EasyChair Preprint

№ 11115

Ensemble Models Based on System Identification and Machine Learning for Downhole Pressure Simulations

Daniel de Marins and Helon Ayala

October 23, 2023

Ensemble Models Based on System Identification and Machine Learning for Downhole Pressure Simulations [⋆]

Daniel B. de Marins [*] **Helon Ayala** [**]

[*] *Department of Mechanical Engineering Pontifical Catholic University of Rio de Janeiro, RJ (e-mail: dboechat.m@gmail.com)*
[**] *Department of Mechanical Engineering Pontifical Catholic University of Rio de Janeiro, RJ (e-mail: helon@puc-rio.br)*

Abstract: The digital era has come, the use of data analytics, cloud applications and environments is growing faster than years ago in oil and gas subject. The industry own a huge quantity of data storage from Instrumentation, logging and sensors. In the other side reservoir simulation has a vital role in petroleum engineering due it application to fluids flow rates forecasting specially for exploitation strategic definition. The complexity of using sophisticated mathematical models can define a specific simulation study scenario and even with simplifications it needs the domain knowledge in different areas such as reservoir, production, engineering and geosciences. Some ideal assumptions can simplify the differential equations making not representative of the complex behavior of the underground fluid and without these assumptions the reservoir complexity increase greatly. In this paper, we use a public data from Volve field corresponding to the production years 2008 to 2018. We built an assemble model based on system identification, with a Non Linear Auto-Regressive (NARX) model and Artificial Neural Network (ANN) architecture to simulate reservoir model and forecast real time the downhole pressure for short-term decisions. The ensemble method did not provide better prediction compared to standalone short-term simulations techniques. We use the initial dynamic data to build a consistency model able to downhole pressure forecasting helping the oil production optimization with significant error decrease.

Keywords: NARX, artificial neural network, hybrid model, ensemble, production simulation, production data, volve field

1. INTRODUCTION

The use of the data driven approach in Oil and gas area in increasing powered by the machine learning techniques and power processing evolution (Koroteev and Tekic (2021)). Reservoir simulation and production engineering request a huge quantity for the decision making strategy (Onalo et al. (2019)). Usually these areas request reservoir information for building complex mathematical models to discover any physical relationship (Li et al. (2019)). In the real case considering all relation between properties require a huge computational power. So some ideal assumptions can make the simplified equations unrepresentative of the complex subsurface fluid behaviors (de Oliveira Werneck et al. (2022)).

Many companies already own a consolidated data storing pipeline from sensors measuring pressures, temperatures, vibrations, flow rates and logging. They are dealing with huge volume of data but, in the most of cases, without processing data capability (Mohammadpoor and Torabi (2020)). The oil industry plant digitalization help to reach higher efficiency, improve the fault detection and controlling and the income, production or flow rate forecasting (Aggoun and Chetouani (2021); Bo et al. (2020); Solanki et al. (2022); Lei et al. (2020)).

In their technical paper, Koroteev and Tekic (2021) conducted a thorough analysis of the potential applications of AI, focusing several examples such as the use of real-time drilling telemetry to detect rock type and predict potential failures through a dedicated tool, and the optimization of production efficiency through a data-driven tool that provides objective forecasts for well treatment campaigns.

Several models based on deep learning and recurrent neural networks were developed with different architectures, input data and combinations of characteristics present in the data set to estimate bottom hole pressure throughout the production period (Alakeely and Horne (2021); Li et al. (2019); Chen et al. (2021)). The impacts of using different pre-processing techniques were analysed, various configurations of stacked Recurrent Neural Networks (RNN) were also varied, windows and time scales arranged for training these models using real data and synthetic benchmark data (de Oliveira Werneck et al. (2022)). The optimization attempts were made with the input variables present in the

[⋆] This work was conducted in association with the ongoing Project registered under human resources program for the oil and gas sector (PRH-10) powered by ANP/FINEP, Brazil.

database, such as temperature, gas-oil ratio, valve opening percentage, gas flow in order to obtain a set of data of minimum inputs that allow the generation of a model that represents the phenomenon of interest well and at the same time has the lowest possible complexity.

Alakeely and Horne (2021) evaluated the impacts of using autoencoders in feature extraction or dimensionality reduction for forecasting well liquid and multiphase restricted flow rate using wellhead surface measurements and demonstrated the application to real field data. They built several architectures using feed forward, recurrent neural networks and auto encoders.

