



Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Passive learning to address nonstationarity in virtual flow metering applications

Mathilde Hotvedt ^{a,b,*}, Bjarne A. Grimstad ^{a,b}, Lars S. Imsland ^a

^a Norwegian University of Science and Technology, Department of Engineering Cybernetics, O. S. Bragstads plass 2D, Elektroblokk D, 7034 Trondheim, Norway

^b Solution Seeker AS, Rådhusgata 24, 0151 Oslo, Norway

ARTICLE INFO

Keywords:

Virtual flow metering
Nonstationarity
Passive learning
Online learning
Periodic batch learning
Neural networks

ABSTRACT

Steady-state process models are common in virtual flow meter applications due to low computational complexity, and low model development and maintenance cost. Nevertheless, the prediction performance of steady-state models typically degrades with time due to the inherent nonstationarity of the underlying process being modeled. Few studies have investigated how learning methods can be applied to sustain the prediction accuracy of steady-state virtual flow meters. This paper explores passive learning, where the model is frequently calibrated to new data, as a way to address nonstationarity and improve long-term performance. An advantage with passive learning is that it is compatible with models used in the industry. Two passive learning methods, periodic batch learning and online learning, are applied with varying calibration frequency to train virtual flow meters. Six different model types, ranging from data-driven to first-principles, are trained on historical production data from 10 petroleum wells. The results are two-fold: first, in the presence of frequently arriving measurements, frequent model updating sustains an excellent prediction performance over time; second, in the presence of intermittent and infrequently arriving measurements, frequent updating in addition to the utilization of expert knowledge is essential to increase the performance accuracy. The investigation may be of interest to experts developing soft-sensors for nonstationary processes, such as virtual flow meters.

1. Introduction

Many real-world, physical processes are nonstationary (Sayed-Mouchaweh & Lughofer, 2012). To various degrees, process conditions and properties change with time. Nevertheless, a common assumption in process modeling is time independence, leading to stationary, or steady-state, models (Granero-Belinchón et al., 2019). Several arguments militate for the utilization of steady-state models. Firstly, many processes are slowly time-varying making the stationary assumption reasonable for short-term applications. Secondly, steady-state models typically reduce the cost of model development and maintenance (Solle et al., 2016). Thirdly, these models are often less computationally heavy, which can increase the suitability in real-time control and optimization applications (Gravdahl & Egeland, 2002). On the other hand, the performance of steady-state models in nonstationary conditions typically degrade with time and necessitates algorithms that improve the handling of nonstationarity.

Virtual flow metering (VFM) is a soft-sensor technology that utilizes process models for continuous prediction of the multiphase flow rate at key locations in a petroleum asset (Toskey, 2012). In Fig. 1, a simplified

illustration of the production system for one petroleum well is given along with typically available sensor measurements for well-equipped wells. A multiphase flow meter (MPFM) measures the phasic flow rates, $q = [q_{\text{gas}}, q_{\text{oil}}, q_{\text{water}}]$, through the production choke valve. Under well-testing, the phasic flow rates can be measured using the test separator. The total multiphase flow rate through the production system is $Q = q_{\text{gas}} + q_{\text{oil}} + q_{\text{water}}$. A typical application of VFM is as a back-up to the MPFM in case of failure (Varyan et al., 2015).

The underlying process of the VFM comprises the reservoir, wells, pipelines, and processing facility. This process is nonstationary with time-varying process conditions and properties (Guo et al., 2007). The multiphase flow rate through the production system has a dynamic nature with both fast and slow transients. Fast transients occur with control changes, which induce pressure waves through the system, such as the opening of the choke valve (Jansen, 2015). These are in the time range of minutes to hours. Slow transients are caused by the reservoir being depleted with time, which in turn results in a pressure declination in the production system and a decreased production flow rate (Foss

* Corresponding author at: Norwegian University of Science and Technology, Department of Engineering Cybernetics, O. S. Bragstads plass 2D, Elektroblokk D, 7034 Trondheim, Norway.

E-mail addresses: mathilde@solutionseeker.no (M. Hotvedt), bjarne.grimstad@solutionseeker.no (B.A. Grimstad), lars.imsland@ntnu.no (L.S. Imsland).

<https://doi.org/10.1016/j.eswa.2022.118382>

Received 8 February 2022; Received in revised form 5 July 2022; Accepted 1 August 2022

Available online 12 August 2022

0957-4174/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

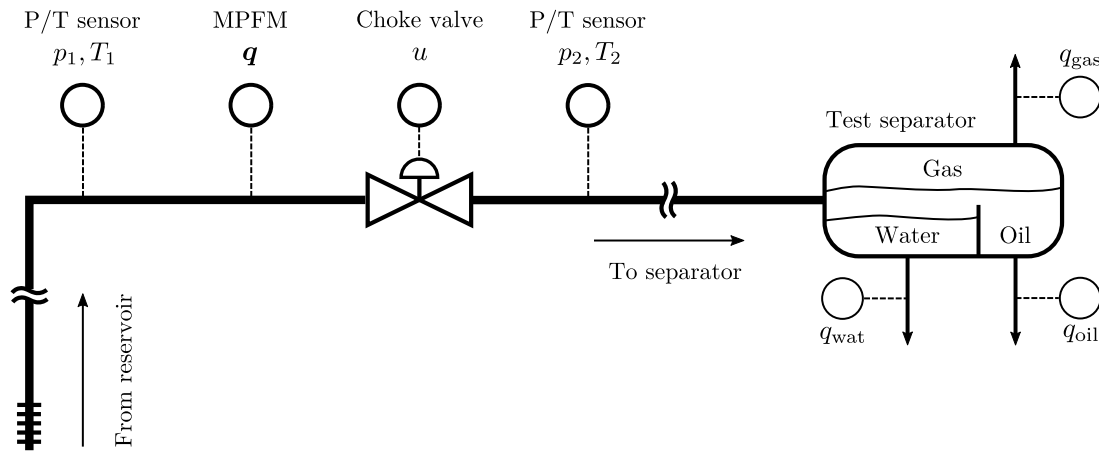


Fig. 1. A simplified illustration of the petroleum production system with typical sensor placements. A multiphase flow meter (MPFM) measures the phasic flow rates through the choke valve. Measurements of the phasic flow rates can also be obtained when the well is tested, using, for instance, a test separator.

et al., 2018). These occur in a time range of months to years, dependent on the size of the reservoir. Furthermore, as the petroleum asset ages, technologies such as artificial lift with gas or water are applied to improve production. Other sporadic changes such as maintenance tasks will also induce transient process behavior. Hence, the natural approach to VFM is nonstationary models. Several commercial VFMs such as Olga and LedaFlow are nonstationary (Amin, 2015), and other examples exist in literature (Holmås & Løvli, 2011; Jordanou et al., 2017). On the other hand, due to the slow dynamics of the reservoir, steady-state reservoir conditions for a certain time interval can often be assumed (Shippen, 2012). Furthermore, considering the inherent complex multiphase flow characteristics, which make it challenging to develop and solve nonstationary VFMs, steady-state VFMs are the most common approach in literature (Bikmukhametov & Jäschke, 2019), both for physics-based models (Shippen, 2012; Varyan et al., 2015) and machine learning (ML) models (AL-Qutami et al., 2017, 2017a, 2017b, 2018; Bikmukhametov & Jäschke, 2020; Grimstad et al., 2021). Nevertheless, studies show that steady-state VFM models should be updated or recalibrated in time to provide adequate long-term prediction accuracy (Hotvedt et al., 2022; Sandnes et al., 2021). Several model learning methods exist that attempt to account for nonstationarity without imposing temporal dependencies in the model. The learning methods can be divided into an active or passive method (Ditzler et al., 2015). In passive learning, the process is assumed to be continuously changing and the model is routinely updated with access to new measurements. In active learning, statistical tests are used to detect significant changes in the process conditions, whereupon model updating is initiated.

For the VFM application, it is not uncommon that new observations arrive infrequently, for example, twice a year or at the most once per month under well-testing (Monteiro et al., 2020). In such an event, active learning is redundant as the process conditions and properties are likely to have changed significantly during the elapsed time, and the model should be updated with each new measurement. For assets with access to continuous flow rate measurements, such as MPFM measurements, the VFM models would likely benefit from updating using these measurements in between well-tests. Nevertheless, in industry, even with frequent access to new measurements, model learning can occur intermittently due to limited resources or manual, non-systematic workflows (Koroteev & Tekic, 2021).

To the authors' knowledge, no studies have investigated the influence of the update frequency on sustaining the prediction accuracy of steady-state VFM models over time, hence, obtaining a high long-term performance. This research contributes in this direction by examining two passive learning methods: periodic batch learning and online learning. Six VFM models are developed for the petroleum production choke valve in 10 petroleum wells on Edvard Grieg, an asset

on the Norwegian Continental Shelf (Lundin Energy Norway, 2020). Real production data spanning five years are used in the development. The long-term predictive performance is expected to increase with the frequency of which the models are updated. The best performance is expected from online learning, for which the models are updated with every new measurement. For periodic batch learning, the performance is expected to drop as the frequency is lowered. The rest of the article is structured the following way: Section 2 presents relevant theory for steady-state modeling of processes in nonstationary conditions. Thereafter, Section 3 describes the available data and the VFM model types. In Section 4, the numerical study examining the learning methods is described and results visualized and discussed. Lastly, Section 5 gives concluding remarks.

2. Steady-state modeling in nonstationary conditions

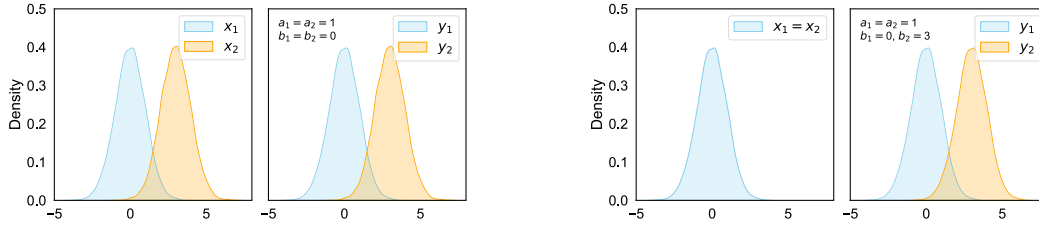
Consider a stream of observations $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_t, y_t), \dots\}$, where $x_t \in \mathbb{R}^d$ represents measured process conditions and $y_t \in \mathbb{R}$ a (dependent) target variable at time t . In general, the set S can be thought of as a realization of a stochastic process \mathcal{P} governed by a generative model (Oliveira et al., 2021)

$$p_t(x, y) = p_t(y | x)p_t(x). \quad (1)$$

In (1), $p_t(x)$ is the marginal distribution of the process conditions, and $p_t(y | x)$ is the conditional distribution of the target, both at time t . The index t indicates that the distributions may be time-variant, and therefore \mathcal{P} may be nonstationary.

