Analysis of the performance of a Crude-oil Desalting System based on historical data

Ehsan Ranaee[1], Hamzeh Ghorbani[2], Sajjad Keshavarzian[1], Pejman Ghazaeipour Abarghoei[3], Monica Riva[4], Fabio Inzoli[1*], Alberto Guadagnini[4]

[1]Dipartimento di Energia, Politecnico di Milano, Via Lambruschini 4, 20156 Milano, Italy

[2]Islamic Azad University, Ahvaz branch, Young Researchers Club, Iran

[3]Iran University of Science and Technology, Narmak, Tehran, Iran

[4]Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Piazza L. Da Vinci 32, 20133 Milano, Italy

* Corresponding author. Tel. +39 02 2399 3883. Fax. +39 02 2399 3913

E–mail address: fabio.inzoli@polimi.it

Abstract

Our study is keyed to the development of a methodological approach to assess the workflow and performance associated with the operation of a crude-oil desalting/demulsification system. Our analysis is data-driven and relies on the combined use of (*a*) Global Sensitivity Analysis (GSA), (*b*) machine learning, and (*c*) rigorous model discrimination/identification criteria. We leverage on an extensive and unique data-set comprising observations collected at a daily rate across a three-year period at an industrial plant where crude oil is treated through a combination of demulsification/desalting processes. Results from GSA enable us to quantify the system variables which are most influential to the overall performance of the industrial plant. Machine learning is then applied to formulate a set of candidate models whose relative skill to represent the system behavior is quantified upon relying on model identification criteria. The integrated approach we propose can then effectively assist to (*a*) modern and reliable interpretation of data associated with performances of the crude oil desalting process and (*b*) robust evaluation of future performance scenarios, as informed by historical data.

Keywords: Crude oil plant assessment, Sensitivity Analysis, Uncertainty Quantification, Principal Component Analysis, Machine Learning.

1. Introduction

Crude oil extracted from reservoirs generally contains water with soluble salts, solid particles, and/or dissolved chemicals, including metals (e.g., vanadium, nickel) [1-4]. The presence of water reduces the total volume of the oil delivered to refineries and increases operational costs. Soluble salts can cause negative impacts on refinery facilities, in the form of, e.g., corrosion, and/ or reduced effectiveness of the catalysts that are used for crude oil refining processes. Such phenomena, in turn, lead to increased maintenance/service expenses, in terms of intensive energy use, reduction of refinement system performance, and uncontrolled changes in oil properties, which can then hamper the effectiveness of the entire energy cycle.

Previous experiences (e.g., [5-6]) document that corrosive properties of water-soluble salts can cause significant damage to operating equipment (e.g., pipes, valves, pumps, tanks, and tanker vessels) as well as internal parts of refinery distillation towers. Impurities may cause fouling on interior surfaces of the equipment, thus potentially leading to clogging of the oil heater pipes and increase of temperature and pressure therein, with critical implications on energy use. Laboratory experiments performed by Hamadi et al. [7] document that impurities in crude oil lead to increased values of density and viscosity for crude oil emulsion. This, in turn, can reduce the API gravity degree of the crude oil emulsion as compared to pure oil, thus yielding an oil of reduced marketable quality.

In this setting, crude oil is typically processed in desalting/demulsification units before being conveyed to refineries. The latter may be operated on the basis of various techniques (e.g., gravity-based, thermal, chemical, electrical, or through approaches grounded on clean water injection, membrane use, ultrasonic waves, and microwave) or a combination of diverse technologies.

Demulsification is typically designed as a pre-treatment to break water/oil emulsions and remove impurities and salty water from the oil. It takes place according to two main stages, i.e., (*i*) emulsification and (*ii*) separation of salt/water from oil. Emulsification relies on mixing fresh water with a salt-water-oil emulsion. Studies have been performed to assess the behavior of this process,

including the evaluation of the size and distribution of the water droplets in crude oil emulsion and droplet breakage and coalescence under turbulent flow conditions [8]. After mixing fresh water with crude oil, salt-water should be separated from the organic phases to complete the demulsification process.

After demulsification, oil is collected from the various desalting units and a blend is formed to comply with the characteristics of the target refinery. Properties of crude oils extracted from diverse sedimentary basins (with differing petrophysical properties) can exhibit marked variabilities, including differences of salinity and impurities content. It is then important to verify if oil batches processed in diverse desalting units (and then collected into a single refining unit) are associated with similar characteristics. Laboratory tests on samples of crude oil are typically performed and the type and amount of demulsifiers to be employed is suggested on the basis of the crude-oil properties [9].

Demulsification may also be implemented jointly with other types of desalting techniques. It is then challenging to have a robust appraisal of the extent at which laboratory tests can be scaled to field conditions to cope with, e.g., (*i*) interaction of a mixture of different types of demulsifiers, and/ or (*ii*) interaction of the demulsification with other desalting techniques (possibly) implemented in the field.

Less et al. [10] study the performance of an electrostatic coalescer by analyzing a set-up based on electrical desalting technique alone and in conjunction with the injection of demulsifying materials into the crude oil. Meidanshahi et al. [11] analyze the feedback between salt concentration and water droplet size when demulsification is performed through electrical desalting. These results suggest that simultaneous use of electrical and chemical techniques for desalting show a markedly higher potential than relying on each technique as a standalone process.

In this broad context, it is then emphasized that optimization of oil treatment processes through a firm characterization of the system functioning stands as an important step towards enhancing sustainability within oil and gas industries with the aim of maximizing general efficiency with a

reduction of energy consumptions. Along these lines, we consider here an industrial plant comprising a set of desalting units where a combination of desalting techniques is implemented in a coherent and structured way (see details in Section 2). We then rely on a modern theoretical approach to the analysis of experimental data collected in this industrial plant to allow evidencing the relative importance and relationship between performance of the industrial plant and monitored parameters driving the system behavior.

Dai et al. [12] recently proposed a data-driven approach for the optimization of a crude oil blending system and improvement of the operating efficiency of a refinery. These authors rely on a case study to illustrate the effectiveness of the proposed methodology. Multivariate probabilistic analyses have been employed to diagnose and manage large datasets associated with a variety of applications in the petroleum industry, and in more general energy industry sector. Qin and Chiang [13] provide a comprehensive review of recent developments and applications of machine learning and artificial intelligence (AI) approaches, as applied to big commercial data sets. These authors remark that uncertainties and possible lack of critical measurements are major aspects to be considered when applying these approaches to industrial settings. Siena et al. [14] rely on Bayesian hierarchical clustering (BHC) and Principal Components Analysis (PCA) techniques to select the most appropriate Enhanced Oil Recovery (EOR) technique amongst a set of available options for a given reservoir. Sad et al. [15] perform a set of laboratory experiments on oil samples produced from various wells in the same field to study compliance of physic-chemical properties of blends of crude oil with the characteristics of a target refinery. These authors implement a hierarchical cluster analysis to classify laboratory data associated with samples collected from different production wells. Nonlinear methods such as artificial neural networks (ANNs) are also applied in a multivariate analysis context to optimize the efficiency of a desalting unit (e.g., [16]). Mohammad et al. [17] develop and apply ANN models to evaluate the performance of a liquid desiccant dehumidifier in terms of the water condensation rate and dehumidifier effectiveness.

Some applications of machine learning in crude oil refinery planning and management can be found in the literature. Mouret et al. [18] introduce a Lagrangian decomposition scheme to simultaneously optimize (*a*) a crude-oil operations scheduling model and (*b*) a coarse refinery planning model. More recently, Gao et al. [19] apply machine learning concepts to introduce a piecewise linear modeling approach to assist refinery scheduling, with an application to the operations associated with a refinery in China. Ochoa-Estopier et al. [20] and Gueddar and Dua [21] consider ANNs for the optimization of heat-integrated crude oil distillation systems and refinery units, respectively.

In this framework, Sensitivity Analysis (SA) may serve as a general framework within which one can assess the way a given quantity of interest is influenced by diverse controlling (or driving) elements. Al-Qahtani and Elkamel [22] consider uncertainty in process parameters to analyze performance of a refinery plan as well as an integration network of petroleum refineries. These authors document that the refinery model analyzed is most sensitive to variations in the pricing of imported crude oil and exported final products as opposed to changes in product demand. Typical applications of SA are concerned with the diagnosis of the behavior of a given model of a system functioning. As such, the relationship between a target quantity (or modeling goal) and the related controlling variables is filtered by the structure of the model considered. A major purpose of a SA is then to ranking and screening the control variables which are most influential to the target quantities according to some predefined sensitivity metrics. When one is interested in the analysis of the way multiple controlling variables influence a set of target quantities that jointly contribute to define the performance of the system considered, performing a SA considering each of the target variables separately might be influenced by the occurrence of possible correlations among these. Lamboni et al. [23] proposed to circumvent this issue by relying on a combined use of (*i*) a principal component analysis (PCA) of the set of multivariate target variables and (*ii*) variance-based sensitivity indices to

characterize the contribution of each controlling quantity on the set of identified principal components.

As such, a distinctive objective of our analysis is the assessment of a workflow based on a suitable theoretical framework leading to the identification of the key quantities which are typically monitored in a desalting plant that can facilitate the control of the optimal performance of the plant. We do so by providing an approach which enables us to effectively employ a combination of SA tools in the absence of a clear model structure relating controlling and target variables. We rely on a unique (in terms of quality and quantity of data) dataset to evaluate the influence of system variables on some desired system state quantity (which would represent a modeling goal). Thus, our analysis is observation-driven and rests on a multivariate analysis framework to assess information collected from field settings involving the use of different demulsifiers in combination with a set of desalting techniques to discuss the way the ensuing results can assist estimation of quantities of interest (i.e., production of wastewater, processed oil and salt from the desalting unit) representing the effectiveness/performance of the system. We do so by first identifying a set of control variables associated with ambient/processing conditions of the joint crude oil demulsification/desalting system considered. We then (*i*) consider jointly various SA approaches to assess the relative influence of the system control variables on the performance of the demulsification/desalting system, as quantified through key target quantities, (*ii*) formulate a suite of linear/ nonlinear models to interpret observations of such target variables, and (*iii*) rely on formal model selection criteria to evaluate the relative skills of the considered models to characterize the system behavior.