Heghedus et al. (2019) developed a model prediction for well flow rate time series based on pressure time series from Permanent Down hole Gauges wells using the nonlinear autoregessive (NARX) and the long short term memory (LSTM) neural networks were assembled and tested on a synthetic data set to compare results of pressure prediction.

Zha et al. (2022) studied the use of combining for forecasting CNN and LSTM models monthly gas field production. In their research the CNN is used for automatic feature extraction to simplify feature extraction, while LSTM is used to learn the sequence dependence of the time series prediction.

In another study, Siavashi et al. (2022) proposed an upscaling approach for macroscopic properties of single and two-phase flow. The approach combines CNNs and downsampling techniques to improve the prediction accuracy.

Lei et al. (2020) employed the NARX model to estimate real-time flow rates in multi-product pipelines. The authors developed data-driven adaptive models at local pressure mutation points and demonstrated that the NARX model outperforms standard recurrent neural networks by maintaining input pressure signals for two to three times longer. This allows the NARX model to effectively simulate the time-delay characteristics of flow processes, making it a promising tool for real-time flow rate estimation in the oil and gas industry.

However, the possibility of developing dynamic models was identified, and no longer static as seen in the literature, for solving problems involving estimation of some target variable such as oil production flow over time or identification of patterns present in a data set for fault and anomaly detection for example. It has already been shown these NARX models allow the detection of patterns of a given time series (Tian and Horne (2017)). This technique uses input and output delays to develop a dynamic model capable of generating a reliable output from estimated parameters and errors between the true value and those obtained.

In the other side, the usual machine learning approach for any measurement foresting like flow rate, oil and gas production rate are based on the soft sensor technology. They are commonly used for determine physical quantities and require minimal process knowledge (Lei et al. (2020)). The main problem of this approach is the need of updated input information and for real time this can be quite challenging.

To overcome the limitations of soft sensor approach, a hybrid method is proposed to improve the accuracy of the oil production estimation. The NARX model and ANN are adopted for use as data-driven adaptive models.

The objective of this paper is to introduce a novel ensemble data-driven approach that can aid short-term decision-making for well production and pressure simulation in the oil and gas industry.

The proposed solution employs a window prediction approach based on the initial record to develop a model that can predict upcoming production rates and pressures for several days without relying on real-time data inputs. This approach aims to enhance the accuracy of short-term predictions, allowing for more effective decision-making.

To develop the model, historical production data was used for calibration and validation purposes. The dataset was divided into a training sample to train the model and a test sample to evaluate its accuracy. The resulting model was able to predict future production rates and pressures with a high level of accuracy, making it a potentially valuable tool for optimizing well production in the oil and gas industry.

This work is organized as follows: in Section 2 is presented the description of the case study, in Section 3 it is explained the adopted methodology, in Section 4 we show the achieved results. Finally, in Section 5 we state the conclusion.

2. CASE STUDY DESCRIPTION

Regarding the dataset, it is from a real-life well in Volve field (one of the latest databases released by Equinor (2018) to the public for research purposes) used to build the models. The details regarding the data will follow later. A portion of the data from the well is employed to develop the models whereas the remaining part of the data is used as the blind case to further verify the predictive performance of the models.

In May 2018, Equinor (2018) provided a dataset from a real-life well in Volve field to the public for research purposes used to build the models. It includes geological, geophysical, and reservoir engineering data from the Volve oil field, which is located in the North Sea, approximately 200 km west of Stavanger, Norway.

The Volve dataset includes well logs, seismic data, production data, and reservoir engineering data, among other types of information. It is intended to provide a comprehensive and realistic dataset for testing and benchmarking subsurface technologies and workflows, and to promote collaboration and innovation in the oil and gas industry.

Regarding the production data, it consists of the data of 7 wells and each well consists of the data, like On stream hours, Average Downhole Pressure and Temperature, Average Choke Size Percentage and others.

In this study, we utilized data from well NO 15/9-F-1 C in the Volve field for illustrative purpose, which has been made available for research purposes by Equinor (2018). For this well, the production period lasts from April 2014 to April 2016, encompassing a range of information relevant to the study, this result in 741 daily records for all mentioned variables.

Although this well have information about several variables we focuses in studding Average Downhole Pressure, Average Downhole Temperature, On stream hours, Oil, Gas and Water Volume from Well and Average Choke Size Percentage only.

