of substantial mechanical wear of equipment with time. It is believed that real drift is less prominent than virtual drift in a petroleum asset and is the reason why real drift is not simulated in this study. In both datasets, $N_2 = N_3 = 5000$ noise-free observations are sampled. The datasets are split into training and test according to time with $N_{2,\text{test}} = N_{3,\text{test}} = 2000$. Hence, the models will be used to predict future process responses. The validation dataset consists of the 600 latter training observations ordered by time. Dataset $\mathcal{D}_2$ mimics the depleting reservoir as described above. This scenario is illustrated in Figure 2. The p_1 is decreased in time using an exponential function, whereas the choke opening is increased in steps of 2.5%. The remaining variables are kept constant for any t: $p_2 = 22$ bar, $T_1 = 50°C$, $\eta_{\text{oil}} = 85\%$, and $\eta_{\text{wat}} = 2\%$. Dataset $\mathcal{D}_3$ mimics a scenario where the gas-to-oil ratio (GOR) increases with time. This phenomenon typically occurs when the reservoir pressure drops below the bubble point pressure such that the gas dissolved in the oil starts to escape (Jahn et al., 2008). Fig. 3 illustrates the resulting flow rate q and the mass fractions of oil η_{oil} (green) and gas η_{gas} (orange) when the GOR is linearly increased from 200 to 1000. The p_1 is the same as for $\mathcal{D}_2$ illustrated in Fig. 2. The remaining variables are kept constant for any t: $p_2 = 22$ bar, $T_1 = 50°C$, $u = 100\%$, and $\eta_{\text{wat}} = 2\%$.

4. MODELS

Five production choke models have been developed: two physics-based, one data-driven, and two gray-box models. The models will be described briefly below. More details can be found in Hotvedt et al. (2022). The first physics-based model is the Sachdeva model, referred to as M, and defined by the short notation

$$\hat{y}_{\text{M}} = f_{\text{M}}(\boldsymbol{x}; \boldsymbol{\phi}_{\text{M}}) \in \mathbb{R}, \tag{4}$$

The true area function is kept unknown, and a linear relationship is utilized instead. Among the $\boldsymbol{\phi}_{\text{M}}$ is the discharge coefficient, which is a multiplicative calibration factor used to change the magnitude of the area function. In industrial VFMs, additional calibration factors exist to change the shape of the function. Here, these are excluded to restrict the capacity of M, enforcing a significant mismatch between $\mathcal{P}$ and M.

The second physics-based model is the advanced Sachdeva model used for $\mathcal{P}$, described in Section 2, referred to as M*. That is, the physical equations of the model are equal to the simulator, and the true area function is known. However, the true values of $\boldsymbol{\phi}$ in $\mathcal{P}$ are kept unknown from M* and $\boldsymbol{\phi}_{\text{M}^\star}$ must be estimated from data. M* is defined by

$$\hat{y}_{\text{M}^\star} = f(\boldsymbol{x}; \boldsymbol{\phi}_{\text{M}^\star}) \in \mathbb{R}. \tag{5}$$

Hence, any process-model mismatch will be a consequence of parameter deviation away from the true values and not structural mismatches as for the M.

The data-driven model is a fully connected, feed-forward neural network and is selected due to its large capacity. The model D is defined by

$$\hat{y}_{\text{D}} = f_{\text{D}}(\boldsymbol{x}; \boldsymbol{\phi}_{\text{D}}) \in \mathbb{R}, \tag{6}$$

where $\boldsymbol{\phi}_{\text{D}} = \{(\boldsymbol{W}_1, b_1), \dots (\boldsymbol{W}_L, b_L)\}$ are the weights and biases in the neural network on each layer $l = 1, ..., L$. The rectified linear unit is used as activation function.

