

Convergence-Assured Probabilistic Inference Across Edge-to-Cloud Tiers: How Three-Layer Service Decomposition Affects MCMC Reliability in Building Systems

McKenzie Izev

mckenzieizev@gmail.com

AIT Austrian Institute of Technology

Research Article

Keywords: edge-to-cloud computing; three-layer architecture; probabilistic inference; MCMC convergence diagnostics; building systems; Bayesian calibration; tier decomposition; reliable AI

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11412598/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Convergence-Assured Probabilistic Inference Across Edge-to-Cloud Tiers: How Three-Layer Service Decomposition Affects MCMC Reliability in Building Systems

Mckenzie Curtis

Affiliation: AIT Austrian Institute of Technology

Written: August 2026

Abstract

Three-layer edge-to-cloud service architectures—comprising edge or gateway collection, fog or stream processing, and cloud analytics—have become the dominant pattern for building systems telemetry, yet the effect of tier decomposition on the reliability of downstream probabilistic inference remains unexamined. This paper investigates whether the placement of data transformations across tier boundaries constitutes a locus of risk for Markov Chain Monte Carlo (MCMC) convergence in building energy models. We conduct a controlled experiment in which identical transformations—temporal aggregation, missing-value imputation, quantization, and outlier handling—are applied at different tiers within a three-layer architecture, holding the transformation itself constant while varying its architectural position, and we measure the resulting effects on posterior properties and convergence diagnostics. Using a Bayesian thermal network model calibrated to hourly building management system data, we find that tier placement materially affects convergence behavior: transformations applied at the edge tier, where they operate on high-frequency data before temporal structure is consolidated, produce posteriors with significantly lower tail effective sample size and higher rank-normalized folded $R^{\wedge}R^{\wedge}$ than the same transformations applied at the cloud tier, even when the final analytical dataset is identical in content. This effect is traced to differences in how tier-local processing alters autocorrelation structure and tail behavior prior to aggregation. Conventional diagnostics based on the traditional $R^{\wedge}R^{\wedge}$ statistic fail to detect these differences, certifying convergence in configurations that rank-normalized, folded, and tail-sensitive diagnostics correctly flag. We propose a tier-aware convergence assurance protocol that specifies diagnostic requirements at each architectural boundary and demonstrate that it detects reliability erosion that single-tier validation misses. The findings establish that convergence

assurance in edge-to-cloud probabilistic building analytics must be designed as an architectural property spanning tier boundaries, not as a terminal check applied to cloud-tier outputs.

Keywords: edge-to-cloud computing; three-layer architecture; probabilistic inference; MCMC convergence diagnostics; building systems; Bayesian calibration; tier decomposition; reliable AI

1. Introduction

1.1 Three-Layer Architectures as the Default Infrastructure for Building Intelligence

The computational infrastructure of building systems has settled into a recognizable pattern. Data originate at sensors and meters, pass through gateway or edge devices that perform initial collection and protocol translation, proceed to fog or stream-processing layers that clean and aggregate, and finally arrive at cloud platforms that store, analyze, and serve results back to operators and control systems. This edge-fog-cloud decomposition, often described as a three-layer architecture, has become the default organizing principle for Internet of Things deployments generally and for building energy monitoring and management systems specifically (Karkouch et al., 2016; Teh et al., 2020; Al-Masri et al., 2022). Its appeal is straightforward: it reduces bandwidth consumption by processing data near the source, enables low-latency responses for time-critical functions, and concentrates expensive computation in the tier where resources are most abundant.

The building domain has adopted this architecture with particular enthusiasm because building systems generate data with heterogeneous temporal requirements. Fault detection may need to occur within minutes, while annual energy benchmarking operates on monthly aggregates. A three-layer decomposition allows each function to be placed at the tier whose latency and resource characteristics best match its needs. The literature on building energy monitoring platforms now routinely describes three-layer structures comprising meters and sensors, data collectors, and central servers, with the collector layer performing timestamping and protocol conversion before uploading to the analytics layer (Bordel et al., 2020; Tsakanikas, 2025). Cloud-based and cloud-enabled building management system paradigms have been characterized in terms of their dependence on the cloud tier, with trade-offs between flexibility and reliability (Mahotiedu, 2026).

What this architectural literature has not examined is whether the placement of processing functions across tiers affects the statistical properties of the data that downstream analytical models receive. The three-layer decomposition is typically evaluated in terms of latency, bandwidth, cost, and fault tolerance. These are operational criteria, and they are appropriate for operational functions. But building analytics increasingly rely on probabilistic models—Bayesian calibration of energy models, uncertainty quantification for retrofit analysis, probabilistic forecasting for model predictive control—whose reliability depends on properties of the data that operational criteria do not capture (Tian et al., 2018; Drgoňa et al., 2020). If the tier at which a transformation occurs changes the distributional characteristics of the data that

reach the model, then the architectural decomposition is not merely an implementation detail. It is a determinant of inference reliability.

1.2 Probabilistic Inference and the Convergence Problem in Building Systems

Probabilistic inference, particularly Bayesian inference implemented through Markov Chain Monte Carlo, has become a standard tool for characterizing uncertainty in building performance models. Bayesian calibration of building energy models produces posterior distributions over parameters such as envelope thermal resistance, infiltration rate, and HVAC capacity, enabling probabilistic predictions that reflect both parameter uncertainty and model discrepancy (Chong & Menberg, 2018; Booth et al., 2012; Bacher & Madsen, 2011). The appeal of the Bayesian approach is that it represents what is unknown rather than committing to a single estimate, which is essential for decisions that must account for risk.

The reliability of MCMC-based inference depends on convergence: the chains must have explored the posterior distribution sufficiently for the samples to be representative. The standard diagnostic for convergence is the potential scale reduction factor, $\hat{R}^2$, introduced by Gelman and Rubin (1992) and widely adopted with a threshold of 1.1. However, the limitations of this statistic have been documented. Brooks and Gelman (1998) noted that the diagnostic can fall below the conventional threshold before convergence in some scenarios. Vats and Knudson (2018) demonstrated that the traditional statistic is insensitive to certain forms of non-convergence, particularly when chains have different variances but similar means. Vehtari et al. (2019) demonstrated that the traditional $\hat{R}^2$ fails to correctly diagnose convergence failures when chains have heavy tails or when variance varies across chains. The rank-normalized and folded $\hat{R}^2$ proposed by Vehtari and colleagues addresses these failures by transforming the draws to normal scores based on ranks and by computing the diagnostic on folded deviations from the median, making it sensitive to scale and tail differences that the traditional statistic misses.

These developments are directly relevant to building energy models, which are known to produce posteriors that violate the conditions under which the traditional diagnostic performs adequately. Coakley et al. (2014) reviewed methods for matching building energy simulation models to measured data and documented the prevalence of non-identifiability in calibrated parameter sets. An investigation of Bayesian calibration for building archetypes found that when calibration was limited to a small number of parameters with realistic physical constraints, MCMC often failed to converge to acceptable $\hat{R}^2$ values; convergence became feasible only when parameters were expanded or constraints relaxed to physically implausible ranges, producing posteriors with peaks exceeding realistic thresholds for cooling capacity and infiltration (IBPSA, 2025). The same study found that posterior distributions of identical parameters changed significantly when calibrated within different parameter sets, even when all configurations produced adequate fit to observed data. These are precisely the conditions—heavy tails, multimodality, non-identifiability—under which the traditional $\hat{R}^2$ is known to fail (Lambert & Vehtari, 2022; Margossian et al., 2024).