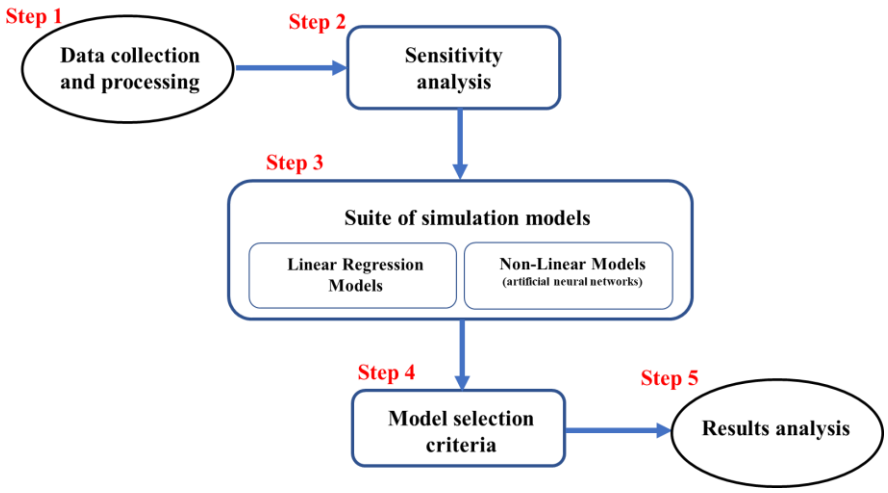


Fig. 1 *Flowchart of the analysis framework.*

Our analyses are structured according to the flowchart depicted in Fig. 1 and comprising the following five sequential steps:

- *step 1*: acquisition of field data associated with a demulsification/desalting industrial plant;
- *step 2*: evaluation (through SA approaches) of the relative influence of the main controlling variables on the selected target quantities, with possible reduction (based on SA results) of the dimensionality of the space identified by the controlling variables;
- *step 3*: application of a collection of linear/nonlinear simulation models to interpret observations of the selected target quantities;
- *step 4*: assessment of the quality of the considered simulation models through model selection criteria;
- *step 5*: analysis of the results to summarize and highlight key relationships between controlling variables and the target quantities of interest considered to evaluate the performance of the demulsification/desalting system.

With reference to the simulation models, we rely on a typical linear regression modeling framework and an artificial neural networks- (ANN-) based non-linear approach (e.g., [24-26]).

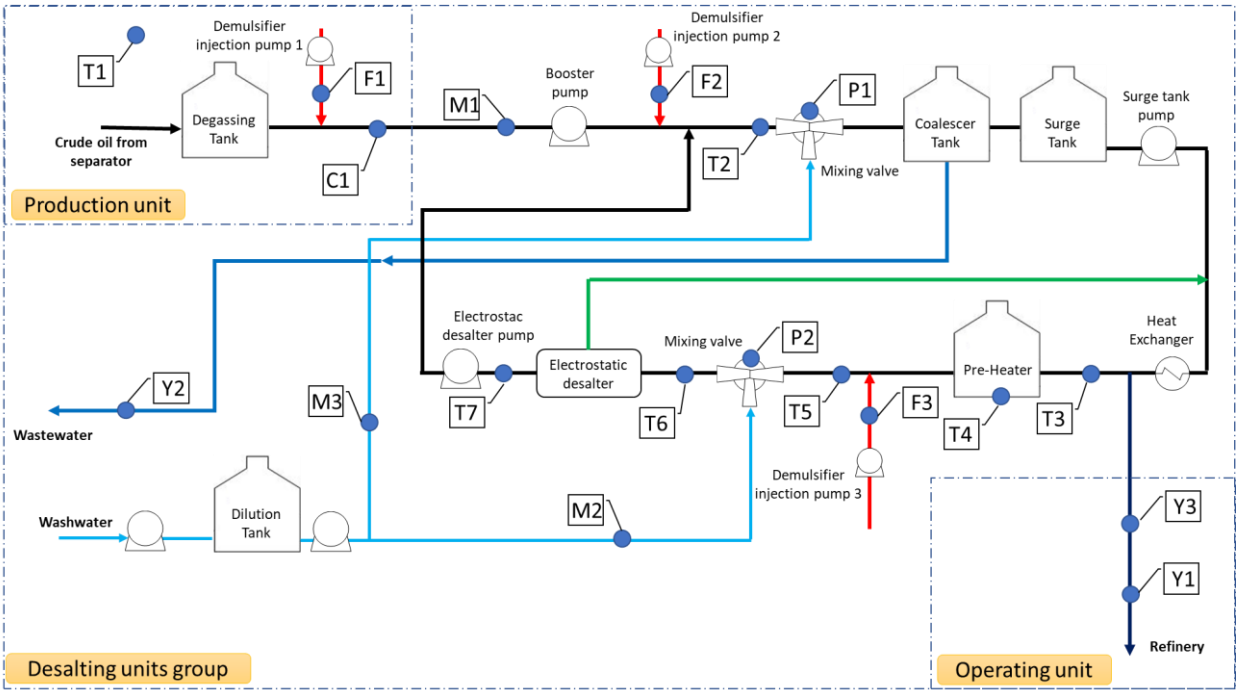
The study is organized as follows. We briefly describe in Section 2.1 the desalting plant analyzed and the collected data. Section 2.2 is devoted to the process of on-site data collection. We

then provide a synthesis of the key theoretical elements underpinning sensitivity analysis and dimensionality reduction technique (Section 2.3) as well as simulation models (Section 2.4). Model selection criteria are introduced in Section 2.5. Implementation of the analysis workflow and the ensuing results are discussed in Section 3.

2. Methodology and application

2.1. Plant analysis and data collection

The investigated plant comprises 3 different modules, serving the same oil field. Fig. 2 depicts a schematic representation of a desalting module, including the types of data collected. These plants are logically structured according to 3 main units, i.e., production, desalting, and operating units.



Collected data					
ID	Ref	description	ID	Ref	description
X_1	M1	Oil volume flow rate @ Production unit	X_{11}	T4	Fluid temperature @ Pre-heater bath
X_2	M2	Fresh water volume flow rate @ Electrostatic desalter	X_{12}	T5	Fluid temperature @ outlet Pre-heater
X_3	M3	Fresh water volume flow rate @ Coalescer tank	X_{13}	T6	Fluid temperature @ inlet Electrostatic desalter
X_4	F1	Demusilfier injection @ Production Unit	X_{14}	T7	Fluid temperature @ inlet Electrostatic desalter
X_5	F2	Demusilfier injection @ Coalescer tank	X_{15}	P1	Pressure drop @ mixing valve (Coalescer tank)
X_6	F3	Demusilfier injection @ Electrostatic desalter	X_{16}	P2	Pressure drop @ mixing valve (Electrostatic desalter)
X_7	C1	Salt amount @ Production unit			
X_8	T1	Ambient temperature	Y_1	Y1	Oil volume flow rate @ Operating unit
X_9	T2	Fluid temperature @ inlet Coalescer tank	Y_2	Y2	Wastewater volume flow rate
X_{10}	T3	Fluid temperature @ inlet Pre-heater	Y_3	Y3	Salt amount @ Operating unit

Fig. 2 *Schematic drawing of the analyzed crude oil desalting module. Solid blue circles correspond*

to locations of probes for data collections. Bottom table lists parameter identifiers (ID) and references to the plant schematics (Ref).

At the early stage, crude oil enters the production unit and is conveyed to the degassing tank for gas separation. A demulsifier is injected into the oil exiting from the degassing tank, before it enters the desalting unit. The salt content of the emulsion is measured at the entrance of the desalting unit. In the next stage, crude oil is pumped to the coalescer tank by means of a crude booster pump. The pressure drop across a mixing valve located before the coalescer tank is recorded, fluid temperature being monitored at the entrance of the coalescer tank. The emulsion remains in the coalescer tank for the time needed to separate part of the wastewater from oil. The internal structure of the coalescer tank enables crude oil to be subject to a rotating motion within the tank (typically for 24 hours, depending on the amount of crude oil) and provides sufficient time for water droplets to be separated from the oil. Water droplet separation in the coalescer tank is expedited by injection of fresh water from the dilution tank and by addition of a demulsifier to break the water-in-oil emulsion. The resulting separated water is then conveyed to the wastewater section, the oil entering the surge tank. Oil exiting from the coalescer tank typically contains around 0.5% to 2% water, depending on process conditions. Oil accumulates in the surge tank and is then pumped at a uniform flow rate to the heat exchanger. Afterwards, oil enters the pre-heater and, after injection of the demulsifier, enters the electrostatic desalter. Fluid temperature is also recorded before, inside and after the pre-heater unit. Pressure drop at a mixing valve positioned before the electrostatic desalter is recorded. A small amount (about 3 - 10%) of fresh water is injected into the mixing valve upstream the electrostatic desalter to decrease salt concentration in the crude oil. Relying on an electrical field and on emulsion breakage increases the rate of deposition of water droplets. Accordingly, the remaining droplets in the salt water are separated from the oil in the electrostatic desalting tank. Fluid temperature data are also collected before and after electrostatic desalter. Finally, the output oil from the electrostatic desalter enters again the heat exchanger to be cooled down, before being conveyed to the operating

unit. Water extracted from the crude oil in the electrostatic desalting tank is pumped (by means of an electrostatic desalter pump) to the coalescer tank to (*i*) increase the fluid temperature and (*ii*) reduce concentration of the salt at the entrance of the coalescer tank. The rate of the produced oil and the amount of separated salt is measured before sending the processed oil to the refinery. The plant performance is then assessed through the evaluation of (*i*) the rate of production oil (Y_1), (*ii*) the rate of wastewater at the plant outlet (Y_2), and (*iii*) the amount of separated salt detected in the oil production stream (Y_3).

We collect across a three-year period onsite measurements of the 19 variables in Fig. 2, from 3 different desalting units (in the same oil field) where demulsification is performed through injection of 5 different demulsifiers in conjunction with (*i*) the gravity force desalting technique, (*ii*) a thermal desalting technique, and (*iii*) the washing water injection approach. A total number of N = 1820 observations are available for each variable during the three-year period.

2.2. Data analysis methodology

Monitored values of the quantities listed in Fig. 2 are collected in matrices $\underline{\underline{\mathbf{X}}}^*$ and $\underline{\underline{\mathbf{Y}}}^*$:

$$\underline{\underline{\mathbf{X}}}^* = \underline{\underline{\mathbf{X}}}^*\left(\mathbf{X}_1^*, \mathbf{X}_2^*, ..., \mathbf{X}_{16}^*\right); \qquad \underline{\underline{\mathbf{Y}}}^* = \underline{\underline{\mathbf{Y}}}^*\left(\mathbf{Y}_1^*, \mathbf{Y}_2^*, \mathbf{Y}_3^*\right) \tag{1}$$

where the entries of each vector $\mathbf{X}_i^*$ (i = 1, 2, ..., 16) and $\mathbf{Y}_k^*$ (k = 1, 2, 3) correspond to N available data.