The two different gray-box models are based on the M. The first is an error model where a data-driven model attempts to capture additive mismatches between $\mathcal{P}$ and M. This model is referred to as H-E:

$$\begin{aligned} \hat{y}_{\text{H-E}} &= f_{\text{H-E}}(\boldsymbol{x}; \boldsymbol{\phi}_{\text{H-E}}) \\ &= f_{\text{M}}(\boldsymbol{x}; \boldsymbol{\phi}_{\text{M}}) + f_{\text{D}}(\boldsymbol{x}; \boldsymbol{\phi}_{\text{D}}) \in \mathbb{R}. \end{aligned} \tag{7}$$

The second hybrid model addresses the unknown area function of $\mathcal{P}$ by multiplying the initial linear function of the M with a neural network: $A = A_{\text{M}} \times A_{\text{D}}$. Hence, both the magnitude and shape of the area function may be adjusted. This model is referred to as H-A:

$$\begin{aligned} \hat{y}_{\text{H-A}} &= f_{\text{H-A}}(\boldsymbol{x}; \boldsymbol{\phi}_{\text{H-A}}) = f_{\text{M}}(\boldsymbol{x}, A_{\text{D}}; \boldsymbol{\phi}_{\text{M}}) \in \mathbb{R} \\ A_{\text{D}} &= f_{\text{D}}(\boldsymbol{x}; \boldsymbol{\phi}_{\text{D}}) \in \mathbb{R}. \end{aligned} \tag{8}$$

As the neural network in H-A is multiplied with a small value (A_{M}), the capacity of the H-A is likely smaller than the capacity of H-E. This can be argued by acknowledging that large outputs from the network in H-A will be less influential on the flow rate predictions than a large output from the network in H-E.

For all models $i \in \{\mathrm{M}^\star, \mathrm{M}, \mathrm{H\text{-}A}, \mathrm{H\text{-}E}, \mathrm{D}\}$, the parameters are estimated using maximum a posteriori (MAP) estimation:

$$\begin{aligned}\hat{\boldsymbol{\phi}}_i &= \arg\max_{\boldsymbol{\phi}} p(\boldsymbol{\phi}_i \mid \mathcal{D}_k) \\ &= \arg\min_{\boldsymbol{\phi}} \Big[\sum_{t=1}^{N_k} \frac{1}{\sigma_\varepsilon^2} \left(y_t - \hat{y}_{i,t}\right)^2 \\ &+ \sum_{j=1}^{m} \frac{1}{\sigma_{i,j}^2} \left(\phi_{i,j} - \mu_{i,j} i\right)^2 \Big].\end{aligned} \tag{9}$$

where m is the number of parameters. The priors on the parameters are assumed normal $\phi_{i,j} \sim \mathcal{N}(\mu_{i,j}, \sigma_{i,j}^2)$. The optimization problem is solved using stochastic, iterative, gradient-based optimization with the optimizer Adam (Kingma and Ba, 2015) and early stopping. Details of the training algorithm are given in Hotvedt et al. (2022).

5. CASE STUDY

Four experiments (Exp. 1-4) have been conducted to answer the four hypotheses in Section 1. Below, each experiment will be described, and the results visualized. Due to stochasticity, the experiments are run several times, called trials. The results of the trials will be visualized in figures with the median (p_{50}) as a solid line and a shaded area to indicate the lower (p_{25}) and upper (p_{75}) quantiles.

5.1 Exp. 1 - decreasing dataset size

Description This experiment examines the performance of the models to a decreasing training dataset size. Dataset $\mathcal{D}_1$ is used for this purpose using the noise-free measurements. The considered training data lengths are $N \in \{2, 4, 8, 20, 40, 80, 800, 4000, 8000\}$. The training data is randomly extracted from $\mathcal{D}_1$ in each trial.

Results The model performance in terms of the mean absolute error (MAE) is visualized as a function of N in Fig. 4.

5.2 Exp. 2 - increasing noise level

Description This experiment investigates the robustness of the models to an increasing noise level. The models will be trained using dataset $\mathcal{D}_1$ and the output measurements with the different noise levels σ_ε.

Results Fig. 5 shows the relative error as a function of the coefficient of variation σ_ε/μ. The relative error is calculated by dividing the MAE obtained at one noise level by the MAE obtained with noise-free measurements. The MAE is calculated using the noise-free q as the true value. A relative error larger than 1.0 means the model performance has decreased.

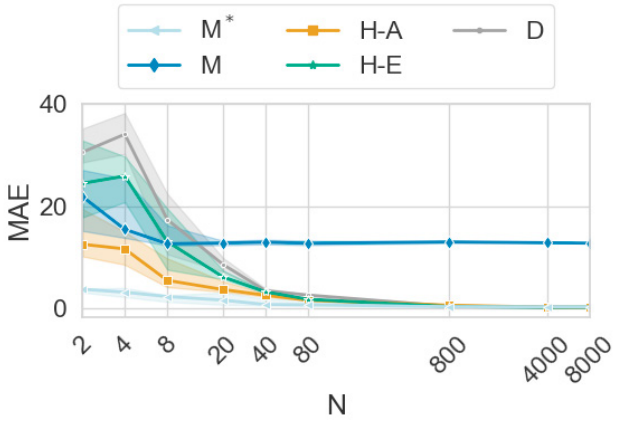

Fig. 4. The mean absolute error as a function of the training set size.