In real-time applications of machine learning, like data-driven virtual flow metering, it is natural to develop models on historical data and test the model performance on future data. Collect in $D_{a:b} = \{(x_t, y_t)\}_{t=a}^b$ the sequence of observations with $t \in [a, b]$, and in $D_a = \{(x_t, y_t)\}_{t=a}$ the single observation at $t = a$. For a model to be developed at time $t = T$, the training dataset is denoted by $D^r = D_{1:T}$ and the test dataset by $D^r = D_{T+1:\infty}$.

Many machine learning models and algorithms are based upon the assumption that the training and test dataset originate from the same probability distribution; the data points in S are independent and identically distributed (i.i.d.) (Hastie et al., 2009). When the stochastic process \mathcal{P} in (1) is nonstationary, the i.i.d. assumption is invalidated, as a dataset shift can occur when moving from the training phase to the test phase. In the following, different types of dataset shifts are explored, and suitable learning methods to alleviate the effect of nonstationarity on predictive performance are discussed.



(a) Virtual drift. The marginal distribution $p(x)$ changes from one time step to another causing a change $p(y | x)$. The parameters remain unchanged.

(b) Real drift. The marginal distribution $p(x)$ does not change but $p(y | x)$ changes as a consequence of the changed model parameters.

Fig. 2. Dataset shifts illustrated with (a) virtual drift and (b) real drift.

2.1. Dataset shifts

When \mathcal{P} is nonstationary, the joint probability distribution can shift in time resulting in $p_t(x, y) \neq p_{t+\tau}(x, y)$ for an arbitrary lapse $\tau > 0$ in time. Using the model in (1) two types of dataset shifts, also called concept drifts, can occur in time: virtual and real drift¹ (Ditzler et al., 2015; Quiñero-Candela et al., 2009). With virtual drift, the marginal distribution shifts in time. That is, $p_t(x) \neq p_{t+\tau}(x)$ for $\tau > 0$. With real drift, the conditional distribution shifts in time, such that $p_t(y | x) \neq p_{t+\tau}(y | x)$ for $\tau > 0$. Real and virtual drift may happen separately or simultaneously, in any case shifting the joint distribution with time. Notice, in (Quiñero-Candela et al., 2009), several other specialized forms of dataset shifts are discussed.

As an example, consider a process with a conditional distribution

$$y_t = a_t x_t + b_t, \quad (2)$$

with parameters $\theta_t = \{a_t, b_t\}$. Two subsequent time instances $t = 1$ and $t = 2$ are examined. The input at $t = 1$ is sampled from $p_1(x) \sim \mathcal{N}(0, 1)$. At $t = 2$, the mean changes such that $p_2(x) \sim \mathcal{N}(3, 1)$. If the model parameters remain unchanged, this is virtual drift, and the response in y changes only as a consequence of changes in the marginal distribution. The scenario is illustrated in Fig. 2(a). In another scenario, consider the input distribution to remain unchanged, but the b parameter of the model to change from $b_1 = 0$ to $b_2 = 3$. The parameter change causes the conditional distribution in (2) to change, thereby causing real drift, illustrated in Fig. 2(b). Notice, in the two scenarios, virtual and real drift cannot be distinguished by analyzing y only.

Virtual drift is commonly seen in the VFM application. For example, in time with the reservoir being depleted the pressure through the production system decreases. At the same time, in the early life of a petroleum asset, the production engineers can often increase the choke openings to maintain a constant production rate, also called plateau production (Jansen, 2015). The VFM application can also experience real drift. Substantial mechanical wear of the equipment in the well can occur with time, for instance, due to sand production, and can result in a change in the flow rate even for unchanged process conditions. It is believed that virtual drift is the major cause of observed dataset shifts in VFMs. However, as Fig. 2 illustrates, it can be difficult to separate between the two types of drifts.

The next section discusses the impact that dataset shifts can have on steady-state VFM models.

2.2. Parameter estimation of steady-state models

A common approach to steady-state modeling is to use an inductive method to learn an approximation of the conditional distribution $p_t(y | x)$ in (1) from a fixed set of steady-state observations $D_{1:T}$. A typical form of the approximation is

$$\hat{y}_t = f_\theta(x_t) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2), \quad (3)$$

where f_θ is a parametric model of the mean, with parameters θ , and ϵ_t is a homoscedastic noise term. The model in (3) is a steady-state model since \hat{y}_t is conditioned on x_t , and the parameters θ and σ_ϵ are time-invariant. The i.i.d. assumption is thus used. Note that, the resulting model is steady-state even though the data used to learn the model originate from a nonstationary process.

Conditional models, like the steady-state model in (3), are commonly trained using maximum a posteriori (MAP) estimation. In MAP estimation, the mode of the posterior distribution $p(\theta | D) \propto p(D | \theta)p(\theta)$ is maximized. Here, the likelihood $p(D | \theta)$ is given by (3) and $p(\theta)$ is a prior on the θ parameters. For a normal prior, $\theta_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $i = 1, \dots, N_\theta$, the optimization problem can be expressed as follows:

$$\begin{aligned} \hat{\theta} &= \arg \max_{\theta} \log p(D | \theta) + \log p(\theta) \\ &= \arg \min_{\theta} \sum_{i=1}^N \frac{1}{\sigma_\epsilon^2} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{N_\theta} \frac{1}{\sigma_i^2} (\theta_i - \mu_i)^2. \end{aligned} \quad (4)$$

where N is the number of data points in the training dataset. From (4), it is seen that MAP estimation is a trade-off between minimizing the squared errors and parameter deviation away from its respective mean value μ_i . By multiplying the objective function by σ_ϵ^2/N , the equivalence of MAP estimation to the familiar minimization of mean squared error with ℓ_2 -regularization is obtained (Goodfellow et al., 2016).

In the machine learning domain, (4) is commonly optimized by first-order gradient descent methods (Bishop, 2006). These methods update the parameters iteratively according to the following scheme:

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} - \gamma^{(k)} \mathcal{M}(\mathcal{B}, \hat{\theta}^{(k)}), \quad k = 1, \dots, E \quad (5)$$

where E is the number of iterations or steps taken towards the optimal value, γ is the learning rate or step-size, and \mathcal{M} is the set of equations calculating the step direction. The \mathcal{B} is a set of observations extracted from the training dataset and can be in the range of one to all observations. Any parameter that is not included in θ is called a hyperparameter, for instance, γ , E , and $|\mathcal{B}|$.

The above approach to steady-state modeling is susceptible to dataset shifts since the estimate (optimum) in (4) likely will change with time, resulting in poor test performance. When applied to VFM, for which the data is generated by a nonstationary process, both virtual and real concept drift will negatively influence the long-term predictive performance. A VFM performance that diminishes with time, has been

¹ Other naming conventions for virtual drift are virtual concept drift, covariate shift, or input drift. Real drift is also known as real concept drift or output drift.

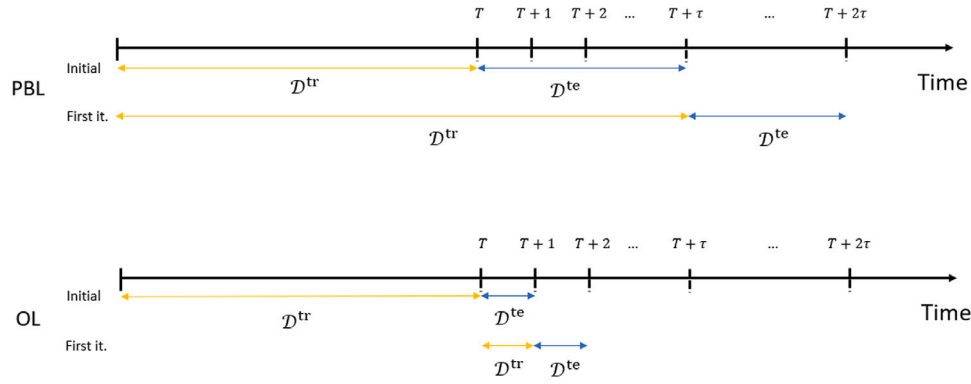


Fig. 3. One iteration of the periodic batch learning and online learning update procedure after obtaining the initial parameter estimate. D^{tr} are training datasets used in the estimation problem and D^{te} are test datasets used to test the predictive capabilities of the model.

documented in several publications (Grimstad et al., 2021; Hotvedt et al., 2022; Sandnes et al., 2021). In the following section, passive learning methods are discussed. These methods can be used to account for dataset shifts in steady-state modeling.

2.3. Passive learning for steady-state models

In passive learning, the process \mathcal{P} is assumed to be continuously changing with time, and model updating is routinely initiated regardless of whether or not dataset shifts occur. Two methods of passive learning are examined: online learning (OL) and periodic batch learning (PBL).

At time $t = T$, an initial parameter estimate is obtained from $D^{tr} = D_{1:T}$ using the approach in Section 2.2. The estimated parameters are referred to as $\hat{\theta}_T$, and the resulting steady-state model is given by

$$\hat{y}_t = f_{\hat{\theta}_T}(x_t) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2). \quad (6)$$

From this point in time, the two learning methods can be applied. These are visualized in Fig. 3 and are explained in the consecutive sections.

2.3.1. Periodic batch learning

In periodic batch learning, the model in (6) is used to make predictions for $\tau > 0$ time steps $D_{T:T+\tau}^{te}$ before it is retrained at $t = T + \tau$. In retraining, the new parameters $\hat{\theta}_{T+\tau}$ are estimated using all data observed at that time as training data $D_{1:T+\tau}^{tr}$ and the approach in Section 2.2. The procedure is repeated with a period of τ , where the posterior parameter distribution can be described with

$$p(\theta | D_{1:T+\tau}) \propto p(D_{1:T+\tau} | \theta)p(\theta), \quad (7)$$

An appropriate τ must be determined and can be accomplished by applying a change or shift detection algorithm offline on historical data. There exist much literature on shift detection algorithms, see for example Raza et al. (2015) and references therein. In this research, Hotelling's T-squared test for two multivariate, independent samples is used to investigate a null hypothesis stating that no virtual drift is present in the dataset. The algorithm for determining τ is described in Appendix A.

