Nonlinear Model Predictive Control of a pilot scale Fed-Batch Ethanol Fermentation Reactor Using Hard, Soft, and Hybrid Temperature-Constraint Strategies

Gabriel Baioni e Silva[a,c], Thais Suzane Milessi[a,b], Luis Ricardez-Sandoval[c], Felipe Fernando Furlan[a,b]

[a] *Universidade Federal de São Carlos, Programa de Pós-graduação em Engenharia Química, São Carlos - SP, Brasil.*
[b] *Universidade Federal de São Carlos, Departamento de Engenharia Química, São Carlos - SP, Brasil.*
[c] *University of Waterloo, Department of Chemical Engineering, Waterloo - ON, Canada.*

ABSTRACT

Ethanol fermentations in industrial units are often cooling-limited, making upper temperature bounds a recurrent operability requirement in fed-batch operation. In non linear predictive control (NMPC) implementations, however, strict hard enforcement can become numerically delicate when optimal trajectories remain close to an active limit, so that small numerical offsets across control cycles may lead to sub-optimal or infeasible decisions and interrupted receding-horizon updates. To address this practical issue, three temperature-constraint handling options were evaluated for NMPC of a fed-batch ethanol fermentation reactor: hard enforcement, soft enforcement via slack penalization, and a hybrid hard-to-soft fallback. All cases used the same nonlinear DAE model, orthogonal-collocation transcription, and IPOPT-based solution pipeline to isolate the effect of constraint handling. Closed-loop simulations, including a baseline and a 27-scenario study, indicate that controlled relaxation substantially reduces failed control updates while maintaining comparable production-oriented performance; the hybrid approach preserves the original objective structure whenever hard-mode solves are reliable, while behaving similarly to a well-tuned soft formulation when the temperature limit is frequently active.

Key words: Nonlinear model predictive control; Fed-batch fermentation; Ethanol production; Temperature constraints.

1 Motivation, Purpose, and Background

Fuel ethanol is a key lever in energy-transition strategies, enabling scalable reductions in fossil-fuel dependence and greenhouse-gas emissions. Brazil is among the largest producers, reaching 30.6 million m^3 in 2024 (EPE, 2025). Sector expansion is further supported by national blending mandates (30% anhydrous ethanol in gasoline) (CNPE, 2025) and by the associated cumulative avoided emissions. Despite industrial maturity, large-scale operation remains constrained by robustness and efficiency limits.

Thermal management is a persistent bottleneck in industrial fermentation. Ethanol fermentation is exothermic and, in plants with legacy equipment and limited retrofit capacity, cooling constraints combined with ambient fluctuations can drive reactors beyond the optimal yeast temperature range, degrading productivity and viability (Basso et al., 2011). Temperature is therefore both a yield-related variable and a safety and operability constraint.

Dynamic optimization can mitigate thermal limitations without major hardware changes. In fed-batch operation, the feed flow rate is the primary manipulated input during most of the run, directly affecting substrate availability, metabolic activity, and heat generation. Shaping the feed profile can therefore modulate biological heat release and reduce temperature excursions (Mantovanelli et al., 2007). After feeding ends, temperature becomes largely dictated by the cooling system, leaving limited corrective capacity.

A practical difficulty arises when the optimizer persistently pushes the trajectory toward an active temperature bound. In that regime, perturbations and model mismatch can place the realized state for the next NMPC cycle marginally outside the admissible set. With a strict bound, this can trigger infeasibility, thus causing a degradation of controller performance and a potential loss in the system.

In many industrial fermentation settings, however, control aims to keep temperature close to a preferred window rather than enforce an absolute wall. Minor deviations (sub-degree fluctuations) can be operationally acceptable and may not compromise product quality or safety. This motivates constraint-handling choices that

prioritize reliable closed-loop operation near active limits.