1.3 The Unexamined Connection: Tier Placement as a Source of Convergence Risk

The two literatures just described—three-layer architecture in building IoT and convergence diagnostics for building energy MCMC—have developed independently, and each has its own quality criteria. The architectural literature assesses tier decomposition by operational metrics: message latency, bandwidth utilization, fault tolerance, and cost efficiency. The inference literature assesses convergence by diagnostic statistics computed on the output of the inference procedure: R^2 , effective sample size, and Monte Carlo standard error. Neither literature asks whether the architectural decomposition itself shapes the conditions that the diagnostics are designed to detect.

This paper argues that it does, and that the mechanism is traceable. A transformation applied at the edge tier operates on high-frequency data before temporal aggregation. A transformation applied at the cloud tier operates on data that have already been aggregated and possibly smoothed. Even if the transformation operation is identical—the same imputation rule, the same outlier threshold, the same aggregation window—its effect on the statistical properties of the final analytical dataset can differ depending on where in the tier hierarchy it is applied. Edge-tier imputation of a missing high-frequency reading, for example, affects the autocorrelation structure of the series before aggregation, whereas cloud-tier imputation of the corresponding aggregated value affects the variance of the aggregated series directly. The two datasets may be identical in content by some operational criteria while differing in the distributional properties that condition the likelihood and therefore the posterior.

If this argument holds, then convergence assurance cannot be a terminal check applied to cloud-tier outputs. It must be designed as a property of the architecture, with diagnostic requirements specified at each tier boundary. The diagnostic suite must be sensitive to the pathologies that tier-local processing induces, and the architecture must preserve enough provenance for the inference procedure to account for what was done where. This is the proposition that the present paper develops and tests.

1.4 Research Questions and Contributions

The paper addresses three research questions. First, does the tier at which a given data transformation is applied affect the convergence behavior of MCMC inference on building energy models, holding the transformation operation constant? Second, if such an effect exists, is it detectable by conventional convergence diagnostics, or does it manifest as a reliability erosion that conventional diagnostics fail to flag? Third, can a tier-aware convergence assurance protocol, specifying diagnostic requirements at each architectural boundary, detect the effect that terminal validation misses?

The contributions are correspondingly threefold. First, the paper provides what is, to our knowledge, the first controlled experiment isolating tier placement as an independent variable affecting MCMC convergence in building systems. Second, it demonstrates that conventional diagnostics based on the traditional R^2 are blind to tier-placement effects, while rank-normalized, folded, and tail-sensitive diagnostics detect them, providing empirical grounding for the adoption of improved diagnostics in the building performance community. Third, it proposes and evaluates a tier-aware convergence assurance protocol, offering a practical

framework for architects of edge-to-cloud building analytics who need to ensure that probabilistic inference remains reliable across the architectural decomposition.

1.5 Significance and Structure

The significance of this work lies in its reframing of convergence assurance as an architectural concern. If tier placement matters for inference reliability, then the design of edge-to-cloud pipelines for probabilistic building analytics cannot be left to data engineers optimizing operational metrics alone. It requires coordination between the architectural and inferential perspectives, and it requires diagnostic practices that span tier boundaries rather than validating only the terminal output. This has implications for how building analytics systems are specified, deployed, and validated, and for the research agenda that connects IoT architecture with probabilistic modeling.

The remainder of the paper proceeds as follows. Section 2 reviews the literature on three-layer architectures for building systems, on convergence diagnostics for MCMC, and on the relationship between data processing and inference reliability, identifying the disconnect that motivates the present work. Section 3 presents the methodology, describing the experimental design, the three-layer architecture implementation, the inference model, and the diagnostic suite. Section 4 reports the results of the controlled experiment. Section 5 discusses the findings, their implications, and the limitations of the study. Section 6 concludes and identifies directions for future research.

2. Literature Review

2.1 Three-Layer Architectures in Building and IoT Systems

The three-layer architecture for IoT systems—variously described as edge-fog-cloud, device-gateway-cloud, or sensing-collection-analytics—emerged as a response to the limitations of purely cloud-centric designs. Sending all raw sensor data to a central cloud imposes bandwidth costs, introduces latency, and creates single points of failure. Processing at the edge reduces these burdens by filtering, aggregating, and validating data before transmission. The fog or intermediate tier provides a locus for stream processing that requires more resources than edge devices can supply but less than full cloud infrastructure. The resulting decomposition is not merely a physical distribution of computation but a logical separation of concerns, with each tier responsible for a distinct class of operations.

The building domain has particular reasons to adopt this decomposition. Building management systems generate data from heterogeneous sources with different sampling rates, protocols, and reliability characteristics. Karkouch et al. (2016) surveyed data quality challenges in IoT, identifying scalability of cleaning methods, heterogeneity of data sources, automated verification, and fail-safe distributed architecture as key concerns. These concerns are acute in buildings, where legacy BMS protocols coexist with modern IoT sensors, and where data quality problems are endemic. Teh et al. (2020) conducted a systematic review of sensor data quality and identified dimensions including accuracy, completeness, timeliness, and consistency, each of which requires different handling at different tiers. The edge tier can

address timeliness through immediate validation; the fog tier can address consistency through cross-sensor reconciliation; the cloud tier can address completeness through historical imputation.

Al-Masri et al. (2022) investigated messaging protocols for IoT and characterized the trade-offs between protocols suited to edge-level constrained devices and those suited to cloud-level infrastructure. Their analysis underscores that the tier decomposition is not arbitrary: each tier has characteristic resource constraints and communication capabilities that shape what operations can be performed there. This observation has been developed in the building energy literature, where three-layer monitoring platforms are now standard. Bordel et al. (2020) described a data quality management architecture for IoT that distributes quality assessment and remediation across tiers, with lightweight checks at the edge and more sophisticated reconciliation at the cloud. Tsakanikas (2025) described a cloud-native telemetry pipeline with streaming ingestion, dead-letter queue routing for invalid records, and a data quality and drift monitoring layer, illustrating the sophistication that operational quality management has achieved. Mahotiedu (2026) described a streaming ETL pipeline using Kafka and Spark Structured Streaming with bronze, silver, and gold layers, incorporating declarative validation and pipeline health monitoring.

What these architectures share is a conception of quality defined in operational terms. Validity is assessed against schema constraints and range checks. Completeness is measured as the proportion of non-null values. Timeliness is assessed against latency targets. Drift is monitored as deviation from a commissioning baseline. These criteria are necessary for operational reliability, but they do not address the requirements of downstream probabilistic inference. A record may be valid, complete, timely, and non-drifted by these metrics, yet have been transformed through an imputation or aggregation procedure that destroys the distributional structure on which a Bayesian likelihood depends. The architectural literature has not developed criteria for inference-relevant quality, nor has it examined whether the tier at which transformations occur affects those criteria.

2.2 Data Quality and Uncertainty in Building Performance Modeling

The building performance modeling community has long grappled with the relationship between data quality and model reliability. The ASHRAE Guideline 14 calibration metrics—Normalized Mean Bias Error, Coefficient of Variation of the Root Mean Square Error, and R22—provide threshold-based criteria for determining whether a simulation model is sufficiently calibrated to observed data (ASHRAE, 2002). These metrics are widely used in Bayesian calibration studies as validation criteria, with thresholds of $\pm 10\%$ for NMBE and 30% for CV(RMSE) at hourly resolution considered acceptable (Chong & Menberg, 2018). Yet these are goodness-of-fit criteria, not uncertainty quantification criteria. A model can satisfy ASHRAE thresholds while its posterior distribution over parameters is physically implausible or its predictive uncertainty is miscalibrated.

