Improving Process Monitoring via Dynamic Multi-Fidelity Modeling

Rastislav Fáber [*] **Marco Vaccari** [**]
Riccardo Bacci di Capaci [**] **Karol Ľubušký** [***]
Gabriele Pannocchia [**] **Radoslav Paulen** [*]

[*] *Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, 812 37 Bratislava, Slovakia*
[**] *Department of Civil and Industrial Engineering, University of Pisa, 561 22 Pisa, Italy*
[***] *Slovnaft, a.s., 824 12 Bratislava, Slovakia*

Abstract: We study real-time process monitoring, where employed online sensors yield inaccurate information. A multi-fidelity (MF) modeling approach is adopted that integrates dynamic information from online, low-fidelity (LF) data with infrequent, high-fidelity (HF) laboratory measurements. The proposed methodology is demonstrated on a composition monitoring problem derived from real oil refinery operations. The developed MF model exhibits a significant improvement in accuracy with respect to both LF data (online sensor) and the HF model (standard soft sensor). The results highlight the potential of MF modeling for improving process monitoring and control through the integration of diverse data sources.

Keywords: Dynamic models, Multi-fidelity models, Gaussian processes, Feature selection

1. INTRODUCTION

Effective monitoring and control are critical for optimizing industrial operations (Colombo et al., 2017; Zhu et al., 2018; Yin and Kaynak, 2015), where a prerequisite is a precise and timely gathering of process information by sensing of the underlying quantities. Direct sensing can be performed by hard sensors via online or lab analyzers, where lab sensorics is used when an online alternative is unavailable, costly, or unreliable. Soft sensors (Kadlec et al., 2009) can be used to obtain the desired information in real time. They employ prediction models combining other online-sensor data. Online hard/soft sensors seemingly circumvent the need for lab analyses, yet on-demand calibration by the lab measurements is still necessary to maintain their accuracy.

Consequently, industrial plants involve several online sensors, which provide large amounts of data, and are even accompanied with infrequent lab measurement records. These datasets only rarely find an advanced use, beyond the single-point calibration. The so-called multi-fidelity (MF) modeling (Giselle Fernández-Godino, 2023) promises to exploit various related or duplicate datasets and fuse the information contained. Our goal here is to find a joint use of laboratory high-fidelity (HF) data and the low-fidelity (LF) data from online sensors to improve the industrial monitoring.

Data-based modeling is at the core of our endeavor. These models leverage historical data and have been applied successfully across several industrial sectors (Bahramian et al., 2023). There are several milestones of designing a data-based model: data processing, selection of model structure (linear/non-linear, static/dynamic, parametric/non-parametric), feature selection, model training. Above all, a crucial importance must be paid to the feature selection, which is a key prerequisite for choosing an appropriate structure of a prediction model. Techniques such as principal component analysis, partial least squares, and many more (Bastos et al., 2022) have been used successfully. Non-parametric approaches can be used in model development, such as Gaussian Process Regression (GPR) (Rasmussen, 2004) that has also shown good performance when applied to industrial tasks (Ge et al., 2011). Gaussian processes also found use in MF modeling (Bradford et al., 2020). When dealing with high-dimensional datasets, feature selection is paramount (Perdikaris et al., 2015).

This study investigates MF models for industrial monitoring by integrating online sensor data with laboratory measurements to boost predictive accuracy. Combining frequent LF data with precise HF data, our approach supports reliable, timely monitoring. Using GPR within the MF framework, this method provides output predictions and quantifies uncertainties, reinforcing decision robustness. Additionally, we study the use of dynamic models within the MF modeling frame-

[⋆] E-mail address: `rastislav.faber@stuba.sk` (R. Fáber).

This work is funded by the Slovak Research and Development Agency under the project APVV-21-0019, by the Scientific Grant Agency of the Slovak Republic under the grant 1/0263/25, and by the European Commission under the grant 101079342 (Fostering Opportunities Towards Slovak Excellence in Advanced Control for Smart Industries). RP acknowledges the financial support by the European Commission under the grant scheme NextGenerationEU project no. 09I03-03-V04-00530.

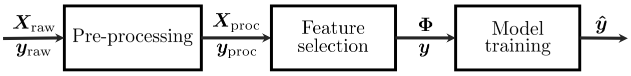
$\boldsymbol{X}_{\text{raw}}$
$\boldsymbol{y}_{\text{raw}}$
Pre-processing
$\boldsymbol{X}_{\text{proc}}$
$\boldsymbol{y}_{\text{proc}}$
Feature selection
$\boldsymbol{\Phi}$
$\boldsymbol{y}$
Model training
$\hat{\boldsymbol{y}}$

Fig. 1. Model development flowchart.

work. We outline the key steps of model development in Fig. 1.