2.3.2. Online learning

In online learning, model updating occur for each new observation that arrives. However, the posterior distribution at the next time step is updated using only the current observation as the training data and the posterior distribution at the previous time step as the prior. For instance, at $t = T + 1$:

$$p(\theta | D_{1:T+1}) \propto p(D_{T+1} | \theta)p(\theta | D_{1:T}). \quad (8)$$

Mathematically, (8) can be derived as follows. With the approach in Section 2.2, the posterior parameter distribution at $t = T + 1$ is given by

$$p(\theta | D_{1:T+1}) = \frac{p(D_{1:T+1} | \theta)p(\theta)}{p(D_{1:T+1})}, \quad (9)$$

where $p(D_{1:T+1})$ is the proportionality constant in Bayes' law. Applying the i.i.d. assumption, the likelihood function of the model and the evidence can be written as

$$p(D_{1:T+1} | \theta) = \prod_{t=1}^{T+1} p(D_t | \theta) = p(D_{1:T} | \theta)p(D_{T+1} | \theta) \quad (10)$$

$$p(D_{1:T+1}) = \prod_{t=1}^{T+1} p(D_t) = p(D_{1:T})p(D_{T+1}),$$

respectively. Note that, while the i.i.d. assumption is likely false for a nonstationary process, it is already used in steady-state modeling. Inserting (10) in (9), the posterior parameter distribution at $t = T + 1$ can be written as

$$p(\theta | D_{1:T+1}) = \frac{p(D_{T+1} | \theta)}{p(D_{T+1})} \cdot \frac{p(D_{1:T} | \theta)p(\theta)}{p(D_{1:T})} = \frac{p(D_{T+1} | \theta)}{p(D_{T+1})} \cdot p(\theta | D_{1:T}) \quad (11)$$

and (8) is obtained.

An issue becomes apparent when deriving the MAP estimate for (8)

$$\hat{\theta}_{T+1} = \arg \max_{\theta} \left[\log p(D_{T+1} | \theta) + \log p(\theta | D_{1:T}) \right]. \quad (12)$$

Ideally, the parameter estimation in the previous time step should have provided both the mean and the variance of the updated posterior parameter distribution $p(\theta | D_{1:T}) \sim \mathcal{N}(\mu_T, \Sigma_T)$. However, MAP estimation gives point estimates of the mode only. When the likelihood and prior is normal, an estimate of the mean $\mu_T = \hat{\theta}_T$ is obtained since the mode and mean coincides, but Σ_T remains unknown. Therefore, the second term in (12) cannot be calculated if MAP estimation is used in each time step. As discussed in Section 2.2, this term is ℓ_2 -regularization of the parameters. According to (Goodfellow et al., 2016), for some cases, the algorithm early stopping has a similar effect as ℓ_2 -regularization. For linear models, the solution obtained with early stopping equals a solution with ℓ_2 -regularization where the regularization term is determined by the number of iterations and step-size in early stopping (Santos, 1996). Therefore, for the OL algorithm implemented in this research, the iterative optimization algorithm in (5) uses the posterior parameter estimate from the previous time step as a starting point but iterates only a few steps k towards the optimal value. In such a sense, the approach is similar to an early stopping approach, and will to some degree include parameter regularization.

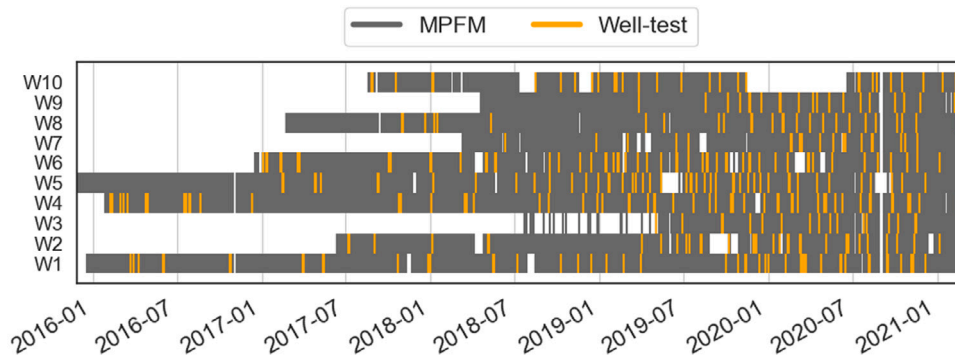


Fig. 4. Visualization of the occurrence of observations for each well against time. Some wells have older historical observations than others. Both multiphase flow meter and well-test measurements are available.

2.4. Comparison of periodic batch and online learning

There are advantages and disadvantages to both passive learning methods. With OL, the model can quickly adapt to changes in process conditions. Further, as only new observations are used, old data may be discarded yielding low memory requirements. However, it has been shown that some machine learning models such as neural networks are prone to catastrophic forgetting when trained using OL (Goodfellow et al., 2013; Kemker et al., 2018; Parisi et al., 2019). Catastrophic forgetting is a situation where the model excessively overfits its parameters to new observations resulting in a decreased performance on previously seen observations. This situation occurs due to the stability–plasticity dilemma (Wickliffe & Robins, 2005). The neural network requires adequate plasticity to adapt to new patterns, but too much can cause the network to forget previously learned patterns. The reverse is true for stability. The stability–plasticity of the models is connected to the hyperparameters of the learning algorithms. With time, the optimal hyperparameters can change. This is a problem for OL in real-time applications as a hyperparameter search in each iteration can be infeasible, dependent on the frequency of arrival of new observations. Another potential issue for the OL is the required complex system integration. The method will require fast processing capabilities of new observations to account for erroneous sensor measurements, and model performance monitoring applications are a necessity to analyze model drift and catastrophic forgetting (Ditzler et al., 2015). Furthermore, the learning method must be automated as manual, although systematic, handling of model updating can be impractical in real-time due to limited resources.

PBL addresses catastrophic forgetting as all available observations are used in model updating. Yet, using this method for each new observation can be impractical in real-time applications due to a larger training time caused by larger datasets (Kemker et al., 2018). Therefore, a longer period (τ in (7)) between model retraining can be required and sudden shifts in the data can be missed. On the other hand, if the underlying process is slowly changing, a lower update frequency can be sufficient to capture dominant changes in process conditions. Correspondingly, a manual yet systematic handling of the learning method including measurement preprocessing, conducting a hyperparameter search, and the actual model learning can be more achievable in each iteration. For VFM applications, studies have indicated that the inclusion of too old data may be redundant and not improve the model performance significantly (AL-Qutami et al., 2018; Grimstad et al., 2021). Thus, a windowing strategy can be applied to discard redundant data (Ditzler et al., 2015).

3. Data and models

In this research, six different VFM model types are considered. The data used to develop the VFMs and examine the effect of the learning

methods on the long-term prediction performance are real production data from 10 wells, W1–W10, on the Edvard Grieg asset (Lundin Energy Norway, 2020). The available data and the VFM model types are described in the below sections.

3.1. Available data

The available process data consists of observations from the $M = 10$ wells indexed by $j \in \{1, \dots, M\}$. The dataset of well j is $\{(x_{t,j}, y_{t,j})\}_{t=1}^{N_j}$, where N_j is the number of observations, explanatory variables are $x_{t,j} = (u, p_1, p_2, T_1, \eta_{oil}, \eta_{gas})_{t,j} \in \mathbb{R}^6$, and target variables are $y_{t,j} = Q_{t,j} \in \mathbb{R}$. The η_{oil} and η_{gas} are the fractions of oil and gas in the fluid mixture. Ideally, the fractions should be estimated using a different model, for instance, a wellbore model as in Kittilsen et al. (2014). For simplification, the fractions are approximated using the measured phasic volumetric flows. Measurements of the target variable, the mixture volumetric flow rate, are from both well-tests conducted with a test separator, and from the multiphase flow meter in each well. Commonly, well-test measurements have higher accuracy than MPFM measurements as MPFM are prone to failure and drift over time (Falcone et al., 2013). The data from all wells is denoted by \mathcal{D} . Each of the datasets is generated using the processing technology in Grimstad et al. (2016). This technology compresses the data by removing fast transients. However, slow transients can still be present. Further, the datasets are passed through a set of filters that remove undesired, illogical measurements, for instance, negative pressures or negative flow rate measurements. The wells have an unequal number of observations spanning a different time range, see Fig. 4. Some wells have historical observations back to 2016 while others have their first observations in late 2018. Further, there are periods where observations are lacking for some of the wells represented by white holes in the data in Fig. 4. Here, the well in question can have been shut down, or the sensors failed. In total, there are 26743 observations from the 10 wells, spanning more than five years of production history. On average, there is less than one day between each measurement. The time between well-tests for a well is varying, with more than one year at the longest, and less than a day at the shortest.

3.2. Virtual flow meter models

The six different VFM models considered range from machine learning, or data-driven, to physics-based, or mechanistic, models:

1. A linear regression model (LR)
2. A fully connected feed-forward neural network (NN)
3. A multi-task learning model (MTL)
4. A hybrid, gray-box error model (HEM)
5. A hybrid, gray-box area function model (HAM)
6. A mechanistic model (MM)

There are advantages and disadvantages with all model types (Hotvedt et al., 2022; Solle et al., 2016). Mechanistic models are built from physical laws and require little process data in development. Yet, simplifications and assumptions are often necessary to make mechanistic models computationally feasible in real-time applications. Hence, model bias or process-model mismatch is typically encountered. Machine learning models are built from available data only and require no prior knowledge about the physics of the process. The capacity of machine learning models vary, where the NN is a typical model with high capacity and the LR a model with low capacity. High capacity models enable adaptation to arbitrarily complex physical relationships as long as these are reflected in the data, commonly reducing model bias. However, due to the inherent bias–variance trade-off of high capacity models, minimizing the bias results in higher variance (Hastie et al., 2009). Therefore, high capacity models are often influenced by poor quality data or data located in the small data regime, a situation not uncommon for the VFM application (Grimstad et al., 2021). Furthermore, higher variance typically decreases the generalization abilities to previously unobserved data, and such models can struggle if used in nonstationary environments where the process experiences dataset shifts. The hybrid models attempt to utilize knowledge from both the mechanistic and the data-driven modeling domain to preserve the advantages but diminish the disadvantages of both methods.