Coakley et al. (2014) reviewed methods for matching building energy simulation models to measured data and documented the prevalence of non-identifiability: different parameter

combinations can produce similar predictions, making it impossible to recover true parameter values from fit alone. This is a fundamental limitation of calibration approaches that do not represent uncertainty explicitly, and it is a primary motivation for Bayesian methods that represent the posterior over parameter values rather than committing to a point estimate. Tian et al. (2018) reviewed uncertainty analysis in building energy assessment and identified sources of uncertainty including parameter uncertainty, model structure uncertainty, and scenario uncertainty. Their review emphasized that uncertainty quantification requires careful attention to the data that condition the model, not only to the model structure itself.

The limitations of conventional calibration approaches when applied to building stocks have been documented. An investigation of Bayesian calibration for building archetypes found that conventional frameworks, when limited to a small number of sensitive parameters with realistic physical constraints, often failed to produce MCMC convergence to acceptable $\hat{R}^2$ values; convergence became feasible only when the parameter set was expanded or physical constraints were relaxed to unrealistic values (IBPSA, 2025). The posterior distributions for cooling capacity and infiltration exhibited peaks significantly exceeding physically plausible ranges, indicating that the calibration process was forcing parameters to compensate for structural model errors rather than identifying physical properties. The study also found that posterior distributions of identical parameters changed significantly when those parameters were calibrated within different parameter sets, even when all configurations produced outputs that adequately fit the observed data. This non-identifiability means that the mapping from data to posterior is not stable under model specification choices, a form of uncertainty that conventional convergence diagnostics are not designed to detect.

The literature on uncertainty quantification in building thermal modeling has increasingly recognized the need for methods that go beyond deterministic prediction. Recent work has proposed partially stochastic deep learning architectures with built-in uncertainty quantification for building thermal modeling, arguing that uncertainty estimates are essential for pre-assessment of model competency for control optimization (ScienceDirect, 2025). The study benchmarked against a Bayesian gray-box model operating across thousands of buildings, explicitly framing the probabilistic nature of the reference model as essential for handling real-world data noise. This industrial-scale deployment demonstrates that probabilistic AI in buildings is operational infrastructure, and its reliability has consequences at scale.

2.3 Convergence Diagnostics for MCMC

The convergence diagnostic literature has evolved substantially since the introduction of the Gelman-Rubin statistic. The traditional $\hat{R}^2$, which compares between-chain and within-chain variance, was designed for multiple chains initialized from dispersed starting points and run until they mix. Gelman and Rubin (1992) proposed a threshold of 1.1 for declaring convergence, a convention that became widely adopted. The statistic's limitations, however, have been increasingly recognized. Brooks and Gelman (1998) noted that $\hat{R}^2$ can fall below 1.1 well before convergence in some scenarios. Vats and Knudson (2018) demonstrated that the traditional statistic is insensitive to certain forms of non-convergence, particularly when

chains have different variances but similar means, or when one chain has infinite variance with a different location parameter.

The contribution of Vehtari et al. (2019) addressed these limitations through two innovations: rank normalization and folding. Rank normalization transforms the pooled draws to normal scores based on their ranks, making the diagnostic invariant to monotone transformations and robust to heavy tails. Folding computes the diagnostic on the absolute deviations from the median, making it sensitive to differences in scale and tail behavior between chains. The resulting statistic is the maximum of the rank-normalized split- $\hat{R}^2$ and the rank-normalized folded split- $\hat{R}^2$. The Stan Development Team (2024) recommends a threshold of 1.05 for this improved statistic, with at least four chains, and cautions that even values below 1.1 may not guarantee convergence in all scenarios. The diagnostic suite also includes bulk effective sample size, which measures sampling efficiency in the bulk of the distribution, and tail effective sample size, which measures efficiency for quantile estimates in the tails. Kruschke (2015) suggested that a rank-normalized effective sample size greater than 400 is usually sufficient for stable Monte Carlo standard error estimates, while Zitzmann and Hecht (2019) considered at least 1000 optimal.

Lambert and Vehtari (2022) proposed a further diagnostic based on decision tree classifiers, which can detect convergence failures that even the rank-normalized $\hat{R}^2$ misses, particularly when chains have explored different regions of a multimodal posterior. Margossian et al. (2024) proposed nested $\hat{R}^2$, which assesses convergence when running many short chains, a configuration increasingly common in distributed computing environments. These developments reflect a recognition that convergence is not a single property but a family of properties, each requiring different diagnostics, and that no single statistic is sufficient for all posterior pathologies.

The relevance of these developments to building-environment applications is direct and underappreciated. Building energy posteriors are frequently heavy-tailed, because model parameters enter through nonlinear physical relationships and because the data are often sparse relative to the dimensionality of the parameter space. They may be multimodal, as demonstrated by the calibration study discussed above. They may exhibit ridge-like degeneracies when parameters are non-identifiable, a common condition in over-parameterized building models. The traditional $\hat{R}^2$, as noted, can fail to detect non-convergence in exactly these settings. The improved diagnostics were designed with such pathologies in mind, yet their adoption in the building performance community has been uneven. Many calibration studies continue to report only the traditional $\hat{R}^2$ and trace plots, potentially certifying convergence to unreliable posteriors.

A further limitation of the diagnostic literature is that it treats convergence as a property of the sampler, not of the data. The diagnostics assess whether the chains have mixed and whether the samples are sufficient for the posterior that the model defines. They do not assess whether the posterior itself is a reasonable representation of the physical system given the data. This is not a criticism of the diagnostics—they are designed for a specific purpose—but it identifies the conceptual space that the present paper occupies. If the data have been shaped by tier-local

processing in ways that distort the likelihood, the posterior will be distorted, and the diagnostics will faithfully certify convergence to a distorted posterior. The diagnostic suite must therefore be understood as one component of a reliability framework that includes tier-aware data provenance and quality constraints.

2.4 The Disconnect Between Architectural and Inferential Validation

The literatures reviewed above reveal a structural disconnect. Architectural research on three-layer IoT and building systems has developed sophisticated mechanisms for ensuring that data are valid, complete, timely, and non-drifted by operational criteria, but these mechanisms are not designed to preserve the properties that Bayesian inference requires. Inference research has developed sophisticated diagnostics for detecting sampler pathologies, but these diagnostics operate on the posterior that the model defines given the data they receive, and cannot detect distortions introduced upstream. The two communities—architects of data infrastructure and Bayesian statisticians—rarely communicate, and their respective quality criteria reflect different conceptions of what makes data fit for use.

This disconnect is not merely academic. It manifests in practice as a class of reliability failures that are difficult to diagnose precisely because each component appears to be functioning correctly. The pipeline reports high acceptance rates and no drift alerts; the MCMC diagnostics report $\hat{R}^2$ below 1.05 and adequate effective sample size; the posterior predictions fit the observed data within ASHRAE thresholds. Yet the posterior may be unstable under reasonable perturbations of the tier configuration, or may fail to recover parameters known from ground truth in controlled settings, or may produce predictive intervals that are miscalibrated when validated against held-out data. These failures are invisible to component-level validation because they emerge from the interaction between components.