Accordingly, this study compares three temperature-handling options within the same NMPC framework: strict hard enforcement, soft enforcement via slack variables with penalty, and a hybrid scheme that attempts the hard formulation and falls back to a soft solve when a reliable hard update is not available. The hybrid logic is particularly appealing when one wishes to preserve the original objective whenever possible, while still maintaining closed-loop continuity near the constraint boundary. At the same time, a purely soft formulation may be sufficient in many cases, offering a simpler implementation provided that the slack penalty is tuned to reflect acceptable thermal flexibility.

2 Methodology

2.1 Computational Environment and Implementation

All simulations were implemented in Python on a Windows workstation using Pyomo.DAE for dynamic optimization modeling and IPOPT (Wächter e Biegler, 2006) as the nonlinear programming solver. Numerical post-processing and analysis were performed with NumPy and pandas, and figures were generated with Matplotlib. NMPC execution, plant simulation, and comparative analyses were carried out in the same software stack to ensure consistent numerical behavior across strategies.

2.2 Process Model and State Description

The fed-batch fermentation reactor was represented by a nonlinear differential-algebraic model with state vector

$$\mathbf{x}(t) = [V(t), C_S(t), C_P(t), C_X(t), T(t)]^\top, \tag{1}$$

where V is reactor volume, C_S substrate concentration, C_P product concentration, C_X biomass concentration, and T broth temperature. The manipulated variable was the feed flow rate $F(t)$, constrained by $0 \le F(t) \le F_{\max}$ and $V(t) \le V_{\max}$.

The specific growth rate was described as

$$\mu(t)\big(K_S + C_S(t)\big) = \mu_{\max} C_S(t) \exp\left(-\frac{C_S(t)}{S_m}\right)\left(1 - \frac{C_P(t)}{P_m}\right), \tag{2}$$

and the dynamic balances were

$$\dot{V}(t) = F(t), \tag{3}$$

$$V(t)\dot{C}_X(t) = \mu(t)C_X(t)V(t) - F(t)C_X(t), \tag{4}$$

$$V(t)\dot{C}_S(t) = F(t)\big(C_{S,in} - C_S(t)\big) - \frac{\mu(t)C_X(t)V(t)}{Y_{X/S}}, \tag{5}$$

$$V(t)\dot{C}_P(t) = \mu(t)C_X(t)V(t)\frac{Y_{P/S}}{Y_{X/S}} - F(t)C_P(t), \tag{6}$$

$$V(t)\rho C_p \dot{T}(t) = F(t)\rho C_p\big(T_{in} - T(t)\big) - \Delta H_r \frac{\mu(t)C_X(t)V(t)}{Y_{X/S}} - UA\big(T(t) - T_c\big), \tag{7}$$

with constant coolant reference temperature T_c.

2.3 Plant-Model Relationship in Closed Loop

The same DAE structure and parameter set were used in both the plant simulator and the NMPC predictor. No deliberate plant-model mismatch was introduced. This choice isolated the effect of constraint handling and numerical convergence from model-uncertainty effects. At each control cycle, the optimizer computed the feed profile over the prediction horizon and only the first control move was applied. The plant was then propagated over one sampling interval with this constant input under a receding-horizon implementation.

2.4 Temperature-Constraint Strategies

Three alternatives were considered for handling the upper temperature limit.

(i) Hard constraint:

$$T(t) \le T_{\max}.$$

(ii) Soft constraint:

$$T(t) \leq T_{\max} + s_T(t), \quad s_T(t) \geq 0,$$

where $s_T(t)$ is penalized in the objective to discourage violations while maintaining feasibility.

(iii) Hybrid hard-to-soft fallback: a hard-constrained solve was attempted when the measured temperature was below the limit. If the solver did not return an acceptable solution, the cycle was re-solved using the soft formulation and the first admissible move was applied. When the measured temperature was at or above the limit, the cycle started directly in soft mode. This design retains the original objective structure whenever the hard problem is numerically well conditioned, while avoiding interruptions near the boundary.