To enhance the readers' comprehension of the production scenario, we have computed and presented the mean, standard deviation (SD) minimum and maximum values of each variable in table 1.

Input and Output	Mean	SD	Min.	Max.
Avg Pressure (bar)	247.33	27.99	207.22	313.87
Avg Temp. (°C)	105.20	3.44	95.87	108.50
On stream hours	13.47	11.64	0.00	25.00
Avg Choke Size (%)	29.79	25.45	0.00	93.63
Gas Volume (m^3)	36.84	44.64	0.00	521.00
Water Volume (m^3)	235.11	309.89	0.00	991.00
Oil Volume (m^3)	219.80	245.35	0.00	245.35

Table 1. Mean, standard deviation, minimum and maximum of input and output parameters of the production case considering all the data points.

In figure 1, we present the correlations between the variables in our dataset. The distribution of average downhole pressure is found to be less concentrated than other variables such as stream hours and average downhole temperature. Notably, we observe a stronger relationship between average choke opening and the other variables.

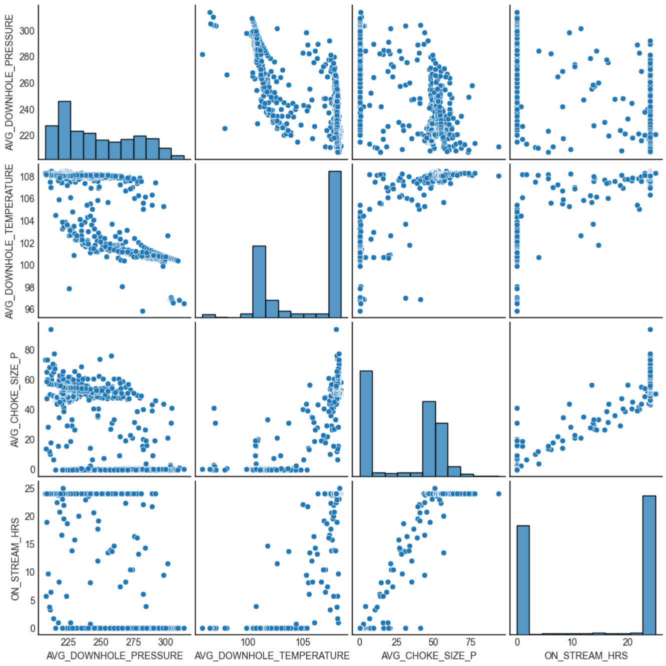

Figure 1. The pairwise relationships and distribution of average temperature, pressure, choke size opening, and stream hours in the volve dataset, highlighting potential correlations and patterns.

3. METHODOLOGY

In this work, we applied different numerical procedures to simulate the downhole temperature. We used the system identification technique ***A***uto***R***egressive with e***X***ogenous input and the ***N***onlinear ***A***uto***R***egressive with e***X***ogenous input (NARX) Billings (2013) and Neural NARX (NNARX) to assemble two simulations strategies.

The ARX model is described as exposed in Eq. 1.

$$\begin{aligned} y(k)+&a_1y(k-1)+\cdots+a_{na}y(k-na) = \\ &b_1u(k-1)+\cdots+b_{nb}u(k-nb)+\boldsymbol{\xi}(k), \end{aligned} \tag{1}$$

where $y(k)$ is the $k-th$ sample from the output signal, $u(k)$ is the $k-th$ sample from the input signal and a and b are the parameters that tune the model. na and nb are the number of model coefficients. We can consider the model in terms of a product of matrices. Therefore, we present the input and the output as vectors, $\boldsymbol{u}$ and $\boldsymbol{y}$. The parameters are sorted in a parameter vector $\boldsymbol{\Theta}$. The signal considered is composed of N samples. Then, we take the auxiliary number $p = 1 + max(na, nb)$. The vectors are defined as exhibited in Eq. (2).

$$\begin{aligned} \boldsymbol{y} &= [y(p), y(p-1), \cdots, y(N)] \\ \boldsymbol{\Theta} &= [a_1, \cdots, a_{na}, b_1, \cdots, b_{nb}] \end{aligned} \tag{2}$$

It is necessary define a regression matrix, $\boldsymbol{\Phi}$, to compose the system. The matrix is filled with the vectors

$\boldsymbol{\phi}(k) = [-y(k-1), \cdots, -y(k-na), u(k-1), \cdots, u(k-nb)]$. Then, the matrix is as follows in Eq. (3).

$$\boldsymbol{\Phi} = [\boldsymbol{\phi}^T(p), \boldsymbol{\phi}^T(p+1), \cdots, \boldsymbol{\phi}^T(N)]^T \tag{3}$$

Therefore, the system may be rewritten as in Eq. (4).