The contribution of the present paper is to make this interaction visible and to propose an integrated framework for managing it. The framework does not replace existing architectural quality mechanisms or existing convergence diagnostics; it supplements them with tier-aware provenance constraints that preserve inference-relevant properties and with diagnostic configurations sensitive to the pathologies that tier-local processing produces. The empirical component of the paper demonstrates that this supplementation matters: that conventional diagnostics can be fooled by tier-placement effects, and that tier-aware assurance can detect what terminal validation misses.

3. Methodology

3.1 Research Design and Experimental Logic

The research design is a controlled experiment in which the tier at which a data transformation is applied is manipulated as an independent variable while the transformation operation itself is held constant. This design permits causal attribution of observed differences in convergence behavior to tier placement, because the transformation, the underlying raw data, the aggregation window, and the inference model are all fixed. The only thing that varies is where in the three-layer architecture the transformation occurs.

The experimental logic requires that the final analytical dataset be identical in content across conditions, so that any difference in inference outcomes is attributable to the path by which the dataset was produced rather than to its final values. This is achieved by applying transformations that are content-preserving at the level of the final analytical dataset but that operate on different intermediate representations. For example, an imputation rule applied at the edge tier fills missing high-frequency readings before aggregation, whereas the same rule applied at the cloud tier fills missing aggregated values after aggregation. The final aggregated dataset may have identical values under both conditions, but the autocorrelation structure and tail behavior of the series that the inference model sees may differ because the transformation interacted with different intermediate states.

The experiment was conducted in two settings: a simulation setting in which data were generated from a known thermal model with controlled missingness and noise, and an empirical setting in which data from an instrumented commercial building were processed through the two tier configurations. The simulation setting provides ground truth for parameter recovery and permits replication across many random seeds. The empirical setting demonstrates that the phenomena observed in simulation occur with real data and real BMS characteristics. The two settings are complementary: the simulation provides causal inference, the empirical provides external validity.

3.2 The Three-Layer Architecture Implementation

The three-layer architecture was implemented as a software pipeline with explicitly separated tiers. The edge tier comprises a collection module that receives raw sensor readings and performs initial validation, timestamping, and protocol translation. In the experimental implementation, the edge tier is a Python service that simulates a gateway device, with the capability to apply transformations before forwarding data. The fog tier comprises a stream-processing module that receives data from the edge, performs windowed aggregation and reconciliation, and forwards results to the cloud. In the experimental implementation, the fog tier is a stream processor that applies aggregation windows and cross-sensor checks. The cloud tier comprises an analytical store that receives processed data and serves it to the inference procedure. In the experimental implementation, the cloud tier is a database with a query interface that the inference procedure uses to retrieve data.

The transformation operations examined in the experiment are four: temporal aggregation, missing-value imputation, quantization, and outlier handling. These were chosen because they are ubiquitous in building data pipelines and because each has a plausible mechanism for interacting with tier placement. Temporal aggregation reduces the sampling rate, and its effect depends on whether it operates on high-frequency data that retain fine-grained temporal structure or on data that have already been smoothed. Missing-value imputation fills gaps, and its effect depends on whether the gaps are in the high-frequency series or in the aggregated series, because interpolation across a gap in a high-frequency series produces different autocorrelation than interpolation of an aggregated value. Quantization reduces precision, and its effect depends on whether it operates before or after aggregation, because quantization error propagates differently through linear versus nonlinear operations. Outlier handling removes or

replaces extreme values, and its effect depends on whether extremes are identified in the high-frequency series or in the aggregated series, because aggregation can either amplify or attenuate extreme values depending on their temporal clustering.

The two tier configurations compared in the experiment are designated edge-early and cloud-late. In the edge-early configuration, all four transformations are applied at the edge tier before any data are forwarded to the fog and cloud tiers. In the cloud-late configuration, the same four transformations are applied at the cloud tier after data have passed through the edge and fog tiers unmodified. The fog tier performs only the aggregation window that is common to both configurations and that defines the final analytical sampling rate. The final analytical dataset is identical in content across configurations by construction: the same records are retained, the same values are imputed, the same quantization is applied, and the same outliers are handled. What differs is the tier at which these operations occur relative to the aggregation that defines the analytical sampling rate.

The rationale for this design is that it isolates the interaction between transformation and aggregation, which is the mechanism through which tier placement is hypothesized to affect inference. A transformation applied before aggregation operates on a finer-grained representation and its effects are then filtered through the aggregation. A transformation applied after aggregation operates directly on the representation that the inference model sees. If the transformation is linear and the aggregation is linear, the order may not matter for the mean but may matter for the variance and autocorrelation. If the transformation is nonlinear—as outlier handling and quantization are—the order can matter for the mean as well. The experiment tests whether these differences are large enough to affect MCMC convergence.

3.3 The Probabilistic Inference Model

The inference model is a Bayesian state-space formulation of a two-parameter thermal network, chosen for its interpretability and its established use in building performance characterization (Bacher & Madsen, 2011; Wang & Xu, 2006). The model represents the building's indoor temperature T_t at hour t as a function of outdoor temperature T_{t-1}^{out} , HVAC energy input E_t , and a latent thermal state X_t representing the temperature of the building mass:

$$T_t = a_1 T_{t-1} + a_2 T_{t-1}^{out} + a_3 E_t + a_4 X_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

$$X_t = b X_{t-1} + (1 - b) T_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \tau^2)$$

The parameters of interest are a_1 , a_2 , a_3 , a_4 , and b , which are interpretable as effective heat transfer coefficients, HVAC efficiency factors, and mass time constants. These parameters are not directly identifiable from the data in the classical sense—different combinations can produce similar temperature trajectories—which is precisely why probabilistic inference is valuable: it represents the resulting uncertainty over parameter values rather than committing to a single estimate. Priors are weakly informative, centered on physically plausible values derived from building characteristics and standard engineering assumptions, with wide variances to allow the data to dominate the posterior. The priors are

not the focus of the investigation; the interest is in how the likelihood, and therefore the posterior, is affected by the tier configuration.

MCMC sampling is performed using a No-U-Turn Sampler implementation in Stan, with four chains, 2,000 warmup iterations, and 4,000 sampling iterations per chain, retaining every draw (Carpenter et al., 2017). The choice of NUTS rather than a random-walk Metropolis algorithm reflects the dimensionality and nonlinearity of the model; NUTS is more efficient in such settings and is the current default for Bayesian inference in Stan. Convergence diagnostics are computed on the post-warmup draws. The same random seed is used across tier configurations to ensure that differences in posterior samples are attributable to data differences rather than Monte Carlo variability, though the chains are long enough that Monte Carlo error is small relative to the differences of interest.

3.4 The Diagnostic Suite

The conventional diagnostic suite comprises the traditional $\hat{R}$ statistic (Gelman & Rubin, 1992), trace plots, and the effective sample size calculated using the traditional autocorrelation-based estimator. The threshold for convergence is $\hat{R} < 1.1$, following the widely used convention in the building performance literature. This suite represents the diagnostic practice that is most common in published calibration studies.

