



ARTICLE

Automated Probabilistic Mechanistic Modeler: A Self-Auditing Bayesian Workflow for Epidemic Forecasting with Layered Uncertainty Quantification

Emmanuel Addai^{*}, Zi Wang, Devika Ambekar and Xiaowei Xu^{*}

Department of Information Science, University of Arkansas at Little Rock, Little Rock, AR, USA

^{*}Corresponding Authors: Emmanuel Addai. Email: eaddail@ualr.edu or papayawewit@gmail.com; Xiaowei Xu. Email: xwxu@ualr.edu

Received: 13 August 2026; Accepted: 15 September 2026; Published: 20 October 2026

ABSTRACT: Mechanistic epidemic models can converge and reproduce observed data while still producing poorly calibrated forecasts or relying on weakly identifiable assumptions. These risks become more consequential when model construction, inference, validation, and reporting are embedded within automated scientific workflows. We present the Automated Probabilistic Mechanistic Modeler (APMM), a self-auditing Bayesian workflow for mechanistic epidemic forecasting that integrates model specification, probabilistic inference, diagnostic evaluation, structural comparison, uncertainty quantification, stress testing, and human oversight within a checkpointed state graph. The current implementation uses deterministic decision rules to route model outputs through predefined validation gates, while versioned scientific tools generate all numerical quantities rather than the workflow controller. APMM distinguishes four sources of uncertainty—parametric, observational, structural, and scenario uncertainty—and evaluates predictive adequacy using convergence diagnostics, proper scoring rules, empirical interval coverage, and rolling-origin forecasting. We demonstrate the workflow as a proof of concept using the Centers for Disease Control and Prevention (CDC) Influenza Hospitalization Surveillance Network (FluSurv-NET) data for the 2023–24 U.S. influenza season. The initial Susceptible–Exposed–Infectious–Recovered (SEIR) specification converged computationally but exhibited severe post-peak undercoverage, with only 0.12 empirical coverage for a nominal 80% predictive interval. The automated revision mechanism subsequently introduced an endemic baseline in the observation model, eliminating the severe undercoverage, although the resulting 1.00 coverage is interpreted cautiously and does not establish exact nominal calibration. Sensitivity and scenario analyses further demonstrate how the workflow records parameter influence and counterfactual transmission effects as separate uncertainty components. The results show that APMM can operationalize diagnostic failures as explicit revision or escalation decisions while maintaining transparent provenance and reproducibility. The present study establishes the workflow architecture rather than claiming state-of-the-art influenza forecasting performance, and provides a foundation for broader cross-season validation, richer model libraries, and future large language model (LLM)-assisted scientific control.

KEYWORDS: Mechanistic epidemic forecasting; Bayesian inference; uncertainty quantification; self-auditing systems; influenza hospitalization; CRPS; model calibration

1 Introduction

Mechanistic epidemic models remain important tools for understanding infectious-disease dynamics and supporting public-health planning. Compartmental models such as susceptible–exposed–infectious–recovered (SEIR) systems provide interpretable representations of transmission, latent infection, recovery,

and intervention effects, and they continue to inform forecasting, hospital-capacity planning, and outbreak-response analysis [1]. Their usefulness, however, depends not only on how well they reproduce observed data but also on whether their assumptions, uncertainty estimates, and out-of-sample predictions are sufficiently reliable.

A central difficulty is that a mechanistic model can appear successful while still being inadequate for forecasting. Bayesian inference may converge numerically, posterior trajectories may closely follow the observed epidemic curve, and point-prediction errors may be small, yet the resulting predictive distribution can remain poorly calibrated or depend strongly on weakly identifiable parameters [1–4]. Structural and practical identifiability studies have shown that several parameter combinations in epidemic models cannot always be recovered reliably from limited surveillance data [4].

