Systems & Control Transactions

Original Research Article
Peer Reviewed Conference Proceeding
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

Forecasting Time-to-Cyclic Steady State in Periodic Bioprocesses via a Multi-Feature k-Nearest Neighbours Framework

Yasser Algoufily[ab*], Foteini Michalopoulou[ab], Maria M. Papathanasiou[ab], Mehmet Mercangöz[ab]

[a] Imperial College London, Department of Chemical Engineering, London, United Kingdom
[b] Sargent Center for Process Systems Engineering, London, United Kingdom
* Corresponding Author: y.algoufily22@imperial.ac.uk.

ABSTRACT

Early and reliable prediction of convergence to cyclic steady state (CSS) is increasingly important in periodic downstream bioprocessing, where switching and cut decisions are tuned for a repeatable cyclic regime. This work addresses time-to-CSS (TCSS) forecasting and CSS-existence classification for multicolumn countercurrent solvent gradient purification (MCSGP) systems under run-to-run feed variability. We propose a Multi-Feature k-Nearest Neighbours (MF-kNN) framework that performs long-horizon one-shot trajectory forecasting from an early run segment. CSS outcomes are inferred by reapplying a peak-based convergence rule to the predicted trajectory, while CSS existence is predicted via neighbour-label voting. The approach uses multivariate, standardised features, run-level splits, and a windowed neighbour search to reduce computation. Hyperparameters are tuned with a CSS-oriented objective function that balances trajectory fidelity, TCSS error, and misclassification penalties. On an in-silico MCSGP dataset (98 runs; Δt = 0.2 s; 6800 steps/run) with varying initial modifier concentration, MF-kNN produces accurate full-run forecasts from early data and enables operationally useful early go/no-go decisions. Across outlets, results support accurate CSS timing inference and high CSS-existence classification accuracy (up to 100% on selected outlets), indicating MF-kNN as a transparent and deployment-ready complement to cycle-to-cycle CSS monitoring and control.

Keywords: cyclic steady state, periodic bioprocessing, MCSGP, time-to-CSS forecasting, k-nearest neighbours, one-shot forecasting

INTRODUCTION

Biomanufacturing is progressively shifting from conventional batch workflows toward increasingly integrated operations that deliver higher productivity with reduced footprint (e.g., via higher volumetric productivity, reduced hold steps, and tighter integration of unit operations) [1, 2]. This evolution is visible in industrial biopharmaceutical manufacturing, particularly in down-stream purification where chromatographic separations have moved from single-column, non-periodic operation to cyclic multicolumn switching schemes, such as capture purification transitioning from batch Protein A chromatography toward periodic counter-current chromatography (PCC) [3], challenging polishing tasks extending from single-column batch gradient operation to multicolumn countercurrent solvent gradient purification (MCSGP) [4], and, more broadly, simulated moving bed (SMB) chromatography employing cyclic switching to sustain periodic operation with improved utilisation relative to single-column batch processing [10]. Because these schemes rely on repeated switching cycles, their analysis and performance assessment hinge on whether the process has reached a repeatable periodic regime, commonly termed cyclic steady state (CSS).

CSS generally refers to the regime in periodic operation where cycle-to-cycle behaviour becomes repeatable within defined tolerances, so that outlet profiles and cycle-level outcomes no longer exhibit systematic drift [5]. In MCSGP, this notion is operationalised by requiring that consecutive-cycle chromatograms and cut outcomes meet repeatability criteria, typically assessed by

comparing feature summaries such as peak maxima, peak areas, retention times, and pool compositions against acceptance thresholds tied to process objectives and analytical variability [5]. This repeatability-based workflow also aligns with process-development and performance-qualification practice, where multiple consecutive cycles are evaluated at a candidate operating point to confirm stable cyclic performance before re-porting yields and purities [6].

With CSS as the target operating regime in such periodic processes, an operationally important question is whether the process will attain a repeatable cyclic pattern within the available operating horizon and, if so, when [5]. The answer affects column switching schedules, fraction collection, solvent usage, and product quality in cyclic downstream operations [5]. Overly conservative actions can waste resources and extend runtime, while premature decisions can degrade purity or yield and disrupt intended continuous operation [5].

These considerations motivate reliable early CSS forecasting and classification to support (i) early run-level go/no-go decisions on continued operation, (ii) anticipatory adjustment of fractionation and recycle policies during the transient approach to CSS, and (iii) improved planning of pooling and resource utilisation (e.g., buffer and hold capacities) in integrated downstream trains.

MCSGP provides a representative case study for periodic CSS decision support: it is a semi-continuous chromatography platform for biomolecule purification that exhibits strongly periodic dynamics under coordinated column switching and solvent gradients [4]. These dynamics are shaped by nonlinear adsorption and mass transfer, while feed variability and initial conditions in-fluence convergence toward CSS [5]. In practice, operating poli-cies such as gradient programs, cut criteria, and recycle routing are typically tuned for the cyclic regime; therefore, uncertainty about CSS convergence and timing propagates into decision un-certainty and disturbances [5]. Accordingly, this work examines whether early-cycle information can enable prospective inference of CSS convergence and timing in MCSGP, shifting beyond retrospective confirmation toward prospective decision support that reduces the latency inherent to cycle-to-cycle monitoring and UV-based control [5].