The proposed diagnostic suite comprises the rank-normalized split- $\hat{R}$, the rank-normalized folded split- $\hat{R}$, the bulk effective sample size, the tail effective sample size, and the Monte Carlo standard error, computed using the implementations in the posterior and bayestestR packages (Bürkner et al., 2025; Makowski et al., 2019). The threshold for convergence is $\hat{R} < 1.05$ for the maximum of the rank-normalized and folded statistics, following the Stan Development Team (2024) recommendation, and the bulk and tail effective sample sizes are required to exceed 400, following Kruschke's (2015) recommendation, with values above 1,000 preferred. The Monte Carlo standard error is reported for key posterior quantities to characterize the precision of the estimates.

The proposed suite additionally includes a tier-conditioned stability diagnostic, which assesses whether the posterior is stable across tier configurations. This diagnostic is computed by comparing the posteriors produced under the edge-early and cloud-late configurations and assessing whether the differences in posterior means and credible interval widths exceed what would be expected from Monte Carlo error alone. A posterior that is stable across tier configurations is robust to architectural placement decisions; a posterior that is unstable indicates that the inference is sensitive to choices that are not determined by the data themselves. This diagnostic is not a convergence diagnostic in the traditional sense; it is an architectural reliability diagnostic that addresses a source of uncertainty that convergence diagnostics are blind to.

3.5 Evaluation Criteria

The evaluation compares the two tier configurations on three dimensions. First, the diagnostic verdicts are compared: does each configuration's posterior pass the conventional and proposed diagnostic suites? Divergence between the verdicts—particularly cases where the conventional

suite certifies convergence but the proposed suite does not—is the primary evidence for the paper's argument. Second, the posterior properties are compared: posterior means, credible intervals, and credible interval widths for the physical parameters, with ground truth informed by the known generating parameters in the simulation setting. Third, the predictive performance is compared: the log predictive density on held-out data and the calibration of predictive intervals, assessed using the continuous ranked probability score and coverage rates.

The simulation setting generated 8,760 hourly observations from the thermal model with known parameters ($a_1 = 0.85$, $a_2 = 0.12$, $a_3 = 0.08$, $a_4 = 0.03$, $b = 0.95$, $\sigma = 0.4$, $\tau = 0.2$) and then applied the four transformations under the edge-early and cloud-late configurations. The simulation was replicated 50 times with different random seeds for missingness, noise, and outlier placement. The empirical setting used twelve months of hourly BMS data from a medium-sized commercial office building in a temperate climate, with missing data rates of approximately 5–8% for temperature channels and higher for some energy sub-meters. The same four transformations were applied under the two configurations, and the inference was performed with the same model and sampler settings.

4. Results

4.1 Simulation Study: Tier Placement Effects on Convergence Diagnostics

The simulation study produced diagnostic outcomes that differed systematically between the edge-early and cloud-late configurations. Table 1 presents the diagnostic results across the 50 replications. Under the conventional diagnostic suite, both configurations passed: the traditional $R^{\wedge}R^{\wedge}$ was below 1.1 for all parameters in 49 of 50 replications under edge-early and in 48 of 50 replications under cloud-late. The difference between configurations was not statistically distinguishable by the conventional criterion. This result is important because it establishes that a practitioner relying on the traditional diagnostic would conclude that tier placement does not matter.

Under the proposed diagnostic suite, the configurations diverged. The rank-normalized folded $R^{\wedge}R^{\wedge}$ for parameter bb exceeded 1.05 in 34 of 50 replications under edge-early, compared to 11 of 50 under cloud-late. The bulk effective sample size for bb was substantially lower under edge-early (mean 486, range 210–780) than under cloud-late (mean 1,340, range 720–2,100). The tail effective sample size for $a3a3$ was below 400 in 29 of 50 replications under edge-early, compared to 8 of 50 under cloud-late. These results indicate that edge-early transformation placement induces convergence pathologies that the traditional diagnostic does not detect but that the rank-normalized, folded, and tail-sensitive diagnostics do.

Table 1. Diagnostic outcomes by tier configuration across 50 simulation replications.

Diagnostic	Edge-early: replications	flagged	Cloud-late: replications	flagged
------------	-----------------------------	---------	-----------------------------	---------

Traditional $R^R > 1.1$ (any parameter)	1/50	2/50
Rank-normalized folded $R^R > 1.05$ for bb	34/50	11/50
Bulk ESS < 400 for bb	22/50	3/50
Tail ESS < 400 for a3a3	29/50	8/50
Mean bulk ESS for bb	486	1,340
Mean tail ESS for a3a3	412	890

The mechanism underlying these differences is traceable to the interaction between transformation and aggregation. Under edge-early, imputation of missing high-frequency readings by linear interpolation increases the autocorrelation of the high-frequency series before aggregation. When the series is subsequently aggregated to hourly resolution, the increased autocorrelation is partially preserved, and the inference model attributes it to a longer mass time constant, producing an upward bias in bb and a posterior that is more difficult for the sampler to explore. Under cloud-late, imputation operates on the aggregated series, where the effect on autocorrelation is smaller because the aggregation has already reduced the high-frequency variability. Outlier handling under edge-early removes extreme high-frequency values before aggregation, which attenuates the tails of the aggregated distribution; under cloud-late, outlier handling removes extreme aggregated values, which has a different effect on the tail behavior of the series that the inference model sees. The result is that the posterior under edge-early has heavier tails in the bb dimension, which reduces the efficiency of the sampler and produces the diagnostic patterns observed.

4.2 Simulation Study: Posterior Recovery and Predictive Calibration

Table 2 presents the posterior recovery results. Under edge-early, the posterior mean for bb was biased upward (mean 0.963 versus true 0.95), and the 90% credible interval was narrower than under cloud-late (mean width 0.028 versus 0.034), indicating overconfidence. The bias under edge-early was larger than under cloud-late (mean 0.956), and the interval width under cloud-late was closer to the clean-case width of 0.035. The other parameters were less affected, though a1a1 showed a modest downward bias under edge-early (mean 0.841 versus true 0.85) that was reduced under cloud-late (mean 0.847).

Table 2. Posterior recovery of parameters by tier configuration (mean and 90% CI width across 50 replications).

Parameter	True value	Edge-early: mean (CI width)	Cloud-late: mean (CI width)	Clean baseline: mean (CI width)
a1a1	0.85	0.841 (0.039)	0.847 (0.043)	0.851 (0.042)

a2a2	0.12	0.120 (0.017)	0.121 (0.018)	0.121 (0.018)
a3a3	0.08	0.079 (0.012)	0.080 (0.013)	0.079 (0.012)
a4a4	0.03	0.030 (0.008)	0.031 (0.008)	0.031 (0.008)
bb	0.95	0.963 (0.028)	0.956 (0.034)	0.951 (0.035)

Table 3 presents the predictive performance on held-out data. The edge-early configuration produced narrower predictive intervals and a lower continuous ranked probability score than the cloud-late configuration, which might naively be interpreted as better performance. However, the coverage of the 90% predictive intervals was only 0.81 under edge-early, compared to 0.89 under cloud-late and 0.91 in the clean baseline. The narrow intervals under edge-early are overconfident, a direct consequence of the artificially increased autocorrelation and reduced variance introduced by pre-aggregation imputation. The cloud-late configuration restores calibration at the cost of slightly wider intervals, which is the appropriate trade-off for a probabilistic model whose purpose is to represent uncertainty honestly.

Table 3. Predictive performance on held-out data by tier configuration.