Probabilistic epidemic forecasting has developed several methods for addressing these problems. Proper scoring rules such as the continuous ranked probability score (CRPS), empirical interval coverage, posterior predictive checks, rolling-origin evaluation, model comparison, and sensitivity analysis provide complementary information about predictive performance and uncertainty [1,5]. Proper scoring rules are particularly valuable because they assess the full predictive distribution rather than only a point estimate [5]. Bayesian model comparison and stacking can further quantify relative predictive support among candidate specifications, while sensitivity analysis can identify which assumptions or parameters most strongly influence quantities of interest. These methods are valuable, but in many conventional workflows they remain diagnostic procedures applied after model fitting. A model may therefore continue through analysis and reporting unless an analyst manually interprets the diagnostics and decides whether revision is necessary.

The increasing automation of scientific workflows creates a related governance problem. Workflow frameworks can coordinate data ingestion, model execution, inference, diagnostics, and reporting, while recent large-language-model (LLM)-based systems have extended automation toward planning, experiment design, tool selection, and scientific communication [6–10]. At the same time, LLM-based systems remain vulnerable to unsupported tool use, hallucinated outputs, and unverified quantitative claims [11,12]. Such systems create opportunities for more efficient scientific modeling, but they also make provenance and quantitative control increasingly important. In particular, numerical quantities reported by an automated system should remain traceable to validated computational procedures, and evidence of model inadequacy should be capable of changing the workflow rather than being recorded only as a warning. The present work therefore focuses on controlled scientific automation, rather than autonomous scientific reasoning.

To address this need, we introduce the Automated Probabilistic Mechanistic Modeler (APMM), a self-auditing Bayesian workflow for mechanistic epidemic forecasting. APMM organizes model specification, prior-predictive checking, Bayesian calibration, convergence assessment, forecast evaluation, structural comparison, sensitivity analysis, scenario analysis, revision, and reporting within a checkpointed state graph. The current implementation uses deterministic and prespecified decision policies. Diagnostic outputs are evaluated against defined decision criteria, and failed checks can route the workflow back to model specification or recalibration, while repeated failure triggers escalation to a human reviewer. Thus, the demonstrated system should be understood as an automated probabilistic modeling workflow rather than as a fully autonomous or LLM-driven scientific agent.

APMM also treats uncertainty as a collection of distinct sources rather than as a single undifferentiated predictive band. The workflow separates parametric uncertainty, arising from uncertainty in inferred model parameters; observational uncertainty, arising from the relationship between the latent epidemic process and reported surveillance measurements; structural uncertainty, arising from dependence on the selected model specification; and scenario uncertainty, arising from explicit counterfactual assumptions about future transmission conditions. Each uncertainty component is linked to the computational procedure from which

it is obtained. This layered representation is intended to make uncertainty statements easier to interpret and audit and to avoid implying that a single posterior interval captures every important source of uncertainty.

The primary methodological contribution of APM is therefore not a new epidemic compartmental model, Bayesian sampler, diagnostic statistic, or autonomous reasoning algorithm. Instead, the contribution is the integration of existing scientific components into a gated and reproducible forecasting workflow in which diagnostic evidence has operational consequences. A failed convergence or forecast-adequacy assessment can prevent automatic reporting and initiate revision; structural alternatives can be compared within the same workflow; assumptions and routing decisions are retained in an audit record; and unresolved failures can be escalated for human review. In this sense, self-auditing refers specifically to the ability of the workflow to compute predefined adequacy checks, use their results to modify its execution path, and preserve a traceable record of those decisions.

The framework is evaluated using U.S. influenza hospitalization surveillance data from the Centers for Disease Control and Prevention (CDC) FluSurv-NET. The initial proof-of-concept experiment demonstrates a specific failure mode: an SEIR specification can converge satisfactorily while producing severely under-covered predictive intervals in the post-peak period. The automated workflow detects this inadequacy and invokes a revised observation specification before reporting. Importantly, elimination of severe undercoverage is not interpreted as evidence of exact probabilistic calibration. In particular, empirical coverage of 1.00 for a nominal 80% predictive interval may indicate conservative or excessively diffuse uncertainty. Forecast adequacy must therefore be assessed jointly through proper scoring rules, interval calibration, predictive sharpness, and rolling-origin evaluation [1,5]. The purpose of the empirical demonstration is consequently to assess whether the workflow can identify, route, record, and respond to model deficiencies, rather than to claim state-of-the-art influenza forecasting performance.