The demulsification process can be described through the following functional relationship

$$\mathbf{Y} = f(\mathbf{X}) \tag{2}$$

where vectors $\mathbf{Y} = (Y_1, Y_2, Y_3)$ and $\mathbf{X} = (X_1, X_2, ..., X_{16})$ include all (generally unknown) variables listed in Table 1 whose observed values are collected in $\underline{\underline{\mathbf{X}}}^*$ and $\underline{\underline{\mathbf{Y}}}^*$, respectively.

Our strategy rests on the study of the way the variability of a set of target state variables Y_k (collected in $\mathbf{Y}$) is influenced by the remaining system variables (also termed as controlling variables,

collected in **X**). We do so through a unique application of a combination of Global Sensitivity Analysis (GSA) approaches (see Section 2.3).

Upon relying on GSA, we assess the possibility of performing a dimensionality reduction of the space identified by the system controlling variables. Accordingly, the functional relationship (2) becomes

$$f_1\left(\mathbf{Y}, \mathbf{X}_R\right) = 0 \tag{3}$$

where $\mathbf{X}_R$ includes the set of R controlling variables associated with the reduced dimensionality space which is eventually identified. The functional relationship (3) is then characterized through linear regression and nonlinear ANN-based modeling techniques (Section 2.4). Parameters of the constructed simulation models are evaluated upon relying on the available data (Section 2.1). We recall that the latter are collected from 3 different desalting units (in the same field) and 5 types of demulsifiers have been used across the considered 3 years temporal frame. As such, the development of an interpretive model should also consider the possibility of including information about (*i*) the specific desalting module from which each measurement is collected and (*ii*) the type of demulsifiers used for a given record. We take this type of information into account in our regression models by following a strategy proposed in prior studies (e.g., [27]) and mapping information on the location of data sampling and the type of employed demulsifiers into a set dummy (categorical) variables collected, respectively, in vectors **U** and **Q**. We set $U_1 = 1$ (while $U_2 = U_3 = 0$) when a measurement record is collected in the first desalting units, corresponding notations being employed to identify data sampled at the second or third desalting units. The same approach is implemented to define **Q**, whose entries are dummy variables Q_l (l = 1, 2, ..., 5) quantifying information on the type of demulsifier used. Variables U_1 - U_3 and Q_1 - Q_5 can then be included in $\mathbf{X}_R$ to characterize the ensuing simulation model in a such way that:

$$\mathbf{X}_R = \mathbf{X}_R\left(\mathbf{X}, \mathbf{Q}, \mathbf{U}\right) \tag{4}$$

2.3. Sensitivity Analysis

We rely on a Sensitivity Analysis (SA) approach to identify the system controlling variables which (amongst those listed in Fig. 2) are most influential to the effectiveness of the demulsification process. As stated in Section 1, our SA is completely data-driven and relies on the available dataset described in Sections 2.1 and 2.2. As such, our results enable us to evaluate the influence of system variables on some desired system state quantity without the filter of an interpretive model. Outcomes of the SA are then employed to drive the construction of a collection of candidate models (see Section 2.4) whose relative interpretive skills are evaluated according to the approach illustrated in Section 2.5.

Our SA starts by considering in (3) the performance of the system as rendered through the following 3 target quantities, which are measured as outputs from the Desalting Unit and form the entries of **Y**, i.e., (*i*) Y_1, oil volume; (*ii*) Y_2, wastewater volume, and (*iii*) Y_3, salt mass.

Various approaches to SA are proposed in the literature to diagnose the behavior of a given model, where a target modeling goal depends on a number of uncertain model parameters. These are essentially framed within the context of (*i*) Local Sensitivity Analysis (LSA) or (*ii*) Global Sensitivity Analysis (GSA). LSA approaches are typically based on the evaluation of local derivatives (also termed elementary effects or local sensitivities) of modeling goals with respect to a control variable at a given location in Γ. Such an approach allows quantifying the sensitivity of a given (dependent) target variable Y_k to perturbations of X_i and information about the importance of X_i is limited to the location where a given elementary effect is assessed [28]. These results can (in principle) be extended to the entire space where model control variables are defined when, e.g., (*i*) the target system state variables (Y_k) display a linear behavior with respect to the control variables (X_i), and (*ii*) an adequate number of trajectories is sampled within Γ for the evaluation of the sensitivities. While the elementary effects are evaluated at a set of points in the parameter space, our study rests on a GSA framework.

Each state variables Y_k (with $k = 1, \ldots, 3$) is considered to depend on quantities X_i (with $i = 1, \ldots, I$). We assess the strength of the dependence between Y_k and X_i through SA techniques which are implemented upon relying on the parameter space $\Gamma = \Gamma_{X_1} \times \ldots \times \Gamma_{X_I}$, $\Gamma_{X_i} = [X_{i,\min}, X_{i,\max}]$ being the support (range of variability) of X_i. This approach is based on the interpretation of each quantity X_i as a random quantity whose variability influences the output Y_k.

Note that we perform our GSA before model formulation, ensuing calibration and performance evaluation. As such, our GSA addresses the possibility of (*i*) improving our understanding of the relevance of each controlling variable on the target quantities, and (*ii*) identifying variables which might be uninfluential in the context of a subsequent model formulation and calibration (e.g., [29-31] and references therein). We rest on the joint use of multiple global sensitivity indices, each providing a particular insight to a given aspect of sensitivity. This strategy yields a comprehensive depiction of the way model controlling parameters influence target state variables and minimizes the risk of classifying as unimportant some variables which might have a non-negligible impact on selected features of the output of interest. We illustrate some details of the sensitivity approaches and metrics we employ in Section 2.3.1, 2.3.2, and 2.3.3.

2.3.1. Global Sensitivity Analysis based on Morris Indices

We rely here on the evaluation of the Morris indices [32], which enable us to extend the concept of derivative-based local sensitivity to the entire space Γ. These indices are evaluated as

$$\mu_{X_i}^{Y_k} = E(|\, d_{X_i}^{Y_k}\,|); \qquad \text{with} \qquad d_{X_i}^{Y_k} = \frac{Y_k(X_1,\ldots,X_i+\delta_i,\ldots,X_I) - Y_k(X_1,\ldots,X_i,\ldots,X_I)}{\delta_i} \tag{4}$$

Here, $\mu_{X_i}^{Y_k}$ is the Morris Index for Y_k associated with variable X_i; $E(\bullet)$ denotes expected value; $d_{X_i}^{Y_k}$ is an approximation of the partial derivative of Y_k with respect to X_i and evaluated at a given point in Γ; and δ_i is a perturbation to X_i.

2.3.2. Moment-based Sensitivity Indices

Dell'Oca et al. [33] propose estimating global sensitivity metrics based on the (statistical) moments of the probability density function (*pdf*) of target variables. This technique provides richer information than the commonly used variance-based GSA [34]. Here, we consider GSA as rendered through the assessment of the *AMA* metrics (termed after the Authors' initials [33])

$$AMAE_{X_i}^{Y_k} = \begin{cases} \dfrac{1}{\left|Y_k^0\right|} E\left[\left|Y_k^0 - E[Y_k \mid X_i]\right|\right] & \text{if} \quad Y_k^0 \neq 0 \\ E\left[\left|E[Y_k \mid X_i]\right|\right] & \text{if} \quad Y_k^0 = 0 \end{cases} \tag{6}$$

$$AMAV_{X_i}^{Y_k} = \frac{E\left[\left|V[Y_k] - V[Y_k \mid X_i]\right|\right]}{V[Y_k]} \tag{7}$$

where $AMAE_{X_i}^{Y_k}$ and $AMAV_{X_i}^{Y_k}$ represent the sensitivity indices associated with the first and the second moment of Y_k, respectively, as linked to variability of X_i; the quantity Y_k^0 in (6) indicates the (unconditional) mean of Y_k, as computed over Γ; $V[\bullet]$ denotes variance; the symbol $Y_k \mid X_i$ indicates conditioning of Y_k to a known value of X_i within Γ; $E(\bullet)$ denotes expected value. Essentially, indices $AMAE_{X_i}^{Y_k}$ and $AMAV_{X_i}^{Y_k}$ quantify the expected change of the mean and variance of Y_k due to our knowledge of X_i (see [33], [35]).

2.3.3. PCA-based Sensitivity Indices

As stated in Section 1, Lamboni et al. [23] perform a GSA of a set of (multivariate) model outputs by combining a principal component analysis (PCA) of these with the analysis of variance (ANOVA). As such, these authors introduce a set of sensitivity indices quantifying the influence of selected quantities controlling the model behavior on the principal components associated with the collection of (multivariate) model outputs of interest.

We start by performing PCA through the eigenvalue decomposition of the covariance matrix of the considered set of system outputs. This enables us to transform the original outputs **Y** into a set

of new orthogonal variables, which are sorted according to their contribution to the overall system variance (see [34] for additional details).

We consider here the first order sensitivity indices

$$SI_{i,k} = \frac{V_{i,k}}{V_k} \tag{8}$$

representing the influence of X_i on the k-th principal component (P_k) of the multivariate vector **Y**. Here, $V_{i,k}$ is the variance of P_k associated with variability of X_i and V_k is the total variance of P_k.

The sensitivity indices (8) focus on the k-th principal component of model outputs rather than on the set of multivariate outputs as a whole. To quantify the contribution of each control variable X_i to the total variance of **Y**, Lamboni et al. [23] introduce a generalized first order sensitivity index expressed as

$$GSI_i = \frac{\sum_{k=1}^{K} \lambda_k SI_{i,k}}{\sum_{k=1}^{K} \lambda_k} \tag{9}$$

where λ_k is the k-th (with $k = \{1,\ldots,K\}$) eigenvalue of the covariance matrix of **Y**. One can see that the generalized sensitivity index (9) corresponds to the weighted average of the sensitivity indices on the principal components, the weight of each term in (9) being proportional to the eigenvalue associated with the corresponding principal component. As such, this quantity expresses the unique contribution of control variable X_i on the total variance of the multivariate outputs encapsulated in **Y**.

2.4. Candidate Simulation Models

The application of the GSA techniques illustrated in Section 2.3 can yield a dimensionally reduced space formed by R control variables. We then model the dependence of the target variables **Y** on the reduced set of controlling variables ($\mathbf{X}_R$) through implementation of a suite of linear/nonlinear models via (3).

We follow Takane and Hunter [36-37] to characterize (4) through a multivariate linear regression technique (e.g., [38-39]). We also test the (possibly) nonlinear nature of relationship (3) by constructing an Artificial Neural Network-based simulation model. We do so upon considering a Multi-Layer Perceptron (MLP) Artificial Neural Network (ANN) and training it to reproduce the observations included in $\underline{\underline{\mathbf{Y}}}^{*}$ through the unsupervised learning rule of Oja [40, 41]. Hidden layer(s) contain nonlinear mapping functions. Training of the ANN is performed trough serving sample batches (taken from the measurements) to the model and optimizing the weights, to minimize mean squared error (*MSE*) between the network responses and related measurement records of target variables.