1.1 Literature review: CSS detection and management in periodic bioprocesses

A substantial body of MCSGP work has therefore focused on CSS management via closed-loop operation rather than purely offline certification. Optimisation and control studies treat MC-SGP as a cyclic, hybrid process in which switching and gradient decisions must remain robust under nonlinear adsorption, mass transfer, and delayed measurements, motivating cycle-to-cycle formulations and multi-rate measurement use [5]. Building on this, UV-based closed-loop strategies have been proposed to regulate key cycle outcomes via online peak-tracking and cut adjustment, demonstrating practical controllers that mitigate drift in chromatographic signatures once operation is near the cyclic regime [5]. More recent implementations extend these concepts to improve robustness under disturbances such as feed variability and retention-time shifts, targeting earlier convergence and stable cyclic operation using limited online information [6]. Overall, the MCSGP literature converges on a consistent operational picture: CSS is detected and maintained primarily through cycle-to-cycle repeatability of chromatographic signatures and control actions that keep cycle-level KPIs in-spec [5, 6].

When CSS is addressed predictively, it is most often through model-based simulation and surrogate acceleration rather than prospective inference from early online data. Mechanistic models have been used to simulate MCSGP start-up, CSS attainment, and shutdown behaviour and to predict performance parameters under specified operating conditions, at the cost of repeated-cycle simulation [7]. To reduce this computational burden, hybrid or data-driven surrogates have been developed that target rapid prediction of CSS behaviour for optimisation and decision support (for example, hybrid formulations that achieve CSS predictions substantially faster than high-fidelity models) [8]. Closely related efforts in cyclic adsorption more broadly also emphasise accelerating cyclic-process prediction and synthesis, for instance by physics-informed neural network surrogates designed for cyclic adsorption calculations [9]. These strands primarily support faster CSS computation given a model or training set, rather than early-cycle, forecast-driven go/no-go decisions using only initial-cycle measurements.

Beyond MCSGP, cyclic steady state is a foundational concept in other periodic separations and bioprocess settings. In simulated moving bed (SMB) and related multicolumn chromatography, CSS is often defined formally via periodicity constraints (end-of-cycle states equal start-of-cycle states) and assessed via convergence of cycle-to-cycle outlet profiles and KPIs [10]. The corresponding literature includes dedicated numerical methods for computing CSS efficiently, reflecting the importance of cycle-level periodicity in SMB design and optimisation [11, 10]. Similar convergence logic appears in other cyclic bioprocesses, for example sequencing batch reactors, where steady cyclic operation is assessed when cycle-aggregated performance indicators become repeatable over successive cycles, enabling consistent operation and control [12, 13]. Across these domains, the dominant CSS handling paradigm remains cycle-to-cycle convergence of either (i) observable signatures (peaks, areas, outlet profiles) or (ii) cycle-aggregated KPIs, with control-oriented studies prioritising robust repeatability

under disturbances [5, 10].

1.2 Research gaps and contributions

In MCSGP, and more broadly in periodic downstream and cyclic bioprocess operations, CSS decision support is still dominated by repeatability-based confirmation rather than prospective pre-diction from early-cycle information. Despite the mature body of CSS monitoring and control work surveyed in the previous section, a central limitation for early decision support is that CSS is commonly confirmed retrospectively via consecutive-cycle repeatability checks. This retrospective paradigm is not ideal because it:

- Provides limited quantitative information about proximity to CSS and remaining stabilisation time, constraining proactive planning during start-up.
- Introduces an inherent confirmation latency, since a candidate CSS convergence point can only be validated after observing subsequent cycles (or peaks) that satisfy the tolerance.
- Ties switching, cutting, and recycle actions to confirmed repeatability, delaying high-value decisions (e.g., collection start, recycle routing).

While mechanistic models and advanced nonlinear surrogates can, in principle, predict CSS behaviour, such approaches are often computationally heavy and difficult to deploy online under run-to-run variability [15]. As a result, operators may delay high-value actions until CSS is verified or take conservative actions that reduce productivity. These limitations motivate forecast-driven methods that deliver early, computationally light decision support in periodic operation.