2. PROBLEM DEFINITION

We assume two distinct datasets: historical LF data from online sensors and HF data from laboratory measurements. The LF data are collected continuously at a high sampling frequency. In contrast, HF data are obtained less frequently through, serving as reference points for calibrating and validating the online sensors. Let us define the time instant sets as follows: $\mathcal{T}_{\text{LF}} = \{t_1, t_2, \ldots, t_{n_{\text{LF}}}\}$ representing the time instants when LF data are taken, and $\mathcal{T}_{\text{HF}} = \{\tau_1, \tau_2, \ldots, \tau_{n_{\text{HF}}}\} \subseteq \mathcal{T}_{\text{LF}}$ representing the time instants of laboratory samples. Typically, $n_{\text{HF}} \ll n_{\text{LF}}$. The associated index sets are denoted as $\mathcal{I}_{\text{LF}}$ and $\mathcal{I}_{\text{HF}}$, respectively. We seek a predictive model that can accurately relate the process variables $\boldsymbol{x} \in \mathbb{R}^{n_{\text{x}}}$ to the desired output $y \in \mathbb{R}$.

A linear model predicting the i^{th} instant reads as:

$$\hat{y}_i = \boldsymbol{\theta}^\intercal \boldsymbol{\varphi}_i + \epsilon_i, \tag{1}$$

where $\hat{y}_i$ is the predicted output, $\boldsymbol{\varphi}_i$ is a vector of regressors, and $\boldsymbol{\theta}$ is a vector of model parameters. The error term $\boldsymbol{\epsilon}_i$ captures discrepancies due to measurement errors or unmodelled variability. A nonlinear model can be defined equivalently as:

$$\hat{y}_i = f(\boldsymbol{\theta}_i^\intercal, \boldsymbol{\varphi}_i). \tag{2}$$

The specific form of $\boldsymbol{\varphi}_i$ depends on whether a *static* or *dynamic* modeling approach is selected:

A) *Static model* with $\boldsymbol{\varphi}_i = \boldsymbol{x}_i$.
B) *Dynamic model* with $\boldsymbol{\varphi}_i$ that includes lagged values, given by $\boldsymbol{\varphi}_i = (\boldsymbol{x}_{i-1}, \ldots, \boldsymbol{x}_{i-j}, y_{i-1}, \ldots, y_{i-k})^\intercal$.

Here, j and k represent the respective numbers of lagged values for the input vector $\boldsymbol{x}$ and output y.

3. METHODOLOGY

3.1 Pre-processing

Raw data denoted as $\boldsymbol{X}_{\text{raw}}$ for the input features and $\boldsymbol{y}_{\text{raw}}$ for the dependent variable, often exhibit significant correlations and contain non-random noise (plant start-ups or shutdowns, various disturbance scenarios), outliers, and missing values. These must be addressed before model training by reducing noise, handling missing data and outliers (Fáber et al., 2024).

Outlier detection can be performed using the Minimum Covariance Determinant (MCD) method (Rousseeuw and Driessen, 1999), a robust statistical technique that leverages Mahalanobis distance to identify anomalous data points by minimizing the determinant of the sample covariance matrix $\boldsymbol{S}$. The distance measure is given by:

$$d_i = \sqrt{(\boldsymbol{v}_i - \boldsymbol{\mu})^\intercal \, \boldsymbol{S}^{-1} \, (\boldsymbol{v}_i - \boldsymbol{\mu})}, \tag{3}$$

where $\boldsymbol{v}_i$ stands for data vector as $\boldsymbol{V} = (\boldsymbol{X} \; \boldsymbol{y})^\intercal$ and $\boldsymbol{\mu}$ is the mean vector of the same dimension. 3σ rule can be applied to classify outliers.

In the context of dynamic model training, standard interpolation methods can be applied to replace outliers while preserving the temporal relationships between variables.

3.2 Feature Selection

The objective of Feature Selection (FS) is to identify the subset of process variables $\boldsymbol{\Phi}$ that contribute most significantly to the accurate and reliable prediction. By including only the most relevant variables, we improve model interpretability and computational efficiency and reduce the likelihood of overfitting (Li et al., 2017). The most common methods for FS are briefly reviewed.

Principal Component Regression (PCR) combines Principal Component Analysis (PCA) with regression to improve predictive performance. PCA reduces dimensionality by transforming the original dataset into a new set of uncorrelated variables, maximizing variance and minimizing information loss (Pearson, 1901).