The use of formal model selection criteria to discriminate amongst the set of candidate models which are formulated to interpret the system behavior is illustrated in Section 2.5.

2.5. Model Selection Criteria

The quality of model calibration (either linear regression or nonlinear ANNs in our case) is typically appraised through the evaluation of the mean squared error, *MSE*, as

$$MSE = \frac{J}{O \times N} \qquad \text{with} \quad J = \sum_{k=1}^{O} \sum_{i=1}^{N} (\hat{Y}_{ki} - Y_{ki}^{*})^2 \tag{10}$$

where $\hat{Y}_{ki}$ is a model-based output corresponding to observation Y_{ki}^{*} (i.e., *i*-th sample record of the *k*-th target quantity, *O* being the total number of target quantities) in matrix $\underline{\underline{\mathbf{Y}}}^{*}$.

We rely here on formal model selection criteria to evaluate (in a relative sense) the skill of each of the models we construct to interpret the available information. Among the various model selection criteria proposed in the literature to discriminate amongst models [42], we rest here on the commonly used information criteria AIC_c ([43]) and *BIC* ([44]).

When multiple models are considered to interpret a physical scenario of interest, one may calibrate each of them by minimizing the negative log likelihood criterion, *NLL* [45]

$$NLL = \frac{J}{\sigma_{\mathbf{Y}}^2} + (O \times N)\ln(2\pi\sigma_{\mathbf{Y}}^2) \tag{11}$$

where $\sigma_{\mathbf{Y}}^2$ corresponds to measurement error variance, its Maximum Likelihood (*ML*) estimate being given by $\hat{\sigma}_{\mathbf{Y}}^2 = {J_{\min}}/{(O \times N)}$ (see [46] and references therein for more details). Once the parameters associated with each model are estimated, the alternative model formulations can be ranked by evaluating the following model selection (or discrimination) criteria

$$AIC_c = NLL + 2m + \frac{2m(m+1)}{(O \times N) - m - 1} \tag{12}$$

$$BIC = NLL + m\ln(O \times N) \tag{13}$$

where *m* represents the number of model parameters. Model discrimination criteria can also be employed to determine posterior model weight (for AIC_c) or posterior model probability (for *BIC*), $p(M_\alpha \mid \underline{\underline{\mathbf{Y}}}^*)$, as ([47, 48] and references therein)

$$p(M_\alpha \mid \underline{\underline{\mathbf{Y}}}^*) = \frac{\exp(-{1}/{2}\,\Delta IC_\alpha)\,p(M_\alpha)}{\sum_{l=1}^{L}\exp(-{1}/{2}\,\Delta IC_l)\,p(M_l)} \tag{14}$$

Here, $\Delta IC_\alpha = IC_\alpha - IC_{\min}$, IC_α being a given model selection criterion (12)-(13); $IC_{\min}$ is the minimum value of IC_α across the L candidate models; and $p(M_\alpha)$ is the prior probability of model M_α.

3. Results

We start our analyses by providing basic statistics of the observations acquired at the investigated plant. These are listed in Table 1 in terms of range of variability (as expressed through

the lower and upper values detected for each quantity), mean, standard deviation (*std*), and the coefficient of variation (*CV*).

Table 1 *Range of variability (identified in terms of the lower, Min, and upper, Max, values detected), mean, standard deviation (std), and coefficient of variation (CV) of each sampled variable.*

Data acquisition							
ID	**description**	**units**	**Min**	**Max**	**Mean**	**Std**	**CV**
X_1	Oil volume flow rate @ Production unit	m^3/day	3703	19070	10611	3505	33%
X_2	Fresh water volume flow rate @ Electrostatic desalter	m^3/day	0	556	249	154	62%
X_3	Fresh water volume flow rate @ Coalescer tank	m^3/day	0	1479	376	330	88%
X_4	Demusilfier injection @ Production Unit	ppm	0	69	25	11	44%
X_5	Demusilfier injection @ Coalescer tank	ppm	0	40	6	10	167%
X_6	Demusilfier injection @ Electrostatic desalter	ppm	0	40	14	7	50%
X_7	Salt amount @ Production unit	gr/m^3	450	21200	6759	4786	71%
X_8	Ambient temperature	°C	0	44	28	10	36%
X_9	Fluid temperature @ inlet Coalescer tank	°C	5	65	34	15	44%
X_{10}	Fluid temperature @ inlet Pre-heater	°C	16	56	42	6	14%
X_{11}	Fluid temperature @ Pre-heater bath	°C	60	793	431	230	53%
X_{12}	Fluid temperature @ outlet Pre-heater	°C	23	65	59	4	7%
X_{13}	Fluid temperature @ inlet Electrostatic desalter	°C	14	64	60	1.4	2%
X_{14}	Fluid temperature @ inlet Electrostatic desalter	°C	14	53	42	1.3	3%
X_{15}	Pressure drop @ mixing valve (Coalescer tank)	bar	0.5	12	11	0.26	2%
X_{16}	Pressure drop @ mixing valve (Electrostatic desalter)	bar	1.2	4.5	1.37	0.22	16%
Y_1	Oil volume flow rate @ Operating unit	m^3/day	3231	17640	9662	3405	35%
Y_2	Wastewater volume flow rate	m^3/day	1.59	2432	1146	469	41%
Y_3	Salt amount @ Operating unit	gr/m^3	4	149	24	15	63%

All data are then rescaled and normalized to the unit interval [0, 1], for ease of comparison. Fig. 3 depicts box-plot representations of the variability of target state variables, i.e., Y_1, Y_2, and Y_3. Box-plot representations of system controlling variables X_i are depicted in Fig. 4.

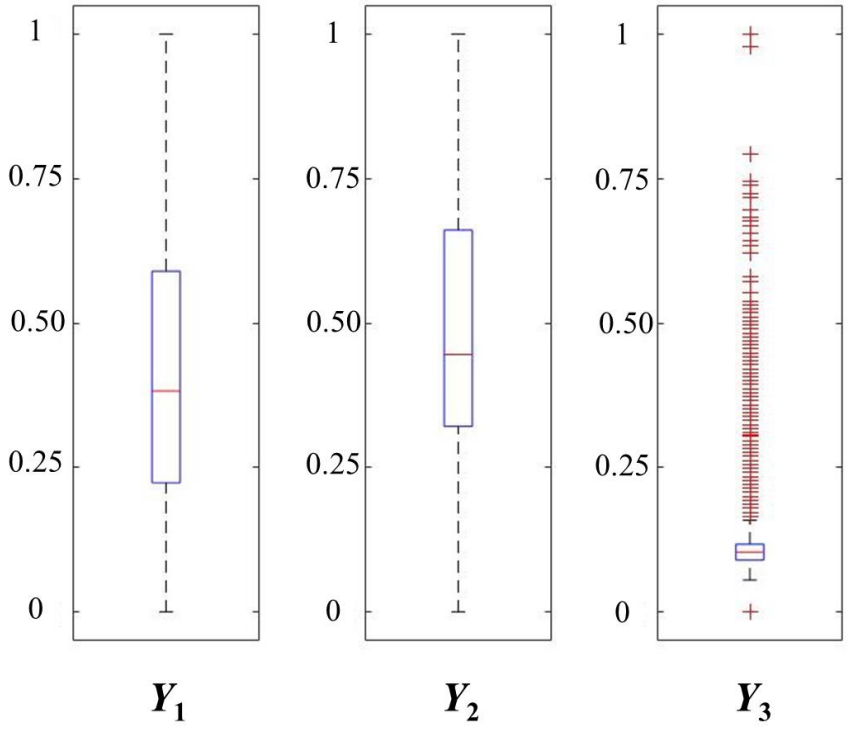
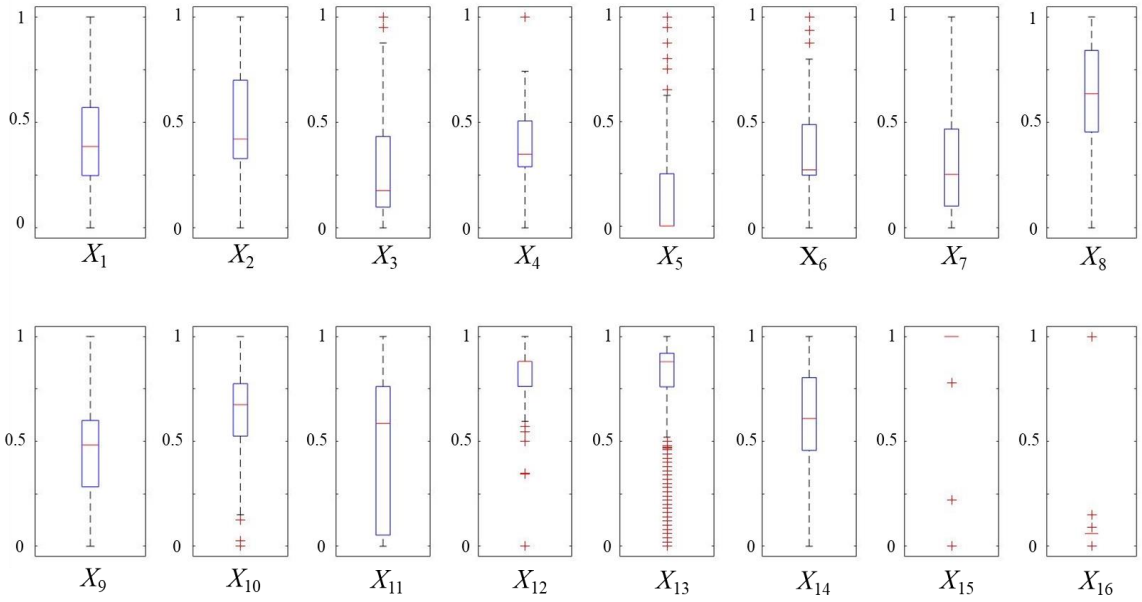


Fig. 3 *Box-plot representation of the variability of target variables (Y_1, Y_2, Y_3). Each box-plot identifies 25th, 50th (median), and 75th percentile scores. Whiskers correspond to the lowest observation comprised within 1.5 the interquartile range (IQR) of the lower quartile and to the highest observation still within 1.5 IQR of the upper quartile; crosses represent outliers.*

Most of the observations of the system performance variables Y_1 and Y_2 are roughly concentrated between the 25th and the 75th percentile of the corresponding distributions. Observed values of the produced amount of salt (Y_3) exhibits a notable number of (high valued) outliers, the highest (normalized) value of the observation still within 1.5 the interquartile range (IQR) of the upper quartile being 0.16.