This work addresses these gaps with a multi-feature k-Nearest Neighbours (MF-kNN) framework [14], developed and evaluated on MCSGP, for TCSS forecasting and CSS-existence classification. By leveraging multivariate, physically meaningful signals, MF-kNN provides long-horizon one-shot predictions from early-cycle data, enabling earlier decisions than retrospective cycle-to-cycle confirmation and UV-triggered actions. Performance is benchmarked against classical and learning-based forecasting models, alongside a univariate kNN model, to contextualise accuracy and runtime. Specifically, MF-kNN is tailored to CSS decision support through:

- CSS-aligned inference: CSS is inferred from forecasts using a peak-based convergence rule yielding both TCSS estimates and a CSS-existence classification.
- Multivariate similarity over measured chromatographic signals: a feature-weighted distance over multiple online variables (e.g., outlet concentration traces and modifier profiles), with standardisation to keep signals comparable and avoid dominance by any single measurement.
- Localised neighbour search: a ±W windowed neighbour search around the forecast start time that improves relevance and reduces runtime, supporting online use.
- CSS-oriented validation and tuning: a validation objective that jointly scores trajectory accuracy and TCSS error, with explicit penalties for incorrect CSS-existence predictions to reduce optimistic false positives.
- Early-cycle feasibility: early, one-shot prediction using only the initial segment of the run.

METHODOLOGY

2.1 Process overview and dataset

We consider periodic MCSGP runs, where coordinated column switching and a solvent-modifier gradient induce repeated elution cycles. The data was generated in silico using a validated gPROMS model of the MCSGP process under run-to-run variability in the initial modifier setting. The dataset comprises N_{run} = 98 runs totalling 666, 400 timesteps, sampled every $\Delta t = 0.2\,s$, with fixed run length T = 6800 timesteps (approximately 22.67 min/run).

For each run $r \in \{1, \ldots, N_{run}\}$, the recorded signals include inlet and outlet concentration traces for the four tracked components, $\{C_{\mathrm{in},c,r}(t), C_{\mathrm{out},c,r}(t)\}_{c=1}^{4}$, where $c = 1$ denotes the modifier (buffer), $c = 2$ weak impurities, $c = 3$ product, and $c = 4$ strong impurities. The initial modifier value is set at the start of each run and varies across runs; we denote this run-level preset by $C_{\mathrm{in},1,r}(0)$. The corresponding modifier trajectory $C_{\mathrm{in},1,r}(t)$ evolves over the run according to the imposed gradient program. This exogenous run-to-run variability influences both whether CSS is reached within the finite run horizon and the resulting TCSS, thereby complicating reliable early detection and forecasting. Figures 1 and 2 illustrate representative periodic concentration profiles and the distribution of $C_{\mathrm{in},1,r}(0)$ across runs, respectively.

2.2 Problem definition

For run r, let $\boldsymbol{x}_r(t)$ denote a designated target concentration profile and let $\{\boldsymbol{u}_{r,j}(t)\}_{j=2}^{p}$ denote auxiliary measured trajectories over $t = 0, \ldots, T$. We define the TCSS for run r as $TCSS_r$. At an early initialisation time $t_0 < T - h$, where h denotes the forecast horizon (timesteps), the objectives are to: (i) generate a one-shot forecast of the remainder of the run, i.e., predict $\boldsymbol{x}_r(t)$ for $t = t0 + 1, \ldots, T$; (ii) classify whether CSS will be achieved within the run horizon; and, if CSS is declared, then (iii) estimate $TCSS_r$.

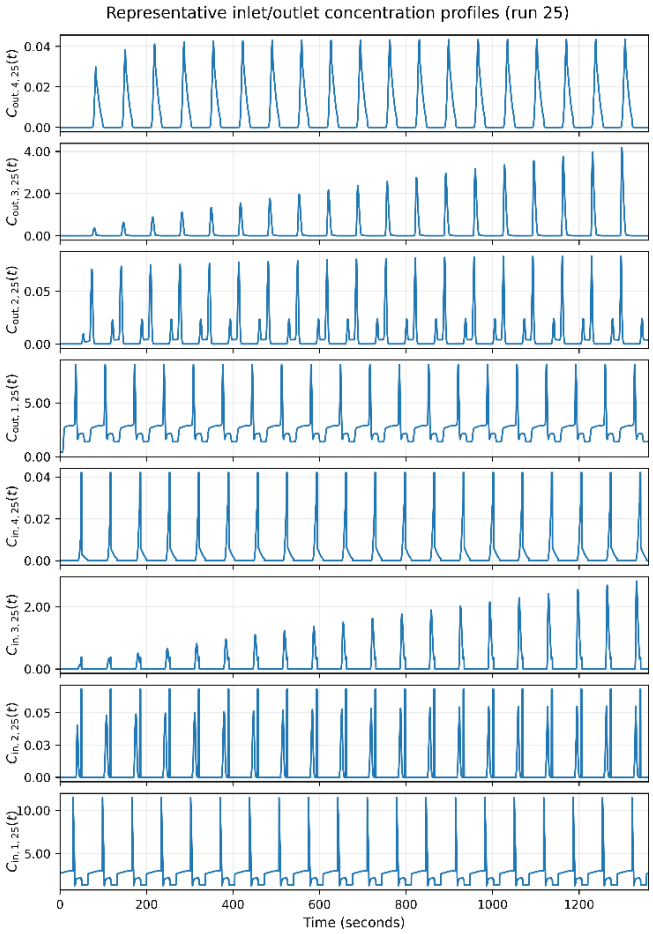


Figure 1. Representative concentration profiles for run 25 showing periodic elution cycles and convergence variability.