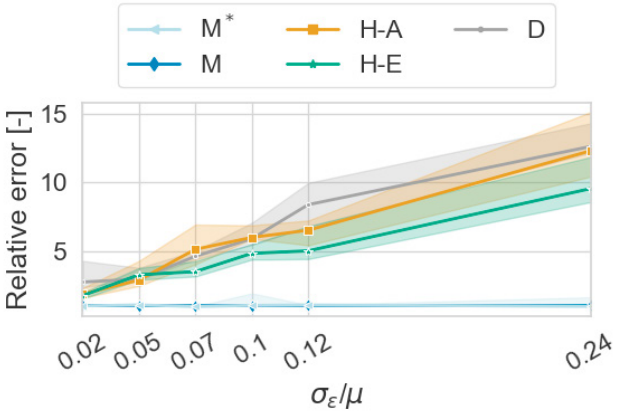

Fig. 5. The relative error as a function of the coefficient of variation for the models.

5.3 Exp. 3 - the depleting reservoir

Description Dataset $\mathcal{D}_2$ is used to analyze the model performances in the nonstationary case of a depleting reservoir.

Results The absolute value of the prediction error (AE) in time is visualized for the different models in Fig. 6. The black, dotted line separates training and test data. Table 1 gives the validation and test MAE for the models.

Table 1. The validation and test mean absolute error in Exp. 3.

	M*	M	H-A	H-E	D
MAE_v	0.1	18.8	2.2	1.3	2.5
MAE_t	1.0	24.7	4.3	2.5	2.8

5.4 Exp. 4 - increasing gas-to-oil ratio

Description Dataset $\mathcal{D}_3$ is used to analyze the model performance in the nonstationary case of an increasing GOR.

Results Fig. 7 shows the absolute error in time separated into training and test data. Table 2 gives the validation and test MAE.

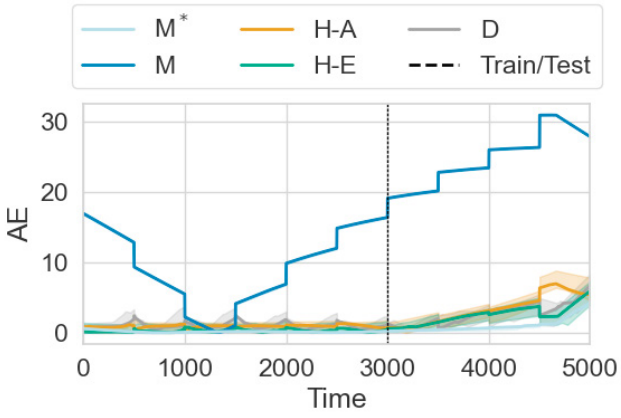

Fig. 6. The absolute error of the model predictions as a function of time for Exp. 3.

Table 2. The validation and test mean absolute error in Exp. 4.

	M*	M	H-A	H-E	D
MAE_v	0.2	2.0	3.0	4.9	6.9
MAE_t	0.3	1.6	3.4	9.0	12.7

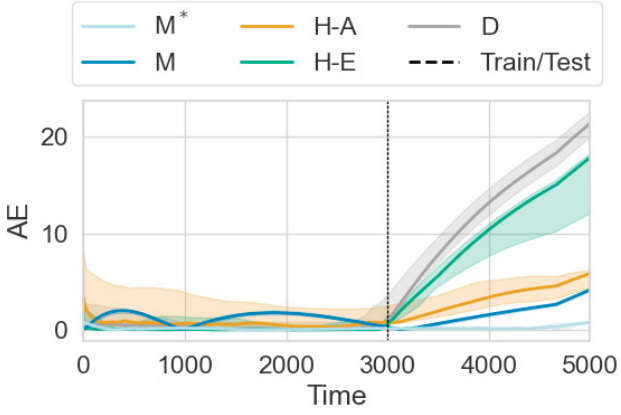

Fig. 7. The absolute error of the model predictions as a function of time in Exp. 4.

