



Journal Website:
<http://sciencebring.com/index.php/ijasr>

Copyright: Original content from this work may be used under the terms of the creative commons attributes 4.0 licence.

 Research Article

Multiscale Numerical Control and High-Performance Simulation for Continuous Casting: Integrating Model Predictive Control, Lattice Boltzmann Methods, and GPU Acceleration

Submission Date: September 01, 2025, **Accepted Date:** September 13, 2025,

Published Date: September 30, 2025

Anil R. Menon

Department of Mechanical and Materials Engineering, University of Melbourne

ABSTRACT

This article presents an integrative, theoretically rigorous, and methodologically detailed exposition on the design, analysis, and numerical implementation of advanced control strategies for continuous casting processes, emphasizing model predictive control for nonlinear parabolic partial differential equation (PDE) systems, lattice Boltzmann method (LBM) based fluid-thermal simulation, and high-performance computing (HPC) implementations on graphics processing units (GPUs). By synthesizing methodological advances from state-of-the-art control theory applied to unsteady PDEs (Yu et al., 2023; Wang et al., 2016; Yu et al., 2018) with kinetic-based fluid modelling and thermodynamics captured by lattice Boltzmann frameworks (d’Humières et al., 2002; He et al., 1998; Lallemand & Luo, 2003), and the practical acceleration strategies using CUDA-enabled GPU computing (Micikevicius, 2009; NVIDIA, 2010; Mudigere, 2009), the paper articulates a comprehensive pipeline: from mathematical problem formulation and discretization strategy, through controller design and stability considerations, to efficient implementation patterns that respect memory access, parallelism, and numerical accuracy on modern heterogeneous architectures. The article places particular emphasis on handling convective terms in unsteady parabolic PDEs, the computational advantages and limitations of multiple-relaxation-time LBM for thermal-acoustic fidelity, and practical considerations when migrating model predictive control algorithms to GPUs for real-time or near-real-time industrial use (Wang et al., 2019). The narrative critically examines the interplay between model fidelity, controller robustness, numerical stability, and computational throughput—highlighting trade-offs, potential failure modes such as longitudinal crack formation in hypoperitectic steels during solidification (Konishi et al., 2002), and pathways for mitigating these through control-informed cooling strategies (Wang et al., 2016). The synthesis culminates in an extended methodological



blueprint suitable for researchers and practitioners seeking to develop publication-ready, production-grade simulation-control systems for continuous casting and analogous thermofluid processes.

KEYWORDS

continuous casting; model predictive control; parabolic PDEs; lattice Boltzmann method; GPU acceleration; thermal-fluid simulation; numerical stability

INTRODUCTION

Continuous casting of steel is a complex, tightly coupled thermofluid and solidification process where spatially distributed thermal fields, phase transformations, and mechanical stress evolution interact over multiple temporal and spatial scales. Addressing control objectives—such as preventing surface and internal defects, optimizing cooling trajectories to manage microstructural outcomes, and maintaining throughput stability—requires a fusion of accurate physical modelling, robust control strategies, and sufficient computational power to achieve timely decision making (Wang et al., 2016; Yu et al., 2018). The consolidation of model predictive control (MPC) frameworks for PDE-governed systems, kinetic-based discretizations like the lattice Boltzmann method (LBM), and GPU-based high-throughput computation forms a promising convergence for managing the competing demands of accuracy, real-time responsiveness, and scalability.

Historically, PDE-constrained control has encountered significant challenges rooted in the infinite-dimensional nature of the governing equations, the nonlinearities associated with convection and phase change, and the discretization-induced trade-offs between numerical stability and control performance (Yu et al., 2023; Wang et al., 2016). Model predictive control is particularly attractive because it

explicitly incorporates future predictions and constraints, enabling the handling of operational limits and multi-variable coupling. However, its application to two-dimensional parabolic PDEs with convective components and nonlinearities—typical of secondary cooling zones in continuous casting—requires advanced numerical solvers and optimization strategies capable of coping with stiff PDEs and large-scale discretizations (Wang et al., 2016; Yu et al., 2018).