2.5 NMPC Formulation and Numerical Transcription

At each sampling instant t_k, NMPC solved a finite-horizon optimization problem with horizon H.

$$\max_{F(\cdot),\mu(\cdot),\mathbf{x}(\cdot),s_T(\cdot)} \quad J = \int_{t_k}^{t_k+H} C_P(t)\,V(t)\,dt - W_T * \int_{t_k}^{t_k+H} s_T(t)\,dt \tag{8}$$

Here, $C_P(t)$ is the product concentration, $V(t)$ is the reactor volume, $s_T(t)$ is the temperature-slack variable, and W_T is the penalty weight associated with temperature-constraint relaxation. The problem was subject to Eqs. (2)–(7), bounds on F, the feed-window logic ($F = 0$ after the feeding deadline), and the volume limitation $V \leq V_{\max}$.

The soft and hybrid formulations included the additional slack variable s_T and its non-negativity constraint. The penalty term $W_T \int s_T(t)\,dt$ was incorporated in the objective function to penalize temperature constraint relaxations while ensuring feasibility.

The dynamic optimization problem was transcribed by orthogonal collocation on Radau points using LAGRANGE - RADAU, with piecewise-constant control over each finite element.

2.6 Failure Handling, Execution Window, and Study Design

If an NMPC step failed, the fallback policy applied the last accepted feed command under hold-last-command to maintain input continuity. The controller was active only during the feeding window. After feed cutoff, plant dynamics continued with $F = 0$ until the end of the simulation horizon.

The baseline setup used total simulation time 12 h, sampling interval 0.2 h, and feed time limit at 8 h. The values $T_{\max} = 32.5\,^{\circ}\mathrm{C}$, $T_{in} = 32.0\,^{\circ}\mathrm{C}$, and $UA = 501\,000\,\mathrm{kJ\,h^{-1}\,^{\circ}C^{-1}}$ were used. The parametric study evaluated 27 scenarios with $T_{\max} \in \{32,33,34\}$, $T_{in} \in \{30,31,32\}$, and $UA \in \{201000, 501000, 801000\}$. The recorded metrics included non-accepted solves, temperature indicators, final volume, average feed, productivity (Ethanol production/volume/time), and computational time.

3 Main Results and Discussion

3.1 Model Basis and Scope

The results reported in this section were obtained with the nonlinear DAE model introduced in Section 2, which coupled kinetic and thermal effects through mass and energy balances. The specific growth rate was defined implicitly by Eq. (2), and the heat-release and heat-removal terms appeared explicitly in the energy balance of Eq. (7). The kinetic and thermal parameter set $(\mu_{\max}, K_S, S_m, P_m, Y_{X/S}, Y_{P/S}, \Delta H_r, UA, T_{in}, T_c, \rho, C_p)$ had been previously adjusted and experimentally validated (Table 1) in a pilot plant for the process conditions of interest. A detailed identification and validation discussion was outside the scope of this paper. The validated parameters were treated as a fixed basis to isolate the effects of NMPC temperature-constraint handling on closed-loop behavior.

3.2 Baseline Comparison of Temperature Constraints

Table 2 reports the baseline closed-loop metrics, and Fig. 1 shows the corresponding trajectories. The hard-constrained NMPC reached 75 L of fermenter volume in less time, but this apparent advantage occurred alongside 10 consecutive non-accepted solves as the temperature trajectory operated near the upper bound. Once the solver stopped producing acceptable updates, the implementation reverted to hold-last-command, so the applied feed was no longer refreshed by the optimizer. As a result, the reported time-to-fill mixes optimized actions with repeated actuation during failure episodes, which limits the interpretability of the faster filling time as a controller performance gain.