403

404 ***Fig. 4*** *Box-plot representation of the variability of controlling variables (X_1 - X_{16}). Each box-plot*
405 *identifies 25th, 50th (median), and 75th percentile scores. Whiskers correspond to the lowest*
406 *observation comprised within 1.5 the interquartile range (IQR) of the lower quartile and to the*
407 *highest observation still within 1.5 IQR of the upper quartile; crosses represent outliers.*

408 Most of the controlling variables are concentrated between the 25th and the 75th percentile of
409 the corresponding distributions, with few notable exceptions: (*i*) X_{15} and X_{16} (i.e., pressure drop at the
410 mixing valves of the Coalescer Tank and Electrostatic Desalter) are typically set to some predefined
411 values (by the operator onsite) and are usually clustered around these across the entire temporal
412 observation window, with only few (high valued) outliers; (*ii*) the temperature monitored after the
413 pre-heater (X_{12}) and before the electrostatic desalter (X_{13}) exhibits several (low valued) outliers; (*iii*)
414 X_5 (i.e., the demulsifier injected at the coalescer tank) shows some high valued outliers; (*iv*) X_6 (i.e.,
415 demulsifier injected at electrostatic desalter) shows only a few high valued outliers.

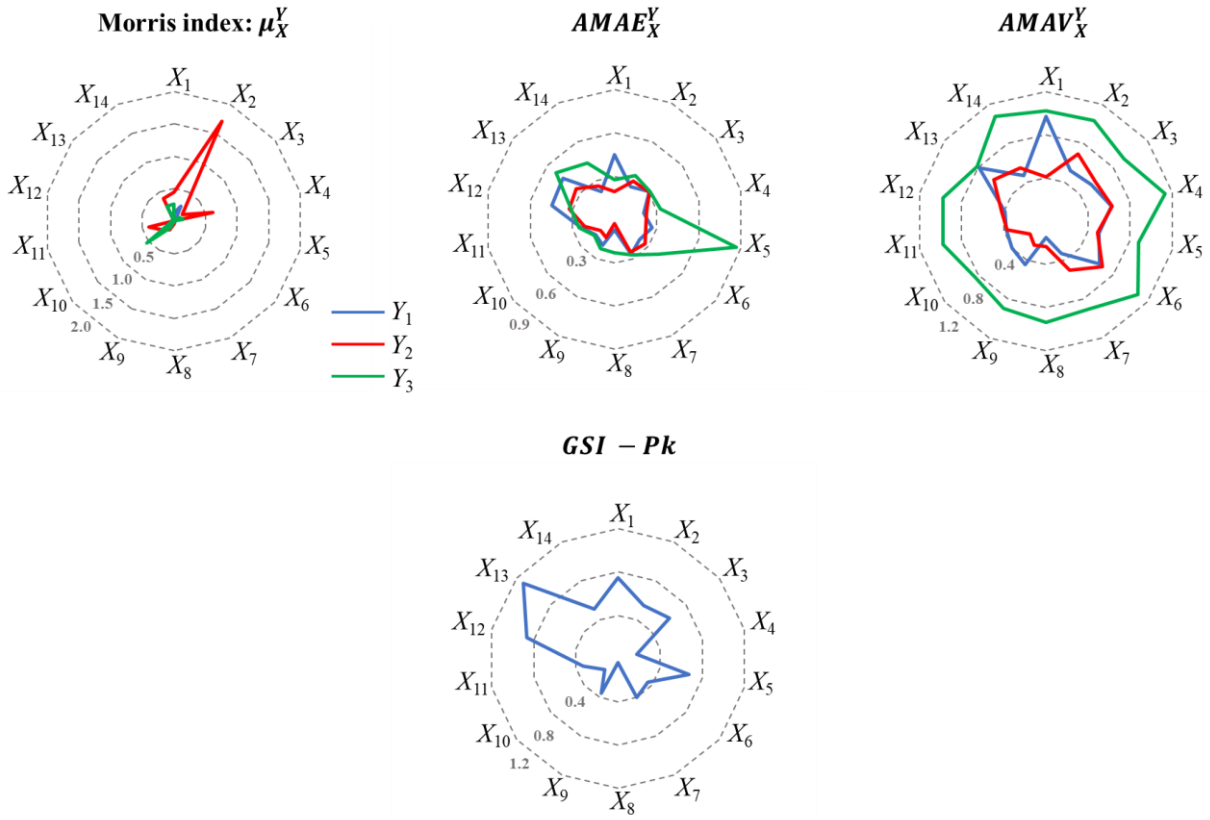
416 The analysis of the bivariate correlation among the system performance variables suggests
417 that (*i*) some degree of correlation can be noted between Y_2 (i.e., waste water volume) and both Y_1
418 (i.e., oil volume) and Y_3 (i.e., amount of salt extracted from the crude oil), as denoted by values of the

correlation coefficient equal to 0.30 and 0.36, respectively; and (*ii*) Y_1 and Y_3 appear to lack correlation (with a low value of the correlation coefficient, equal to -0.06). These documented correlations suggest the opportunity of considering also a PCA-based sensitivity analysis of the kind illustrated in Section 2.3.3 to analyze the sensitivity of the set of multivariate target variables Y_1, Y_2, and Y_3 to the state variables X_1 - X_{16}.

3.1. Global Sensitivity Analysis

Following the approach outlined in Section 2.3, we quantify sensitivity of the target state variables Y_k to each input variable X_i through evaluation of Morris indices $\mu_{X_i}^{Y_k}$ (5), the moment-based $AMAE_{X_i}^{Y_k}$ and $AMAV_{X_i}^{Y_k}$ indices (6)-(7), and the PCA-based sensitivity index (9). A depiction of the ensuing results is offered in Fig. 5.

We evaluate $\mu_{X_i}^{Y_k}$ upon relying on the available dataset and identifying trajectories in the parameter space Γ along which all values of the controlling variables are (approximately) constant with the exception of X_i. Fig. 5a illustrates $\mu_{X_i}^{Y_k}$ evaluated for the available measurement records. We note that in some cases there is only a limited number of trajectories that can be extracted from the measurement records to satisfy the conditions required for the evaluation of the local derivatives at given observation points in Γ. As such, Morris indices cannot unambiguously evaluate sensitivity of Y_k to all X_i variables and we complement this set of results through the $AMAE_{X_i}^{Y_k}$ (Fig. 5b) and $AMAV_{X_i}^{Y_k}$ (Fig 5c) indices.



437

438 ***Fig. 5*** *Radar diagram representations of: (a) Morris ($\mu_{X_i}^{Y_k}$), (b) $AMAE_{X_i}^{Y_k}$, and (c) $AMAV_{X_i}^{Y_k}$ indices*

439 *of Y_k with respect to X_i; (d) PCA-based sensitivity indices, GSI_i , quantifying the effects of X_i on the*

440 *target multivariate set of quantities.*

441 When assessed jointly on the basis of these three indices, none of the target state variables Y_k

442 appears to be influenced by X_{15} and X_{16}. Thus, we omit hereafter any additional comments on possible

443 relationships between Y_k and these controlling variables.

444 Comparing Fig. 5b - 5c, it is clear that all variables Y_k show non negligible sensitivities (in

445 terms of $AMAE_{X_i}^{Y_k}$ and $AMAV_{X_i}^{Y_k}$) to X_1 - X_{14}. In other words, the mean and variance of each Y_k are in

446 general influenced, albeit with differing extents, by all state variables X_1 - X_{14}. We also note that (*i*)

447 Y_1 and Y_2 do not show significant sensitivity to X_8 (i.e., ambient temperature) which is however

448 influential to Y_3; (*ii*) the variability of Y_1 is mainly affected by X_1 (i.e., volume of crude oil served to

449 the system) and X_{13} (i.e., temperature of the mixture fluid monitored before the Electrostatic Desalter),

(*ii*) the average change of the first two statistical moments of Y_2 is mainly affected by knowledge on X_2-X_6 (i.e., injected volume of fresh water and demulsifier amount) and X_{13}, while (*iii*) the mean and variance of Y_3 is almost uniformly influenced by all controlling variables.

We complete our sensitivity analysis by exploring the results stemming from the PCA-based GSA described in Section 2.2.3. While a general criterion to establish if a given set of principal components provides a satisfactory representation of the full variability of a dataset is still lacking, a percentage larger than 75% of the total variance described by (a given number of) principal components is typically deemed sufficient to grasp the main features encapsulated in the original data. Results of our PCA (details not shown) suggest that almost 60.5% of the variance of $\underline{\underline{\mathbf{Y}}}^*$ is explained by considering the first principal component, P_1, the second (P_2) and third (P_3) principal components explaining almost 31.5% and 8.25% of the total variance of $\underline{\underline{\mathbf{Y}}}^*$, respectively. These results indicate that relying on the first two principal components (P_1 and P_2) enables us to retain a significant amount of information about the set of target variables. Considering Fig. 5d, variable X_8 is characterized by negligible value of the associated PCA-based sensitivity index, suggesting that the variability of this quantity is not significantly influential to the target multivariate set of state variables.

Juxtaposing all of these results suggests that we can reduce the dimensionality of the space of the controlling variables, **X**. Accordingly, we consider model development and testing in Section 3.2 by relying on the reduced parameter space formed by variables X_1 - X_{14} and include these in vector $\mathbf{X}_R$.

3.2. Candidate Interpretive Models and Model Selection Criteria

We construct a set of 12 candidate models (listed in Table 2) to be potentially employed for (*a*) the interpretation of the data associated with the state variables considered and representing effectiveness of the crude oil desalting/demulsification process and (*b*) assessment of future

performance scenarios of the plant. Four of these models (M_1 - M_4) are constructed through linear regression, the formulation of the remaining eight nonlinear models (M_5 - M_{12}) being grounded on training of ANNs (see Section 2.3). Models (*i*) M_1, M_5, and M_9 are constructed upon relying on the reduced set of variables included in $\mathbf{X}_R$, following the sensitivity analysis outlined in Section 3.1 (i.e., for a total of 14 controlling variables); (*ii*) M_2, M_6, and M_{10} are based on $\mathbf{X}_R$ and the 3 dummy variables U_1 - U_3 described in Section 2.2 (resulting in a total of 17 controlling variables) with the aim of including information about the desalting unit from which observation records are collected from; (*iii*) M_3, M_7, and M_{11} are based on $\mathbf{X}_R$ and the 5 dummy variables Q_1 - Q_5 described in Section 2.2 (resulting in a total of 19 controlling variables), with the specific aim of including information about the kind of demulsifier injected to the system while being agnostic about the type of desalting unit associated with collected information records; and finally (*iv*) M_4, M_8, and M_{12} are grounded on $\mathbf{X}_R$ and the 8 dummy variables introduced in Section 2.2 (i.e., U_1 - U_3 and Q_1 - Q_5) (for a total of 22 controlling variables). We structure the ANNs either according to a single (i.e., models M_5, M_6, M_7, and M_8) or double (i.e., models M_9, M_{10}, M_{11}, and M_{12}) hidden layer(s). Quantities H_1 and H_2 (in Table 2 indicate the number of neurons in the first and second hidden layer of the ANNs, respectively. The number (m) of parameters to be estimated for each candidate model is also listed in Table 2. Model assessment is performed upon relying on normalized (to the unit interval [0, 1]) values of the data.