On the simulation front, lattice Boltzmann methods present an alternative discretization paradigm to classical finite-difference or finite-element approaches. Originating from kinetic theory, LBM provides a mesoscopic perspective, modeling the evolution of particle distribution functions whose macroscopic moments reproduce Navier–Stokes behaviour and thermal transport in certain limits (McNamara & Zanetti, 1988; d’Humières et al., 2002). The multiple-relaxation-time (MRT) variants and thermally consistent models (He et al., 1998; Lallemand & Luo, 2003) offer improved numerical stability and control over transport coefficients—qualities useful when carefully resolving acoustic and thermal properties relevant to casting flows and heat extraction.

Implementing these advanced models in a way that supports MPC’s real-time optimization requirements motivates the move to GPUs. GPUs offer significant parallel throughput for structured



grid computations and data-parallel kernels characteristic of LBM and explicit PDE solvers (Micikevicius, 2009; Dongarra et al., 2008). Yet, obtaining practical speedups demands careful attention to memory hierarchy, coalesced accesses, and algorithmic resilience to floating-point and boundary-condition idiosyncrasies (Mudigere, 2009; Lee et al., 2010). The computational and algorithmic design space becomes a tripartite optimization problem: controller performance, numerical fidelity, and computational efficiency.

This article integrates these threads into a cohesive narrative: it delineates mathematical formulations for unsteady parabolic PDEs with convection in the context of continuous casting, explicates modern MPC schemes adapted for such systems (including the DY-HS hybrid conjugate gradient approach for constrained optimization problems in unsteady PDEs), explores LBM-based thermal-fluid simulation choices, and details GPU implementation strategies that maintain numerical and control integrity while maximizing throughput (Yu et al., 2023; Wang et al., 2019). In doing so, the article critically synthesizes literature findings, exposes open gaps, and provides detailed methodological guidance for future experimental and computational work.

Methodology

The methodological exposition unfolds across three interconnected components: mathematical and control problem formulation; numerical discretization and simulation strategy; and high-performance implementation and optimization.

Problem formulation and control objectives

At the heart of the continuous casting control problem lies the management of the temperature field and its spatiotemporal evolution, which can be compactly represented by a nonlinear parabolic PDE incorporating diffusive, convective, and source/sink terms that represent cooling sprays and latent heat release during solidification. Practically, the control objective is to shape the boundary and distributed cooling inputs to ensure desired thermal trajectories, avoid defect-prone thermal gradients, and comply with operational constraints (Wang et al., 2016; Yu et al., 2018). The controller must operate in a context where measurements are spatially sparse (e.g., thermocouples, infrared surface measurements) and subject to noise and latency. Model predictive control becomes an appropriate choice because it can incorporate system constraints and anticipate future system evolution over a prediction horizon.

MPC for PDE systems entails embedding the discretized PDE within an optimization problem that minimizes a cost functional reflecting deviations from reference temperatures, control effort, and possibly regularization terms to penalize undesirable spatial gradients or rapid control variations. Constraints include actuator bounds (spray intensities, flow rates), state constraints (maximum allowable temperature differentials), and possibly inequality constraints related to material properties. The optimization problem is typically large-scale due to spatial discretization, non-convex if nonlinearities are present, and time-coupled due to the PDE dynamics. Efficient solvers such as hybrid conjugate gradient methods that combine descending directions from different algorithmic philosophies (e.g., the DY-HS hybrid) are valuable

for navigating the optimization landscape while respecting computational budgets (Yu et al., 2023).

Discretization strategy: combining classical and kinetic approaches

Two broad discretization paradigms are considered: (a) classical finite-difference/finite-element discretizations tailored for parabolic PDEs with convection and (b) lattice Boltzmann methods that model mesoscopic distribution functions whose moments yield macroscopic thermal-fluid behaviour. The choice between these is not merely technical; it interacts with controller design and computational implementation.

Classical discretizations: When employing finite-difference or finite-element schemes, the primary concerns are numerical diffusion introduced by upwinding schemes used to handle convection and the stability constraints associated with explicit time-stepping (Courant–Friedrichs–Lewy (CFL) conditions) versus the computational complexity of implicit solvers. Implicit solvers afford larger time steps and improved stability for stiff diffusion-dominated regimes but increase per-step computational cost and require solving large linear or nonlinear systems—an expense that must be reconciled with the real-time demands of MPC (Wang et al., 2016).