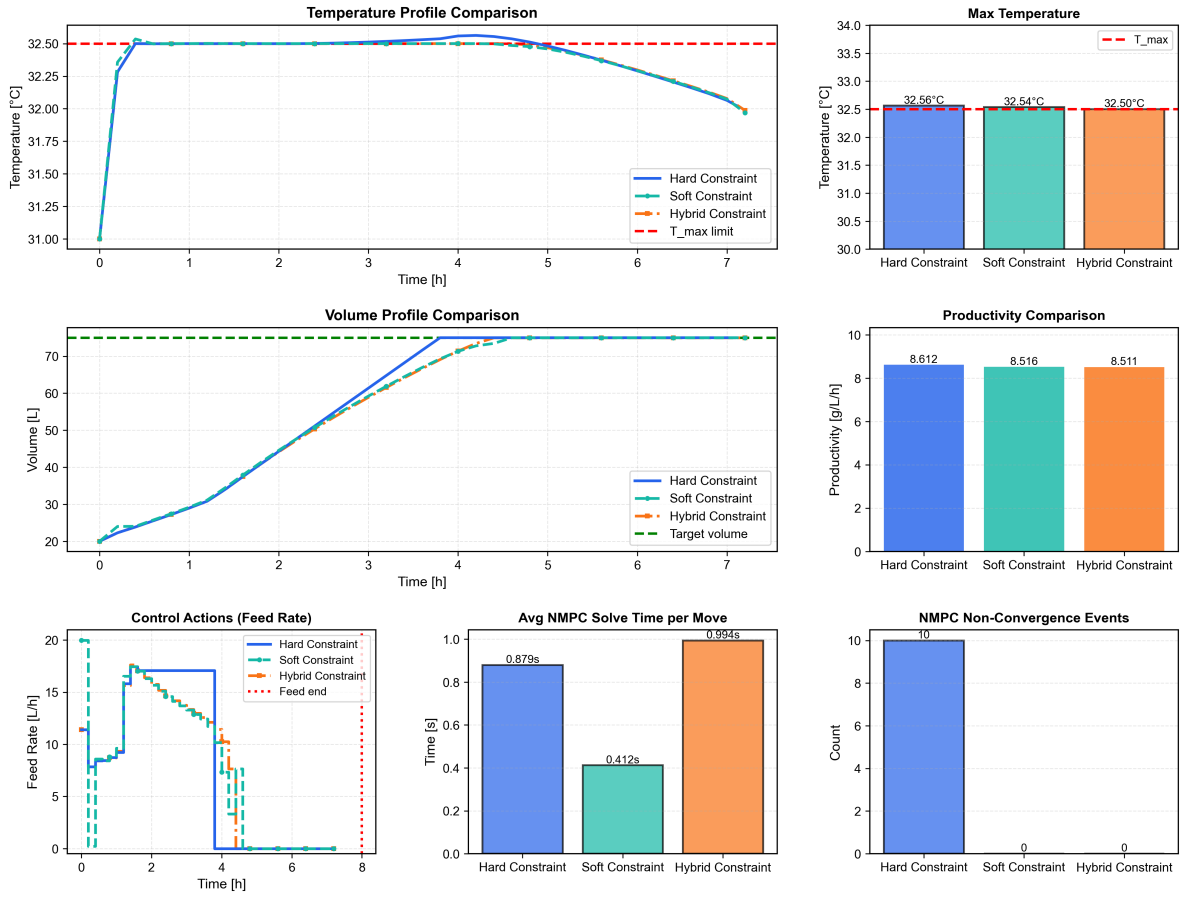
In the same baseline, the soft and hybrid formulations completed the feeding window without interruptions and maintained regular receding-horizon updates. Both achieved the target volume with ethanol production values

Table 1: Kinetic and thermal parameters used in the DAE fermentation model for all simulations.

Parameter	Value	Unit
$\mu_{\max}$	0.12347	h^{-1}
K_S	0.14654	$g\,L^{-1}$
S_m	1323.3	$g\,L^{-1}$
P_m	84.795	$g\,L^{-1}$
$Y_{X/S}$	0.12188	$g\,g^{-1}$
$Y_{P/S}$	0.50079	$g\,g^{-1}$
$C_{S,in}$	167.0	$g\,L^{-1}$
T_{in}	32.0	°C

Parameter	Value	Unit
T_c	31.0	°C
ρ	1000.0	$kg\,m^{-3}$
C_p	4.1868	$kJ\,kg^{-1}\,°C^{-1}$
ΔH_r	-561.85	$kJ\,kg^{-1}$
UA	501000	$kJ\,h^{-1}\,°C^{-1}$
$V_{\max}$	75.0	L
$F_{\max}$	25.0	$L\,h^{-1}$