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\Theta} + \boldsymbol{\xi} \tag{4}$$

The solution $\hat{\boldsymbol{\theta}} = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}\boldsymbol{y}$ defines the ARX model Billings (2013).

Lastly, a Neural Nonlinear AutoRegressive eXogenous (NNARX) was used to simulate the error as present in (5).

$$\boldsymbol{error(k)} = y(k) - \hat{y} \tag{5}$$

The NNARX model consists of several layers of neurons, including input, output, and hidden layers. In this model, past outputs and input variables are used as inputs to predict the future output of the system. The model learns the relationship between past inputs and outputs to predict future outputs. The neural network adjusts its weights during the training process to minimize the error between the predicted and actual outputs. NNARX models have been shown to be effective in a variety of applications, including flow rate estimation in pipelines (Lei et al. (2020)), stock price prediction (Agung (2022)), and power grid forecasting (Sharkawy et al. (2023)).

4. RESULTS

We first attempted to simulate the average downhole pressure using an ARX model varying the variables used as inputs, as well as the number of model coefficients *na* and *nb*. The ARX and NARX estimation has been done by the SysIdentPy python library (Junior et al. (2020)).

Although the relative good performance for the ARX, it was not capable to incorporate all the system dynamic verified in the validation tests as plotted in figure 2. The developed model was built using average downhole temperature as input and $na = nb = 15$ and result in RMSE = 0.3305 and R^2 = 0.9082.

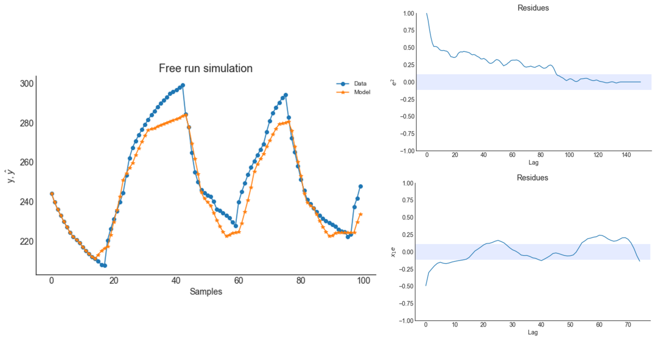

Figure 2. Free Run (FR) simulation for the best developed ARX model.

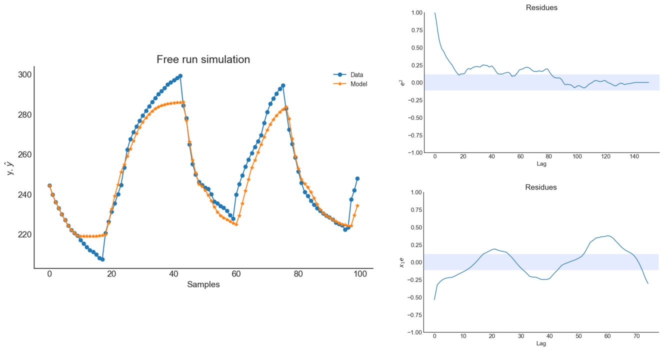

Figure 3. Free run simulation for the best developed NARX model.

To enhance the NARX model, we implemented a polynomial NARX model varying the degrees of the model. The model utilized the average downhole temperature as input with optimal hyper parameters of $na = nb = 10$, as shown in Figure 3. We set *ylag* and *xlag* to the same order number value of 10 in this case. Despite the ability of the model to incorporate nonlinear behavior, the results were light better than those obtained from the ARX model, as evidenced by the RMSE of 0.2841 and R^2 of 0.9264.

The Error Reduction Ratio (ERR) for the regressors of that responsible for the biggest reduction are shown in Table 2.

Figure 4 shows that using output feedback resulted in good performance in both one-step-ahead and free run simulations. One-step-ahead simulation involves predicting the output of a system at each time step using historical data, while free run involves predicting the output over a longer period without any knowledge of the inputs or initial conditions. Verification of the simulations was done by using raw inputs.

