Computational Fluid Dynamics as a Strategic Enabler for AI-Driven Data Centers

Improving PUE, WUE, and Operational Resilience Through Predictive Modeling

Written by:
Sam S. Khalilieh, P.E., LEED AP
National Director
Advanced Manufacturing & Mission Critical
sam.khalilieh@tetratech.com

Reviewed by:
Marvin Weiss, Ph.D., P.Eng.
Vice President
Business Development
Marvin.Weiss@coanda.ca

As data centers evolve to support higher rack densities and AI workloads, while contending with increasingly constrained energy and water resources, traditional thermal management approaches based on conventional rules of thumb and lagging efficiency metrics are becoming insufficient. This paper examines the application of Computational Fluid Dynamics as an operational and planning tool for understanding airflow, heat transfer, and cooling system behavior in modern data centers.

At its core, CFD addresses a simple question: how do air, heat, and liquids move through real space under real operating conditions? CFD quantifies and visualizes these behaviors using physics-based models rather than inference or suppositions, thus providing insight into situations that are difficult or impractical to measure directly.

By integrating CFD with commonly used performance metrics such as Power Usage Effectiveness (PUE) and Water Usage Effectiveness (WUE), operators have been able to identify inefficiencies, evaluate failure and transition scenarios, and improve capacity utilization while managing operational risk.

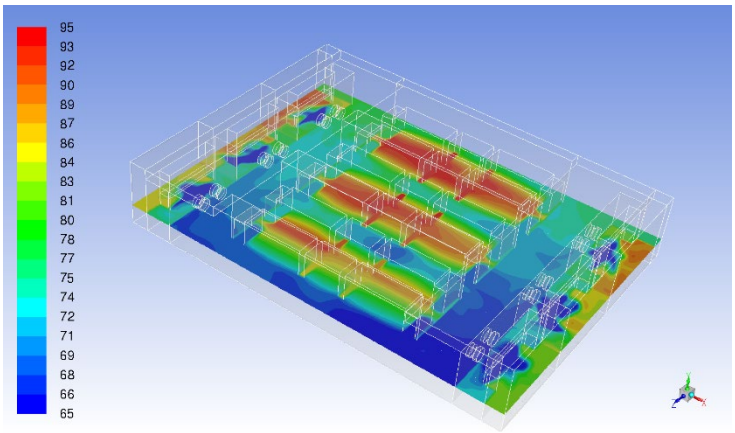


Figure 1: CFD simulation visualizing temperature patterns in hot and cold aisles of a data center

The paper presents a closed-loop operational framework, discusses practical CFD applications in air- and liquid-cooled environments, and summarizes representative performance outcomes from applied studies. In addition, it outlines how CFD functions as a leading indicator for data center performance, why it is increasingly critical in hybrid air- and liquid-cooled environments, and how operators can use it to align AI growth plans with physical infrastructure realities.

The Challenge Facing Modern Data Centers

AI and high-performance computing workloads are fundamentally reshaping data center infrastructure. Rack densities are increasing rapidly, cooling architectures are becoming more complex, and sustainability targets are tightening under regulatory and stakeholder pressure.

At the same time, many facilities still rely on lagging operational indicators such as PUE, WUE, temperature alarms, and incident response to manage risk. These metrics confirm outcomes after inefficiencies or failures have already occurred, often prompting conservative responses such as excessive cooling, lower supply temperatures, and increased water use. As margins shrink, this reactive approach becomes increasingly costly and unsustainable.

What CFD Delivers

Applied to data center environments, CFD provides a detailed, spatial understanding of how airflow, heat transfer, and pressure interact under real operating conditions. It accounts for equipment heat loads, airflow paths, containment effectiveness, cooling unit operation, and heat rejection mechanisms across the facility.

By resolving these interactions at a granular level, CFD allows operators to identify conditions that are not captured by point sensors alone, including localized recirculation, bypass airflow, pressure imbalance, and uneven cooling distribution. This insight enables precise identification of inefficiencies, emerging risks, and unused capacity that would otherwise remain hidden.

Scope of Discussion

This paper focuses on the application of CFD in enterprise, colocation, and hyperscale data center environments supporting mixed IT and AI workloads. The discussion emphasizes operational and planning use cases rather than detailed solver methodologies or vendor-specific implementations.