Figure 1: Closed-loop baseline trajectories under hard, soft, and hybrid NMPC temperature constraints. Curves show feed flow rate, volume, temperature, and concentration states over the feeding window and post-feed phase.



comparable to the hard case, albeit with slightly longer times associated with a slower approach to the substrate-depletion target (<= 1 g/L). The hybrid configuration also yielded the lowest peak temperature, 32.50 °C, consistent with tighter thermal behavior while maintaining solver acceptance throughout the run.

The baseline highlights a numerical vulnerability of strict enforcement when the temperature constraint becomes effectively active. Even with the same model used for prediction and plant propagation, the closed-loop cycle does not reproduce identical states at machine precision. Finite NLP tolerances, small inconsistencies between collocation-based transcription and forward simulation, and round-off effects can shift the state used to

Table 2: Baseline closed-loop performance under hard, soft, and hybrid NMPC temperature constraints.

Metric	Hard	Soft	Hybrid
NMPC failures	10	0	0
Max temperature [°C]	32.56	32.54	32.50
Time to 75 L [h]	3.80	4.60	4.40
Final volume [L]	75.00	74.99	74.97
Total ethanol production [g]	4645.74	4657.69	4654.25
Operational time [h]	7.10	7.20	7.20
Average solve time per move [s]	0.8793	0.4119	0.9938
Hybrid hard uses	0	0	2
Hybrid soft uses	0	0	20

initialize the next NMPC problem. Near $T_{\max}$, a small upward shift is enough to place the initial point outside the admissible set, at which point the hard-constrained NLP may terminate without an acceptable solution. In real-world operation, process perturbations, such as unexpected variations in feed composition, ambient temperature fluctuations, or sensor drift, can similarly push the temperature beyond the feasible region, triggering the same failure mechanism. This mechanism explains why solver acceptance can degrade abruptly precisely when operation sits close to the limit, and why consecutive non-accepted solves can occur despite an otherwise consistent modeling setup.

Slack-based formulations mitigate this failure mode by keeping the NLP feasible under small boundary-crossing offsets. Rather than forcing an infeasible initialization, the optimizer absorbs the mismatch through the slack variable and continues to produce usable control moves, preventing disruption of the receding-horizon loop. The hybrid formulation provides a compromise in which the optimizer retains the unmodified objective in Eq. (8) whenever the hard problem remains numerically well-behaved, while switching to the relaxed version only when a hard-mode update cannot be accepted. In the baseline, the mode counts (2 hard and 20 soft solves) indicate that the hybrid controller operated predominantly in relaxed mode once the constraint region was repeatedly encountered, which explains the close similarity between hybrid and soft trajectories.

For the soft formulation, the slack penalty weight sets the practical trade-off between temperature adherence and production-oriented optimization. If chosen too large, the penalty can dominate the objective, making the Hessian of the objective function poorly conditioned and biasing the solution toward slack minimization (Murray, 1971); if too small, violations may exceed what is operationally acceptable. Consequently, tuning should be guided by the tolerated magnitude and duration of temperature excursions, so that relaxation improves solver acceptance without distorting the intended control priority.

3.3 27-Scenario Study and Practical Implications

Table 3 reports mean performance over the 27-scenario study spanning $T_{\max} \in \{32, 33, 34\}$ °C, $T_{in} \in \{30, 31, 32\}$ °C, and $UA \in \{201000, 501000, 801000\}$ kJ h^{-1} °C^{-1}. Across these conditions, strict enforcement displayed the highest incidence of non-accepted solves and the most unfavorable worst-case behavior, consistent with the sensitivity expected when operation repeatedly approaches a tight feasibility boundary. *NMPC failures* denotes the mean number of non-accepted solves per simulation (averaged over the 27 scenarios), while *Worst-case failures* denotes the largest number of failures observed within a single simulation.