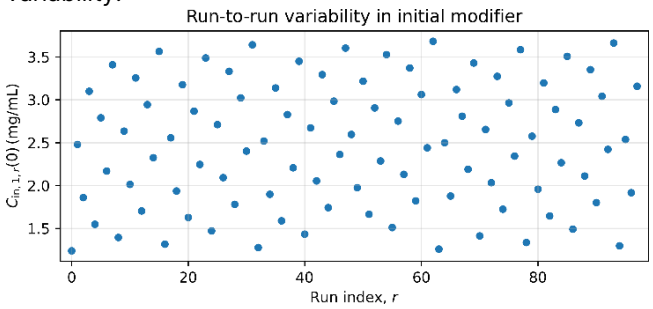


Figure 2. Run-to-run variability in the initial modifier concentration

2.3 Peak-based CSS detection

Peaks are detected on the target series $\boldsymbol{x}_r(t)$ and filtered to retain only the large peaks. Smaller peaks with amplitudes below a fixed cutoff α are discarded, and only peaks with $\boldsymbol{x}_r(t_{peak}) \geq \alpha$ are retained. Let $\{t_{\mathrm{peak},i}, A_i\}_{i=1}^{N_{\mathrm{peak}}}$ denote the retained peak times and amplitudes, where A_i is the amplitude of the i-th large peak. CSS is declared when consecutive large-peak amplitudes remain within tolerance for q successive peak pairs:

$$|A_{i+\ell} - A_{i+\ell+1}| < \varepsilon, \qquad \ell = 0, \ldots, q-1 \tag{1}$$

where ε is the peak-amplitude tolerance and q is the required number of consecutive satisfactions. If Eq. (1) is first satisfied at peak index $i^\star$, then the TCSS is defined as:

$$T_{\mathrm{CSS},r} := t_{\mathrm{peak},i^\star} \tag{2}$$

A binary CSS-existence label $z_r \in \{0,1\}$ is assigned based on whether CSS is reached within the finite run horizon. As illustrated in Figure 3, the TCSS is taken as the time index of the first peak that satisfies the convergence criterion.

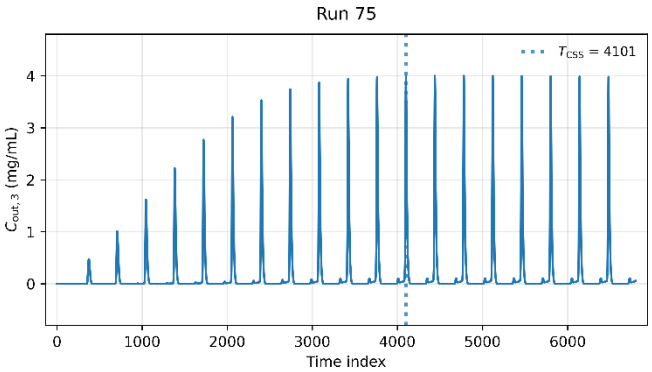


Figure 3. Peak-based CSS definition

2.4 MF-kNN one-shot forecasting

For run r, we construct a $p \times (T+1)$ data matrix by stacking the target and auxiliary signals:

$$Y_r(T) = \begin{bmatrix} x_r(0) & x_r(1) & \cdots & x_r(T) \\ u_{r,2}(0) & u_{r,2}(1) & \cdots & u_{r,2}(T) \\ \vdots & \vdots & \ddots & \vdots \\ u_{r,p}(0) & u_{r,p}(1) & \cdots & u_{r,p}(T) \end{bmatrix} \tag{3}$$

Here, feature index $j = 1$ corresponds to the target $\boldsymbol{x}_r(t)$, and $j = 2, \ldots, p$ correspond to auxiliary trajectories $\boldsymbol{u}_{r,j}(t)$. At an initialisation time $t_0 < T - h$, we represent the recent evolution of each feature using a unified embedding length m (i.e., $m_j \equiv m$ for $j = 1, \ldots, p$). For the target ($j = 1$), we define the embedded evolution segment:

$$\boldsymbol{x}_r(t_0) := [x_r(t_0 - m + 1), \ldots, x_r(t_0)]^{\mathsf{T}} \in R^m \tag{4}$$

Similarly, for each auxiliary feature $j = 2, \ldots, p$, we define:

$$\boldsymbol{u}_{r,j}(t_0) := \left[u_{r,j}(t_0 - m + 1), \ldots, u_{r,j}(t_0)\right]^{\mathsf{T}} \in R^m \tag{5}$$