Lattice Boltzmann methods: The LBM offers a convenient explicit, local-update structure well suited to GPUs. In MRT-LBM formulations, relaxation rates are tuned to adjust physical viscosity and thermal diffusivity and to enhance numerical stability (d’Humières et al., 2002). The thermally consistent LBM models developed by He et al. (1998) and analysed by Lallemand and Luo (2003) provide a framework to capture thermal

transport with desirable acoustic and thermal fidelity. However, applying LBM to strongly convective, phase-changing flows requires attention to boundary conditions, forcing terms modelling external cooling, and coupling strategies that incorporate latent heat effects without destabilizing the kinetic solver. Another central advantage of LBM is its locality: updates at a lattice node depend only on neighbouring node values, making it amenable to parallelization across thousands of GPU threads (Micikevicius, 2009).

Controller-simulation coupling: Ensuring consistent models

A key methodological consideration is model fidelity mismatch: MPC requires a model that is both tractable for optimization and sufficiently representative of reality. Using a high-fidelity LBM or fine-grained finite-element model inside the MPC loop may be computationally prohibitive. A pragmatic strategy is model hierarchy: a reduced-order model or surrogate (e.g., projection-based reduced models, Gaussian process surrogates, or coarser-grid PDE discretizations) can be used within the optimization loop, while high-fidelity simulations verify and refine control strategies offline or asynchronously during extended production runs (Wang et al., 2019). However, when GPU acceleration provides sufficient throughput, more detailed discretizations can be used directly, reducing surrogate-model error. The methodological challenge is to maintain stability and control performance despite approximate gradients or model mismatches—necessitating robust optimization formulations and possibly the inclusion of feedback laws or constraint tightening to account for uncertainty.



Optimization solver choices and the DY-HS hybrid algorithm

Solving the large-scale optimization problem arising from MPC for PDE systems calls for iterative solvers that exploit problem structure. Conjugate gradient methods are attractive for symmetric positive-definite linear subproblems; however, nonlinearities and nonquadratic cost functionals require variants that can handle changing Hessian approximations. The DY-HS hybrid algorithm (Yu et al., 2023) synthesizes two conjugate-gradient update schemes—each with complementary properties regarding descent direction selection and numerical stability—to accelerate convergence on unsteady PDE-constrained problems. The hybrid approach maintains conjugacy properties essential for efficient descent while adapting to ill-conditioning typical of discretized PDE operators with convective dominance. The methodological implementation couples the optimization routine with adjoint-based gradient computations: adjoint PDEs, discretized consistently with forward models, compute sensitivity information with respect to control variables efficiently, avoiding direct differentiation of large discretized systems.

Numerical integration and boundary treatments

Boundary conditions in continuous casting are complex: convective outflow, conduction into rollers, and localized cooling sprays induce mixed Neumann and Robin-type conditions, while moving boundaries and phase-change interfaces introduce further complexity. Numerically, consistent discretization of boundary fluxes is vital to avoid artificial reflections or spurious heat sources that can mislead control algorithms. LBM boundary

schemes must be carefully chosen—bounce-back, non-equilibrium extrapolation, or immersed boundary techniques—as each has implications for mass and energy conservation. In classical discretizations, ghost nodes, weak enforcement techniques in finite elements, or characteristic-based treatments are used to incorporate convective fluxes without destabilizing the scheme.

High-performance implementation considerations

When translating these algorithms to GPUs, attention must focus on memory bandwidth, thread divergence, and fine-grained synchronization. LBM's streaming-and-collision paradigm maps naturally to GPU kernels: collision updates perform joint computations on local distribution sets, while streaming can be implemented as memory moves that benefit from coalesced access patterns. For PDE-based implicit solvers and adjoint computations, sparse linear algebra kernels and preconditioners must be GPU-aware; otherwise, data movement between CPU and GPU can nullify throughput gains (Micikevicius, 2009; Mudigere, 2009). Further, numerical reproducibility and floating-point behaviour change across architectures, requiring testing and error-analysis to ensure control policies derived from GPU-enabled simulations remain valid when moved to plant controllers.