The principal contributions of this work are summarized as follows:

1. **A self-auditing automated Bayesian workflow.** APM integrates model specification, probabilistic calibration, diagnostic evaluation, model comparison, stress testing, revision, and human escalation within a checkpointed state graph in which diagnostic failures can alter the execution path.
2. **Layered and traceable uncertainty quantification.** Parametric, observational, structural, and scenario uncertainty are represented separately and connected to the computational procedures from which they arise.
3. **Operational forecast validation.** Convergence diagnostics, proper probabilistic scoring, empirical calibration assessment, rolling-origin evaluation, structural comparison, and sensitivity analysis are treated as workflow decision evidence rather than solely as post-hoc summaries.
4. **Explicit provenance and controlled computation.** Numerical scientific quantities are generated by versioned computational tools, while the workflow-control layer evaluates their structured outputs and records routing decisions, providing a reproducible foundation for future LLM-assisted extensions.

The remainder of the paper is organized as follows. [Section 2](#) reviews related work on automated scientific workflows, Bayesian mechanistic modeling, probabilistic forecast evaluation, structural uncertainty, and model governance. [Section 3](#) presents the APM workflow architecture and its checkpointed control structure. [Section 4](#) describes the mechanistic models, Bayesian inference procedure, uncertainty decomposition, forecast-evaluation criteria, model comparison, and stress-testing protocol. [Section 5](#) reports the empirical results. [Section 6](#) discusses the implications, limitations, and extensions of the framework, including broader cross-season validation, richer model libraries, and future LLM-assisted control policies. [Section 7](#) concludes the paper.

2 Related Work

2.1 Automated Scientific Workflows and Bayesian Epidemic Modeling

Scientific computing increasingly relies on automated workflows that coordinate data ingestion, model execution, parameter estimation, diagnostic evaluation, comparison, and reporting. Workflow orchestration frameworks provide mechanisms for representing these processes as directed graphs or state machines with persistent state, conditional routing, checkpointing, and tool execution [9,10]. Such mechanisms are well established in software engineering and workflow automation and are not, by themselves, a contribution of the present work. Mechanistic epidemic models remain central to infectious-disease analysis because they encode assumptions about transmission, susceptibility, progression, recovery, and intervention effects as interpretable dynamical systems. Compartmental models derived from the classical SIR framework provide the foundation for widely used extensions such as SEIR, SEIRS, age-structured, stochastic, and time-varying transmission models [13].

Bayesian inference provides a natural framework for calibrating such models because uncertainty in parameters, latent states, observation processes, and forecasts can be propagated through the model [14,15]. Modern probabilistic programming libraries, including NumPyro, facilitate posterior inference for dynamical systems using algorithms such as Hamiltonian Monte Carlo and the No-U-Turn Sampler (NUTS) [14,15]. Diagnostic tools such as ArviZ further support posterior summarization, posterior predictive checking, effective-sample-size assessment, and convergence diagnostics [16]. Rank-normalized $\hat{R}$ provides an additional criterion for assessing whether multiple Markov chains have mixed adequately [17].

Successful Bayesian computation, however, does not establish that an epidemiological model is scientifically adequate. A model may achieve apparently satisfactory convergence while remaining structurally or practically non-identifiable, misspecified, or unreliable for out-of-sample prediction [18,19]. Structural and practical identifiability analyses demonstrate that several epidemic model parameters cannot always be uniquely or precisely estimated from typical surveillance data [18,19]. Consequently, Bayesian calibration should be treated as one component of a broader model-evaluation process rather than as sufficient evidence that a forecast should be accepted.