The multi-feature evolution segment is then formed by stacking the embedded segments as rows:

$$Y_r(t_0) = \begin{bmatrix} \boldsymbol{x}_r(t_0)^{\mathsf{T}} \\ \boldsymbol{u}_{r,2}(t_0)^{\mathsf{T}} \\ \vdots \\ \boldsymbol{u}_{r,p}(t_0)^{\mathsf{T}} \end{bmatrix} \in R^{p \times m} \tag{6}$$

Multi-feature distance

Given a query run r at initialisation time t_0, candidate neighbours are drawn from memory runs $r' \in R_{mem}$. For each memory run r' and admissible endpoint τ, the candidate segment $Y_{r'}(\tau) \in R^{p\times m}$ is constructed identically to $Y_r(t_0)$ by taking the most recent m samples of each feature up to τ, with τ constrained to $m-1 \le \tau \le T-h$. Similarity between the query segment $\boldsymbol{Y_r}(t_0)$ and a candidate segment $\boldsymbol{Y_{r'}}(\tau)$ is computed via a feature-weighted Euclidean distance. Defining $\Delta\mathbf{Y} := \boldsymbol{Y_r}(t_0) - \boldsymbol{Y_{r'}}(\tau)$, we compute:

$$D\big(\boldsymbol{Y_r}(t_0), \boldsymbol{Y_{r'}}(\tau)\big) = \sqrt{\sum_{j=1}^{p} w_{f,j} \left\|\Delta \boldsymbol{Y_{j,:}}\right\|_2^2} \tag{7}$$

where $\boldsymbol{Y_{j,:}}$ denotes the j-th row and $w_{f,j} \ge 0$ with $\sum_{j=1}^{p} w_{f,j} = 1$.

Neighbour selection

The k candidates with smallest distance D are selected as neighbours, indexed by pairs $\{r_n, \tau_n\}_{n=1}^{k}$. These neighbours are then weighted to favour candidates that are (i) close in the embedded signal space and (ii) close in the run-level initial modifier setting. First, we define the distance proximity score:

$$s_n^{(D)} := \frac{1}{2}\left(1 - \frac{D_n}{\sum_{\ell=1}^{k} D_\ell}\right) \tag{8}$$

Here $D_n := D\big(Y_r(t_0), Y_{r_n}(\tau_n)\big)$. Next, define the modifier proximity score:

$$s_n^{(C)} := \frac{1}{2}\left(1 - \frac{\Delta C_n}{\sum_{\ell=1}^{k} \Delta C_\ell}\right) \tag{9}$$

Here $\Delta C_n := |C_{\text{in},1,r}(0) - C_{\text{in},1,r_n}(0)|$. If a normalising denominator is zero, the corresponding score is set uniformly over the k neighbours. The final aggregation weights are obtained by summing the two scores and normalising to sum to one:

$$\omega_n := \frac{s_n^{(D)} + s_n^{(C)}}{\sum_{l=1}^{k}\left(s_\ell^{(D)} + s_\ell^{(C)}\right)}, n = 1, \ldots, k \tag{10}$$

One-shot forecast

Finally, the one-shot forecast of the target trajectory is obtained by weighted aggregation of the neighbours' future continuations:

$$\hat{\boldsymbol{x}}_{\boldsymbol{r}}(t_0+1{:}\,t_0+h) = \sum_{n=1}^{k} \omega_n \, x_{r_n}(\tau_n+1{:}\,\tau_n+h) \tag{11}$$

CSS inference

Because historical runs are labelled with CSS existence, the neighbour set supports classification without an additional model. Let $z^{(n)}$ be the CSS-existence label for neighbour n. The CSS-existence decision is then:

$$\hat{z} = I\left(\sum_{n=1}^{k} \omega_n \, z^{(n)} \ge \eta\right) \tag{12}$$

with voting threshold η (default $\eta = 0.5$). When $\hat{z} = 1$, the predicted $\hat{T}_{\text{CSS}}$ is obtained by applying the same peak rule (Eq. (1)) to the full reconstructed trajectory; i.e. using measured values for $t < t_0$ and the one-shot forecast for $t \ge t_0$.

2.5 Hyperparameter tuning objective

We tune k, feature weights w_f, and any embedding or window parameters, to balance trajectory fidelity, TCSS accuracy, and CSS-existence correctness. Let $\mathcal{L}_{traj}$ denote a trajectory loss over the forecast horizon. We define:

$$\text{Obj}_1 = \frac{1}{N}\sum_{r=1}^{N} \mathcal{L}_{traj}(\boldsymbol{x_r}, \hat{\boldsymbol{x}}_{\boldsymbol{r}}) \tag{13}$$

$$\text{Obj}_2 = \frac{1}{N}\sum_{r=1}^{N} \begin{cases} |T_{\text{CSS},r} - \hat{T}_{\text{CSS, r}}| & if\ z_r = \hat{z}_r = 1 \\ 0 & if\ z_r = \hat{z}_r = 0 \\ penalty & if\ z_r \ne \hat{z}_r \end{cases} \tag{14}$$