Verification, validation, and coupling to defect modelling

Verification against analytical solutions or benchmark problems ensures numerical correctness, while validation against experimental measurements from actual casting lines assesses model fidelity. A critical motivation for advanced

control is defect suppression—longitudinal facial cracks in hypoperitectic steels being a central example where thermal gradients and solidification front morphology interact to create tensile stresses that produce cracking (Konishi et al., 2002). Incorporating phenomenological or microstructure-aware models into the control loop informs the cost functional, enabling direct penalization of defect-prone conditions. However, these couplings increase model complexity and demand more computational power, further stressing the high-performance implementation.

Results

This section describes, in descriptive terms, the expected computational and control outcomes of the integrated methodology when properly implemented and validated. Given the article's theoretical orientation and constraint against numerical tables, the results are presented as qualitative and quantitative expectations grounded in literature findings and method analysis.

Controller performance and stability

Hybrid conjugate-gradient solvers embedded within MPC frameworks have been reported to significantly reduce iteration counts and improve convergence robustness for unsteady PDE-constrained optimizations compared with conventional methods (Yu et al., 2023). In practice, this translates to reduced solution times per MPC cycle when using compatible discretizations and adjoint-based gradient evaluations. The practical outcome is that, for moderate grid resolutions (where the discretized state dimension is on the order of 10^4 to 10^6), the hybrid approach can deliver near-real-time control updates when coupled with GPU-accelerated PDE solves and

efficient linear algebra kernels. Stability analysis of MPC laws derived from such optimizations typically relies on terminal cost and constraint formulations; enforcing dissipativity through cost design and ensuring accurate adjoint computations are central to maintaining closed-loop stability in the face of model mismatch (Wang et al., 2016; Yu et al., 2018).

Numerical fidelity of LBM and classical methods

Multiple-relaxation-time LBM variants offer better control over numerical dispersion and dissipation compared with single-relaxation-time schemes, which is beneficial when resolving thermal waves and acoustic phenomena that can influence defect formation (d'Humières et al., 2002; Lallemand & Luo, 2003). The mesoscopic approach reduces artificial numerical diffusion common with upwind finite-difference schemes, particularly in advection-dominated regimes. However, the LBM's explicit nature places constraints on time-step sizes tied to lattice speeds—constraints that can be mitigated by using coarser lattices supplemented by subgrid models or employing operator-splitting techniques for stiff source terms (He et al., 1998).

Computational throughput and GPU acceleration

Benchmarking studies in related fields indicate that GPU implementations of structured-grid solvers, including LBM and well-optimized finite-difference codes, can achieve order-of-magnitude throughput improvements over single-core CPU implementations and substantial improvements over multi-core CPU runs when architects and coders exploit memory coalescing and shared-memory usage patterns (Micikevicius, 2009; Lee et

al., 2010). However, the realized acceleration depends heavily on arithmetic intensity, memory bandwidth, and algorithmic locality. For MPC loops incorporating adjoint PDEs, the need to solve forward and backward in time can roughly double the computational burden—but this cost remains amenable to parallelization since both forward and backward solves consist of data-parallel time-step computations.

Impact on defect mitigation strategies

Integrating high-fidelity thermal models into MPC enables proactive cooling adjustments that reduce detrimental thermal gradients, thereby lowering the risk of longitudinal cracking as characterized by Konishi et al. (2002). In practice, optimized spray patterns and staged cooling intensities derived from MPC tend to flatten near-surface temperature gradients and reduce thermal shock during phase transitions. The multidimensional control afforded by distributed actuators (multiple spray zones) allows the controller to trade off increased control effort against defect probability—optimizations that are achievable only when simulations are sufficiently predictive and the controller can operate on relevant timescales (Wang et al., 2016).

Sensitivity and robustness outcomes

Adjoint-based gradient computations are sensitive to discretization choices and round-off error, especially in long-horizon MPC problems. Sensitivity analysis reveals that control solutions are robust to moderate measurement noise when cost functionals include state-regularization terms and when constraints account for uncertainties through robust optimization techniques or constraint tightening. The DY-HS hybrid algorithm

shows resilience to ill-conditioning introduced by high Peclet numbers (convection dominance) in the spatial discretization, which improves the robustness of computed controls compared to naive gradient-descent methods (Yu et al., 2023).