Partial Least Squares (PLS) Regression identifies latent variables that explain the most variance in both the dependent variable $\boldsymbol{y}_{\text{proc}}$, and independent variables $\boldsymbol{X}_{\text{proc}}$, effectively handling multicollinearity. The PLS approach can be formulated as:

$$\boldsymbol{X}_{\text{proc}} = \boldsymbol{T}\,\boldsymbol{P}^\intercal + \boldsymbol{E}, \tag{4a}$$

$$\boldsymbol{y}_{\text{proc}} = \boldsymbol{U}\,\boldsymbol{q} + \boldsymbol{r}. \tag{4b}$$

Here, $\boldsymbol{T}$ and $\boldsymbol{U}$ denote the score matrices, $\boldsymbol{P}$ and $\boldsymbol{q}$ represent the weight matrix and vector, respectively, and $\boldsymbol{E}$ and $\boldsymbol{r}$ are the error terms (Geladi and Kowalski, 1986).

LASSO regression incorporates an ℓ_1 penalty to promote model sparsity. It solves the following problem:

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^{N} (y_i - \boldsymbol{\theta}^\intercal \boldsymbol{\varphi}_i)^2 + \lambda \|\boldsymbol{\theta}\|_1, \tag{5}$$

where λ is a tuning parameter that controls the strength of the penalty (Santosa and Symes, 1986).

Stepwise Regression (SR) systematically adds or removes variables in a multilinear model based on their statistical significance. At each step, the p-value of an F-statistic determines whether to add or remove a regressor.

Alternatively, criteria such as the Akaike Information Criterion (AIC), which minimizes information loss; the Corrected Akaike Information Criterion (AICc), which adjusts AIC for small sample sizes; and the Bayesian Information Criterion (BIC), which penalizes model complexity more strictly to favor simpler models for larger datasets, can also guide the FS process (Efroymson, 1960).

In addition to the aforementioned methods, the model structure can be enhanced by expert process knowledge.

3.3 Model Training

Model (1) can be fitted to the available data via:

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|\boldsymbol{y}_l - \hat{\boldsymbol{y}}_l\|_2^2 \equiv \min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^{n_l} (y_{l,i} - \hat{y}_{l,i})^2 \quad (6a)$$

$$\text{s.t. } y_{l,i} = \boldsymbol{\theta}^\intercal \boldsymbol{\varphi}_{l,i} \text{ or } y_{l,i} = f(\boldsymbol{\theta}^\intercal, \boldsymbol{\varphi}_{l,i}),\ \forall i \in \mathcal{I}_l \quad (6b)$$

where $l \in \{\mathrm{HF}, \mathrm{LF}\}$ distinguishes the HF or LF model.

3.4 Multi-fidelity Model Training

The MF combines trained LF and HF models. The regressor matrix $\boldsymbol{\Phi}_{\mathrm{MF}}$ integrates all data:

$$\boldsymbol{\varphi}_{\mathrm{MF},i} = \left(\boldsymbol{\varphi}_{\mathrm{HF},i}^\intercal, \boldsymbol{\varphi}_{\mathrm{LF},i}^\intercal, \hat{y}_{\mathrm{LF},i}, y_{\mathrm{LF},i}\right)^\intercal,\ \forall i \in \mathcal{I}_{\mathrm{HF}}. \quad (7)$$

After an obligatory FS step, the MF model can be trained by GPR. The GP can be defined as:

$$\hat{\boldsymbol{y}}_{\mathrm{MF}} \sim \mathcal{GP}(m(\boldsymbol{\kappa}), k(\boldsymbol{\kappa}, \boldsymbol{\kappa}')), \quad (8)$$

where $\boldsymbol{\kappa}$ represents the predictors derived from the feature-selected dataset $\boldsymbol{\Phi}_{\mathrm{MF}}$, $m(\boldsymbol{\kappa})$ is the mean function representing the expected value of the output, and $k(\boldsymbol{\kappa}, \boldsymbol{\kappa}')$ is the covariance function (kernel) that defines the correlation between the output data.