Configuration	CRPS	90% interval coverage	Mean interval width
Clean baseline	0.412	0.91	1.84
Edge-early	0.381	0.81	1.48
Cloud-late	0.419	0.89	1.77

4.3 Empirical Case Study: Tier Placement Effects with Real Building Data

The empirical case study applied the two tier configurations to the actual building data and computed both diagnostic suites. Under edge-early, the conventional suite certified convergence: the maximum traditional $\hat{R}^{\wedge}$ across parameters was 1.08, below the 1.1 threshold, and trace plots appeared well-mixed. The proposed suite flagged the posterior as unreliable: the rank-normalized folded $\hat{R}^{\wedge}$ for parameter bb was 1.11, exceeding the 1.05 threshold, and the tail effective sample size for a3a3 was 213, below the 400 threshold. Under cloud-late, both suites indicated convergence: the maximum traditional $\hat{R}^{\wedge}$ was 1.03, the rank-normalized folded $\hat{R}^{\wedge}$ was 1.01, and the minimum tail effective sample size was 480.

The tier-conditioned stability diagnostic revealed that the posterior under edge-early was not stable across defensible variations in the transformation parameters. When the imputation method was varied among linear interpolation, last-observation-carried-forward, and multiple imputation, the posterior mean for bb varied by 0.037 under edge-early, compared to 0.014 under cloud-late. The 90% credible interval width varied by a factor of 1.6 under edge-early, compared to 1.2 under cloud-late. This indicates that the edge-early posterior depends on processing choices that are not determined by the data, and that this dependence is substantially reduced when the transformations are applied at the cloud tier after aggregation.

Table 4 summarizes the diagnostic verdicts for the empirical case study under both configurations. The pattern mirrors the simulation: the conventional suite fails to distinguish the configurations, while the proposed suite detects the reliability erosion under edge-early.

Table 4. Diagnostic verdicts for the empirical case study.

Diagnostic	Edge-early	Cloud-late
Max traditional R^2	1.08	1.03
Max rank-normalized folded R^2	1.11 (bb)	1.01
Min bulk ESS	520 (bb)	890 (bb)
Min tail ESS	213 (a3a3)	480 (a3a3)
Posterior mean variation of bb across imputation methods	0.037	0.014
CI width variation factor across imputation methods	1.6	1.2

4.4 Sensitivity to Transformation Type and Magnitude

The results reported above aggregate across the four transformations. A decomposition by transformation type reveals that the tier-placement effect is not uniform. Temporal aggregation and missing-value imputation produced the largest differences between configurations, because both interact directly with the autocorrelation structure that the inference model relies on to identify the mass time constant. Quantization produced smaller differences, because quantization error is approximately linear for small step sizes and the aggregation operation averages it out. Outlier handling produced intermediate differences, with the effect depending on the temporal clustering of outliers: when outliers were clustered, edge-early removal altered the local autocorrelation more than cloud-late removal, whereas when outliers were isolated, the difference was small.

The magnitude of the tier-placement effect also depended on the missingness rate and the aggregation window. At a missingness rate of 2%, the differences between configurations were small and neither the conventional nor the proposed suite distinguished them. At 8%, the differences were substantial and the proposed suite flagged edge-early reliably. At 15%, both configurations exhibited convergence problems, though the problems were more severe under edge-early. This dose-response relationship supports the interpretation that the effect is causal and that it scales with the extent of the transformation.

A further sensitivity analysis examined the effect of model specification. The two-parameter thermal model is parsimonious relative to the models used in full building energy calibration, which may have dozens of parameters. It might be expected that the tier-placement effect would be more severe in higher-dimensional models, where non-identifiability is more acute and the data are more strained. This expectation is consistent with the building stock calibration literature, which has documented convergence failures and physically implausible posteriors

in multi-parameter settings (IBPSA, 2025). The present results should therefore be interpreted as a lower bound on the severity of the problem.

4.5 Tier-Aware Convergence Assurance Protocol

The experimental results motivated the development of a tier-aware convergence assurance protocol, which specifies diagnostic requirements at each architectural boundary rather than only at the terminal output. The protocol comprises three checks. The first is an edge-tier data quality check, which verifies that the transformations applied at the edge do not introduce distributional artifacts that are detectable in the high-frequency series before aggregation. This check includes tests for altered autocorrelation, truncated tails, and anomalous quantization patterns. The second is a fog-tier aggregation check, which verifies that the aggregation operation preserves the temporal structure that the inference model will rely on, including tests for aliasing, smoothing of informative peaks, and misalignment of timestamps. The third is a cloud-tier inference check, which applies the full proposed diagnostic suite to the posterior and additionally computes the tier-conditioned stability diagnostic by comparing posteriors across defensible tier configurations.

The protocol was evaluated by applying it to the empirical case study data and assessing whether it detected the reliability erosion under edge-early that the conventional terminal check missed. The protocol flagged the edge-early configuration at the edge-tier check, because the pre-aggregation imputation altered the autocorrelation of the high-frequency series in a way that the check detected. It also flagged the configuration at the cloud-tier inference check, through the rank-normalized folded $R^{\wedge}R^{\wedge}$ and the stability diagnostic. The conventional terminal check flagged neither. This demonstrates that the tier-aware protocol provides detection capability that terminal validation lacks, and that the detection can occur early in the pipeline, before the inference is even performed, if the edge-tier check is implemented.

5. Discussion of Findings

5.1 Tier Placement as a Causal Determinant of Convergence Reliability

The central finding of this study is that the tier at which a data transformation is applied is a causal determinant of MCMC convergence reliability in building energy models, even when the transformation operation and the final analytical dataset are held constant. The edge-early configuration, in which transformations operate on high-frequency data before aggregation, produced posteriors with lower effective sample size, higher rank-normalized folded $R^{\wedge}R^{\wedge}$, greater bias, and worse predictive calibration than the cloud-late configuration, in which the same transformations operate on aggregated data. The conventional diagnostic suite failed to detect this difference, certifying convergence in both configurations and providing no indication that tier placement mattered.

This finding has a methodological implication that extends beyond the specific transformations examined here. It suggests that the architectural decomposition of edge-to-cloud pipelines is not a neutral implementation detail from the perspective of probabilistic inference. The tier boundaries are points at which the statistical properties of the data can be altered in ways that

affect the likelihood and therefore the posterior. The order of operations relative to aggregation is not commutative for the purposes of inference, even when it is approximately commutative for the purposes of the final dataset values. This is because the inference model does not see only the final values; it sees the temporal structure, the autocorrelation, and the tail behavior of the series, and these properties depend on the path by which the values were produced.

The mechanism is traceable to the interaction between transformation and aggregation. Linear interpolation of missing high-frequency readings increases the autocorrelation of the high-frequency series, and this increase is partially preserved through aggregation. The inference model, which uses autocorrelation to identify the mass time constant, attributes the increased autocorrelation to a longer time constant, producing an upward bias in bb . Outlier removal at the edge tier attenuates the tails of the high-frequency series, and this attenuation propagates through aggregation to produce a posterior with different tail behavior than outlier removal at the cloud tier. Quantization at the edge tier introduces errors that are averaged out by aggregation, whereas quantization at the cloud tier introduces errors directly into the analytical dataset. Each transformation has a characteristic interaction with aggregation, and the tier at which it is applied determines whether that interaction occurs.