Let N denote the number of validation runs. We combine the trajectory term Obj_1 and the CSS-aligned term Obj_2 into a single scalar objective:

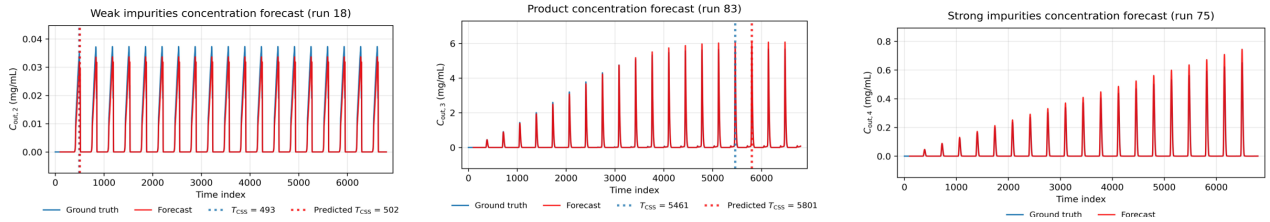
$$\text{Obj}_{\text{overall}} = (1-\beta)\text{Obj}_1 + \beta\text{Obj}_2 \tag{15}$$

where $\beta \in [0, 1]$ controls the trade-off.

2.6 Experimental protocol and implementation notes

We use a run-wise split of $80/10/10\%$ for memory/validation/test. For N_{run} = 98, this yields N_{mem} = 78, N_{val} = 10, and N_{test} = 10. The features are standardised to prevent feature scale dominance. Forecasts are computed one-shot using only the first 100 timesteps of the unfolding run.

Hyperparameters are tuned on validation runs using grid search. Tuned parameters include the neighbour count $k \in \{1, 3, 5, \ldots, k_{max}\}$, feature weights $w_{f,j} \in [0, 1]$, the neighbour aggregation scheme (uniform vs. distance-weighted), and window/embedding settings. Candidate combinations are scored using the composite objective in Eq. (15). For computational efficiency, the neighbour search is restricted to a local temporal window around the forecast initialisation time such that when forecasting a run at t_0, candidate historical segments are only drawn from memory runs with segment endpoints $\tau \in [max(0, t_0 - W), min(T-1, t_0 + W)]$, where W is a window half-width in timesteps. Distances are computed



(a) $C_{\text{out},2,18}$ (b) $C_{\text{out},3,83}$ (c) $C_{\text{out},4,75}$

Figure 4. Representative one-shot forecasts for the outlet concentrations initialised at t_0 = 100.

between the query segment ending at t_0 and each candidate segment ending at τ, and the k nearest neighbours are selected by which a one-shot forecast is then generated as defined in Eq. (11). We report trajectory error in MSE, TCSS error in RMSE, and CSS-existence classification accuracy in percentage.

2.7 Benchmark models and comparative evaluation criteria

To contextualise MF-kNN performance, a comparative one-shot forecasting study was conducted against representative classical and learning-based baselines, alongside a univariate kNN model. All models were evaluated under a common run-wise protocol, where forecasts were initialised at the same t_0 timestep, and the prediction horizon was set to the remainder of the run. The forecasting target is the outlet signal $C_{\text{out},3,r}(t)$. Hyperparameters were tuned on the validation runs using model-specific searches.

Benchmarked models. The benchmark suite comprises: (i) MF-kNN; (ii) Univariate kNN; (iii) ARMA as a classical stochastic time-series baseline for periodic signals; (iv) Theta as a lightweight statistical forecasting baseline; and (v) LSTM as a representative recurrent neural forecasting model.

Tuning and scoring. Hyperparameters were tuned via a fixed number of trials on the validation runs. MF-kNN tunes the neighbour count and feature weights; univariate kNN tunes the neighbour count; ARMA tunes (p, q); Theta tunes θ; and the LSTM tunes hidden-unit configuration. Evaluation considers both (a) trajectory prediction error for $C_{\text{out},3,r}(t)$ over the forecast horizon and (b) error on the derived CSS timing metric from the corresponding predictions, so that models are compared on the same decision-support objective.

RESULTS AND DISCUSSION

Qualitative performance. Figure 4 illustrates representative one-shot forecasts for outlet concentration profiles using only the first 100 timesteps. Across these examples, MF-kNN reproduces the dominant periodic dynamics over the full horizon, correctly distinguishes CSS from no-CSS runs, and yields TCSS estimates that are consistent with the peak-based rule.