4. AN INDUSTRIAL CASE STUDY

The alkylation process is essential in refineries for producing high-octane branched isoparaffins, i.e., alkylate, a key component of clean gasoline. The production involves the reaction of C_3–C_4 olefins with isobutane (i-C_4) using an acid catalyst. The reaction pathway initiates with the protonation of the olefin, leading to the formation of carbonium cations that react further to produce C_8 isomers (Speight, 2020; Pall, 2018). Maintaining an optimal ratio of reactants is critical for efficient alkylate production. Online analyzers, strategically placed in the plant (see Fig. 2), provide real-time data by measuring concentrations of key components in the feed and recycle streams, enabling effective process control. A comprehensive dataset, $\boldsymbol{X}_{\mathrm{raw}}$, was collected over six months, containing more than 1085 process variables from online sensors, analyzers, and laboratory samples, yielding $N \times 1085$ data points. The study particularly focuses on the Analyzer A3 and its associated laboratory analysis, $\boldsymbol{y}$. The analyzer monitors the i-C_4 concentration in the recycle stream but has exhibited inconsistent performance. These inconsistencies often lead to excess i-C_4 recycling, which imposes additional downstream load due to the need for further heating and treatment. Therefore, accurately predicting and correcting deviations in analyzer data is essential to maintaining efficient production and reducing by-products.

5. IMPLEMENTATION

We apply the data pre-processing methods outlined in Section 3.1 to the raw data $\boldsymbol{X}_{\mathrm{raw}}$ and the corresponding raw output data $\boldsymbol{y}_{\mathrm{raw}}$, reducing the dataset from $N \times 1085$ to $N \times 256$ by eliminating constant and highly correlated variables. Subsequently, we split the pre-processed data, $\boldsymbol{X}_{\mathrm{proc}}$ and $\boldsymbol{y}_{\mathrm{proc}}$, into training and testing sets using a $60/40\,\%$ ratio. This approach helps achieve a well-distributed dataset, ensuring that both sets capture similar trends while preserving chronological date-time sequences. The chosen time frame spans four months for training and two months for testing, which is particularly relevant as seasonal variations during this period can significantly influence process dynamics, including the characteristics of processed olefins and isobutane.

To select relevant, non-redundant features $\boldsymbol{\Phi}$, we apply the statistical FS methods of Section 3.2 on the pre-processed dataset $\boldsymbol{X}_{\mathrm{proc}}$, which can be derived from either LF or HF sources (see Fig. 1). Additionally, in collaboration with our industrial partner, we review the top-ranked variables to ensure their practical relevance. If a selected variable is unsuitable (e.g., an alarm or a non-maintained sensor), we replace it with the next highest-ranked option, maintaining the intended number of features while improving interpretability.

In this stage, we train a static HF model to predict the laboratory-based outputs $\hat{\boldsymbol{y}}_{\mathrm{HF}}$ using the pre-processed input data $\boldsymbol{\Phi}_{\mathrm{HF}}$. The model structure is based on the approach outlined in Eq. (1). The objective is to accurately capture the relationship between the inputs and the laboratory measurements.

Similarly, we train a dynamic LF model to predict the outputs from the online analyzer $\hat{\boldsymbol{y}}_{\mathrm{LF}}$. We apply additional dimensionality reduction on the dataset after the FS highlighted in Fig. 1. Following the dimensionality reduction, we perform dynamic system identification on the LF data ($\boldsymbol{\Phi}_{\mathrm{LF}}$) using the open-source SIPPY - Systems Identification Package for Python (Armenise et al., 2018). We use the PARSIM-K robust identification method (Pannocchia and Calosi, 2010), which is particularly suitable for closed-loop data, to ensure that the dynamic model $\hat{\boldsymbol{y}}_{\mathrm{LF}}$ accurately represents the system dynamics with the fewest parameters necessary. To balance model accuracy with simplicity, we adopt three selection criteria (AIC, AICc, and BIC from Section 3.2) to determine the optimal model order.

To develop the MF model, we use the `scikit-learn` library (Pedregosa et al., 2011) in Python. Before training, we apply an additional dimensionality reduction step to extract PCs from the input features, incorporating both the LF model inputs and its predictions ($\boldsymbol{\Phi}_{\mathrm{LF}}$ and $\hat{\boldsymbol{y}}_{\mathrm{LF}}$). Selecting an appropriate kernel function, $k(\boldsymbol{\kappa}, \boldsymbol{\kappa}')$, is critical for capturing process trends effectively. We evaluate multiple kernel configurations, refining the structure iteratively based on model performance and data characteristics. The final composite kernel is

$$k(\boldsymbol{\kappa}, \boldsymbol{\kappa}') = C \cdot \sigma^2 e^{-\frac{M^2}{2\ell^2}} + \sigma^2 e^{-\frac{2\sin^2(\pi M/p)}{\ell^2}} + \sigma^2 \delta_{\boldsymbol{\kappa},\boldsymbol{\kappa}'}, \quad (9)$$

where $M = \|\boldsymbol{\kappa} - \boldsymbol{\kappa}'\|_2$. The 1st component represents a Constant kernel C; the 2nd component is the Radial Basis Function (RBF) kernel, capturing smooth variations; the 3rd component is a Periodic kernel, incorporating sinusoidal patterns; and the 4th component is a White kernel, accounting for noise through the Kronecker delta function $\delta_{\boldsymbol{\kappa},\boldsymbol{\kappa}'}$. The parameter ℓ is the length scale, determining the influence range, the