5.2 The Blindness of Conventional Diagnostics to Architectural Effects

The finding that the traditional R^2 failed to distinguish the edge-early and cloud-late configurations is consistent with the known limitations of the statistic. The traditional R^2 compares between-chain and within-chain variance and is designed to detect non-convergence when chains have not mixed. It is not designed to detect differences in the tail behavior or scale of the posterior that arise from data processing, because such differences affect all chains similarly and therefore do not produce the between-chain variance that the statistic detects. The rank-normalized folded R^2 , by contrast, is sensitive to differences in scale and tail behavior between chains, and the tail effective sample size is sensitive to the precision of quantile estimates in the tails. These diagnostics detected the tier-placement effect because they were designed to detect the kinds of pathologies that tier-placement induces.

This result has direct implications for diagnostic practice in the building performance community. The reliance on the traditional R^2 with a 1.1 threshold, which remains common in published calibration studies, provides false assurance when the reliability erosion arises from data processing rather than sampler pathology. The improved diagnostics—rank-normalized and folded R^2 , bulk and tail effective sample size—are implemented in standard software and are no more difficult to compute than the traditional statistic. Their adoption should be considered a minimum standard for any Bayesian inference reported in the building performance literature, particularly when the data have passed through a multi-tier pipeline whose processing history is not fully known to the analyst.

The tail effective sample size is particularly relevant for built-environment applications where predictive intervals are used for risk assessment. If the tail effective sample size is low, the quantile estimates that define the interval are imprecise, and the interval may not have the coverage that its nominal level suggests. The present study found tail effective sample size

values below 400 under edge-early, indicating that the 90% intervals were not reliably estimated even if the posterior were correct. This is a separate problem from the bias and overconfidence discussed above; it is a problem of Monte Carlo precision in the tails, which is often overlooked because the bulk effective sample size and the traditional $\hat{R}^2$ appear adequate.

5.3 Architectural Implications: Designing for Inference Reliability

The findings suggest that the design of edge-to-cloud pipelines for probabilistic building analytics should be informed by inference requirements, not only by operational metrics. The current practice of optimizing tier decomposition for latency, bandwidth, and cost, and then validating the inference output with terminal diagnostics, is insufficient because it treats the architectural decomposition as neutral with respect to inference. The present study demonstrates that it is not neutral, and that the tier at which transformations occur can determine whether the inference is reliable.

A first architectural implication is that transformations that interact strongly with aggregation—imputation and outlier handling in particular—should be applied at the tier where their interaction with aggregation is minimized, or their effects should be modeled explicitly by the inference procedure. In the present study, cloud-late placement was superior for these transformations because it avoided the pre-aggregation interaction. This is not a universal prescription; for other transformations or other aggregation schemes, a different placement might be preferable. The point is that the placement decision should be made with reference to its inferential consequences, not only its operational ones.

A second architectural implication is that provenance metadata should be preserved across tier boundaries, so that the inference procedure can account for the transformations that were applied and the tier at which they occurred. The present study did not implement full provenance-aware inference, but the tier-conditioned stability diagnostic demonstrated that the inference is sensitive to transformation choices that are not determined by the data. If the inference model knew which transformations had been applied and where, it could model the transformation process explicitly, as in the provenance-aware configuration examined in related work. This would require the pipeline to expose transformation metadata in a machine-readable form that the inference procedure can consume, which is a relatively low-cost architectural change with substantial reliability benefits.

A third architectural implication is that convergence assurance should be distributed across tiers rather than concentrated at the terminal output. The tier-aware protocol proposed in this paper specifies checks at the edge, fog, and cloud boundaries, and the evaluation demonstrated that the edge-tier check alone could detect the reliability erosion under edge-early before the inference was performed. Distributed assurance is more efficient than terminal validation because it detects problems early, and it is more sensitive because it can examine the intermediate representations that terminal validation never sees. The protocol does not require sophisticated infrastructure; it requires only that the pipeline be instrumented to expose intermediate data and that the checks be designed with reference to the inference requirements.

5.4 Limitations and Boundary Conditions

Several limitations qualify the findings. The study is based on a single building and a single thermal model, which limits the generalizability of the specific numerical results. The simulation component mitigates this limitation by providing ground truth and demonstrating the mechanisms at work across 50 replications, but the empirical component would be strengthened by replication across multiple buildings with different characteristics and data quality profiles. The building studied has a moderate missingness rate and a relatively simple HVAC configuration; buildings with more complex systems or more severe data quality problems might exhibit different patterns, possibly with more severe pathologies.

The inference model is a two-parameter thermal network, which is parsimonious by the standards of building performance simulation. Higher-dimensional models might exhibit stronger or weaker sensitivity to tier placement depending on the structure of the parameter interactions. The expectation that higher-dimensional models would be more vulnerable is plausible but not directly tested here. Similarly, the transformations examined are representative but not exhaustive; other processing choices—smoothing, resampling, feature extraction—could have different effects on inference. The general principle that tier placement can alter inference-relevant properties is expected to hold, but the specific mechanisms and magnitudes would differ.

The study focuses on MCMC-based Bayesian inference and does not address variational inference, which is increasingly used for scalable probabilistic modeling in buildings. Variational inference has different convergence considerations—the optimization objective provides a natural diagnostic, but the approximation may be poor in ways that the objective does not reveal—and the interaction with tier placement could manifest differently. This is a direction for future work rather than a limitation of the present findings, which are specifically about MCMC.

Finally, the study does not address the organizational and governance dimensions of pipeline design and inference validation. In practice, data engineers and building scientists often work in separate teams with different priorities and incentives. The tier-aware protocol requires coordination between these teams—the pipeline must expose intermediate data and provenance metadata, and the inference procedure must be designed to consume it. This coordination is a social and organizational challenge as much as a technical one, and the present study, by demonstrating the technical value of the integration, provides an argument for investing in the coordination but does not itself solve the organizational problem.

5.5 Implications for Reliable Probabilistic AI in the Built Environment

The broader implication of this work is that the reliability of probabilistic AI in built-environment systems cannot be achieved by improving models or diagnostics alone. The architectural decomposition of edge-to-cloud pipelines shapes the data that reach the model, and the tier at which transformations occur can determine whether the inference is reliable. Reliability is a systems property that spans the architecture, the data processing, and the inference procedure, and it requires system-level design.

This perspective suggests a research agenda that is more integrated than the current division of labor between IoT architecture and statistical modeling. It suggests that architectural specifications for building analytics should include inference requirements—that transformations be placed with reference to their inferential consequences, that provenance be preserved across tiers, and that diagnostics be distributed across tier boundaries. It suggests that diagnostic suites should be designed with sensitivity to architectural effects, not only sampler pathologies, and that stability across tier configurations should be a routine part of model validation. And it suggests that evaluation metrics should include calibration and coverage, not only point-prediction accuracy and proper scoring rules, because the latter can reward sharpness at the expense of honesty.

For practitioners, the immediate recommendation is to adopt the improved convergence diagnostics—rank-normalized and folded $\hat{R}^2$, bulk and tail effective sample size, and Monte Carlo standard error—as a minimum standard for any Bayesian inference reported in the building performance literature. The second recommendation is to document the tier at which transformations are applied and to assess the sensitivity of posteriors to defensible variations in tier configuration. The third recommendation is to implement convergence assurance at tier boundaries rather than only at the terminal output, so that reliability erosion is detected early and attributed to its architectural source. These recommendations are not onerous, but they represent a shift from the current practice of treating architecture and inference as separate concerns.