The quantified performance ranges cited below reflect typical outcomes observed in applied CFD studies under steady-state and selected transitional conditions. These ranges are intended to illustrate representative results rather than guarantee performance, as actual outcomes depend on facility design, operating discipline, climate, cooling architecture, and load characteristics.

From Lagging Metrics to Leading Indicators

PUE and WUE are *lagging outcome metrics*. They describe how efficiently a facility operated, but they do not explain why inefficiencies occurred or where future risk lies. CFD acts as a *leading indicator* by modeling current and future operating conditions before those conditions manifest in measured performance. This distinction becomes increasingly important as AI workloads compress thermal tolerance and reduce operational margin for error. In practice, CFD enables operators to:

- Predict thermal risk before alarms occur.
- Identify inefficiencies before energy or water is consumed.
- Validate operating envelopes before the IT load is deployed.
- Evaluate failure modes before redundancy is tested in production.

Beyond day-to-day operations, the same physics-based modeling is increasingly applied during planning to evaluate thermal behavior before infrastructure decisions are finalized, reducing reliance on conservative rules of thumb and helping avoid design choices that later constrain efficiency, scalability, or resilience. By simulating airflow, heat transfer, and interactions among cooling systems under a range of “what-if” scenarios, CFD allows planners to test assumptions, explore alternative layouts or cooling strategies, and identify potential constraints early in the design process.

Using PUE, WUE, and CFD Together: A Closed-Loop Performance Model

Figure 2 summarizes how efficiency metrics and physics-based analysis are integrated into a closed-loop operating model. PUE and WUE provide standardized visibility into energy and water efficiency outcomes and remain essential for performance monitoring and benchmarking.

CFD complements these metrics by diagnosing the physical drivers behind observed trends and by evaluating the impact of potential changes before they are implemented. Used together, these tools support a continuous cycle of monitoring, analysis, action, and verification.

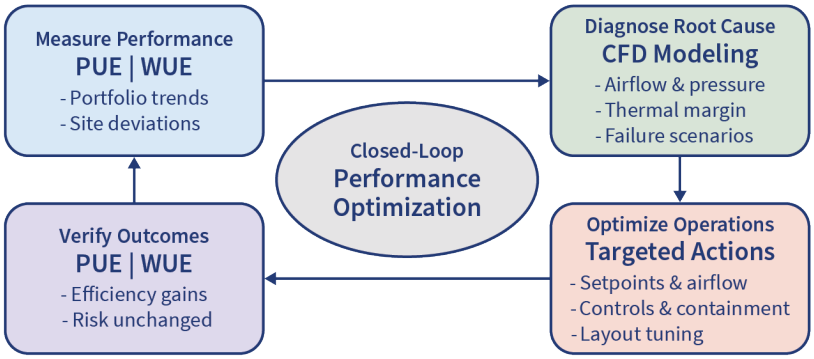


Figure 2

High-performing data center operators use PUE, WUE, and CFD together as a closed-loop operating model:

- PUE and WUE identify performance deviations and trends.
- CFD diagnoses the underlying physical drivers and predicts the impact of change.
- Targeted operational actions are implemented based on modeled behavior.
- PUE and WUE then confirm whether efficiency improvements were achieved without increasing risk.

Together, PUE, WUE, and CFD enable predictive, evidence-based data center operations. This closed-loop approach is especially critical in high-density AI and hybrid air- and liquid-cooled environments, where narrow operating tolerances make reactive efficiency management both costly and risky.

How CFD Improves Power Usage Effectiveness

CFD enables operators to safely improve PUE by identifying efficiency opportunities that are invisible to point sensors and dashboards. It reveals how airflow and cooling capacity interact across the entire facility. Key improvement mechanisms include:

- **Eliminating Overcooling**

  Many data halls contain large zones that are significantly colder than required due to bypass air, excess fan pressure, or poorly tuned containment. CFD quantifies these conditions precisely, allowing operators to safely raise supply air temperatures and reduce chiller and fan energy without increasing thermal risk.

- **Reducing Fan Energy**

  Fan power is often one of the most significant contributors to non-IT energy use. CFD reveals excess airflow, pressure imbalances, and recirculation paths, enabling fan speeds to be reduced while maintaining adequate cooling where needed.

- **Unlocking Stranded Capacity**

  CFD frequently reveals cooling capacity that exists but cannot be used due to poor airflow distribution. Correcting these issues allows higher rack densities to be deployed without new mechanical infrastructure.