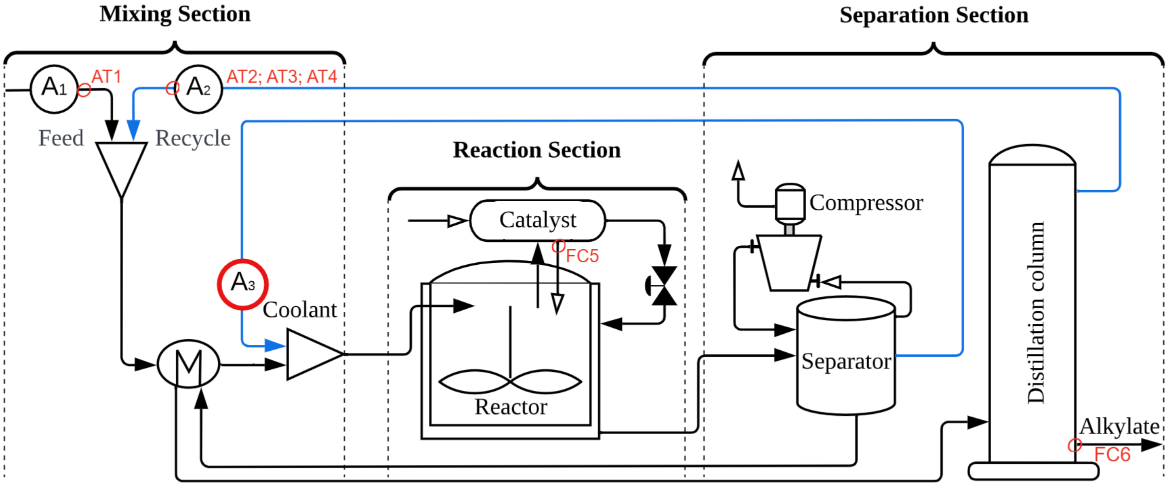

Fig. 2. Simplified schematic of the alkylation unit, highlighting Analyzer A3 (red) as well as the six selected variables.

period p determines the distance between repetitions of the function, and σ^2 is the noise variance.

6. RESULTS

We standardize the dataset to ensure consistent variable scaling across all analyses and present the results accordingly to maintain confidentiality. FS is performed on the LF dataset as it contains continuous input data across the entire timescale. In contrast, HF FS resulted in overfitting and a suboptimal feature set, as confirmed by the industrial partner. By using the full LF dataset, rather than the sparsely sampled HF data, we ensure that the selected features capture temporal variations and process trends.

Table 1 presents the Root Mean Square Error (RMSE) comparison for various FS methods applied to the preprocessed LF dataset ($\boldsymbol{X}_{\text{proc}}$). The results indicate that PCR offers the best performance in terms of accuracy, achieving RMSE values of 0.16 for training and 0.21 for testing. However, this method relies on 75 uncorrelated principal components (PCs), making it less practical for industrial applications due to sensor maintenance challenges. PLS achieves slightly worse RMSE values of 0.28 for training and 0.32 for testing. LASSO yields RMSE values of 0.30 for both training and testing sets, promoting sparsity in the model, though it does not perform as well as PCR or PLS in this instance. SR emerges as the most suitable option for balancing predictive accuracy and simplicity, achieving RMSE values of 0.21 for training and 0.24 for testing. The final verified inputs $\boldsymbol{\Phi}$ ($N \times 6$), shown in Fig. 2, include four concentrations measured by online analyzers (olefin feed propylene — AT1, deisobutanizer recycle propane — AT2, isobutane — AT3, and n-butane — AT4) and two flow rates (recycled fresh acid — FC5 and alkylate to storage — FC6).

Table 2 evaluates predictive performance of all trained models using the RMSE metric against the HF data $\boldsymbol{y}_{\text{HF}}$ on the standardized dataset. Firstly, the RMSE values for the current online analyzer data $\boldsymbol{y}_{\text{LF}}$ compared to lab-measured data $\boldsymbol{y}_{\text{HF}}$ are 0.63 for training and 0.80 for testing, indicating poor performance.