Discussion

The methodological synthesis articulated above intersects several important themes: the balancing act between model fidelity and computational tractability, the role of discretization choice in determining both numerical and control outcomes, and the practicalities of deploying high-throughput computing resources to support near-real-time optimization.

Trade-offs between fidelity and performance

High-fidelity LBM models capture mesoscopic transport phenomena with desirable numerical properties, yet their explicit structure and fine spatial requirements can inflate computational cost. Conversely, reduced models or coarser finite-element discretizations shrink optimization problem sizes and make MPC cycles feasible on limited hardware but incur model mismatch risks. The pragmatic path lies in multilevel model hierarchies: using faster surrogates within the control loop while periodically recalibrating them with high-fidelity GPU-accelerated simulations. Such hierarchical approaches preserve real-time responsiveness without sacrificing long-term accuracy (Wang et al., 2019). Crucially, any surrogate must be constructed with an eye toward preserving control-relevant dynamics—modes and frequencies that significantly influence the cost or constraint satisfaction.

Numerical stability and controller robustness



Convection-dominated PDEs, common in continuous casting due to transport of enthalpy by moving material and sprays, challenge numerical schemes and adjunct optimization solvers. Artificial numerical diffusion from stabilizing discretizations can distort gradient information used by the optimizer, leading to suboptimal or even destabilizing control actions. The MRT-LBM and properly designed finite-element stabilization techniques can mitigate these effects, but one must carefully calibrate relaxation parameters and stabilization coefficients so that they reflect physical dissipation rather than numerical artefacts (d’Humières et al., 2002; Lallemand & Luo, 2003). Controller robustness can be improved by incorporating model-form uncertainty into the MPC formulation (e.g., min–max or stochastic MPC), which cushions performance against residual model errors.

GPU-specific implementation challenges

GPU acceleration is not a panacea. Achieving reliable, portable, and maintainable code requires careful software engineering. Kernel fusion to reduce global memory traffic, use of shared memory to accelerate local stencils, and attention to numerical reproducibility when parallel reductions are performed are all necessary for robust deployments (Micikevicius, 2009; Mudigere, 2009). Additionally, the hardware landscape evolves rapidly; achieving long-term maintainability may require abstraction layers that can target different backends (CUDA, HIP, or vendor-specific accelerators) without rewriting core numerical kernels. Furthermore, debugging and verifying GPU-accelerated adjoint computations can be significantly harder than for

CPU code due to limited debuggers and nondeterministic thread scheduling behaviors.

Practical deployment considerations

Bringing these techniques to production casting lines introduces a range of non-technical constraints: actuator latencies, sensor placement limitations, safety regulations, and operator acceptance. The controller must not only be mathematically stable and computationally efficient but also integrate with legacy control systems and adhere to safety-critical response times. A human-in-the-loop design is often prudent—initial validation phases benefit from supervisory human operators who can override control actions or provide high-level guidance while system performance is evaluated.

Open research questions and future directions

Several avenues remain open for deeper exploration. First, rigorous proofs of closed-loop stability for MPC schemes using surrogate models warrant attention, especially when surrogate errors are non-negligible and time-dependent. Second, the incorporation of phase-change kinetics and microstructure evolution into controllers in a computationally tractable way remains challenging: models that capture defect nucleation and growth at a fidelity necessary for control-based mitigation are presently too expensive to run in real time. Third, exploring mixed-precision computing strategies that trade some numerical precision for throughput on GPUs may unlock new performance regimes; this requires quantifying the effect of rounding errors on adjoint sensitivities and control decisions (Lee et al., 2010; Micikevicius, 2009). Finally, machine learning-based surrogates, trained from GPU-accelerated

high-fidelity simulations, present an attractive path to combine predictive power with low evaluation cost—but their integration into safety-critical MPC frameworks requires careful constraints enforcement and interpretability measures.

Limitations

This article has focused primarily on the conceptual and methodological synthesis rather than on reporting experimental or production-line data. The absence of explicit numerical experiments in this manuscript limits the ability to provide measured speedups or quantified control improvements. Nevertheless, the methodological guidance draws directly on established literature and reported benchmarks, offering a principled pathway toward empirical evaluation.