Enabling Meaningful Reductions in Water Usage Effectiveness

Water availability is increasingly a first-order constraint, particularly in drought-prone regions and water-stressed markets. CFD supports WUE reduction by:

- Maximizing economizer effectiveness.
- Reducing evaporative cooling enabled by conservative or poorly tuned control logic
- Eliminating water use driven by conservative safety margins.
- Optimizing airflow so that evaporative cooling and adiabatic pre-cooling systems operate only when required.

When airflow and heat rejection paths are understood with precision, water becomes a controlled variable rather than an insurance policy.

Beyond efficiency metrics, CFD reduces unnecessary energy and water consumption driven by uncertainty. By providing a physics-based understanding of airflow and heat rejection behavior, CFD allows operators to operate closer to actual thermal limits without relying on conservative safety margins. This precision reduces overcooling, limits excess evaporative water use, and improves overall resource efficiency while maintaining reliability.

Why this matters:

As AI workloads increase, thermal density and operating margin decrease; decisions based solely on lagging efficiency metrics often lead to overcooling, excessive water use, and stranded capacity. Physics-based modeling provides a way to evaluate risk and performance before changes are implemented.

For operators tracking carbon performance at scale, these efficiency gains also affect Carbon Usage Effectiveness (CUE). By reducing unnecessary cooling energy driven by uncertainty and conservative operating margins, CFD lowers total energy consumption for a given IT load. In regions where grid carbon intensity is non-zero, this reduction translates directly into lower associated emissions and improved CUE. In this way, CFD indirectly supports carbon-efficiency objectives by minimizing avoidable energy demand rather than by influencing energy sourcing or procurement strategies.

Quantified Performance Benchmarks

The following benchmarks reflect typical outcomes observed in operational CFD studies. Actual results vary by facility design and operating conditions.

Energy and Cooling Performance Fan energy reduction: 10-30% Chiller energy reduction: 5-15% PUE improvement: 0.03-0.10	**Water Usage** Evaporative or adiabatic water reduction: 10-25%
Capacity and Density Stranded cooling capacity recovery: 10-25% Safe rack density increase: 10-40% in targeted zones	**Risk and Reliability** Hot-spot recurrence reduction: 50-90% after corrective action. Improved confidence in failure and transition scenarios, particularly for liquid cooling retrofits.

CFD Applications That Matter for Data Center Operations

CFD is often associated with airflow visualization, but its real value for owners and operators lies in how it informs performance, capacity, and risk decisions that directly influence Power Usage Effectiveness and Water Usage Effectiveness.

At an operational level, CFD supports several critical applications.

- **Understanding Actual Airflow and Thermal Behavior**

  CFD reveals how supply air, return air, and heat move through the data hall, including bypass airflow, recirculation, and pressure imbalance that are not captured by temperature sensors alone. Figure 3 shows the CFD-predicted temperature distribution in the data hall, highlighting localized hot spots and uneven cooling driven by airflow imbalance and recirculation. These conditions frequently drive overcooling, which inflates both energy and water consumption.

  Typical outcome: Eliminating localized overcooling identified by CFD often enables increases in supply air temperature of approximately 2–6°F, reducing fan and cooling energy use and improving PUE without increasing thermal risk.

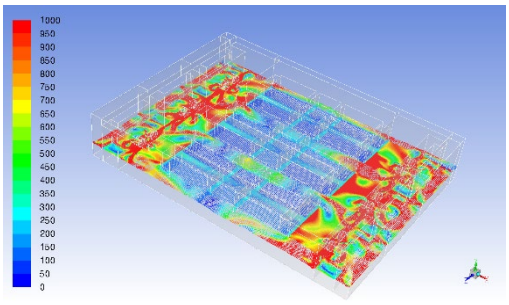


Figure 3: CFD simulation visualizing velocity patterns in hot and cold aisles of a data center

- **Quantifying Containment and Air Management Effectiveness**

  Rather than assuming containment performs as designed, CFD quantifies leakage, short-circuiting, and edge effects under real operating conditions. These losses directly degrade cooling efficiency and reduce usable capacity.

  Typical outcome: Correcting containment leakage identified through CFD has been observed to recover 10–25% of previously stranded cooling capacity, allowing higher rack densities without additional mechanical infrastructure.