Regressors	Parameters	ERR
0	y(k-1)	1.1837E+00
1	y(k-9)y(k-3)	1.8132E-07
2	x1(k-1)y(k-2)	1.8155E-03
3	y(k-2)y(k-1)	-6.4768E-04
4	x1(k-2)ˆ2	1.9229E-03
5	x1(k-3)	-1.1813E-01
6	x1(k-20)x1(k-18)	-7.1323E-04
7	y(k-19)y(k-6)	-2.5525E-03
8	x1(k-10)x1(k-5)	2.2329E-03
9	x1(k-1)y(k-10)	-2.0845E-03

Table 2. ERR according to each selected regressor

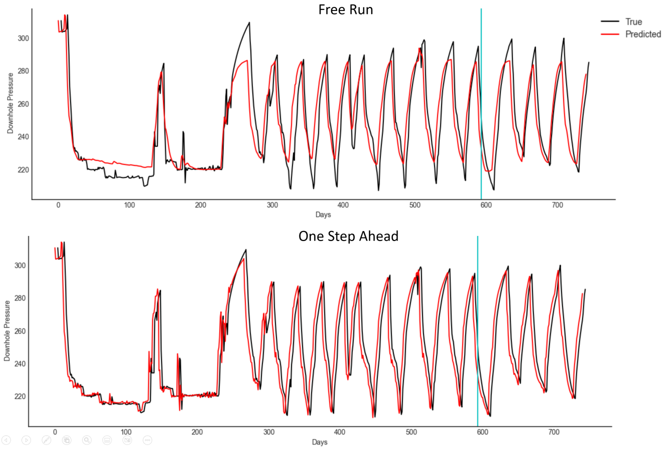

Figure 4. Comparison of one-step ahead simulation and free run methods for system identification.

One final attempt was made to simulate the NARX model error using the NNARX model, using the choke size opening as the input variable. However, despite the increased ability of the neural network to recognize the system dynamics, this did not result in an improvement in accuracy as plotted in figure 5 and 6. In this study, it was found that mapping the error signal was not a trivial task, and the observed increase in noise and floating point error was likely attributed to the inherent chaotic behavior of the system.

For this study, we used a neural network with 30 neurons in the input layer and three hidden layers, each consisting of 100 neurons with dropout layers set to 0.2. We employed the hyperbolic tangent (*tanh*) function as the activation function and a learning rate of $1x10^{-3}$. Early stopping was used with a patience parameter of 16. The neural network was trained using 128 epochs and a batch size of 32.

5. CONCLUSION

In this study, we developed ARX, NARX, and NNARX hybrid models to simulate the average downhole pressure of a well in the Volve field. The models were trained and tested to ensure that they accurately captured the input-output relationship before conducting blind validation.

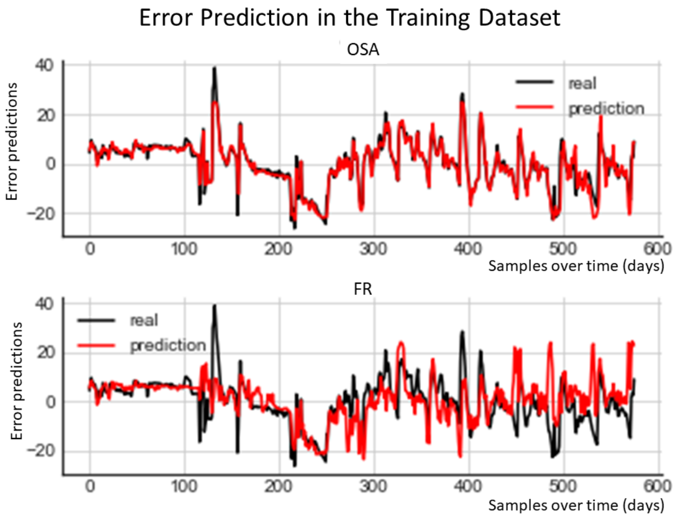

Figure 5. Comparison of one-step ahead simulation and free run methods including the NARX error simulation model in the train dataset