2.2 Probabilistic Forecast Evaluation and Uncertainty

Evaluation of probabilistic epidemic forecasts requires more than point-prediction accuracy. A predictive distribution should ideally balance calibration with sharpness: observations should be statistically consistent with the predictive distribution, while prediction intervals should remain as informative as possible. Proper scoring rules provide a principled means of evaluating these properties because they reward probabilistic forecasts for assigning appropriate probability to observed outcomes [5].

Large-scale epidemic forecasting studies have emphasized evaluation across multiple forecast origins, seasons, and models rather than a single terminal hold-out period [20,21]. Rolling-origin evaluation is particularly useful because a model is repeatedly fitted using only information available up to a given time and then evaluated on subsequent observations. This allows forecast performance to be examined during different phases of an epidemic, including growth, peak, decline, and possible secondary waves.

These considerations are directly relevant to automated forecasting workflows. A model that fails probabilistic evaluation should not automatically proceed to reporting merely because its posterior sampler converged. In APMM, forecast diagnostics are therefore intended to act as workflow evidence: poor predictive performance can trigger model revision, reconsideration, or human escalation. The empirical coverage is interpreted together with proper scoring and predictive sharpness rather than as an isolated acceptance criterion.

Uncertainty in mechanistic epidemic forecasting extends beyond posterior uncertainty in fitted parameters. Different mathematical structures may provide plausible explanations of the same observations while implying different future trajectories. Such structural uncertainty can arise from assumptions regarding waning immunity, time-varying transmission, population heterogeneity, observation processes, reporting delays, or the number of epidemic waves.

Predictive model-comparison methods provide one way to assess relative support among competing structures. Leave-one-out cross-validation and related information-based approaches estimate out-of-sample predictive performance [22]. Bayesian predictive stacking can then construct an ensemble or summarize the relative predictive contribution of candidate models [23]. Importantly, stacking weights should be interpreted as relative predictive support within the candidate set, not as independent evidence that any particular model is scientifically correct. Their interpretation therefore depends strongly on the diversity and adequacy of the candidate model library.

This distinction is important for mechanistic forecasting. Comparison between only minor variants of the same SEIR structure provides limited evidence about structural uncertainty. A stronger evaluation requires genuinely different candidate mechanisms, such as SEIR and SEIRS structures, time-varying transmission models, alternative observation models, age-structured systems, or multi-wave/sub-epidemic formulations. The APMM framework therefore treats model comparison as conditional on the set of available candidate structures rather than as proof of structural validity.

2.3 Position of the Present Work

The present work lies at the intersection of automated scientific workflow orchestration, Bayesian mechanistic epidemic modeling, probabilistic forecast evaluation, structural model comparison, and uncertainty quantification. Existing research provides mature methods for each of these individual tasks. Graph-based workflow frameworks coordinate computational components [9,10]; probabilistic programming systems perform Bayesian inference [14,15]; convergence and identifiability diagnostics evaluate inference quality [17–19]; proper scoring and rolling-origin evaluation assess probabilistic forecasts [5,20,21]; and model comparison and sensitivity analysis characterize structural dependence and model response [22–24].

APMM does not claim novelty for these individual statistical methods, for graph-based workflow orchestration itself, or for the separation of software components into routing and numerical functions. Instead, the contribution is their integration into a self-auditing, checkpointed Bayesian forecasting workflow in which quantitative diagnostics have explicit operational consequences. A failed diagnostic can prevent automatic reporting, route the workflow toward a revised model specification, or trigger human escalation. Model assumptions, diagnostic outputs, routing decisions, and workflow states are retained as part of the audit record.

The present implementation is deliberately deterministic. Model selection and revision operate over a validated candidate library using prespecified decision rules; numerical quantities are produced by versioned scientific tools; and the workflow controller acts only on their structured outputs. The term self-auditing therefore refers to automated diagnostic evaluation, routing, and provenance rather than to autonomous scientific intelligence.