The MTL models are somewhat different from the other model types. This model type enables learning from a plurality of wells, where each well presents a learning task. Instead of separately training a model for each well, which can be considered as single-task learning, the models are simultaneously trained. The advantage of using multi-task learning is two-fold. First, it allows for parameter sharing among models which can drastically improve data efficiency and predictive performance in the small-data regime. This is analogous to an MM whose equations are shared among wells. Second, compared to single-task learning, simultaneous training can lessen the effort and computational cost of developing models when the number of wells becomes large.

In the following sections, a mathematical description of the six VFM models is introduced. In addition to these, a benchmark model used to compare the performance of the models is described.

3.2.1. Benchmark model

A simple benchmark model predicts the flow rate to be the same as the last observed flow rate. Consider chronologically ordered observations $\{y_{1,j}, \dots, y_{N_j,j}\}$ for well j so that $y_{t,j}$ is observed after $y_{t-1,j}$. The prediction from the benchmark model is

$$\hat{y}_{t,j} = y_{t-1,j}, \quad t = 1, \dots, N_j, \quad j = 1, \dots, M. \quad (13)$$

Note that with this model the prediction is independent of the explanatory variables $x_{t,j}$. Further, if the petroleum production is on plateau, resulting in each new observation deviating little from the previous, the benchmark model has the potential of high accuracy.

3.2.2. Linear regression model

The linear regression model fits a multidimensional line to the observed data. The functional form is given by $f_{\theta}^{(LR)} : \mathbb{R}^d \rightarrow \mathbb{R}$ and is evaluated for a given x as

$$\hat{y} = \mathbf{w}^T \mathbf{x} + b. \quad (14)$$

The model parameters consist of a weight vector $\mathbf{w} \in \mathbb{R}^d$ and a bias $b \in \mathbb{R}$, $\theta = \{(\mathbf{w}, b)\}$.

3.2.3. Feed-forward neural network model

In general, the feed-forward neural network is a set of nonlinear regression lines. It has a functional form $f_{\theta}^{(NN)} : \mathbb{R}^d \rightarrow \mathbb{R}$. For a neural network with L hidden layers and one output layer, the parameters are $\theta = \{(W^{(l)}, \mathbf{b}^{(l)})\}_{l=1}^{L+1}$, where $W^{(l)}$ and $\mathbf{b}^{(l)}$ are the weights and biases

of layer l , respectively. The dimensions of $W^{(l)}$ and $\mathbf{b}^{(l)}$ determine the width of layer l .

In this work, the rectified linear unit (ReLU) activation function is used as the nonlinearity in the hidden layers (Glorot et al., 2011). The ReLU function is denoted by $a : \mathbb{R}^d \rightarrow \mathbb{R}^d$, $a(\mathbf{z})_i := \max(0, z_i)$, where the max operator is applied element-wise for $i = 1, \dots, d$. This makes the neural network a set of piecewise linear regression lines. The evaluation of model $f_{\theta}^{(NN)}(\mathbf{x})$ for a given \mathbf{x} is

$$\begin{aligned} \mathbf{z}^{(1)} &= \mathbf{x} \\ \mathbf{z}^{(l+1)} &= a(W^{(l)}\mathbf{z}^{(l)} + \mathbf{b}^{(l)}), \quad l = 1, \dots, L \\ \hat{y} &= W^{(L+1)}\mathbf{z}^{(L+1)} + \mathbf{b}^{(L+1)}. \end{aligned} \quad (15)$$

3.2.4. Multi-task learning model

A MTL formulation introduces a new semantics of the model parameters compared to the NN in Section 3.2.3. Let α denote parameters that are shared among tasks (here wells), and let $\beta_j \in \mathbb{R}^P$ be P task-specific parameters for wells $j = 1, \dots, M$. The parameters of the MTL model for M wells are collected in $\theta = \{\alpha, \beta_1, \dots, \beta_M\}$.

When processing a data point $x_{t,j}$ of well j , the model must select the corresponding task-specific parameters, β_j . The selection can be made by introducing an encoding of tasks. Let e_j be an indicator vector of dimension M , with all zeros, except for a one in position j . By stacking the task-specific parameters in a matrix B with columns $B_{*,j} = \beta_j$, a selection can be made by performing the multiplication $\beta_j = B e_j$.

A simple MTL model is obtained by utilizing the selection mechanism described above. First, β_j is selected using the encoding e_j . Next, $x_{t,j}$ and β_j are fed through a residual neural network with shared parameters α . In this work, a residual neural network with pre-activation is used to allow for an identity mapping of the task-specific parameters (He et al., 2016). The resulting model is a simplified version of the MTL choke model introduced in (Sandnes et al., 2021).

The functional form of the MTL model is $f_{\theta}^{(MTL)} : \mathbb{R}^d \times \{0, 1\}^M \rightarrow \mathbb{R}$, where the second argument is the task encoding vector. The evaluation of $f_{\theta}^{(MTL)}(\mathbf{x}, e_j)$ for a data point \mathbf{x} of well j , is performed as follows:

$$\begin{aligned} \beta_j &= B e_j, \\ \hat{y} &= g_{\alpha}(\mathbf{x}, \beta_j), \end{aligned} \quad (16)$$

where g_{α} is a residual neural network with L residual blocks given by

$$\begin{aligned} \mathbf{z}^{(1)} &= W^{(0,1)}\mathbf{x} + W^{(0,2)}\beta_j + \mathbf{b}^{(0)}, \\ \mathbf{r}^{(l)} &= W^{(l,2)}a(W^{(l,1)}a(\mathbf{z}^{(l)} + \mathbf{b}^{(l,1)}) + \mathbf{b}^{(l,2)}), \quad l = 1, \dots, L, \\ \mathbf{z}^{(l+1)} &= \mathbf{r}^{(l)} + \mathbf{z}^{(l)}, \quad l = 1, \dots, L, \\ \hat{y} &= W^{(L+1)}\mathbf{z}^{(L+1)} + \mathbf{b}^{(L+1)}. \end{aligned} \quad (17)$$

The weights and biases in (17) are collected in α and are shared among the M wells. These parameters can be learned from all the data in D .

3.2.5. Mechanistic model

The mechanistic choke model is taken from Sachdeva et al. (1986). The equations are developed from the steady-state mass and momentum balance equations for one-dimensional flow along a streamline. In short notation, the mechanistic model is given by $f_{\theta}^{(MM)} : \mathbb{R}^d \rightarrow \mathbb{R}$ with parameters $\theta = \{\rho_{oil}, \rho_{wat}, \kappa, M_{gas}, p_{cr}, C_D\}$, and the equation for the volumetric flow rate through the choke is given by:

$$\begin{aligned} \hat{y} &= Q = \frac{\dot{m}}{\rho_{SC}} \\ &= \frac{C_D A_2(u)}{\rho_{SC}} \\ &\times \sqrt{2\rho_2^2 p_1 \left(\frac{\kappa}{\kappa - 1} \eta_{gas} \left(\frac{1}{\rho_{gas,1}} - \frac{p_r}{\rho_{gas,2}} \right) + \left(\frac{\eta_{oil}}{\rho_{oil}} + \frac{\eta_{wat}}{\rho_{wat}} \right) (1 - p_r) \right)}, \end{aligned} \quad (18)$$

Details regarding the model are found in Hotvedt et al. (2022).

3.2.6. Hybrid error model

This model uses the mechanistic model in Section 3.2.5 as a baseline but inserts a neural network as introduced in Section 3.2.3 to capture the error between the mechanistic model output and measurements, or the process-model mismatch. The functional form of the model is given by $f_{\theta}^{(\text{HEM})} : \mathbb{R}^d \rightarrow \mathbb{R}$ with parameters $\theta = \{\theta_{\text{MM}}, \theta_{\text{NN}}\}$, where the physical model parameters are the same as given in Section 3.2.5: $\theta_{\text{MM}} = \{\rho_{\text{oil}}, \rho_{\text{wat}}, \kappa, M_{\text{gas}}, p_{\text{cr}}, C_D\}$, and the neural network parameters are the weights and biases on each layer of the network as described in Section 3.2.3: $\theta_{\text{NN}} = \{(W^{(l)}, b^{(l)})\}_{l=1}^{L+1}$. The evaluation of HEM for a data point \mathbf{x} is described by

$$\hat{y} = f_{\theta}^{(\text{HEM})}(\mathbf{x}) = f_{\theta_{\text{MM}}}^{(\text{MM})}(\mathbf{x}) + f_{\theta_{\text{NN}}}^{(\text{NN})}(\mathbf{x}) \quad (19)$$

3.2.7. Hybrid area function model

This model also uses the mechanistic model in Section 3.2.5 as a baseline. However, the mechanistic relation for the area function $A_2(u)^{(\text{MM})}$ is manipulated by multiplying with a neural network. This may be interpreted as replacing the discharge coefficient C_D from the MM with a neural network. Accordingly, $f_{\theta}^{(\text{HAM})} : \mathbb{R}^d \rightarrow \mathbb{R}$ with parameters $\theta = \{\theta_{\text{MM}}, \theta_{\text{NN}}\}$, where $\theta_{\text{MM}} = \{\rho_{\text{oil}}, \rho_{\text{wat}}, \kappa, M_{\text{gas}}, p_{\text{cr}}\}$ and $\theta_{\text{NN}} = \{(W^{(l)}, b^{(l)})\}_{l=1}^{L+1}$. The evaluation of $f_{\theta}^{(\text{HAM})}$ for data point \mathbf{x} is as follows:

$$\begin{aligned} A_2 &= A_2(u)^{(\text{MM})} \times f_{\theta_{\text{NN}}}^{(\text{NN})}(\mathbf{x}) \\ \hat{y} &= f_{\theta}^{(\text{HAM})}(\mathbf{x}) = f_{\theta_{\text{MM}}}^{(\text{MM})}(\mathbf{x}, A_2) \end{aligned} \quad (20)$$

Note, the complete vector of explanatory variables is used as input to the area function network and not just the choke opening u . This is due to the expectation of the effective flow area being dependent on the characteristics of the fluid flowing through the choke, which cannot be captured with just u .

3.3. Prior parameter distribution

All the VFM models except the benchmark model need specification of the prior parameter distributions $\theta_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. For the data-driven model parameters θ_{NN} , He-initialization is utilized, which is recommended for neural networks with ReLU as activation function (He et al., 2015). For the mechanistic model parameters θ_{MM} , typical values for the mean μ_i is commonly known. For instance, a typical value for the density of freshwater is 1000 kg/m³. The variance may be estimated using the known bounds of the parameter in question. Details on prior parameter specification in gray-box models may be found in Hotvedt et al. (2022).