Characterization of the linear models is straightforward and relies on estimating model parameters through minimization of (10) (or, equivalently, (11)). The training process of the ANNs includes feeding sample batches (taken randomly from the available measurement records) as input vectors to the network, evaluating prediction errors (in terms of mean square error, *MSE* (10), or, equivalently, *NLL* (11)) with respect to measured values of the desired model output, and then tuning the weights of the network to minimize such errors [49].

Table. 2 *Characteristics of the linear regression models (M_1 - M_4) and ANN-based nonlinear (M_5 - M_{12}) candidate models to represent desalting process. Quantities H_1, and H_2 indicate the number of neurons in the first and second hidden layer of the ANNs, respectively; R and m denote the number of controlling variables included in a given candidate model and the associated number of parameters to be estimated, respectively.*

	Linear regressions				**Nonlinear ANNs**							
Candidate Model	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}
R	**14**	**17**	**19**	**22**	**14**	**17**	**19**	**22**	**14**	**17**	**19**	**22**
H_1	-	-	-	-	28	34	38	44	28	34	38	44
H_2	-	-	-	-	0	0	0	0	14	17	19	22
M	42	51	57	66	392	578	722	968	826	1207	1501	2002
MSE (N=1200)	0.0046	0.0041	0.0043	0.0039	0.0037	0.0034	0.0034	**0.0030**	0.0038	0.0033	0.0033	**0.0030**
MSE (N=1820)	0.0046	0.0043	0.0044	0.0041	0.0041	0.0036	0.0038	0.0034	0.0039	0.0037	0.0034	**0.0033**

Table 2 lists MSE values (10) associated with the set of 12 candidate models employed to represent the desalting process and subject to calibration on the available N data. In order to assess the robustness of the models and of the calibration procedure, all models have been calibrated by (i) considering all N = 1820 available data and (ii) considering only a subset of N = 1200 data, randomly selected within the collection of 1820 samples.

As expected, the value of *MSE* is generally lower when information about the type of desalting unit or demulsifier is included in the model or when a non-linear model structure is considered. Our results reveal that models M_8 and M_{12} are characterized by an identical value of $MSE(N = 1200)$ and very close values of $MSE(N = 1820)$. This suggests a weak improvement of the results (in terms of *MSE*) by inclusion of a second hidden layer in the construction of the ANN models. In general, knowledge about the desalting unit from which observations are collected is more relevant (for reducing *MSE*) than information about the injected demulsifier.

The ability of the calibrated models to interpret the process considered is assessed upon relying on the model selection criteria (12) - (13). Fig. 6 illustrates AIC_c, and BIC evaluated for all calibrated candidate models (M_1 - M_{12}) and upon relying on the selected subset or on the complete set of measurement records. Results in Fig. 6 clearly show that the performance of the model (in terms of both BIC and AIC_c) decreases when model complexity is increased, i.e., transitioning from a linear to a non-linear model. The value of AIC_c is less dependent on the degree of model complexity (as evaluated in terms of linear vs. non-linear structure and/or number of model parameters) than BIC. Such a discrepancy between BIC and AIC_c is linked to the different weighting of the number m of model parameters and of the number of data $(O \times N)$ appearing in formulations (12) and (13). Our results show that model M_4 (linear model considering reduced input parameter space while including whole set of dummy variables **U** and **Q**) provides the best performance in terms of BIC and AIC_c, with a value of posterior model probability ($p(M_4 \mid \underline{\underline{\mathbf{Y}}}^*)$, evaluated through (14)) larger than 98%. On the basis of these results, we select M_4 for the additional analyses described in the following, as it provides (*i*) adequate capability of reproducing measurement records, and (*ii*) simplicity of model structure.

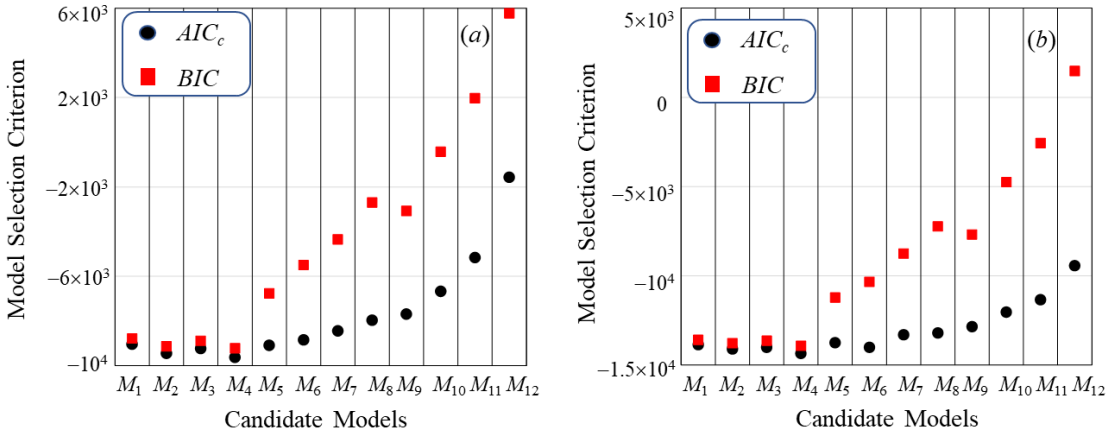


Fig. 6 *Values of model selection criteria (AIC_c and BIC) evaluated for the candidate models (M_1 - M_{12} introduced in Table 3) calibrated on N = (a) 1200 and (b) 1820 measurement records.*

The estimated values of the (linear regression) coefficients associated with the terms appearing in model M_4 when using the subset of 1200 data are depicted in Fig. 7. Similar results have been obtained when model construction relies on the complete data set. Our results show that variable X_1 is associated with the highest coefficient value (i.e., = 0.96), this result being consistent with the observation that the amount of processed oil (Y_1) should closely correspond to the volume of crude oil served to the desalting unit. Evaluation of the amount of produced wastewater (Y_2) is mainly driven by the weights evaluated for X_{12} and X_{13} (i.e., fluid temperature at the Pre-heater and Electrostatic Desalter), as well as information on the particular desalting unit employed in the plant (as rendered through U_1, U_2, and U_3). The negative sign of the coefficient associated with X_1 indicates that an increase of the amount of crude oil served to the system correspond to a decrease (for fixed values of X_2 - X_{14}) of extracted wastewater (Y_2) from the crude oil. The separated amount of salt (Y_3) at the output of the unit is rendered by this model via a high weight of (*a*) the volume of the fresh water (X_3) injected at Coalescer tank to the system, (*b*) the fluid temperature at the Electrostatic Desalter (X_{14}), and (*c*) the amount of demulsifier (X_5) injected to the system at the Coalescer Tank. Considering that the model is assessed as a best estimate of the system behavior, this result is consistent with the sensitivity displayed by these parameters to the mean of Y_3, as expressed through $AMAE_{X_i}^{Y_3}$ (see Fig. 5b). The positive sign and the value of the model coefficients associated with X_3, X_5 and X_{14} indicate a beneficial influence of combining (*i*) injection of fresh water, (*ii*) injecting demulsifier and (*iii*) desalting at the Electrostatic Desalter for extraction of salt from crude oil.

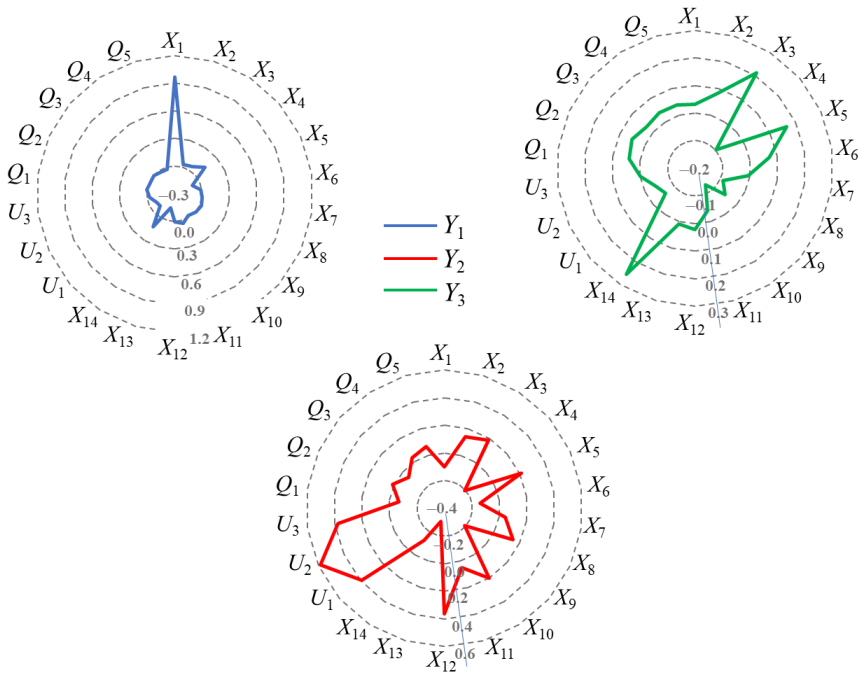


Fig. 7 *Estimated values of the linear regression coefficients associated with model M_4 and controlling variables (i.e., X_1 - X_{14}, U_1 - U_3, and Q_1 - Q_5) after calibration and referring to target variables: Y_1 (produced oil), Y_2 (produced wastewater), and Y_3 (amount of extracted salt from the analyzed system).*

The inclusion of available information on the desalting units from which measurements are collected (as embedded in U_1 - U_3) and the type of demulsifier injected to the system (i.e., Q_1 - Q_5) is not effective to represent the behavior of Y_1 and Y_3. Otherwise, the impact of information content associated with U_1 - U_3 is notable to model Y_2. One can also note that (*i*) the amount of injected fresh water (X_2) and demulsifier (X_6) at the Electrostatic Desalter; and (*ii*) ambient temperature (X_8), fluid temperature before the Coalescer Tank (X_9), and fluid temperature before the Pre-heater bath (X_{10}) exert some degree of secondary influence to estimate Y_1 and Y_3. Information on salt content of the fluid at the entrance of desalting unit (X_7) is only weakly influential to Y_1, Y_2, and Y_3. In this context, all of the results embedded in Fig. 7 are consistent with the GSA findings included in Fig. 5.