Quantitative performance. We summarise quantitative results on the held-out test set using three primary metrics. Table 1 reports trajectory accuracy as MSE, TCSS error as RMSE, and CSS-existence classification accuracy (%). Table 2 reports the CSS-existence confusion matrix for $C_{\text{out},3,r}(t)$, from which precision and recall can be computed. As shown in Table 1, MF-kNN achieves consistently low trajectory errors across the outlet signals, indicating that the one-shot forecasts reproduce the dominant periodic structure and amplitude of the concentration profiles. The CSS-existence classification performance is strong overall (average accuracy 96.7%), with perfect classification reported for $C_{\text{out},3,r}(t)$ and $C_{\text{out},4,r}(t)$ and a small degradation for $C_{\text{out},2,r}(t)$ at 90%, suggesting that early segments contain sufficient information to discriminate convergent from non-convergent behaviour in most cases. In contrast, the TCSS RMSE exhibits a clearer dependence on the outlet signal: $C_{\text{out},2,r}(t)$ and $C_{\text{out},3,r}(t)$ show substantially lower timing errors than $C_{\text{out},4,r}(t)$, indicating that CSS timing is harder to infer for $C_{\text{out},4,r}(t)$, likely due to weaker observability of peak stabilisation or greater sensitivity of its peak features to transient dynamics and noise. Overall, this trend highlights that accurate CSS existence prediction can be achieved even when the precise timing of convergence is more uncertain for specific outlets.

Benchmark comparison. Under the common one-shot protocol described in Section 2.7, Table 3 shows that MF-kNN provides the most reliable decision-support performance for $C_{\text{out},3,r}(t)$ among the benchmarked methods. Beyond achieving the lowest CSS timing error, MF-kNN is the only method in the benchmark set that simultaneously maintains low trajectory error and high CSS-existence accuracy, which is the key requirement for early operational decisions. The univariate kNN baseline degrades both timing and classification performance relative to MF-kNN, indicating that leveraging additional measured signals in the distance metric improves neighbour relevance and reduces ambiguity in early-cycle matching. The LSTM and ARMA baselines perform poorly in this setting, which is consistent with the difficulty of

learning stable long-horizon dynamics from limited early segments under run-to-run variability without extensive model capacity and careful regularisation. The Theta model performs worst on both timing and classification, suggesting that lightweight univariate statistical structure is insufficient to capture the nonlinear, multivariate transient CSS behaviour of MCSGP. Overall, the benchmark results support the central premise of this work: multi-feature similarity enables robust early matching to historical trajectories, yielding more reliable long-horizon forecasts and CSS timing inference than both univariate similarity and representative statistical or learning-based baselines in the one-shot regime.

Table 1: Quantitative performance on the test set.

Signal	TCSS RMSE (steps)	Time-series MSE ($\times 10^{-3}$)	CSS classification accuracy (%)
$C_{\mathrm{out},2}(t)$	157.3	0.02	90.0
$C_{\mathrm{out},3}(t)$	340.0	1.15	100.0
$C_{\mathrm{out},4}(t)$	1356.0	0.04	100.0
Avg.	**617.8**	**0.40**	**96.7**

Table 2: CSS-existence confusion matrix for $C_{\mathrm{out},3}(t)$.

	Predicted CSS	Predicted No-CSS
Actual CSS	20.0%	0.0%
Actual No-CSS	0.0%	80.0%

Table 3: Benchmark performance for $C_{\mathrm{out},3}(t)$.

Model	TCSS RMSE (steps)	Time-series MSE ($\times 10^{-3}$)	CSS classification accuracy (%)
MF-kNN	340.0	1.15	100.0
kNN	537.90	14.95	40.0
LSTM	2917.74	57.86	30.0
ARMA	2937.48	60.62	10.0
Theta	4455.30	98.34	10.0

CONCLUSION

This paper introduced a MF-kNN framework for early, one-shot forecasting CSS behaviour in periodic downstream bioprocesses, demonstrated on MCSGP system runs subject to run-level variability in the initial modifier concentration. Using only an early segment of each run, MF-kNN forecasts the full-run outlet trajectories, infers TCSS, and predicts CSS existence via neighbourhood voting. The methodology is designed for practical deployment through run-wise splitting, feature standardisation, and a windowed neighbour search that reduces runtime while preserving local relevance.

Across the evaluated outlet signals, MF-kNN yields consistent qualitative agreement with the true periodic profiles and supports accurate early decision support for CSS convergence. In the benchmark study, MF-kNN outperformed representative statistical and learning-based baselines on the CSS timing objective under a common one-shot protocol, reinforcing the value of similarity-based forecasting when long-horizon predictions must be derived from limited early-cycle information.

From an operational perspective, the proposed framework enables earlier switching and cut decisions, earlier initiation of collection, and early termination or alternative actions when convergence is unlikely, with the potential to reduce solvent usage and wasted runtime during non-convergent start-up.

Future work

Future work will (i) validate the proposed MF-kNN framework on experimental/industrial MCSGP datasets, beyond the current in-silico study; (ii) quantify operational benefits under realistic decision policies, e.g., solvent usage, time savings, and yield/purity impacts from earlier go/no-go and collection-start decisions; (iii) integrate MF-kNN predictions into closed-loop operation, for example as a supervisory layer that adapts cut/collection logic or informs receding-horizon MPC for cyclic operation; and (iv) extend the feature set used for neighbour retrieval to improve robustness and to provide calibrated confidence estimates for TCSS and CSS-existence predictions.