Table 1. RMSE comparison for the used FS methods on the LF dataset.

Method	Training RMSE	Testing RMSE
PCR	0.16	0.21
PLS	0.28	0.32
LASSO	0.30	0.30
SR	0.21	0.24

Table 2. Comparison of model performance metrics across predictive approaches.

Comparison	Training RMSE	Testing RMSE
$\boldsymbol{y}_{\text{HF}}$ to $\boldsymbol{y}_{\text{LF}}$	0.63	0.80
$\boldsymbol{y}_{\text{HF}}$ to $\hat{\boldsymbol{y}}_{\text{HF}}$	0.18	0.91
$\boldsymbol{y}_{\text{HF}}$ to $\hat{\boldsymbol{y}}_{\text{LF}}$	0.52	0.91
$\boldsymbol{y}_{\text{HF}}$ to $\hat{\boldsymbol{y}}_{\text{MF}}$	0.31	0.38

Focusing on the HF model $\hat{\boldsymbol{y}}_{\text{HF}}$, we observe a notable contrast between its training and testing performance. Despite achieving a low RMSE of 0.18 during training, the error escalates to 0.91 when applied to the testing set. This degradation in predictive capability is visually apparent in Fig. 3, where the HF model predictions are depicted in purple. It is apparent that the selected features corroborate with the signal changes and that an application of bias correction based on HF data would improve model performance. This improvement would though be lagged because of HF data sparsity.

The system identification procedure led to the selection of the dynamic LF model $\hat{\boldsymbol{y}}_{\text{LF}}$. To train this model, we first apply PLS on the processed feature set $\boldsymbol{X}_{\text{proc}}$, reducing the six selected variables into PCs. The explained variance (EV) was calculated as a function of the number of PCs, and the "elbow" point on the variance plot was selected, where adding additional PCs offered minimal gain in information. This approach led to the selection of four PCs, which accounted for approximately 95 % of the EV. The model order was determined using the AIC, which suggested a 4th order model, AICc recommended a 10th order, and the BIC proposed a 3rd order. Based on the trade-off between model accuracy and complexity, the 4th order model was selected. The resulting dynamic LF model $\hat{\boldsymbol{y}}_{\text{LF}}$ produced RMSE values of 0.52 for training and 0.91 for