Fig. 8 depicts scatterplots of observed and modeled values of the target variables, as obtained with M_4 after parameter estimation relying on the subset of 1200 observations (red symbols). The

remaining 620 available observations (not employed during the calibration) are also depicted in the scatterplot (blue symbols) to provide an appraisal of the robustness of the model. These results show the remarkable capability of the model to reproduce observed values of Y_1. Otherwise, model calibration results show acceptable correspondence with data when considering Y_2 and Y_3.

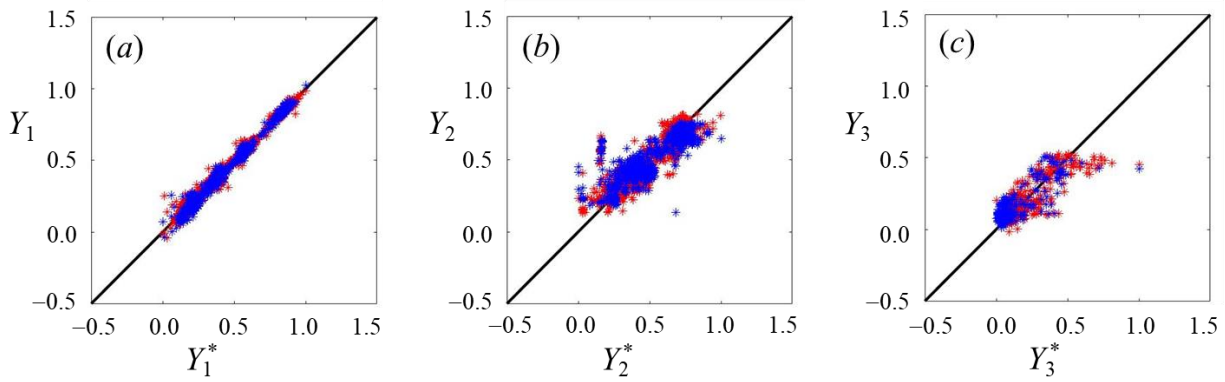


Fig. 8 *Scatterplot of observed and predicted values of Y_k obtained through the linear regression model M_4. Results are depicted for model calibration based on the subset comprising 1200 measurements (red symbols); results associated with the remaining 620 available observations (not employed during calibration) are also reported (blue symbols).*

Finally, Fig. 9 juxtaposes empirical distributions (i.e., *sample pdf*) of the observed values of Y_k and those rendered by the corresponding values evaluated through model M_4. The Kullback-Leibler divergence (*KLD*) [50] is used here as a metric to quantify the difference between the two distributions. Results obtained by considering model calibration against the subset of 1200 observation records (blu lines) and the complete set are depicted (red lines). Results for Y_1 can be considered of good quality (*KLD* = 0.02 and 0.06 for 1200 or 1820 measurement records, respectively). Otherwise, moderate values of discrepancy can be observed for Y_2 (*KLD* = 0.11 and 0.35, for 1200 or 1820 measurement records, respectively), Y_3 being characterized by the largest differences between the observed and model-based distributions (*KLD* = 5.1 and 1.8, for 1200 or 1820 measurement records, respectively). Note that the high values of *KLD* associated with Y_3 are

linked to the higher peak displayed by the *sample pdf* of Y_3^*, as compared to the corresponding *pdf* of $\hat{Y}_3$ (see Fig. 9c and 9f). The small values of *KLD* obtained for Y_1 are consistent with the results depicted in Fig. 8.

Additionally, we note that multiple peaks of the *sample pdf* for Y_1 and Y_2 in Fig. 9 are related to the fact that data are collected from differing desalting units. Otherwise, measurement records of the extracted salt (Y_3) collected from different desalting units are associated with a unique pronounced peak, a very small secondary peak being barely visible (Fig 9c and 9f).

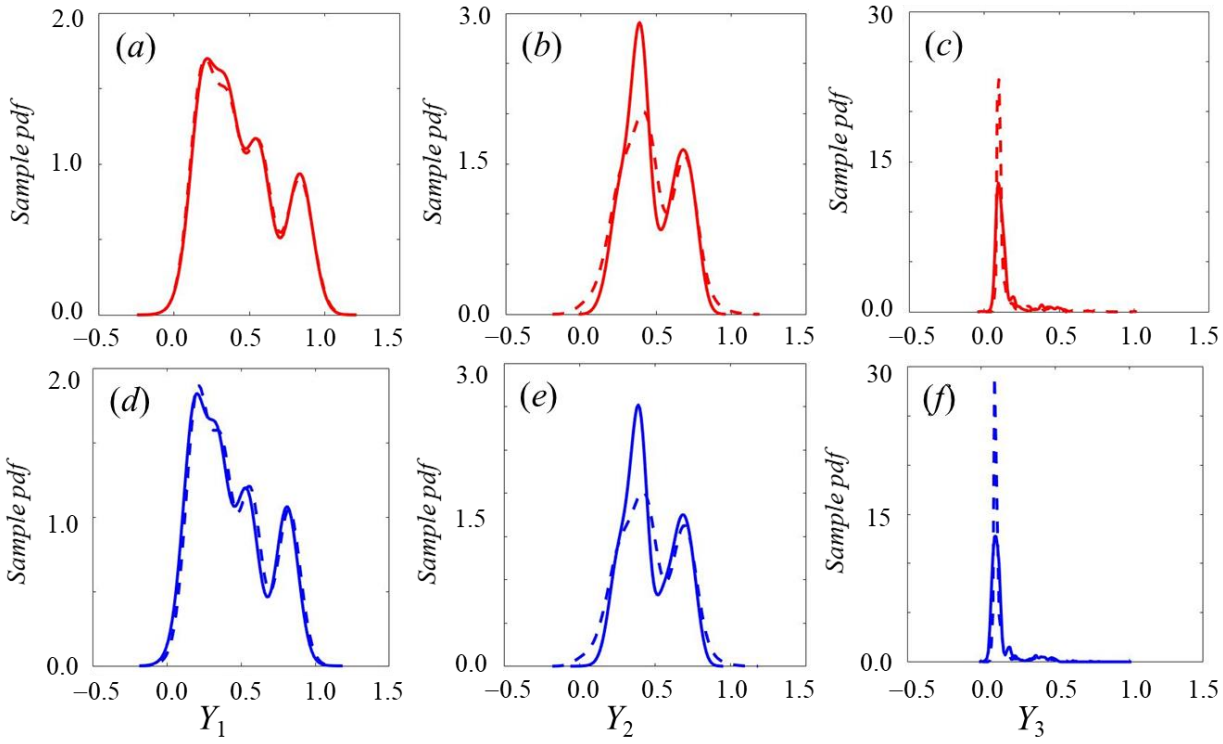


Fig. 9 *Empirical distributions (i.e., sample pdf) of the observed (dashed curves) and modeled (continuous curves) values of* Y_k. *Results are obtained by considering model calibration against the subset of 1200 (a - c) observations and the complete set of 1820 (d - f) records*.

4. Conclusions

We analyze the performance of a system of desalting/demulsification units to process crude oil before conveying it to a refinery upon relying on an original combination of (*a*) various Global Sensitivity Analysis (GSA) approaches, (*b*) machine learning techniques, and (*c*) model discrimination criteria. Our analysis is completely data-driven, as it rests on a unique data base collected across a three-year period through daily measurements from three desalting units (in the same oil field) where demulsification is performed through injection of 5 different demulsifiers in conjunction with three desalting techniques. The overall performance of the system is evaluated in terms of three main quantities, expressing (*i*) the output volume of the oil from Desalting Unit (Y_1), (*ii*) the volume of the wastewater extracted from Desalting Unit (Y_2), and (*iii*) the separated amount of salt at the output of the unit (Y_3). These are seen to depend on a total of 16 controlling variables (see Table 1), representing ambient and operational conditions in the plant. Our study leads to the following major conclusions.

1. Joint application of GSA techniques based on the Morris [32], Moment-based [33], as well as PCA-based [40] sensitivity indices enables us to clearly identify a reduced set of 14 variables that exert the highest influence on the overall performance of the industrial plant. All of these indices are evaluated numerically, upon relying on the available extensive database. We find that the variability of Y_1 (i.e., output volume of the oil from Desalting Unit) is mainly affected by the volume of crude oil served to the system, while Y_2 (i.e., volume of the wastewater extracted from Desalting Unit) is mainly influenced by the volume of injected water to the system and the amount of injected demulsifier. These quantities do not show significant sensitivity to the variation of the ambient temperature, which is in turn influential to the extracted amount of salt from crude oil at the desalting unit (Y_3).

2. A suite of 12 linear and nonlinear regression models are derived and considered as candidates for (*a*) the interpretation of the data associated with the performance of the crude oil desalting process and (*b*) the evaluation of future performance scenarios of the plant. The nonlinear models are grounded on training of Artificial Neural Networks (ANNs) on the available observations of the reduced set of controlling variables identified through the GSA. The relative skill of each of these models is evaluated via formal model identification criteria. We find that a simple linear regression model shows remarkable capabilities (with posterior model probability higher than 98%) to evaluate the performance of the crude oil demulsification process. Results from the combination of modeling techniques employed suggest that: (*a*) the volume of processed oil is mainly driven by the volume of crude oil served to the desalting unit; (*b*) the volume of extracted wastewater is mainly controlled by fluid temperature at Pre-eater and Electrostatic Desalter; (*c*) the amount of extracted salt is mainly controlled by the volume of the fresh water injected to the system, amount of demulsifier injected to the system at the Coalescer Tank, and the fluid temperature at the Electrostatic Desalter; and (*d*) the benefit of including in the modeling effort available information of the desalting units from which measurements are collected is notable to model the amount of extracted wastewater while being less marked to represent the volume of processed oil and extracted salt.
3. Our results confirm the effectiveness of a combination of demulsification and desalting techniques for the extraction of salt from crude oil. The methodological and operational approach we present is general and ready to be engineered and transferred to industrial plants where identification of the key variables governing the quality of the system behavior enables optimal design and control through an appropriate monitoring campaign.

References

[1] Hu G, Li J, Zeng G. Recent development in the treatment of oily sludge from petroleum industry: a review. J Hazard Mater 2013;261:470–90. doi:10.1016/J.JHAZMAT.2013.07.069.

[2] Olajire AA. The petroleum industry and environmental challenges. J Pet Env Biotechnol 2014;5:2157–7463. doi:10.4172/2157-7463.1000186.