- **Supporting Hybrid Air and Liquid Cooling Operation**

  In hybrid environments, CFD helps operators balance residual air-cooling needs with liquid-cooled loads. This avoids the typical operational response of overcooling the entire room to protect a small subset of air-cooled components, which negatively impacts both PUE and WUE.

Across these applications, CFD provides the physical context needed to interpret observed performance trends and support informed operational decisions.

Failure Precursors in High-Density AI Racks

High-density AI racks operate with significantly narrower thermal and electrical margins than traditional IT equipment. Power density is higher, heat generation is more concentrated, and acceptable operating ranges are tighter. As a result, localized deviations in airflow, temperature, or cooling delivery that would be inconsequential in conventional racks can lead to throttling, instability, or protective shutdowns in AI environments.

These failures are rarely caused by a uniform temperature rise across a rack. More often, they originate from localized hot spots, uneven airflow distribution, transient load changes, or degraded cooling conditions that develop faster than alarms can respond. Understanding these precursors is essential for managing AI infrastructure reliably, particularly as rack densities continue to increase and operating margins shrink.

CFD does not "prevent failures."
It reveals the physical conditions that lead to failure early enough to intervene.

Using CFD to Model Failure, Transition, and Resilience

The following discussion focuses on the use of CFD as an operational analysis tool, rather than as a one-time design validation exercise. It is important to distinguish between design-stage and operational use of CFD. While CFD is often associated with one-time design validation, its operational application is fundamentally different. Operational CFD, however, is not a real-time monitoring tool and does not run continuously. Instead, it is applied periodically to evaluate thermal margin, sensitivity, and "what-if" behavior under current and proposed operating conditions, enabling proactive decisions before changes are made or risks materialize.

Aspect	Design CFD	Operational CFD
Primary purpose	Validate design feasibility	Evaluate robustness, margin, and risk
When it is used	During design or major retrofit	During live operation, as conditions change
Typical frequency	One-time or limited iterations	Periodic, revisited as needed
Focus	Steady-state performance	Sensitivity, transitions, and degraded scenarios
Questions Answered	"Will this design work?"	"How will the system behave if conditions change?"
Inputs	Planned layouts, assumed loads, setpoints	Actual rack loads, configurations, operating conditions
Relationship to monitoring	Independent of live data	Informed by observed operational trends
Role in decision-making	Supports initial design decisions	Supports proactive operational and planning decisions
Does it run continuously?	No	No

Most data centers are designed with redundancy, but few are evaluated under realistic degraded or transitional conditions before those conditions occur. CFD enables this evaluation safely and proactively.

From an operational risk perspective, this capability directly supports uptime, contractual obligations, and reputational risk management.

While CFD does not directly prevent hardware failures, it supports AI rack reliability by identifying thermal and airflow conditions that often precede failure. By resolving localized hot spots, uneven cooling, and loss of thermal margin under both steady-state and transitional conditions, CFD allows operators to address emerging risks before alarms are triggered or protective limits are reached. This kind of early insight is especially valuable for high-density AI racks operating close to their thermal limits, where even slight deviations can quickly become operational issues.

- **Modeling Equipment Failures and Degraded Operation**

  CFD can simulate the loss of cooling units, fans, or pumps and show how thermal margin erodes spatially and over time. This identifies which areas fail first and whether redundancy behaves as intended.
  Typical outcome: CFD-based failure analysis often reveals that only 20–30% of the data hall experiences meaningful thermal stress during single-equipment failures, enabling targeted mitigation rather than global overcooling that would otherwise degrade PUE.

- **Evaluating Liquid Cooling Degradation and Fallback Scenarios**

  As liquid cooling is deployed for AI workloads, tolerance for error decreases. CFD allows operators to understand how partial loop degradation or air-system fallback affects residual air-cooled components and overall thermal stability.

Typical outcome: facilities using CFD to validate fallback scenarios have reduced reliance on conservative, water-intensive safety margins, resulting in 10–25% reductions in evaporative or adiabatic water use under partial-load conditions.

- **Assessing Control System Transitions**

  CFD can be used to evaluate thermal behavior during economizer transitions, load ramps, and emergency operating modes–conditions that often cause instability and conservative operator intervention.

By enabling physics-based evaluation of failure and transition scenarios, CFD reduces the likelihood of reactive responses that protect up time at the expense of long-term efficiency.