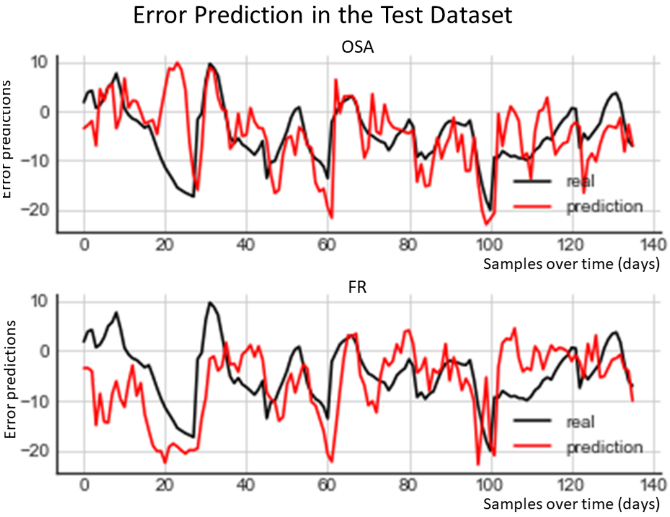

Figure 6. Comparison of one-step ahead simulation and free run methods including the NARX error simulation model in the test dataset

This study demonstrate the potential benefits of using system identification techniques for production oil optimization by providing accurate and timely insights into the behavior of the production system. This approach can help operators to improve the efficiency, reliability, and profitability of their operations.

Although we attempted to improve the simulation results by using a more complex model to estimate the NNARX estimation error, it did not yield the expected improvements. The simulation error proved to be challenging to manage using the neural NNARX approach for system identification due to the inherent chaotic behavior of the system.

Our analysis revealed that the use of system identification techniques can improve production oil optimization by accurately predicting the behavior of the production system, the model was able to identify optimal production strategies that resulted in higher production rates and reduced operational costs. The model also provided valuable insights into the performance of the production system, highlighting areas for improvement and potential areas of failure.

Future research in this area could focus on developing more sophisticated models that incorporate additional data sources and advanced analytics techniques, as well as exploring the potential benefits of real-time optimization and control.

ACKNOWLEDGEMENT

The authors gratefully appreciate the financial support from CAPES, Pontifical Catholic University of Rio de Janeiro (PUC-RIO) and the Brazilian National Agency for Petroleum, Natural Gas and Biofuels (ANP) through the Human Resources Program for the Petroleum, Natural Gas and Biofuels sector - PRH-ANP that made this research possible.

REFERENCES

Aggoun, L. and Chetouani, Y. (2021). Fault detection strategy combining narmax model and bhattacharyya distance for process monitoring. *Journal of the Franklin Institute*, 358(3), 2212–2228. doi: https://doi.org/10.1016/j.jfranklin.2021.01.001. URL `https://www.sciencedirect.com/science/article/pii/S0016003221000016`.

Agung, I. (2022). Input parameters comparison on narx neural network to increase the accuracy of stock prediction. *JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING*, 6, 82–90. doi:10.31289/jite.v6i1.7158.

Alakeely, A. and Horne, R. (2021). Application of deep learning methods to estimate multiphase flow rate in producing wells using surface measurements. *Journal of Petroleum Science and Engineering*, 205, 108936. doi:https://doi.org/10.1016/j.petrol.2021.108936. URL `https://www.sciencedirect.com/science/article/pii/S0920410521005970`.

Billings, S.A. (2013). *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains.* John Wiley & Sons.

Bo, L., Jiannan, G., and Xiangdong, X. (2020). The digital twin of oil and gas pipeline system. *IFAC-PapersOnLine*, 53(5), 710–714. doi: https://doi.org/10.1016/j.ifacol.2021.04.162. URL `https://www.sciencedirect.com/science/article/pii/S2405896321003025`. 3rd IFAC Workshop on Cyber-Physical Human Systems CPHS 2020.

Chen, H., Liu, H., Chu, X., Liu, Q., and Xue, D. (2021). Anomaly detection and critical scada parameters identification for wind turbines based on lstm-ae neural network. *Renewable Energy*, 172, 829–840. doi:https://doi.org/10.1016/j.renene.2021.03.078. URL `https://www.sciencedirect.com/science/article/pii/S0960148121004341`.

de Oliveira Werneck, R., Prates, R., Moura, R., Gonçalves, M.M., Castro, M., Soriano-Vargas, A., Ribeiro Mendes Júnior, P., Hossain, M.M., Zampieri, M.F., Ferreira, A., Davólio, A., Schiozer, D., and Rocha, A. (2022). Data-driven deep-learning forecasting for oil production and pressure. *Journal of Petroleum Science and Engineering*, 210, 109937. doi:https://doi.org/10.1016/j.petrol.2021.109937. URL https://www.sciencedirect.com/science/article/pii/S0920410521015515.