The methodological objective is consequently to investigate whether mechanistic epidemic forecasting can be organized as a reproducible process in which inadequate forecasts are identified before reporting, different sources of uncertainty remain distinguishable, alternative model structures can be compared, and unresolved failures are escalated rather than silently accepted. This architecture also provides a controlled foundation for future LLM-assisted extensions, but such extensions are outside the scope of the current implementation.

3 System Architecture

APMM is implemented as a checkpointed state-graph workflow operating on a shared, versioned modeling artifact (Fig. 1). The graph coordinates the progression from data ingestion and model specification through Bayesian calibration, diagnostic evaluation, model comparison, stress testing, rolling-origin evaluation, and final human review. The APMM implementation uses deterministic and prespecified decision policies. Each transition is therefore triggered by explicit outputs from validated scientific components rather than by autonomous reasoning or free-form model generation.

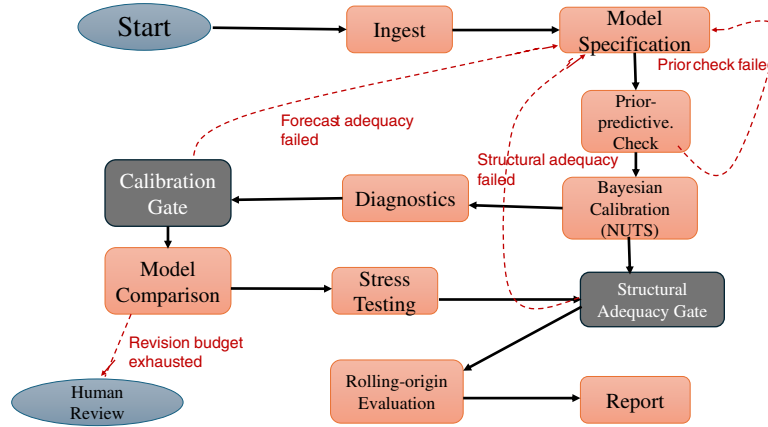


Figure 1: APMM automated self-auditing workflow. Solid edges indicate the primary computational sequence, while dashed edges indicate failure-triggered revision or escalation paths. The calibration and forecast gate evaluates convergence and predictive adequacy, while the structural adequacy gate evaluates relative model support and stress-test outcomes. Failed checks return the workflow to model specification for revision. If the predefined revision budget is exhausted, the workflow escalates to human review rather than automatically producing a final report.

3.1 Shared Workflow State

The workflow is organized around a shared state object that records the information required to reproduce and audit each modeling decision. The state contains the data bundle and its provenance identifier, the current model specification, references to posterior outputs, diagnostic summaries, forecast-evaluation metrics, model-comparison results, stress-test results, assumption records, the current routing decision, and the revision counter.

To keep the workflow state lightweight, computationally intensive objects such as posterior samples, posterior predictive draws, and large simulation outputs are stored in an external artifact registry. The shared state contains only compact references and summary statistics required for routing. This design reduces checkpointing overhead while preserving a traceable connection between every workflow decision and the scientific outputs on which it was based.

3.2 Self-Auditing Decision Process

The term self-auditing refers to the ability of the workflow to evaluate predefined adequacy criteria, use the resulting evidence to modify its execution path, and retain a traceable record of those decisions. APMM does not merely calculate diagnostics for later inspection; selected diagnostic outputs become explicit routing variables.

After Bayesian calibration, the Calibration and Forecast Gate evaluates both posterior sampling quality and predictive adequacy. Sampling quality is assessed using the maximum rank-normalized $\hat{R}$, effective

sample size, and the number of divergent transitions. Predictive adequacy is evaluated using the continuous ranked probability score (CRPS) and empirical coverage of the nominal 80% predictive interval. Coverage is not interpreted as an isolated acceptance criterion. In particular, complete empirical coverage is not treated as evidence of exact nominal calibration because it may also reflect conservative predictive uncertainty.