4. Numerical study

Online learning and periodic batch learning as described in Section 2.3, are used to train the six models in Section 3.2, for the 10 wells, using the data described in Section 3.1. All VFM models except the LR are implemented using the Python framework *PyTorch* (Paszke et al., 2019). The LR is implemented with the Python framework *scikit-learn* (Pedregosa et al., 2011) using the stochastic gradient descent linear regressor to allow for training the model with online learning. As mentioned in Section 2.3.2, the ℓ_2 -regularization term cannot be calculated for the online learning method. However, for the hybrid and mechanistic models, an important factor is that the model parameters with a physical interpretation θ_{MM} stay within feasible bounds. Therefore, ℓ_2 -regularization with the initial priors is applied for these parameters.

The numerical study considers two cases. In Case 1, all available data, both MPFM and well-test measurements are utilized in training. The initial parameter estimate is obtained with historical data before the 1st of January 2019, while the data after this point in time is used to test the learning methods, see Fig. 4. This split of data is referred to as the initial split. In Case 2, the models are trained using well-test

measurements only. To ensure a sufficient amount of training data, the initial split is applied on the 1st of January 2020, see Fig. 4.

Two analyses are conducted before the learning methods can be applied: (1) estimation of the PBL update frequency and (2) a search for optimal hyperparameters in the learning methods. These analyses are given in Sections 4.1 and 4.2, respectively, and are applied on the initial training data. From the outcome of the analyses, the models are trained with the learning methods, and the result for the two cases is given in 4.3 and 4.4, respectively.

4.1. Update frequency estimation for periodic batch learning

To estimate a suitable update frequency, Algorithm 1 in Appendix A with significance level $\alpha = 0.05$ is used on the initial training data from Case 1. This data is split into two new datasets at time 01.07.2018. The six months of observations leading up to 01.01.2019 are used as the test dataset. From Fig. 4, it is seen that W3 does not have observations in the time range suggested. Therefore, the well is excluded from the analysis. In Fig. 5, the HT^2 statistic for each observation in the test dataset is illustrated for four of the wells. The coloring indicates whether or not a shift is detected for the observation. W1 and W2 are the two wells of the nine examined with the longest period before a shift is detected, approximately after five months. W8 and W9 are the wells with the shortest period before a shift is detected, approximately after two weeks. Accordingly, different wells can have different optimal update frequencies, and it is likely to change during the lifetime of the petroleum asset. To simplify model learning, all wells are trained using the same update frequency. Therefore, the PBL is tested with a two weeks update frequency. The results are compared to a PBL with an update frequency of 6 months to examine the potential benefit of more frequent updating.

4.2. Hyperparameter search

For the periodic batch learning approach, a hyperparameter grid search for the learning rate is conducted testing $\gamma \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. Early stopping is applied to determine the appropriate number of iterations E . For all VFM models except the LR, the optimizer Adam is applied. This optimizer have shown results in previous research on VFM modeling (Grimstad et al., 2021; Hotvedt et al., 2020, 2021, 2022). For the LR, Adam is not an option and the model is trained with SGD, yet, with the learning rate scheduler

$$\gamma^{(k)} = \frac{\gamma^{(0)}}{k^a} \quad (21)$$

where $\gamma^{(0)}$ is the initial learning rate, k is the iteration number, and a is a constant, see (5).

For the online learning approach, the hyperparameter grid search is extended to

$\gamma \in \{5 \times 10^{-1}, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-10}\}$. Notice, $\gamma = 10^{-10}$ means close to negligible updating. As online learning processes only one sample at a time, early stopping cannot be applied. Therefore, the hyperparameter search includes experimentation with the number of iterations $E \in \{1, 10, 20\}$. For all models except the LR, the optimizers SGD and Adam are examined. For the LR, the learning rate scheduler (21) along with a constant learning rate is investigated.

The best combination of hyperparameters is chosen as the set that minimized the mean absolute percentage error (MAPE) across the wells for each model type. The resulting hyperparameters for Case 1 and Case 2 can be seen in Tables B.1 and B.2, respectively.

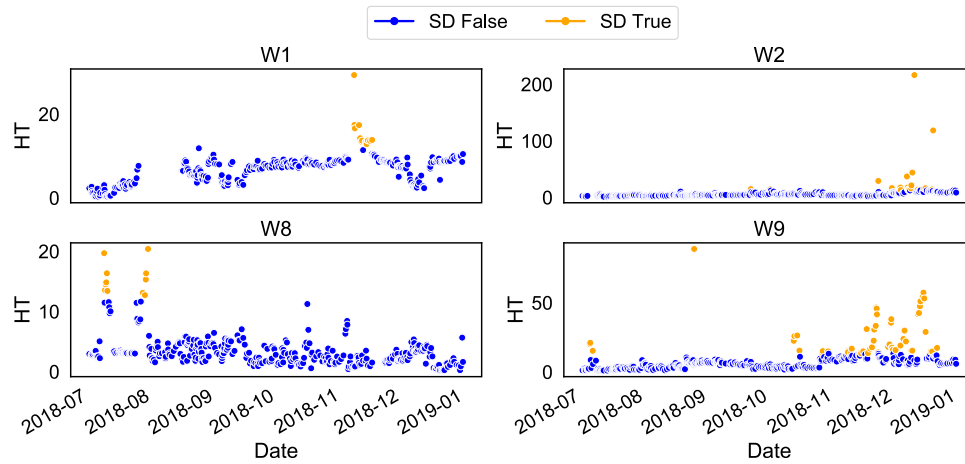


Fig. 5. The Hotelling's T-squared statistic with $\alpha = 0.05$ for each of the observation in the test dataset. The coloring indicates if the test observation is detected as a shift (SD: shift detection).

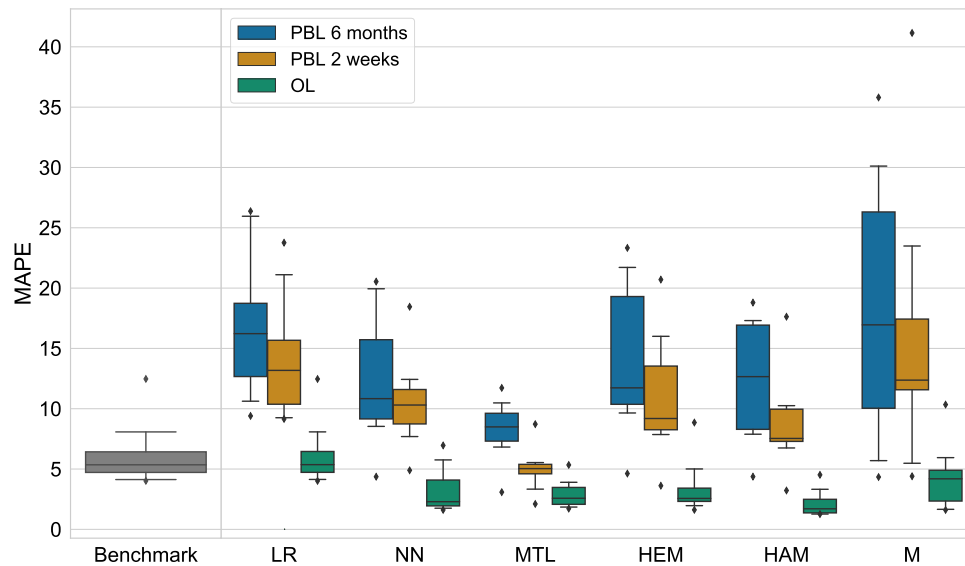


Fig. 6. The distribution of average error for each well, grouped for the models and learning methods. The models are trained with all available measurements. Compared to the performance of the benchmark model. The boxes show the P_{25} , P_{50} (median), and P_{75} percentiles. The whiskers show the P_{10} and P_{90} percentiles.

4.3. Results of case 1

In this case, both MPFM and well-test measurements are utilized in training. The box plot in Fig. 6 shows the distribution of performances for the wells in terms of the MAPE grouped on the model type and learning method. The reported MAPE for one well is calculated using the predictions on all observations in the initial test set. The models are compared to the benchmark model. Table 1 gives an overview of the average MAPE across the wells for each model, and the last column presents the average MAPE of the learning methods across all wells and models. For the interested reader, Table C.1 gives a detailed overview of the MAPEs for each well and model. There are several interesting observations.

Firstly, as expected, the results clearly show that the model error decreases with an increased update frequency. On average, all models achieve a lower prediction error with PBL every second week compared to PBL every six months. With the OL, the average error decreases further with all models achieving an average error of less than 7%. The overall best average performance across wells is achieved with OL on the HAM. The low MAPEs indicate that with access to frequently arriving measurements such as MPFM measurements, and allowed to learn continuously from them, the learning problem is relatively simple

Table 1

Average mean absolute percentage error across the wells for the models and learning methods trained on both MPFM and well-test measurements. The last column is the average MAPE across all wells and models.

Learning method	LR	NN	MTL	HEM	HAM	M	All
PBL 6 months	16.8	12.4	8.3	14.2	12.4	18.1	13.7
PBL 2 weeks	14.2	10.5	5.0	10.9	8.7	15.7	10.8
OL	6.2	3.2	2.9	3.4	2.1	4.2	3.7

and a complex model is not necessary to achieve high VFM accuracy. This is supported by the good performance of the Benchmark which outperforms nearly all models trained with PBL. On the other hand, a disadvantage with the Benchmark is that it cannot be used for sensitivity analyses or in production optimization.

Secondly, from Table C.1 it is observed that there are large differences in the error reduction for each well when the update frequency is increased. For instance, for W9 and most models, the error is greatly reduced going from the PBL 6 months to the OL. On the other hand, for W1 the reduction is not as prominent. This is likely related to whether or not the data generating distribution shifts with time. In Fig. 7, the Hotelling's T-squared statistic is plotted for W1 and W9 using Algorithm

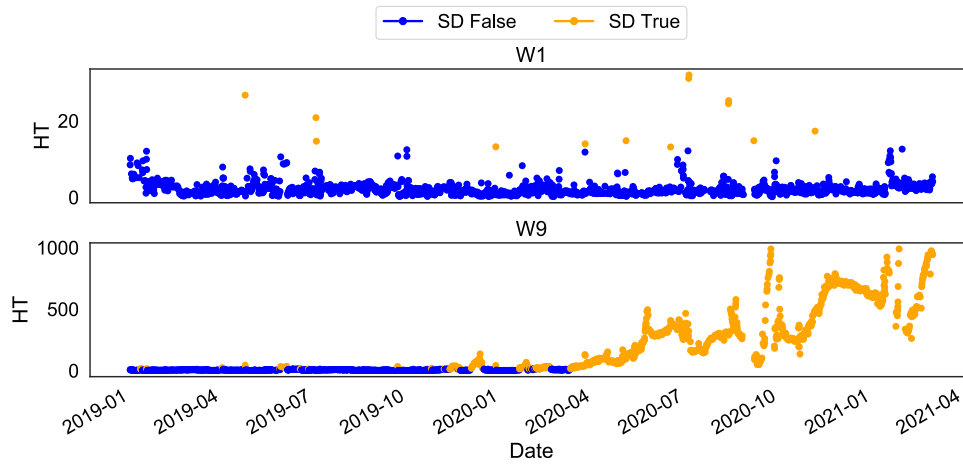


Fig. 7. The Hotelling's T-squared statistic for W1 and W9 comparing each observation from 01.01.2019 and forward with time to the training data containing the historical data before 01.01.2019. As seen, for W1 most observations are not detected as shifts. Whereas for W9, all observations towards to end are marked as a shift.