Equinor, E. (2018). Volve data village. *Equinor Data Portal Beta*.

Heghedus, C., Shchipanov, A., and Rong, C. (2019). Advancing deep learning to improve upstream petroleum monitoring. *IEEE Access*, 7, 106248–106259. doi:10.1109/ACCESS.2019.2931990.

Junior, W.R.L., da Andrade, L.P.C., Oliveira, S.C.P., and Martins, S.A.M. (2020). Sysidentpy: A python package for system identification using narmax models. *Journal of Open Source Software*, 5(54), 2384. doi:10.21105/joss.02384. URL https://doi.org/10.21105/joss.02384.

Koroteev, D. and Tekic, Z. (2021). Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future. *Energy and AI*, 3, 100041. doi:https://doi.org/10.1016/j.egyai.2020.100041. URL https://www.sciencedirect.com/science/article/pii/S2666546820300410.

Lei, H., Kai, W., Changchun, W., Jing, G., and Xie, P. (2020). Hybrid method based on particle filter and NARX for real-time flow rate estimation in multi-product pipelines. *Journal of Process Control*, 88, 19–31. doi:https://doi.org/10.1016/j.jprocont.2020.02.004. URL https://www.sciencedirect.com/science/article/pii/S0959152419304342.

Li, Y., Sun, R., and Horne, R. (2019). Deep Learning for Well Data History Analysis. *OnePetro*, Day 1 Mon, September 30, 2019. doi:10.2118/196011-MS. URL https://doi.org/10.2118/196011-MS. D011S008R002.

Mohammadpoor, M. and Torabi, F. (2020). Big data analytics in oil and gas industry: An emerging trend. *Petroleum*, 6(4), 321–328. doi:https://doi.org/10.1016/j.petlm.2018.11.001. URL https://www.sciencedirect.com/science/article/pii/S2405656118301421. SI: Artificial Intelligence (AI), Knowledge-based Systems (KBS), and Machine Learning (ML).

Onalo, D., Oloruntobi, O., Adedigba, S., Khan, F., James, L., and Butt, S. (2019). Dynamic data driven sonic well log model for formation evaluation. *Journal of Petroleum Science and Engineering*, 175, 1049–1062. doi:https://doi.org/10.1016/j.petrol.2019.01.042. URL https://www.sciencedirect.com/science/article/pii/S092041051930049X.

Sharkawy, A.N., Ali, M.M., Mousa, H.H.H., Ali, A.S., Abdel-Jaber, G.T., Hussein, H.S., Farrag, M., and Ismeil, M.A. (2023). Solar pv power estimation and upscaling forecast using different artificial neural networks types: Assessment, validation, and comparison. *IEEE Access*, 11, 19279–19300. doi:10.1109/ACCESS.2023.3249108.

Siavashi, J., Najafi, A., Ebadi, M., and Sharifi, M. (2022). A cnn-based approach for upscaling multiphase flow in digital sandstones. *Fuel*, 308, 122047. doi:https://doi.org/10.1016/j.fuel.2021.122047. URL https://www.sciencedirect.com/science/article/pii/S0016236121019232.

Solanki, P., Baldaniya, D., Jogani, D., Chaudhary, B., Shah, M., and Kshirsagar, A. (2022). Artificial intelligence: New age of transformation in petroleum upstream. *Petroleum Research*, 7(1), 106–114. doi:https://doi.org/10.1016/j.ptlrs.2021.07.002. URL https://www.sciencedirect.com/science/article/pii/S2096249521000491.

Tian, C. and Horne, R.N. (2017). Recurrent Neural Networks for Permanent Downhole Gauge Data Analysis. *OnePetro*, Day 1 Mon, October 09, 2017. doi:10.2118/187181-MS. URL https://doi.org/10.2118/187181-MS. D011S008R007.

Zha, W., Liu, Y., Wan, Y., Luo, R., Li, D., Yang, S., and Xu, Y. (2022). Forecasting monthly gas field production based on the cnn-lstm model. *Energy*, 260, 124889. doi:https://doi.org/10.1016/j.energy.2022.124889. URL https://www.sciencedirect.com/science/article/pii/S0360544222017923.