Conclusion

Integrating model predictive control for unsteady parabolic PDEs, lattice Boltzmann-based thermal-fluid simulation, and GPU-accelerated numerical implementations offers a compelling framework for addressing the complex control challenges of continuous casting. The confluence of advanced optimization algorithms (such as the DY-HS hybrid conjugate gradient), numerically robust discretizations (MRT-LBM and stabilized finite-element schemes), and careful GPU-aware implementation strategies can produce practical controllers that are both predictive and responsive. Key to success are consistent adjoint computations, hierarchical model strategies that balance fidelity and computational cost, and software engineering practices that preserve numerical accuracy while exploiting hardware parallelism. Future work

should prioritize rigorous closed-loop stability analysis with surrogate-in-the-loop MPC, experimental validation on industrial casting lines, and exploration of mixed-precision and machine learning surrogates to further enhance computational tractability and control effectiveness.

References

1. Yu, Y., Wang, Y., Deng, R., & Yin, Y. (2023). New DY-HS hybrid conjugate gradient algorithm for solving optimization problem of unsteady partial differential equations with convection term. *Mathematics and Computers in Simulation*, 208, 677–701.
2. Wang, Y., Luo, X., & Wang, H. (2019). GPU-Based model predictive control of nonlinear parabolic partial differential equations system and its application in continuous casting. *IEEE Access*, 7, 79337–79353.
3. Konishi, J., Miltzer, M., Samarasekera, I. V., et al. (2002). Modeling the formation of longitudinal facial cracks during continuous casting of hypoperitectic steel. *Metallurgical & Materials Transactions B*, 33(3), 413–423.
4. Yu, Y., Luo, X., Liu, Q., et al. (2018). Model predictive control of a dynamic nonlinear PDE system with application to continuous casting. *Journal of Process Control*, 65, 41–55.
5. Wang, Y., Luo, X. C., Yu, Y., et al. (2016). Optimal control of two-dimensional parabolic partial differential equations with application to steel billets cooling in continuous casting secondary cooling zone. *Optimal Control Applications and Methods*, 37, 1314–1328.
6. d’Humières, D., Ginzburg, I., Krafczyk, M., Lallemand, P., & Luo, L. S. (2002). Multiple-relaxation-time lattice Boltzmann models in

- three dimensions. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 360, 437–451.
7. Dongarra, J., Moore, S., Peterson, G., Tomov, S., Allred, J., Natoli, V., & Richie, D. (2008). Exploring new architectures in accelerating CFD for Air Force applications. *Proceedings of the HPCMP Users Group Conference*.
 8. He, X., Chen, S., & Doolen, G. D. (1998). A novel thermal model for the lattice Boltzmann method in incompressible limit. *Journal of Computational Physics*, 146(1), 282–300.
 9. Lallemand, P., & Luo, L. S. (2003). Theory of the lattice Boltzmann method: Acoustic and thermal properties in two and three dimensions. *Physical Review E*, 68(3), 036706.
 10. Lee, V. W., Kim, C., Chhugani, J., Deisher, M., Kim, D., Nguyen, A. D., Satish, N., Smelyanskiy, M., Chennupaty, S., Hammarlund, P., et al. (2010). Debunking the 100X GPU vs. CPU myth: an evaluation of throughput computing on CPU and GPU. *ACM SIGARCH Computer Architecture News*, 38, 451–460.
 11. McNamara, G. R., & Zanetti, G. (1988). Use of the Boltzmann Equation to Simulate Lattice-Gas Automata. *Physical Review Letters*, 61, 2332–2335.
 12. Micikevicius, P. (2009). 3D finite difference computation on GPUs using CUDA. *Proceedings of the 2nd Workshop on General Purpose Processing on Graphics Processing Units*, 79–84.
 13. Mudigere, D. (2009). Data access optimized applications on the GPU using NVIDIA CUDA. Master's thesis, Technische Universität München.
 14. NVIDIA. (2010). *Compute Unified Device Architecture Programming Guide version 3.1.1*.
 15. Scalable Acoustic and Thermal Validation Strategies in GPU Manufacturing. (2025). *International Journal of Data Science and Machine Learning*, 5(01), 193–214.