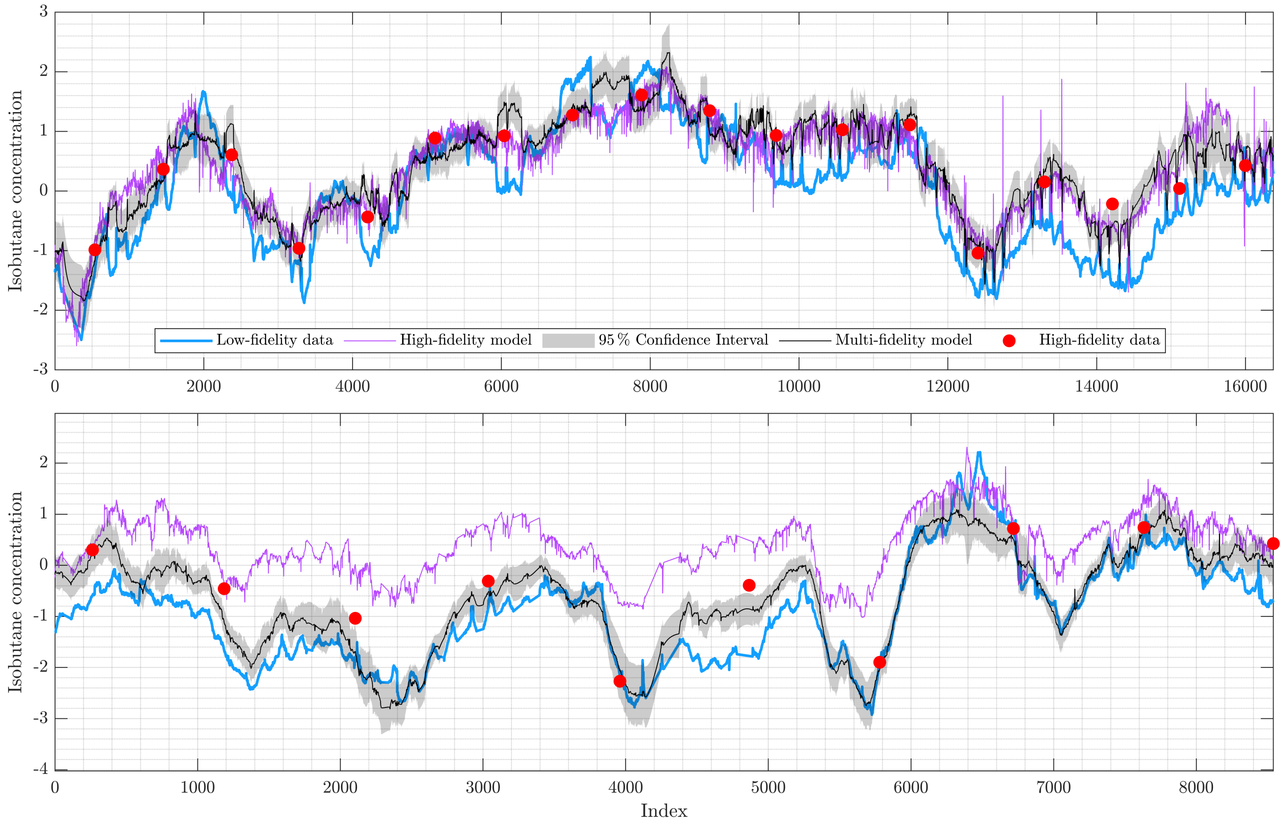

Fig. 3. Predictions of the normalized i-C_4 concentration for training (top), and testing (bottom) sets as data series.

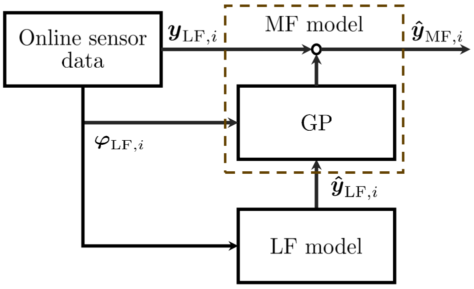
Online sensor data
$\boldsymbol{y}_{\mathrm{LF},i}$
MF model
$\hat{\boldsymbol{y}}_{\mathrm{MF},i}$
GP
$\boldsymbol{\varphi}_{\mathrm{LF},i}$
$\hat{\boldsymbol{y}}_{\mathrm{LF},i}$
LF model

Fig. 4. Workflow of the MF model for prediction.

testing, showcasing moderate improvements over the LF data but still limited generalization. Although the dynamic model showed a slightly higher RMSE (0.91) when compared to the current online analyzer (0.80), it better captured the trends of $\boldsymbol{y}_{\mathrm{HF}}$, which were not included in its training data. The maximum and minimum errors between $\boldsymbol{y}_{\mathrm{LF}}$ and $\boldsymbol{y}_{\mathrm{HF}}$ were 3.38 and 0.003, respectively, while for $\hat{\boldsymbol{y}}_{\mathrm{LF}}$ compared to $\boldsymbol{y}_{\mathrm{HF}}$, they were 1.30 and 0.056. This indicates that, despite the RMSE difference, the dynamic model captures time-dependent patterns more reliably.

The multi-fidelity model ($\hat{\boldsymbol{y}}_{\mathrm{MF}}$) is trained using the GP to predict analyzer deviations from the HF measurements $\Delta_{\mathrm{HF},i} = y_{\mathrm{HF},i} - y_{\mathrm{LF},i},\ \forall i \in \mathcal{I}_{\mathrm{HF}}$ with $\hat{\boldsymbol{\Delta}}_{\mathrm{HF}}(\boldsymbol{\Phi}_{LF}, \hat{\boldsymbol{y}}_{\mathrm{LF}})$. This structure (a particular form of feature selection from (7)) can provide clearer insight for the plant operators. We use PCA to reduce the dimensionality of the seven input variables and to suppress potential collinearity between $\boldsymbol{\Phi}_{\mathrm{LF}}$ and $\hat{\boldsymbol{y}}_{\mathrm{LF}}$. The resulting four PCs serve as inputs to the GP model. This approach (shown in Fig. 4) adjusts the online analyzer outputs to align more closely with HF laboratory measurements. As shown in Fig. 3, the blue solid line represents the original LF data ($\boldsymbol{y}_{\mathrm{LF}}$), the black line shows the adjusted predictions ($\hat{\boldsymbol{y}}_{\mathrm{MF}}$), and the red markers indicate the HF laboratory data. The grey shaded region highlights the 95 % confidence interval of the GP model. This alignment is quantitatively reflected in the MF model accuracy, with RMSE values of 0.31 for training and 0.38 for testing, outperforming the LF and HF models.