6. DISCUSSION

Firstly, notice from Fig. 4 that for large dataset sizes, only M yields a high process-model mismatch. For M*, this was expected as there are no structural mismatches between M* and $\mathcal{P}$. For the other models, the negligible MAE indicates a sufficient capacity. Observe, only a few observations ($N > 80$) were required for the D and Hs to obtain negligible MAE, which suggests that the process is simple to learn. With real-life, complex processes, a higher number of observations would likely be required to remove the bias. Secondly, Fig. 4 shows that the error increases the most for the D when the dataset size decreases, followed by H-E and H-A. This implies that the D has the largest variance, followed by H-E and H-A, and adapts the most to the training data, thus, decreasing the generalizability to the unobserved test data. Fig. 5 shows that the M and M* are robust against an increasing noise level, whereas the Hs and D are not. This confirms that the Hs and D have a larger variance. On the other hand, Fig. 5 shows that the Hs barely achieve a better performance than the D. Moreover, it seems that the H-E has a lower variance than the H-A, which is conflicting with the results in Fig. 4. However, H-E is designed to capture additive mismatches, which is the only considered noise influence and may explain the slightly better performance.

The results from Exp. 1-2 indicates that gray-box models may yield lower variance than a data-driven model and reduce bias in physics-based models with structural process-model mismatches. Therefore, in nonstationary conditions, the expectation is that the Hs will perform better than the D and the M. Figs. 6-7 and Tables 1-2 do show that at least one H performs better than the D in both experiments and that it is advantageous with an H when the process-model mismatch of the M is large as in Exp. 3. The large mismatch in Exp. 3 is a consequence of the available measurements of u making the assumed linear shape of the area function in M of greater influence than in Exp. 4 where $u = 100\%\ \forall t$. It should be noted, the U-shaped curve of the M on the training data in Fig. 6 is due to the objective function in (9), and the performance on the test data can likely be improved by weighing the recent observations the most. On the other hand, in Exp. 3, the performance of the D is comparable with the Hs. In Exp. 4, the discrepancy in performance between the Hs is large, where the H-A and H-E yield good and poor performance, respectively. Ideally, the best model could be deduced a priori to training by examining known process-model mismatches and the capacity of the models. Nevertheless, this showed nontrivial even for these idealized experiments. For instance, in Exp. 3, the H-A was expected to perform best as it targets the discrepancy between the linear and true area function. Nevertheless, H-E yields the best performance, closely followed by the D. Therefore, model selection must be performed posterior to training using the performance on the validation dataset. Accordingly, the importance of extracting the validation dataset representatively increases, for instance, by time for nonstationary processes. Positively, the results in Tables 1-2 indicate that the errors on the validation data are illustrative for the model performances on the test data as the best model yields the lowest error in both experiments. A disadvantage is the increase of overhead on model development and testing. The observant reader notices that the model performances in Figs. 6-7 decrease with time. This is a typical scenario for steady-state modeling in nonstationary conditions. Utilization of learning methods for frequent model updating would likely improve the long-term performances. Such approaches could also handle the existence of both virtual and real drift.

7. CONCLUDING REMARKS

Overall, the results in this research show that a gray-box approach to VFM can reduce both model bias and variance compared to a physics-based and data-driven approach, respectively. From the results and discussions, Hypotheses 1 and 2 from Section 1 are confirmed. However, the gray-box and data-driven models have comparable performances for an increasing data noise level and Hypothesis 3 cannot be confirmed. The results from experiments in nonstationary conditions showed that a gray-box model can improve the

performance of a data-driven model, hence, confirming Hypothesis 4. Moreover, the gray-box model can significantly improve the performance of a physics-based model under large process-model mismatches. On the other hand, the results also show that it is challenging to determine prior to model training which model yields the best performance in different scenarios, and overhead on model development and testing is unavoidable.

Certainly, the hypotheses were only investigated on synthetic data and generalization to real life is challenging. In real life, there may be other undesired and unknown characteristics of the process complicating model development, for instance, increasingly complex and rare physical phenomena or heteroscedastic measurement noise. Moreover, this work only considers two scenarios of nonstationary process behavior, although possible scenarios are numerous. Additionally, other gray-box model variants may yield different results in different scenarios. Nevertheless, the results from this work indicate that gray-box modeling is advantageous for virtual flow metering in certain scenarios and can hopefully act as a guide in modeling real processes.