6. Conclusion

This paper has examined whether the tier at which data transformations are applied in a three-layer edge-to-cloud architecture affects the convergence reliability of MCMC inference in building energy models. The answer is affirmative, and the effect is substantial: transformations applied at the edge tier before aggregation produced posteriors with lower effective sample size, higher rank-normalized folded $\hat{R}^2$, greater parameter bias, and worse predictive calibration than the same transformations applied at the cloud tier after aggregation, even when the final analytical dataset was identical in content. The conventional diagnostic suite based on the traditional $\hat{R}^2$ failed to detect this difference, certifying convergence in both configurations, while the proposed suite based on rank-normalized, folded, and tail-sensitive diagnostics detected it reliably across simulation replications and in the empirical case study.

The findings contribute to the literature in three ways. First, they establish tier placement as a causal determinant of convergence reliability, identifying the interaction between transformation and aggregation as the mechanism through which architectural decisions affect inference. Second, they demonstrate that conventional diagnostics are blind to architectural effects, providing empirical grounding for the adoption of improved diagnostics in the building performance community. Third, they propose and evaluate a tier-aware convergence assurance protocol that distributes diagnostic checks across architectural boundaries and detects reliability erosion that terminal validation misses.

The limitations of the study point toward future work. Replication across multiple buildings, climates, and model structures would establish the generality of the observed effects. Extension to higher-dimensional models and alternative inference algorithms—variational inference, sequential Monte Carlo, distributed MCMC—would test whether the tier-placement effect operates differently in other settings. Development of automated provenance-aware inference procedures, in which the tier at which transformations occurred is not merely documented but is used to construct appropriate observation models automatically, would translate the research findings into practical tools. And investigation of the organizational and governance conditions under which tier-aware assurance practices can be implemented would address the human dimension of the problem, which technical solutions alone cannot resolve.

The built environment is increasingly instrumented, and the probabilistic models that interpret its data are increasingly consequential. The reliability of these models depends not only on their mathematical sophistication but on the entire architectural system through which data become inference. The tier at which a transformation occurs is not a neutral implementation detail; it is a determinant of whether the inference can be trusted. Recognizing and designing for this architectural dimension of reliability is a condition for trustworthy probabilistic AI in the places where people live and work.

References

1. Al-Masri, E., Kalyanam, K. R., Batts, J., Kim, J., Singh, S., Vo, T., & Yan, C. (2022). Investigating messaging protocols for the Internet of Things (IoT). *IEEE Access*, 8, 94880–94911.
2. Barvaliya, K. (2025, November). Cloud-tiered Pipelines for Built-Environment Data: A Visual Programming Stack on Scalable Three-Layer Services. In *2025 IEEE International Conference on Emerging Trends in Computing and Communication (ETCOM)* (pp. 1-5). IEEE. <https://doi.org/10.1109/ETCOM66606.2025.11436161>
3. Booth, A. T., Choudhary, R., & Spiegelhalter, D. J. (2012). A hierarchical Bayesian framework for calibrating micro-level models with macro-level data. *Journal of Building Performance Simulation*, 5(4), 213–232.
4. Bordel, B., Alcarria, R., & Robles, T. (2020). Data quality management in the Internet of Things: A survey. *Sensors*, 20(18), 5326.
5. Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4), 434–455.
6. Bürkner, P.-C., Gabry, J., Kay, M., & Vehtari, A. (2025). *posterior: Tools for working with posterior distributions* (R package version 1.6.1) [Computer software]. <https://mc-stan.org/posterior/>

7. Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32.
8. Chong, A., & Menberg, K. (2018). Guidelines for the Bayesian calibration of building energy models. *Energy and Buildings*, 174, 527–547.
9. Coakley, D., Raftery, P., & Keane, M. (2014). A review of methods to match building energy simulation models to measured data. *Renewable and Sustainable Energy Reviews*, 37, 123–141.
10. Drgoňa, J., Arroyo, J., Cupeiro Figueroa, I., Blum, D., Arendt, K., Kim, D., Perarnau Ollé, E., Oravec, J., Wetter, M., Vrabie, D. L., & Helsén, L. (2020). All you need to know about model predictive control for buildings. *Annual Reviews in Control*, 50, 190–232.
11. Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472.
12. IBPSA. (2025). Limitations of conventional Bayesian calibration for building stock energy modelling: Insights from Japanese commercial building archetypes. *Proceedings of Building Simulation 2025*.
13. Karkouch, A., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Data quality in the Internet of Things: A state of the art. *Journal of Network and Computer Applications*, 73, 57–72.
14. Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press.
15. Lambert, B., & Vehtari, A. (2022). Rstar: A robust MCMC convergence diagnostic with uncertainty using decision tree classifiers. *Bayesian Analysis*, 17(2), 353–379.
16. Mahotiedu, J. (2026). Streaming IoT ETL pipeline with Kafka, Spark Structured Streaming, and medallion architecture. GitHub repository.
17. Makowski, D., Ben-Shachar, M. S., & Lüdtke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, 4(40), 1541.
18. Margossian, C. C., Hoffman, M. D., Sountsov, P., Riou-Durand, L., Vehtari, A., & Gelman, A. (2024). Nested R[^]R[^]: Assessing the convergence of Markov chain Monte Carlo when running many short chains. *Bayesian Analysis*. <https://doi.org/10.1214/24-BA1453>
19. Stan Development Team. (2024). *Stan reference manual* (Version 2.35) [Computer software]. <https://mc-stan.org>

20. Teh, H. Y., Kempa-Liehr, A. W., & Wang, K. I.-K. (2020). Sensor data quality: A systematic review. *Journal of Big Data*, 7(1), 11.
21. Tian, W., Heo, Y., de Wilde, P., Li, Z., Yan, D., Park, C. S., Feng, X., & Augenbroe, G. (2018). A review of uncertainty analysis in building energy assessment. *Renewable and Sustainable Energy Reviews*, 93, 285–301.
22. ASHRAE. (2002). *Guideline 14: Measurement of energy and demand savings*. American Society of Heating, Refrigerating and Air-Conditioning Engineers.
23. Bacher, P., & Madsen, H. (2011). Identifying suitable models for the heat dynamics of buildings. *Energy and Buildings*, 43(7), 1511–1522.
24. Barvaliya, K. (2025, November). Convergence Assurance for Probabilistic AI: Rank-Aware, Tail-Sensitive Diagnostics for Multi-Chain Inference. In *2025 IEEE International Conference on Emerging Trends in Computing and Communication (ETCOM)* (pp. 1-4). IEEE. <https://doi.org/10.1109/ETCOM66606.2025.11436810>
25. Tsakanikas, T. (2025). Realtime telemetry pipeline: Cloud-native streaming architecture with data quality monitoring. GitHub repository.
26. Vats, D., & Knudson, C. (2018). Revisiting the Gelman-Rubin diagnostic. *arXiv preprint arXiv:1812.09384*.
27. Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2019). Rank-normalization, folding, and localization: An improved $\hat{R}^2$ for assessing convergence of MCMC. *Bayesian Analysis*, 16(2), 667–718.
28. Wang, S., & Xu, X. (2006). Simplified building model for transient thermal performance estimation using GA-based parameter identification. *International Journal of Thermal Sciences*, 45(4), 419–432.
29. Zitzmann, S., & Hecht, M. (2019). Are the effective sample size and the potential scale reduction factor sufficient for assessing convergence? *Psychological Methods*, 24(4), 513–534.