Fig. 8. The rolling absolute mean error across the wells against time for the models and learning methods. The window size used to calculate the error is 14 days. The shaded region illustrates the 25 and 75 percentiles of the errors across the wells. The vertical lines illustrate where the models are updated for the PBL 6 months and PBL 2 weeks.

1 on the initial training and test data. Fig. 7 indicates that it is unlikely that W1 experiences dataset shifts. On the other hand, for W9 it can be observed that the data likely shifts with time. Therefore, the results in Table C.1 indicate that OL is better at tracking the local optimum of the learning problem when it changes with time.

Another figure that illustrates the benefit of updating the model more often is Fig. 8, where the prediction error is visualized against time. The error is calculated as a rolling absolute mean error with a window size of 14 days. The shaded regions visualize the 25 and 75 percentiles of the errors across the wells. Notice that the PBL seems to yield a cyclic high and low accuracy. The average error increases with time up until model updating where the average error is reduced, naturally after some delay due to the rolling window. This is best observed for the PBL 6 months, but also to some extent for the PBL 2 weeks.

Table 2

The average mean absolute percentage error across the wells for the models and learning methods trained on well-test measurements. The last column is the average MAPE across all wells and models.

Learning method	LR	NN	MTL	HEM	HAM	M	All
PBL 6 months	43.2	40.7	12.1	17.9	18.3	21.8	25.7
PBL 2 weeks	37.1	31.1	9.6	24.5	16.7	20.5	23.3
OL	44.3	31.4	10.5	18.7	11.6	17.7	22.4

4.4. Results of case 2

In this case, the models are trained on well-test measurements only, see the observations colored orange in Fig. 4. Fig. 9 illustrates the distribution of MAPEs for the wells. Table 2 gives an overview of the

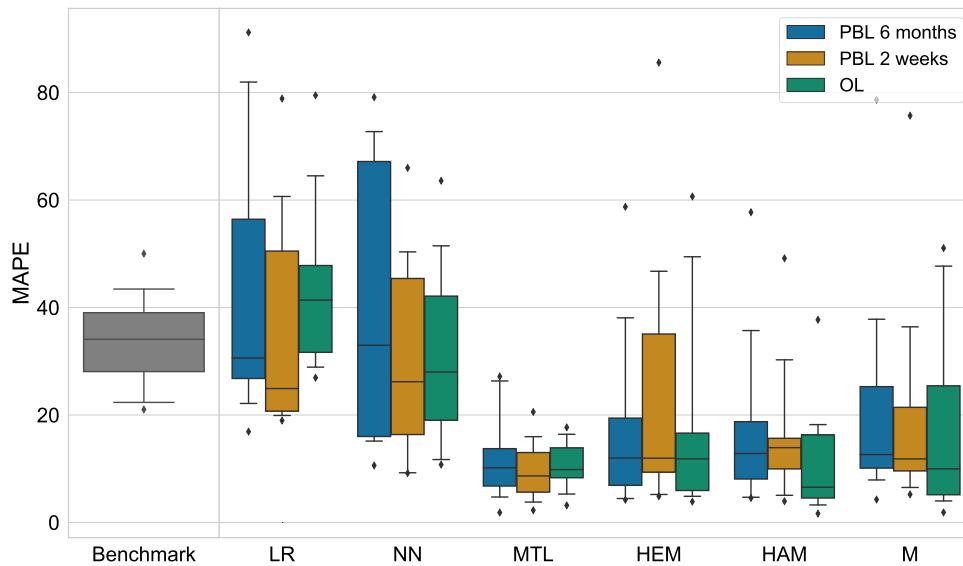


Fig. 9. The distribution of mean absolute percentage error (MAPE) for each well, grouped for the models and learning methods. Here the models are trained using well-test measurements only. Compared to the performance of the benchmark model. The boxes show the P_{25} , P_{50} (median), and P_{75} percentiles. The whiskers show the P_{10} and P_{90} percentiles.

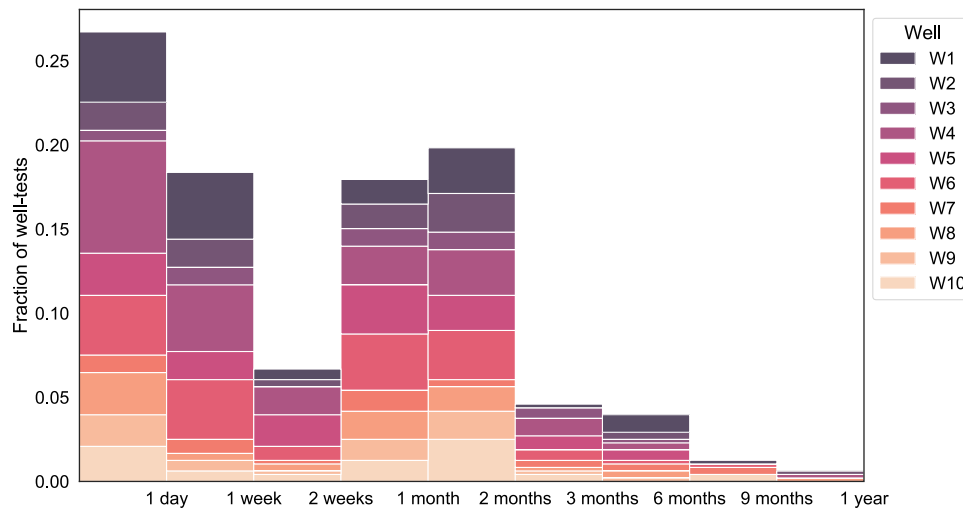


Fig. 10. The elapsed time between new well-tests for each well present in the dataset. Keep in mind that the bins in the histogram have different sizes, from one day to three months of elapsed time. Observe that many of the wells have measurements occurring more than two weeks apart.

average MAPE across the wells. Table C.2 reports the MAPE for each well, model, and method. First of all, notice the significantly different results obtained for this case compared to Case 1. In Case 1, a trend of decreased error for increased update frequency is observed. Here, the difference in performance is negligible for many models and for other models the error increases going from PBL to OL. The observed results are likely related to the elapsed time between each new well-test, illustrated in Fig. 10 by a stacked histogram. Notice that many of the wells have several tests that are more than a month apart. Furthermore, eight of ten wells have the majority of tests occurring with a frequency lower than 14 days, see Table C.3. In such situations, the frequency of model updating is equal for PBL 2 weeks and OL, and the only difference between the two is how the updating is executed. The low frequency of well-tests is also likely the cause of the decreased Benchmark performance compared to Case 1. With a lower frequency, the process conditions can have changed significantly in-between well-tests and two chronological flow rate measurements are likely uncorrelated. The intermittent time between well-tests also makes it challenging to obtain good hyperparameters. If well-tests occur frequently, the model

will likely require small parameter updates, and opposite for infrequently arriving well-tests. Non-optimal hyperparameters can explain the overall poorer average performance for all models and methods than for Case 1.

The large MAPEs in Table 2 show that in the presence of infrequent and intermittent measurements, the learning problem is not trivial and a more complex model than, for instance, the Benchmark, is required to obtain an adequate performance. Nevertheless, the comparable performance of the LR and NN indicate that choosing a data-driven model with higher complexity is not the solution to increased performance in this case. Likely, the amount of data available is too small for high-capacity data-driven models to exploit their capacity. An observation that supports this is that the NN obtained a significant improved performance in Case 1 where the amount of data is higher. On the other hand, having physical considerations in the model structure does seem to be advantageous for VFM when the amount of data is small. From Fig. 9, the MTL, HEM, HAM, and M, all achieve a median MAPE below 20% whereas the NN and LR are well above 20%. Another interesting observation is that learning from several wells as for the MTL seems to yield a more robust approach as the spread in performances is low.

Table B.1

The training algorithm settings as a result of the hyperparameter search for training on both MPFM and well-test measurements. For the batch learning approaches, only the value of the learning rate γ is experimented with. The number of iterations E for PBL is found with early stopping (E.S.). The (s.) and (c.) for the LR refers to using the learning rate scheduler in (21) and constant learning rate, respectively.

Model	PBL 6 months			PBL 2 weeks			OL		
	γ	E	O	γ	E	O	γ	E	O
LR	(s.) 10^{-2}	E.S.	SGD	(s.) 10^{-1}	E.S.	SGD	(c.) 0.5	20	SGD
NN	10^{-4}	E.S.	Adam	10^{-3}	E.S.	Adam	10^{-5}	20	Adam
MTL	10^{-4}	E.S.	Adam	10^{-3}	E.S.	Adam	10^{-6}	20	Adam
HEM	10^{-3}	E.S.	Adam	10^{-3}	E.S.	Adam	10^{-2}	20	SGD
HAM	10^{-3}	E.S.	Adam	10^{-3}	E.S.	Adam	10^{-5}	20	SGD
M	10^{-3}	E.S.	Adam	10^{-3}	E.S.	Adam	10^{-2}	10	Adam

Table B.2

The training algorithm settings as a result of the hyperparameter search for training on only well-test measurements. For the batch learning approaches, only the learning rate γ is experimented with. The number of iterations E for PBL is found with early stopping (E.S.). The (s.) and (c.) for the LR refers to using the learning rate scheduler in (21) or constant learning rate, respectively.