[3] Adeniyi OD, Adediran AA, Adeniyi MI, Yahya MD, Afolabi EA, Adebayo IT, et al. Evaluation of the impact of Kaduna refinery effluent on river Romi. Niger J Technol Res 2017;12(1):17–21. doi:10.4314/njtr.v12i1.4.

[4] JRC. Best available techniques (BAT) reference document for the refining of mineral oil and gas industrial emissions : industrial emissions directive 2010/75/EU (integrated pollution prevention and control). Publications Office of Europian Union; 2015.

[5] Manning FS, Thompson RE. Oilfield processing of petroleum: crude oil. Pennwell books; 1995.

[6] Farrokhi F, Jafari Nasr MR, Rahimpour MR, Arjmand M, Vaziri SA. An investigation on simultaneous effects of several parameters on the demulsification efficiency of various crude oils. Asia-Pacific J Chem Eng 2017;12(6):1012-1022. doi:10.1002/apj.2142.

[7] Hamadi AS, Mahmood LH. Demulsifiers for Simulated Basrah Crude Oil. Eng Technol J 2009;28(1):54–64.

[8] Becher P. Emulsions: Theory and Practice. Oxford University Press, New York, NY (2001).

[9] Fortuny M, Oliveira CBZ, Melo RLFV, Nele M, Coutinho RCC, Santos AF. Effect of Salinity, Temperature, Water Content, and pH on the Microwave Demulsification of Crude Oil Emulsions. Energy and Feuls 2007;21(3) 1358–1364. doi:10.1021/EF0603885.

[10] Less S, Hannisdal A, Bjørklund E, Sjöblom J. Electrostatic destabilization of water-in-crude oil emulsions: Application to a real case and evaluation of the Aibel VIEC technology. Fuel 2008;87:2572–81. doi:10.1016/j.fuel.2008.03.004.

[11] Meidanshahi V, Jahanmiri A, Rahimpour MR. Modeling and optimization of two stage AC electrostatic desalter. Sep Sci Technol 2012;47:30–42. doi:10.1080/01496395.2011.614316.

[12] Dai X, Wang X, He R, Du W, Zhong W, Zhao L, Qian F. Data-driven robust optimization for

crude oil blending under uncertainty. Comp. and Chem. Eng. 2019. doi:10.1016/j.compchemeng.2019.106595

[13] Qin SJ, Chiang LH. Advances and opportunities in machine learning for process data analytics. Comp. and Chem. Eng. 2019;126:465-473. doi:10.1016/j.compchemeng.2019.04.003

[14] Siena M, Guadagnini A, Della Rossa E, Lamberti A, Masserano F, Rotondi M. A novel enhanced-oil-recovery screening approach based on Bayesian clustering and Principal-Component Analysis. SPE Reserv Eval Eng 2016;19:382–90. doi:10.2118/174315-PA.

[15] Sad CMS, Da Silva M, Dos Santos FD, Pereira LB, Corona RRB, Silva SRC, et al. Multivariate data analysis applied in the evaluation of crude oil blends. Fuel 2019;239:421–428. doi:10.1016/j.fuel.2018.11.045.

[16] Gueddar T, and Dua V. Novel model reduction techniques for refinery-wide energy optimisation. Applied Energy 2012:89(1) 117-126. doi:10.1016/j.apenergy.2011.05.056

[17] Mohammad AT, Mat SB, Sulaiman MY, Sopian K, Al-abidi AA. Implementation and validation of an artificial neural network for predicting the performance of a liquid desiccant dehumidifier Energy Convers. Manage., 2013; 67: 240-250. doi/10.1016/j.enconman.2012.12.005

[18] Mouret S, Grossmann IE, Pestiaux P. A new Lagrangian decomposition approach applied to the integration of refinery planning and crude-oil scheduling. Comp. and Chem. Eng. 2011;35:2750-2766. doi:10.1016/j.compchemeng.2011.03.026

[19] Gao X, Jiang Y, Chen T, Huang D. Optimizing scheduling of refinery operations based on piecewise linear models. Comp. and Chem. Eng. 2015;75:105-119. doi:10.1016/j.compchemeng.2015.01.022

[20] Ochoa-Estopier LM, Jobson M, Smith R.Operational optimization of crude oil distillation systems using artificial neural networks. Comp. and Chem. Eng. 2013;59:178-185. doi:10.1016/j.compchemeng.2013.05.030

[21] Gueddar T, Dua V. Disaggregation-aggregation based model reduction for refinery-wide

optimization. Comp. and Chem. Eng. 2011;35(9):1838-1856. doi:10.1016/j.compchemeng.2011.04.016

[22] Al-Qahtani K, Elkamel A. Robust planning of multisite refinery networks: Optimization under uncertainty. Comp. and Chem. Eng. 2010;34(6):985-995. doi:10.1016/j.compchemeng.2010.02.032

[23] Lamboni, M. Monod, H. and, Makowski, D. [2011] Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. Reliability Engineering and System Safety 2011;96(4):450–459. doi:10.1016/j.ress.2010.12.002

[24] Vapnik VN. The nature of statistical learning theory. Springer press, New York, 2000. doi:10.1007/978-1-4757-3264-1.

[25] Choubineh A, Ghorbani H, Wood DA, Moosavi SR, Khalafi E, Sadatshojaei E, Improved predictions of wellhead choke liquid critical-flow rates: Modelling based on hybrid neural network training learning based optimization. Fuel, 2017;207:547-560. doi:10.1016/j.fuel.2017.06.131.

[26] Ghorbani H, Wood DA, Choubineh A, Tatar A, Ghazaeipour Abarghoyi P, Madani M, Mohamadian N. Prediction of oil flow rate through an orifice flow meter: Artificial intelligence alternatives compared. Petroleum 2018: In press. doi:10.1016/j.petlm.2018.09.003.

[27] Bingöl D, Xiyili H, Elevli S, Kılıç E, Çetintaş S. Comparison of multiple regression analysis using dummy variables and a NARX network model: an example of a heavy metal adsorption process. Water and Environmental Journal 2018;32(2): 186/196 doi: 10.1111/wej.12314

[28] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Salsana M, Tarantola S. Global sensitivity analysis - The primer. Wiley, 2008.

[29] Liu H, Chen W, Sudjianto A. Relative entropy based method for probabilistic sensitivity analysis in engineering design. ASME Journal of Mechanical Design 2006; 128(2):326–336. doi:10.1115/1.2159025.

[30] Hutcheson RS, McAdams DA. A hybrid sensitivity analysis for use in early design. Journal of Mechanical Design 2010;132(11). doi:10.1115/DETC2009-87120.

[31] Dell'Oca A, Riva M, Guadagnini A. Global sensitivity analysis for multiple interpretive models with uncertain parameters. Water Resour Res 2020;55(2). https://doi.org/10.1029/2019WR025754

[32] Morris MD. Factorial sampling plans for preliminary computational experiments. technometrics 1991;33(2): 161–174

[33] Dell'Oca A, Riva M, Guadagnini A. Moment-based metrics for global sensitivity analysis of hydrological systems. Hydrol Earth Syst Sci 2017;21:6219–34. doi:10.5194/hess-21-6219-2017.

[34] Sobol' IM. Sensitivity estimates for nonlinear mathematical models. MMCE 1993; 1(4):407–414.

[35] Bianchi Janetti E, Guadagnini L, Riva M, Guadagnini A. Global sensitivity analyses of multiple conceptual models with uncertain parameters driving groundwater flow in a regional-scale sedimentary aquifer. *J. Hydrol*. 2019;574:544–556. doi:10.1016/j.jhydrol.2019.04.035

[36] Takane Y, Hunter MA. A new family of constrained principal component analysis (CPCA). Linear Algebra Appl 2011;434:2539–2555. doi:10.1016/J.LAA.2011.01.002.

[37] Reinsel GC, Velu RP. Multivariate reduced-rank regression. 136. Springer New York; 1998. doi:10.1007/978-1-4757-2853-8.

[38] Scholz M, Vigário R. Nonlinear PCA: a new hierarchical approach. Presented at 10th Eurorean Symposium on Artificial Neural Networks, Bruges, Belgium, April 24-26, 2002.

[39] Ranaee E, Porta G, Riva M, Guadagnini A. Investigation of saturation dependency of oil relative permeability during WAG process through linear and non-linear PCA. Presented at ECMOR XIV–14th European Conference on the Mathematics of Oil Recovery, Catania, Italy, September 2014. doi:10.3997/2214-4609.20141800.

745 [40] Oja E. A simplified neuron model as a principal component analyzer. Springer-Verlag 1982.
746 [41] Neurosolution version 5.07 (2007) Neuro dimension. Gainesville Inc., Florida, (USA). 2007.
747 [42] Ranaee E, Riva M, Porta GM, Guadagnini A. Comparative assessment of three-phase oil
748 relative permeability models. Water Resour Res 2016;52:5341–56.
749 doi:10.1002/2016WR018872.
750 [43] Hurvich CM, Tsai CL. Regression and time series model selection in small samples.
751 Biometrika 1989;76(2):297–307. doi:10.1093/biomet/76.2.297
752 [44] Schwarz G. Estimating the dimension of a model. Ann Stat 1978;6:461–464.
753 doi:10.1214/aos/1176344136.
754 [45] Carrera J, Neuman SP. Estimation of aquifer parameters under transient and steady state
755 conditions: 1. Maximum likelihood method incorporating prior information. Water Resour.
756 Res. 1986;22(2):199-210. doi:10.1029/WR022i002p00199.
757 [46] Ranaee E, Porta GM, Riva M, Blunt MJ, Guadagnini A. Three-phase oil relative permeability
758 prediction through a sigmoid-based model, Journal of Petroleum Science and Engineering
759 2015;26:190–200. doi:10.1016/j.petrol.2014.11.034.
760 [47] Ranaee E, Moghadasi L, Inzoli F, Riva M, Guadagnini A. Identifiability of parameters of three-
761 phase oil relative permeability models under simultaneous water and gas (SWAG) injection.
762 Petroleum Science and Engineering 2017;159:942–951. doi: 10.1016/j.petrol.2017.09.062.
763 [48] Ye M, Meyer PD, Neuman SP. On model selection criteria in multimodel analysis. Water
764 Resour. Res. 2008;44(3):W03428, 10.1029/2008WR006803
765 [49] Moghadasi L, Ranaee E, Inzoli F, Guadagnini A. Petrophysical well Log analysis through
766 intelligent methods. Presented at SPE Bergen One Day Seminar, Bergen, Norway, April 2017.
767 doi:10.2118/185922-MS.
768 [50] Kullback S, Leibler RA. On information and sufficiency. Ann. Math. Statistics 1951;22(1): 79-
769 86. doi:10.1214/aoms/1177729694.