ACKNOWLEDGEMENTS

The authors acknowledge Imperial College London for providing the research environment and computational resources that supported this work. We also thank colleagues in the Department of Chemical Engineering.

AUTHOR IDENTIFIERS

Author ORCIDs:
Algoufily Y: 0009-0000-8945-6028
Michalopoulou F: 0009-0003-0335-3178
Papathanasiou M: 0000-0002-8886-0624
Mercangöz M: 0000-0002-4449-0414

REFERENCES

1. Xu J, Xu X, Huang C, Angelo J, Oliveira CL, Xu M, Xu X, Temel D, Ding J, Ghose S, Borys MC, Li ZJ. Biomanufacturing evolution from conventional to intensified processes for productivity improvement: a case study. mAbs 12: (2020). https://doi.org/10.1080/19420862.2020.1770669
2. Shukla AA, Wolfe LS, Mostafa SS, Norman C. Evolving trends in mab production processes.

Bioengineering & Transla Med 2:58-69 (2017). https://doi.org/10.1002/btm2.10061
3. Godawat R, Brower K, Jain S, Konstantinov K, Riske F, Warikoo V. Periodic counter?current chromatography – design and operational considerations for integrated and continuous purification of proteins. Biotechnology Journal 7:1496-1508 (2012). https://doi.org/10.1002/biot.201200068
4. Müller?Späth T, Aumann L, Melter L, Ströhlein G, Morbidelli M. Chromatographic separation of three monoclonal antibody variants using multicolumn countercurrent solvent gradient purification (MCSGP). Biotech & Bioengineering 100:1166-1177 (2008). https://doi.org/10.1002/bit.21843
5. Papathanasiou MM, Steinebach F, Morbidelli M, Mantalaris A, Pistikopoulos EN. Intelligent, model-based control towards the intensification of downstream processes. Computers & Chemical Engineering 105:173-184 (2017). https://doi.org/10.1016/j.compchemeng.2017.01.005
6. Eisenhuth R, Müller-Späth T. Process characterization and performance qualification of MCSGP. Processes 13:3950 (2025). https://doi.org/10.3390/pr13123950
7. Müller-Späth T, Ströhlein G, Aumann L, Kornmann H, Valax P, Delegrange L, Charbaut E, Baer G, Lamproye A, Jöhnck M, Schulte M, Morbidelli M. Model simulation and experimental verification of a cation-exchange igg capture step in batch and continuous chromatography. Journal of Chromatography A 1218:5195-5204 (2011). https://doi.org/10.1016/j.chroma.2011.05.103
8. Michalopoulou F, Papathanasiou MM. Accelerated process optimization of chromatographic separations using a hybrid modeling approach. IFAC-PapersOnLine 59:427-432 (2025). https://doi.org/10.1016/j.ifacol.2025.07.183
9. Subraveti SG, Li Z, Prasad V, Rajendran A. Physics-based neural networks for simulation and synthesis of cyclic adsorption processes. Ind. Eng. Chem. Res. 61:4095-4113 (2022). https://doi.org/10.1021/acs.iecr.1c04731
10. Rajendran A, Paredes G, Mazzotti M. Simulated moving bed chromatography for the separation of enantiomers. Journal of Chromatography A 1216:709-738 (2009). https://doi.org/10.1016/j.chroma.2008.10.075
11. Minceva M, Pais LS, Rodrigues AE. Cyclic steady state of simulated moving bed processes for enantiomers separation. Chemical Engineering and Processing: Process Intensification 42:93-104 (2003). https://doi.org/10.1016/s0255-2701(02)00038-7
12. Ibrahim G, Abasaeed AE. Modelling of sequencing batch reactors. Water Research 29:1761-1766 (1995). https://doi.org/10.1016/0043-1354(94)00317-z
13. Dionisi D, Rasheed AA, Majumder A. A new method to calculate the periodic steady state of sequencing batch reactors for biological wastewater treatment: model development and applications. Journal of Environmental Chemical Engineering 4:3665-3680 (2016). https://doi.org/10.1016/j.jece.2016.07.032
14. Algoufily Y, Borghesan F, Mercangöz M. Predictive modelling of desiccant drying processes using multi-feature k-nearest neighbours algorithm. IFAC-PapersOnLine 59:205-210 (2025). https://doi.org/10.1016/j.ifacol.2025.07.146
15. Daoutidis P, Lee JH, Rangarajan S, Chiang L, Gopaluni B, Schweidtmann AM, Harjunkoski I, Mercangöz M, Mesbah A, Boukouvala F, Lima FV, del Rio Chanona A, Georgakis C. Machine learning in process systems engineering: challenges and opportunities. Computers & Chemical Engineering 181:108523 (2024). https://doi.org/10.1016/j.compchemeng.2023.108523