Model	PBL 6 months			PBL 2 weeks			OL		
	γ	E	O	γ	E	O	γ	E	O
LR	(s.) 10^{-4}	E.S.	SGD	(s.) 0.5	E.S.	SGD	(s.) 10^{-3}	1	SGD
NN	10^{-4}	E.S.	Adam	10^{-3}	E.S.	Adam	10^{-4}	20	SGD
MTL	10^{-5}	E.S.	Adam	10^{-5}	E.S.	Adam	10^{-5}	20	Adam
HEM	10^{-3}	E.S.	Adam	10^{-4}	E.S.	Adam	10^{-10}	20	SGD
HAM	10^{-5}	E.S.	Adam	10^{-5}	E.S.	Adam	10^{-5}	20	Adam
M	10^{-3}	E.S.	Adam	10^{-3}	E.S.	Adam	10^{-2}	10	Adam

5. Concluding remarks

The key takeaways from this study are two-fold. Firstly, a high update frequency of steady-state VFM models is essential to sustain high performance over time under nonstationary conditions. The results from Case 1 in Fig. 6 clearly document the benefit of frequent model updating when the frequency of measurement arrival is high. Of the two passive learning methods analyzed, online learning achieves the best average performance with an error of 3.7% across all wells and model types. Periodic batch learning with an update frequency of six months obtained the worst average performance of 13.7%. The benefit of frequent updating is less evident for Case 2 in Fig. 9 where the measurements arrive intermittently and infrequently. Yet, Table 2 illustrates an increasing VFM average performance when the update frequency increases. The online learning approach yields a 13% error reduction compared to the periodic batch learning every six months. The decrease in average performance for all models and methods for Case 2 is likely due to the smaller amount of available data to learn from.

Secondly, in the presence of intermittently and infrequently arriving measurements, as for Case 2, Fig. 9 distinctly illustrates that knowledge regarding the process physics is fundamental to obtaining a high performance of the steady-state VFM. The best performing is the multi-task learning model with periodic batch learning every second week with an error of 9.6% on average. The best data-driven model was the neural network with periodic batch learning every second week with an average error of 31.1%. Keep in mind, the learning methods are only investigated for ten offshore wells on the Norwegian continental shelf. Therefore, the results can change for production wells with contrasting multiphase flow characteristics.

The overall conclusion of this study is that steady-state VFMs operating in nonstationary conditions can benefit greatly from passive learning with frequent model updating. As discussed in Section 1, many industrial VFMs are based on steady-state models, and these models are typically calibrated manually and infrequently. Considering this, the results demonstrate a large potential for improved performance of existing VFMs by implementing passive learning for automating

the model calibration procedure. Automation of new or existing VFMs would require an infrastructure for data collection and processing, model training, model deployment, and model performance monitoring to ensure a frequent and robust model updating process. To the authors' knowledge, few, if any, industrial VFM solutions have achieved this level of automation. Therefore, the results and considerations in this study can be of interest to experts developing new or existing VFMs, or other applications of soft sensors in nonstationary environments.

Case-specific considerations must be made when choosing a passive learning method. Online learning offers the highest frequency of model calibration, and correspondingly demands the most of the infrastructure in terms of fast processing capabilities. An alternative is to use periodic batch learning with an updating period matching the expected frequency of significant process change. In this study, model calibration every second week was found to give good performance. Finally, this study found the tuning of hyperparameters of the periodic batch learning method more straightforward than the online learning method. Hence, practitioners are recommended to implement periodic batch learning before attempting to automate model calibration with online learning.

CRedit authorship contribution statement

Mathilde Hotvedt: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Bjarne A. Grimstad:** Conceptualization, Methodology, Software, Data curation, Writing – review & editing, Supervision. **Lars S. Imsland:** Writing – review & editing, Supervision.

Declaration of competing interest

One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to <https://doi.org/10.1016/j.eswa.2022.118382>. The authors Mathilde Hotvedt and Bjarne Grimstad are affiliated with and hold stocks in Solution Seeker AS. Solution Seeker develop and delivers VFM solutions to the petroleum industry. There is a conflict of interest in that the authors may gain financially from publishing good results and may therefore be inclined to manipulate or provide a biased interpretation of the results. We would therefore like to firmly state that we believe in open research and are against such unethical research practice. On the other side, we understand that it cannot and should not remove all doubt. To remove bias from our work, we have a co-author: Prof. Lars S. Imsland at the Norwegian University of Science and Technology, who is not affiliated with Solution Seeker. Prof. Imsland has reviewed and edited the interpretation of the results and conclusions to this study.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors would like to thank Lundin Energy Norway for allowing them to work with production data from a real petroleum field. They would further like to thank Solution Seeker AS for the contribution with data collection and pre-processing.

Funding

This research is supported by Lundin Energy Norway and 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). Lundin Energy Norway has not taken part in data collection and analysis, nor in writing of the report. However, the article is approved by Lundin before submission for publication.

Table C.1

Mean absolute percentage errors for all wells and models. The triple of numbers reported is the error for periodic batch learning every 6 months, periodic batch learning every 2 weeks, and online learning, respectively.

Well	LR	NN	MTL	HEM	HAM	M
1	9.4, 9.3, 12.5	4.4 4.9, 1.6	3.1, 2.1, 1.7	4.6, 3.6, 2.0	4.4, 3.2, 1.3	4.3, 4.4, 1.7
2	26.0, 21.1, 5.4	16.4, 18.5, 5.8	10.5, 8.7, 5.3	19.4, 20.7, 8.9	18.8, 17.6, 4.5	35.8, 41.2, 10.3
3	12.6, 9.8, 4.0	10.0, 11.7, 7.0	9.8, 4.9, 3.9	10.9, 8.8, 5.0	14.7, 10.2, 2.7	30.1, 11.7, 5.9
4	12.8, 12.0, 8.1	10.4, 9.8, 1.8	7.2, 5.0, 1.8	10.5, 9.2, 2.4	8.1, 7.3, 1.3	5.7, 5.5, 2.3
5	18.6, 16.0, 6.7	8.9, 8.5, 3.2	7.5, 3.3, 2.1	12.6, 9.2, 1.6	10.6, 7.6, 1.4	13.5, 11.5, 1.6
6	17.8, 14.8, 5.3	13.7, 11.1, 2.4	8.6, 5.4, 3.6	19.0, 16.0, 2.3	17.1, 10.3, 1.7	27.4, 23.5, 4.1
7	26.4, 23.8, 5.7	20.5, 12.4, 4.4	11.7, 5.1, 3.1	23.3, 11.9, 2.9	17.3, 9.1, 3.3	20.4, 16.5, 4.3
8	14.7, 12.6, 4.1	11.2, 10.8, 1.9	6.8, 5.5, 2.5	9.6, 7.9, 2.7	7.9, 6.7, 1.7	11.2, 12.8, 5.0
9	10.6, 9.2, 5.3	8.5, 7.7, 2.2	8.3, 5.3, 2.1	10.3, 8.1, 1.7	8.9, 7.4, 1.9	9.7, 12.0, 2.6
10	18.8, 13.7, 4.5	19.9, 9.4, 2.2	9.2, 4.5, 2.7	21.7, 14.1, 3.6	16.4, 7.3, 1.9	23.0, 17.8, 4.5

Table C.2

Mean absolute percentage errors for all wells and models trained on only test separator measurements. The triple of numbers reported is the error for periodic batch learning every 6 months, periodic batch learning every 2 weeks, and online learning, respectively.

Well	LR	NN	MTL	HEM	HAM	M
1	26.5, 19.9, 28.9	10.6, 9.3, 10.8	1.9, 2.3, 3.2	4.2, 5.2, 7.9	4.6, 4.0, 1.7	4.3, 6.5, 1.9
2	91.2, 78.8, 64.5	62.0, 66.0, 30.3	27.2, 16.0, 17.7	58.7, 85.6, 49.4	57.7, 49.2, 37.7	78.7, 75.7, 47.7
3	22.2, 20.0, 46.9	79.1, 25.4, 63.6	11.1, 10.7, 14.7	21.6, 46.8, 17.6	14.1, 14.7, 16.5	28.9, 23.8, 51.1
4	32.9, 24.4, 34.8	16.6, 14.7, 18.5	11.0, 10.1, 8.4	12.8, 10.6, 5.3	10.8, 9.9, 7.0	9.9, 10.5, 5.8
5	28.3, 25.4, 26.9	15.1, 9.2, 11.7	4.7, 3.8, 5.3	4.5, 4.9, 3.9	7.2, 10.2, 4.3	7.9, 5.2, 4.9
6	82.0, 60.7, 48.1	72.7, 47.0, 43.7	14.6, 13.8, 11.4	8.6, 11.8, 13.3	11.6, 13.9, 5.4	14.5, 13.1, 8.8
7	43.5, 50.8, 79.5	68.9, 50.4, 51.5	9.3, 6.4, 11.2	11.7, 18.2, 13.7	19.1, 16.0, 18.2	10.8, 14.2, 11.6
8	16.9, 19.0, 30.6	15.8, 21.3, 20.8	6.3, 5.4, 8.3	6.4, 12.2, 10.4	4.7, 5.1, 6.2	12.8, 9.5, 11.1
9	27.7, 22.9, 36.7	27.1, 27.0, 25.7	8.1, 7.2, 8.5	12.3, 9.0, 4.9	17.6, 14.0, 3.3	12.5, 9.9, 4.0
10	60.7, 49.8, 46.1	38.9, 40.5, 37.8	26.3, 20.6, 16.4	38.1, 40.7, 60.7	35.7, 30.3, 15.6	37.8, 36.4, 30.1

Table C.3

The percentage of well-tests for a well where the number of days between two chronological tests resides in the given bins. d=day, w=week, m=month, y=year.