CFD in Hybrid Air- and Liquid-Cooled Environments

As direct-to-chip liquid cooling is adopted for AI and HPC workloads, some assume airflow becomes less critical. Liquid cooling creates a hybrid thermal system that increases complexity rather than reducing it.

- **Residual Air-Cooling Requirements**

  Even in liquid-cooled racks, components such as power supplies, memory, networking equipment, voltage regulation modules, and cabling still rely on air cooling. CFD ensures sufficient airflow for these components without overcooling the space, a common efficiency failure in early liquid-cooled deployments.

- **CFD and the Limits of Residual Air Cooling**

  CFD modeling is frequently applied to evaluate high-density direct-to-chip deployments, particularly to understand airflow distribution, recirculation, and localized temperature behavior associated with residual air-cooled heat loads. While CFD can demonstrate that acceptable average temperatures are achievable under steady-state conditions, it also reveals that airflow requirements and fan energy increase nonlinearly as residual air heat per rack rises into the tens of kilowatts.

  In this scheme, CFD results consistently show that system performance becomes increasingly sensitive to small disturbances, including minor changes in rack impedance, containment leakage, perforated tile placement, or partial equipment outages. These effects emerge not because average temperatures rise uniformly, but because localized recirculation and bypass paths form as airflow rates increase. As a result, CFD often confirms that high-density air-assisted cooling designs can function under controlled conditions, while simultaneously revealing reduced robustness under realistic operating variability.

  As operators evaluate alternatives to air-assisted cooling for managing very high rack densities, immersion cooling has emerged as an approach that changes how heat is removed from IT equipment. By eliminating the need to transport residual heat through air, immersion fundamentally alters the thermal problem rather than attempting to optimize around airflow constraints. As immersion adoption increases, the role of CFD shifts from IT heat removal to facility-level airflow considerations, including ventilation, safety, and personnel comfort, while continuing to support overall environmental and operational analysis.

- **Complex Heat Rejection Paths**

  Liquid cooling introduces additional thermal interfaces, including cold plates, liquid loops, coolant distribution units, heat exchangers, and hybrid air- and water-based rejection systems. CFD allows these interactions to be visualized and validated, particularly in areas where thermal behavior is often misunderstood.

- **Reduced Margin for Error**

  Liquid cooling enables higher temperature differentials and warmer supply conditions, improving efficiency while narrowing error tolerance. CFD allows operators to model pump failures, partial-loop degradation, control-valve faults, and air-system fallback scenarios before they occur in production.

Understanding CFD Software Capabilities

Not all CFD tools support the same objectives, and selecting the wrong class of software can undermine both efficiency and risk management goals. From an owner's perspective, the distinction is not about software features but about decision support.

In practice, design-focused and operational CFD typically rely on the same underlying physics-based simulation tools; the distinction lies in how the models are configured, updated, and used to support different types of decisions.

- **Design-Focused CFD Tools**

  These tools are typically used for new construction or major retrofits. They offer high fidelity but longer turnaround times and are less suited for rapid operational decisions. Their value lies in validating capital design assumptions rather than improving day-to-day performance.

- **Operational Use of CFD**

  Operational platforms are designed to support live facilities. They allow faster scenario analysis and are better aligned with capacity planning, efficiency optimization, and resilience evaluation.

  Typical outcome: facilities using operational CFD to guide airflow and setpoint adjustments have demonstrated PUE improvements of 0.03–0.10, depending on baseline efficiency and cooling architecture.

- **Simplified or Reduced-Order Models**

  Simplified tools can provide quick insights but should not be relied upon for high-density AI or liquid-cooled environments. Used improperly, they can create false confidence and drive decisions that increase risk or degrade WUE.

Regardless of tool category, the impact of CFD on PUE and WUE depends on calibration, validation against real operating data, and disciplined integration into operational decision-making.

Hyperscalers and Colocation Providers: Different Needs, Same Tool

Hyperscale Operators

For hyperscalers, CFD enables repeatable deployment of high-density AI infrastructure, confident operation closer to thermal limits, and capital avoidance by maximizing the use of existing assets. It supports portfolio-wide standardization without sacrificing resilience and validates that AI roadmaps are physically executable at scale.

Colocation Providers

For colocation operators, CFD supports predictable thermal performance across mixed-tenant loads, safe onboarding of AI tenants without over-engineering, fair allocation of cooling capacity, and reduced operational and commercial risk from tenant variability. In this context, CFD becomes a risk management tool as much as an engineering one.