ACKNOWLEDGEMENTS

This research is a part of BRU21 - NTNU Research and Innovation Program on Digital and Automation Solutions for the Oil and Gas Industry (`www.ntnu.edu/bru21`) and supported by Lundin Energy Norway.

REFERENCES

AL-Qutami, T., Ibrahim, R., Ismail, I., and Ishak, M.A. (2018). Virtual multiphase flow metering using diverse neural network ensemble and adaptive simulated annealing. *Expert Systems With Applications*, 93, 72–85.

AL-Qutami, T.A., Ibrahim, R., and Ismail, I. (2017a). Hybrid neural network and regression tree ensemble pruned by simulated annealing for virtual flow metering application. In *IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 304–309.

AL-Qutami, T.A., Ibrahim, R., Ismail, I., and Ishak, M.A. (2017b). Development of soft sensor to estimate multiphase flow rates using neural networks and early stopping. In *International Journal on Smart Sensing and Intelligent Systems*, volume 10, 199–222.

AL-Qutami, T.A., Ibrahim, R., Ismail, I., and Ishak, M.A. (2017c). Radial basis function network to predict gas flow rate in multiphase flow. In *Proceedings of the 9th International Conference on Machine Learning and Computing*, 141–146.

Alsafran, E.M. and Kelkar, M.G. (2009). Predictions of two-phase critical-flow boundary and mass-flow rate across chokes. *SPE Production & Operations*, 24, 249–256.

Amin, A. (2015). Evaluation of Commercially Available Virtual Flow Meters (VFMs). In *Proceedings of the Annual Offshore Technology Conference*, 1293–1318.

Bikmukhametov, T. and Jäschke, J. (2019). Oil production monitoring using gradient boosting machine learning algorithm. *IFAC-PapersOnLine*, 52 (1), 514–519.

Bikmukhametov, T. and Jäschke, J. (2020). Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models. *Computers and Chemical Engineering*, 138.

Ditzler, G., Roveri, M., Alippi, C., and Polikar, R. (2015). Learning in nonstationary environments: A survey. In *IEEE Computational Intelligence Magazine*, 12–25. doi: 10.1109/MCI.2015.2471196Date.

Grimstad, B., Hotvedt, M., Sandnes, A.T., Kolbjørnsen, O., and Imsland, L.S. (2021). Bayesian neural networks for virtual flow metering: An empirical study. *Applied Soft Computing*, 112.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York, USA.

Hotvedt, M., Grimstad, B., and Imsland, L. (2020). Developing a hybrid data-driven, mechanistic virtual flow meter - a case study. *IFAC-PapersOnLine*, 53, 11692–11697.

Hotvedt, M., Grimstad, B., and Imsland, L. (2021). Identifiability and interpretability of hybrid, gray-box models. *IFAC-PapersOnLine*, 54, 389–394.

Hotvedt, M., Grimstad, B., Ljungquist, D., and Imsland, L. (2022). On gray-box modeling for virtual flow metering. *Control Engineering Practice*, 118.

ISO (1996). Natural gas - Standard reference conditions. Standard, International Organization for Standardization.

Jahn, F., Cook, M., and Graham, M. (2008). *Hydrocarbon exploration and production, 2nd edition*. Elsevier.

Kingma, D. and Ba, J.L. (2015). Adam: a method for stochastic optimization. *International conference on learning representations*.

Oerter, R. (2006). *The Theory of Almost Everything: The Standard Model, the Unsung Triumph of Modern Physics*. Pi Press.

Roscher, R., Bohn, B., Duarte, M., and Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *arXiv:1905.08883v3*, 1–29.

Sachdeva, R., Schmidt, Z., Brill, J.P., and Blais, R. (1986). Two-phase flow through chokes. *Society of Petroleum Engineers, Annual Technical Conference and Exhibition*.

Solle, D., Hitzmann, B., Herwig, C., Remelhe, P.M., Ulonska, S., Wuerth, L., Prata, A., and Steckenreiter, T. (2016). Between the poles of data-driven and mechanistic modeling for process operation. *Chemie Ingenieur Technik*.

Toskey, E. (2012). Improvements to deepwater subsea measurements rpsea program: Evaluation of flow modeling. *Offshore Technology Conference*.

Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V. (2020). Integrating physics-based modeling with machine learning: A survey. *arXiv:2003.04919v4*, 1–34.