Well	< 2w	2w-1 m	1 m-2 m	2 m-3 m	3 m-6 m	6 m-9 m	9 m-1y
1	60.9%	10.1%	18.8%	1.5%	7.2%	1.5%	0%
2	46.2%	17.9%	28.2%	0%	5.1%	0%	2.6%
3	36.4%	22.7%	22.7%	13.6%	4.5%	0%	0%
4	64.8%	12.1%	14.3%	5.5%	2.2%	0%	1.1%
5	48.3%	23.3%	16.7%	6.7%	5.0%	0%	0%
6	52.1%	21.9%	19.2%	4.1%	1.4%	1.4%	0%
7	41.7%	25.0%	8.3%	8.3%	8.3%	8.3%	0%
8	45.7%	22.9%	20.0%	2.9%	5.7%	0%	2.9%
9	44.8%	20.7%	27.6%	3.4%	3.4%	0%	0%
10	40.5%	16.2%	32.4%	5.4%	0%	5.4%	0%

Appendix A. Estimation of the update frequency in periodic batch learning

Consider the null hypothesis \mathcal{H}_0 to state that there is no virtual drift present in the data such that input distribution $p(\mathbf{x})$ does not shift with time. This \mathcal{H}_0 is also called the stationary hypothesis. The \mathcal{H}_1 hypothesis is the alternative hypothesis that there is a shift in the data. Mathematically:

$$\begin{aligned} \mathcal{H}_0 : & p_t(\mathbf{x}) = p_{t+\tau}(\mathbf{x}) \text{ for all } \tau > 0 \\ \mathcal{H}_1 : & p_t(\mathbf{x}) \neq p_{t+\tau}(\mathbf{x}) \text{ for any } \tau > 0. \end{aligned} \tag{A.1}$$

Consider two disjoint datasets \mathcal{D}_1 and \mathcal{D}_2 with size N_1 and N_2 and inputs observations $\mathbf{X}_1 \in \mathbb{R}^{d \times N_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{d \times N_2}$, respectively. The Hotelling’s T-squared statistic calculates the probability of equal means of the two multivariate input distributions at a significance level α . The statistic is calculated as

$$HT^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \left(\frac{\boldsymbol{\Sigma}_1}{N_1} + \frac{\boldsymbol{\Sigma}_2}{N_2} \right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T, \tag{A.2}$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the sample mean vector and the $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is the sample covariance matrix of the input. The Hotelling’s T-squared statistic follows the F-distribution $F(d, N_1 + N_2 - d - 1)$ (Hårdle & Simar,

2012). To estimate an appropriate update frequency in PBL, the two-sample Hotelling’s T-squared test can be used on available training data using Algorithm 1

Algorithm 1 Estimation of the update frequency τ in periodic batch learning

```

Require: data  $1 : T = \{(\mathbf{x}_t, y_t)\}_{t=1}^T$ , significance level  $\alpha$ 
1: Set  $l = 1 : T_1$  where  $1 < T_1 < T$ .
2: for  $k = 1, \dots, T - T_1$  do
3:    $2, k = T_1 + k$ 
4:   Calculate  $HT^2$  using (A.2) with  $l$  and  $2, k$ 
5:   Calculate F-statistic  $F_k$  for  $HT^2$ 
6:   Calculate the critical value  $F_{crit}$  at significance level  $\alpha$ 
7:   if  $F_k < F_{crit}$  then
8:     Reject  $\mathcal{H}_0$ , shift detected
9:     return  $\tau = T_1 + k$ 
10:  end if
11: end for
    
```

The algorithm is subject to false shift detections, or type II error, for instance, if the observation is faulty or noisy. A workaround is to test additional observations following with time. If shifts are detected on several subsequent observations, virtual drift has likely occurred. If the following observations are not detected as shifts, likely, the detection is falsely reported.

Appendix B. Hyperparameter search

See Tables B.1 and B.2.

Appendix C. Additional results from the numerical study

See Tables C.1 and C.2.

References

AL-Qutami, T. A., Ibrahim, R., & Ismail, I. (2017). 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* (pp. 304-309).

- AL-Qutami, T. A., Ibrahim, R., Ismail, I., & Ishak, M. A. (2017a). Development of soft sensor to estimate multiphase flow rates using neural networks and early stopping. *International Journal on Smart Sensing and Intelligent Systems*, 10, 199–222.
- AL-Qutami, T. A., Ibrahim, R., Ismail, I., & Ishak, M. A. (2017b). Radial basis function network to predict gas flow rate in multiphase flow. In *Proceedings of the 9th international conference on machine learning and computing* (pp. 141–146).
- AL-Qutami, T., Ibrahim, R., Ismail, I., & Ishak, M. A. (2018). Virtual multiphase flow metering using diverse neural network ensemble and adaptive simulated annealing. *Expert Systems with Applications*, 93, 72–85.
- Amin, A. (2015). Evaluation of commercially available virtual flow meters (VFM). In *Proceedings of the annual offshore technology conference* (pp. 1293–1318).
- Bikmukhametov, T., & Jäschke, J. (2019). Oil production monitoring using gradient boosting machine learning algorithm. *IFAC-PapersOnLine*, 52(1), 514–519.
- Bikmukhametov, T., & Jäschke, J. (2020). Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models. *Computers and Chemical Engineering*, 138.
- Bishop, C. (2006). *Pattern recognition and machine learning*. Springer.
- Ditzler, G., Roveri, M., Alippi, C., & Polikar, R. (2015). Learning in nonstationary environments: A survey. In *IEEE computational intelligence magazine* (pp. 12–25).
- Falcone, G., Hewitt, G., Alimonti, C., & Harrison, B. (2013). Multiphase flow metering: current trends and future developments. In *SPE annual technical conference and exhibition*.
- Foss, B., Knudsen, B. R., & Grimstad, B. (2018). Petroleum production optimization - a static or dynamic problem. *Computers and Chemical Engineering*, 114, 245–253.
- Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *14th international conference on artificial intelligence and statistics*, Vol. 15 (pp. 315–323).
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, Massachusetts: London, England: The MIT Press.
- Goodfellow, I., Mirza, M., Xiao, D., Courville, A., & Bengio, Y. (2013). An empirical investigation of catastrophic forgetting in gradient-based neural networks. ArXiv: 1312.6211.
- Granero-Belinchón, C., Roux, S. G., & Garnier, N. B. (2019). Information theory for non-stationary processes with stationary increments. *Entropy*, 21.
- Gravdahl, T., & Egeland, O. (2002). *Modeling and simulation for automatic control*. Marine Cybernetics AS.
- Grimstad, B., Gunnerud, V., Sandnes, A., Shamlou, S., Skrondal, I. S., Uglane, V., Ursin-Holm, S., & Foss, B. (2016). A simple data-driven approach to production estimation and optimization. In *SPE intelligent energy international conference and exhibition*.
- Grimstad, B., Hotvedt, M., Sandnes, A. T., Kolbjørnsen, O., & Imsland, L. S. (2021). Bayesian neural networks for virtual flow metering: An empirical study. *Applied Soft Computing*, 112.
- Guo, B., Lyons, W., & Ghalambor, A. (2007). *Petroleum production engineering - a computer assisted approach*. Elsevier Science and Technology Books.
- Härdle, W., & Simar, L. (2012). *Applied multivariate statistical analysis* (3rd Ed.). Springer.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York, USA: Springer.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In *Computer vision - ECCV 2016* (pp. 630–645).
- Holmås, K., & Løvli, A. (2011). FlowManager dynamic: A multiphase flow simulator for online surveillance, optimization and prediction of subsea oil and gas production. In *BHR Group - 15th international conference on multiphase production technology* (pp. 241–254).
- Hotvedt, M., Grimstad, B., & Imsland, L. (2020). Developing a hybrid data-driven, mechanistic virtual flow meter - a case study. *IFAC-PapersOnLine*, 53, 11692–11697.
- Hotvedt, M., Grimstad, B., & Imsland, L. (2021). Identifiability and interpretability of hybrid, gray-box models. *IFAC-PapersOnLine*, 54, 389–394.
- Hotvedt, M., Grimstad, B., Ljungquist, D., & Imsland, L. (2022). On gray-box modeling for virtual flow metering. *Control Engineering Practice*, 118.
- Jansen, J.-D. (2015). *Nodal analysis of oil and gas wells - theory and numerical implementation*. TU Delft, The Netherlands: Delft University of Technology.
- Jordanou, J., Antonelo, E. A., Camponogara, E., & Aguiar, M. (2017). Recurrent neural network based control of an oil well. In *XIII simpósio brasileiro de automação inteligente* (pp. 924–931).
- Kemker, R., McClure, M., Abitino, A., Hayes, L., & Kanan, C. (2018). Measuring catastrophic forgetting in neural networks. In *The thirty-second AAAI conference on artificial intelligence*, Vol. 1 (pp. 3390–3398).
- Kittilsen, P., Fjalestad, K., & Aasheim, R. (2014). Stabilized and increased well production using automatic Choke control. *Society of Petroleum Engineers*.
- Koroteev, D., & Tekic, Z. (2021). Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future. *Energy and AI*, 3, 1–10.
- Lundin Energy Norway (2020). Edvard grieg. <https://lundin-energy-norway.com/edvard-grieg/> (Accessed: 18 Jan 2020).
- Monteiro, D. D., Duque, M. M., Chaves, G. S., Filho, V. M. F., & Baioco, J. S. (2020). Using data analytics to quantify the impact of production test uncertainty on oil flow rate forecast. *IIFP Energies Nouvelles*, 75, 1–15.
- Oliveira, G., Minku, L., & Oliveira, A. (2021). Tackling virtual and real concept drifts: An adaptive Gaussian mixture model approach. *IEEE Transactions on Knowledge and Data Engineering*.
- Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, 113, 54–71.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimselshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Álchê Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems*, Vol. 32 (pp. 8024–8035). Curran Associates, Inc..
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). *Dataset shift in machine learning*. Cambridge, Massachusetts: London, England: MIT Press.
- Raza, H., Prasad, G., & Li, Y. (2015). EWMA model based shift-detection methods for detecting covariate shifts in non-stationary environments. *Pattern Recognition*, 48, 659–669.
- Sachdeva, R., Schmidt, Z., Brill, J. P., & Blais, R. (1986). Two-phase flow through chokes. In *Society of petroleum engineers, annual technical conference and exhibition*.
- Sandnes, A. T., Grimstad, B., & Kolbjørnsen, O. (2021). Multi-task learning for virtual flow metering. *Knowledge-Based Systems*, 232.
- Santos, R. J. (1996). Equivalence of regularization and truncated iteration for general ill-posed problems. *Linear Algebra and its Applications*, 236, 25–33.
- Sayed-Mouchaweh, M., & Lughofer, E. (2012). *Learning in non-stationary environments: methods and applications*. Springer.
- Shippen, M. (2012). Steady-state multiphase flow - past, present, and future, with a perspective on flow assurance. *Energy and Fuels*, 26, 4145–4157.
- Solle, D., Hitzmann, B., Herwig, C., Remelhe, P. M., Ulonka, S., Wuerth, L., Prata, A., & 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. In *Offshore technology conference*.
- Varyan, R., Haug, R., & Fønnes, D. (2015). Investigation on the suitability of virtual flow metering system as an alternative to the conventional physical flow meter. In *SPE, APOGCE* (pp. 2–10).
- Wickliffe, C. A., & Robins, A. (2005). Memory retention - the synaptic stability versus plasticity dilemma. *Trends in Neurosciences*, 28, 73–78.