From Static Model to Living System

CFD delivers its greatest value when integrated with control systems, Building Management Systems, and Data Center Infrastructure Management platforms. This integration enables:

- Validation of control sequences against physical behavior.
- Setpoint optimization based on actual airflow patterns.
- Predictive tuning of economizers and fan control.
- Scenario testing before operational changes are implemented.

At this stage, CFD evolves from a static analysis into a living digital model that informs both real-time operations and long-term strategy.

CFD in Planning, Validation, and Capacity Management

Beyond analysis of normal operating conditions, CFD is widely used to support data center planning, validation, and capacity management. By modeling airflow and heat transfer under existing and proposed conditions, CFD allows operators to understand why localized hot spots occur, evaluate mitigation strategies, and avoid corrective actions that rely on trial-and-error.

In both new design and retrofit scenarios, CFD has been used to identify design choices that can introduce unnecessary capital cost, reduce operational margin, or increase sensitivity to disturbances. Evaluating these conditions early allows operators to avoid latent constraints that can later manifest as stranded capacity or elevated risk of downtime.

For operating facilities, CFD supports capacity planning by estimating available cooling headroom under current configurations and by simulating expansion scenarios before equipment is deployed. This approach enables comparison of alternative growth paths and operating strategies without disrupting production environments, reducing reliance on conservative assumptions, and minimizing the risk associated with incremental expansion.

Machine Learning as a Complement to Physics-Based Analysis

In more mature operating environments, machine learning techniques are increasingly used alongside CFD to identify patterns, anomalies, and early indicators of change in thermal and airflow behavior. By analyzing historical operating data, sensor trends, and prior CFD results, machine learning models can help flag emerging conditions that warrant deeper physics-based evaluation.

Importantly, machine learning does not replace CFD or physical modeling. Instead, it helps prioritize when and where CFD analysis is most valuable, while CFD provides the explanatory and predictive insight needed to validate decisions under changing conditions. Used together, these approaches support earlier intervention and more informed operational planning without relying on trial-and-error or purely statistical inference.

CFD as a Core Enabler of AI Roadmap Planning

Many AI roadmaps focus on compute, chips, networking, and software while underestimating physical infrastructure constraints. These plans often fail when thermal, water, or control limitations are discovered too late. CFD provides a physics-based reality check that brings these risks forward into the planning phase.

With CFD, owners and operators can answer critical questions such as:

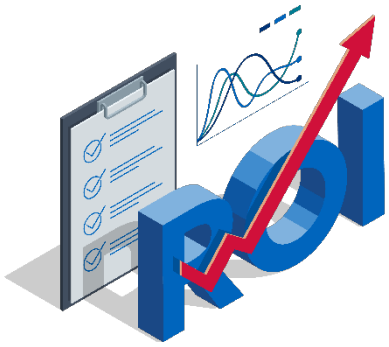
- What rack densities are achievable today and in three to five years?
- Where does airflow, rather than power, become the limiting factor?
- When does liquid cooling become mandatory rather than optional?
- How incremental AI load impacts PUE and WUE.
- Which halls are AI-ready, and which require infrastructure transition?

This shifts AI planning from educated guesswork to evidence-based decision-making.

Operational Return on Investment

While CFD is valuable during design, its highest return on investment is often realized post-occupancy. Common applications include:

- AI and GPU retrofit planning.
- Rack placement and migration strategy.
- Containment optimization.
- Seasonal operating strategy development.
- Failure scenario validation.



At this stage, CFD functions as a decision-support platform that informs both daily operations and long-term infrastructure planning.

Conclusion

Power Usage Effectiveness and Water Usage Effectiveness provide essential visibility into data center efficiency outcomes, but they do not explain the physical drivers behind those outcomes. Computational Fluid Dynamics has increasingly been applied to bridge this gap by providing physics-based insight into airflow, heat transfer, and cooling system interactions.

When used together, PUE, WUE, and CFD form a closed-loop operational framework that enables a shift from reactive efficiency management toward predictive, evidence-based control. As AI workloads increase thermal density and narrow operating tolerances, this integrated approach has become an essential tool for improving energy efficiency, managing water use, and evaluating resilience under both steady-state and transitional conditions.

The growing use of CFD beyond design-phase analysis reflects a broader industry shift toward data-driven, physics-informed operation of high-density data center environments.