Both slack-based and hybrid approaches improved closed-loop continuity and achieved higher mean final volumes and ethanol production values. The soft formulation delivered the lowest average failure count and the best worst-case reliability with shorter average solve times. The hybrid approach reduced interruptions relative to strict enforcement, but it did not consistently outperform the soft approach in reliability or computational effort. In Table 3, *NMPC failures* denotes the mean number of non-accepted solves per simulation (averaged over the 27 scenarios), while *Worst-case failures* denotes the largest number of failures observed within a single simulation.

These results suggest a practical decision rule. When small temperature excursions are acceptable, a soft formulation with a well-chosen penalty can be sufficient. This is especially attractive during startup and transient operation, and in industrial settings with robust downstream quality tolerances. When preserving an unmodified primary objective is desirable whenever possible, a hybrid fallback can be attractive, provided that hard-mode

Table 3: Mean closed-loop performance across the 27-scenario study for hard, soft, and hybrid NMPC temperature constraints.

Metric (mean)	Hard	Soft	Hybrid
NMPC failures	5.6296	0.0741	0.2963
Max temperature [°C]	32.5340	32.7647	32.7902
Final volume [L]	68.1490	71.2980	71.8878
Total ethanol production [g]	4069.05	4334.83	4379.53
Average solve time [s]	0.6272	0.4166	0.5899
Worst-case failures	24	1	4

solves occur frequently enough to justify the added switching logic.

4 Conclusion

Hard, soft, and hybrid temperature-constraint handling approaches were evaluated for NMPC of a fed-batch ethanol fermentation reactor under an upper temperature limit using the same nonlinear DAE model, transcription method, and solver pipeline. Strict enforcement was more prone to non-accepted solves near the active bound, which can interrupt the intended receding-horizon behavior. Slack-based relaxation substantially reduced these interruptions while maintaining temperature close to the limit and preserving comparable production-oriented outcomes. The hybrid fallback retained the original objective structure whenever hard-mode solves were reliable, while providing continuity near the boundary through soft re-solves. Overall, the results indicate that soft constraints can be sufficient in many temperature-limited settings with acceptable small excursions, whereas a hybrid strategy becomes more compelling when maintaining an unmodified primary objective is a key design preference.

5 Acknowledgements

This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Funding Code 001; the São Paulo Research Foundation (FAPESP), grant number #2024/02416-4 and #2025/12302-9; and the National Council for Scientific and Technological Development (CNPq), grant number 310753/2025-2. This work was carried out in partnership with the company Fermentec with the support of the Financier of Studies and Projects (FINEP) and the Ministry of Science, Technology, Innovation, and Communications (MCTIC), with funds from FNDCT (UFSCar 14.708 - ProEx No. 012827/2022-18).

Referências

Basso, L. C., Basso, T. O., Rocha, S. N., Basso, L. C., Basso, T. O., & Rocha, S. N. (2011, setembro). Ethanol Production in Brazil: The Industrial Process and Its Impact on Yeast Fermentation. Em *Biofuel Production - Recent Developments and Prospects*. IntechOpen. https://doi.org/10.5772/17047

CNPE. (2025, junho). Resolução N°9, de 25 de junho de 2025.

EPE. (2025). *BRAZILIAN ENERGY BALANCE 2025* (Relatório Anual). Empresa de Pesquisa Energética. Rio de Janeiro (RJ).
titleTranslation: BALANÇO ENERGÉTICO NACIONAL.

Mantovanelli, I. C. C., Rivera, E. C., Costa, A. C. D., & Filho, R. M. (2007). Hybrid neural network model of an industrial ethanol fermentation process considering the effect of temperature. *136*.

Murray, W. (1971). Analytical expressions for the eigenvalues and eigenvectors of the Hessian matrices of barrier and penalty functions. *Journal of Optimization Theory and Applications*, *7*(3), 189–196. https://doi.org/10.1007/BF00932477

Wächter, A., & Biegler, L. T. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, *106*(1), 25–57. https://doi.org/10.1007/s10107-004-0559-y