Notably, configuring the HF model in the same way as the GP model to predict $\Delta_{\mathrm{HF},i}$ results in a test RMSE of 0.65. This demonstrates that correction strategies alone provide limited improvement. While static models offer simplicity, dynamic models capture complex process behavior more effectively. Our results show that MF models significantly enhance predictive accuracy. Additionally, effective input selection, supported by quantitative analysis and domain knowledge, plays a key role in aligning predictions with real-world process behavior.

7. CONCLUSION

In this work, we developed a multi-fidelity model aimed at improving the accuracy of online analyzers, specifically for monitoring isobutane concentration in alkylation. Our results demonstrate that the model effectively corrects discrepancies between online measurements and true laboratory values, achieving a significant improvement in accuracy of 52.50 %. This enhancement notably elevates monitoring quality, enabling more reliable operational decisions. The findings highlight the power of multi-fidelity modeling in refining process control, showcasing its potential to integrate dynamic, low-fidelity data with high-fidelity laboratory measurements for more precise and effective real-time process monitoring. Future work will focus on enhancing the Gaussian process implementation by addressing process non-linearities to improve the model accuracy and robustness.

REFERENCES

Armenise, G., Vaccari, M., Bacci di Capaci, R., and Pannocchia, G. (2018). An open-source system identification package for multivariable processes. In *UKACC 12th International Conference on Control*, 152–157.

Bahramian, M., Dereli, R.K., Zhao, W., Giberti, M., and Casey, E. (2023). Data to intelligence: The role of data-driven models in wastewater treatment. *Expert Systems with Applications*, 217, 119453.

Bastos, P.D.A., Galinha, C.F., Santos, M.A., Carvalho, P.J., and Crespo, J.G. (2022). Predicting the concentration of hazardous phenolic compounds in refinery wastewater—a multivariate data analysis approach. *Environmental Science and Pollution Research*, 29(1), 1482–1490.

Bradford, E., Imsland, L., Zhang, D., and del Rio Chanona, E.A. (2020). Stochastic data-driven model predictive control using gaussian processes. *Computers & Chemical Engineering*, 139, 106844.

Colombo, A.W., Karnouskos, S., Kaynak, O., Shi, Y., and Yin, S. (2017). Industrial cyberphysical systems: A backbone of the fourth industrial revolution. *IEEE Industrial Electronics Magazine*, 11(1), 6–16.

Efroymson, M.A. (1960). Multiple regression analysis. In A. Ralston and H.S. Wilf (eds.), *Mathematical Methods for Digital Computers*. Wiley, New York.

Fáber, R., Mojto, M., Ľubušký, K., and Paulen, R. (2024). From data to alarms: Data-driven anomaly detection techniques in industrial settings. In *ESCAPE34 - PSE24*.

Ge, Z., Chen, T., and Song, Z. (2011). Quality prediction for polypropylene production process based on CLGPR model. *Control Engineering Practice*, 19(5), 423–432.

Geladi, P. and Kowalski, B. (1986). Partial least square regression: A tutorial. *Anal. Chim. Acta*, 35, 1–17.

Giselle Fernández-Godino, M. (2023). Review of multifidelity models. *Advances in Computational Science and Engineering*, 1(4), 351–400.

Kadlec, P., Gabrys, B., and Strandt, S. (2009). Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, 33(4), 795–814.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., and Liu, H. (2017). Feature selection: A data perspective. *ACM Comput. Surv.*, 50(6).

Pall (2018). Refineries: Application focus h2so4 alkylation process description. Technical report, Pall Corp.

Pannocchia, G. and Calosi, M. (2010). A predictor form parsimonious algorithm for closed-loop subspace identification. *J. Process Control*, 20(4), 517–524.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *London Edinburgh Philos. Mag. & J. Sci.*, 2(11), 559–572.

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

Perdikaris, P., Venturi, D., Royset, J.O., and Karniadakis, G.E. (2015). Multi-fidelity modelling via recursive co-kriging and gaussian–markov random fields. *Proc. of the Royal Society A*, 471(2171), 20150018.

Rasmussen, C.E. (2004). *Gaussian Processes in Machine Learning*, 63–71. Springer Berlin Heidelberg.

Rousseeuw, P. and Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212–223.

Santosa, F. and Symes, W.W. (1986). Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4), 1307–1330.

Speight, J.G. (2020). *The refinery of the future.* Gulf Professional Publishing, Elsevier.

Yin, S. and Kaynak, O. (2015). Big data for modern industry: challenges and trends [point of view]. *Proceedings of the IEEE*, 103(2), 143–146.

Zhu, J., Ge, Z., Song, Z., and Gao, F. (2018). Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data. *Annu. Rev. Control*, 46, 107–133.