If the sampling or predictive-adequacy requirements are not satisfied, the workflow returns to the Model Specification Manager, which selects the next prespecified candidate from the validated model library. The candidate is then recalibrated and re-evaluated.

After a candidate satisfies these checks, the workflow considers relative predictive support from model comparison together with the stress-test results. Because the present candidate set contains closely related SEIR observation-model specifications, these comparisons are interpreted conditionally on the available model library rather than as evidence of comprehensive structural validity.

The revision process is bounded by a predefined maximum number of attempts. If the revision budget is exhausted without satisfying the required criteria, the workflow does not automatically produce a final forecast. Instead, the case is escalated to human review. This bounded-failure behavior prevents indefinite automated cycling and keeps unresolved model inadequacy visible in the audit record.

3.3 Auditability and Human Escalation

Every transition in the workflow is checkpointed. Each checkpoint records the component that executed, the relevant input state, the diagnostic or evaluation output, the resulting decision, and the updated revision counter. The complete sequence can therefore be replayed to determine why a particular model was selected, revised, rejected, or allowed to proceed.

An assumption ledger is maintained alongside the workflow state. The ledger records modeling choices such as fixed epidemiological periods, observation-model assumptions, candidate structures, prior specifications, calibration criteria, and scenario definitions. Consequently, the final report is accompanied not only by numerical forecast outputs but also by a traceable history of the assumptions and workflow decisions that produced them.

This auditability is central to the interpretation of APM. The system is not claimed to autonomously discover the scientifically correct model. Rather, it provides a reproducible mechanism for ensuring that predefined statistical and computational checks are performed consistently, that failed checks have operational consequences, and that unresolved deficiencies are escalated instead of being silently carried into the final report.

4 Methods

4.1 Data, Mechanistic Model, and Bayesian Inference

We use weekly laboratory-confirmed influenza hospitalization rates from CDC FluSurv-NET for the 2023–24 U.S. influenza season. The series contains 52 epidemiological weeks and reports hospitalization rates per 100,000 population, aggregated across age, sex, race, and influenza virus type. The study is intentionally a single-season proof-of-concept of the APM workflow rather than an assessment of cross-season forecasting generalizability. The final eight weeks are retained as the original post-peak diagnostic window, while additional within-season evaluation is performed using rolling forecast origins. At each origin t , only observations available through that week are used to generate one- to four-week-ahead forecasts, preserving temporal ordering [20,21].

The mechanistic core is a susceptible–exposed–infectious–recovered (SEIR) model [13]. Let S , E , I , and R denote the corresponding compartments in a population $N = 10^5$. With $\beta = R_0\gamma$, latent progression rate σ , and recovery rate γ ,

$$\dot{S} = -\frac{\beta SI}{N}, \quad \dot{E} = \frac{\beta SI}{N} - \sigma E, \quad (1)$$

$$\dot{I} = \sigma E - \gamma I, \quad \dot{R} = \gamma I. \quad (2)$$

Weekly incident infections are $\Delta C_t = (N - S_t) - (N - S_{t-1})$. Two prespecified observation specifications are evaluated. The first uses

$$\mu_t = \rho \Delta C_t, \quad (3)$$

where ρ is the combined hospitalization/reporting fraction. The second adds a non-negative background hospitalization component,

$$\mu_t = b + \rho \Delta C_t. \quad (4)$$

For both specifications, $y_t \sim \text{NegBinomial2}(\mu_t, \phi)$, where ϕ is the negative-binomial concentration parameter. The background term represents persistent low-level hospitalization observations after the mechanistic epidemic trajectory approaches zero; it is not an additional epidemic compartment.

Because a single hospitalization series does not reliably identify all mechanistic parameters [18,19], the latent and infectious periods are fixed at 1.5 and 3.5 days, respectively. The inferred quantities are R_0 , ρ , I_0 , ϕ , and b when applicable. The two candidates share the same SEIR dynamics and differ principally in the observation model; consequently, their comparison is interpreted as observation-model sensitivity rather than comprehensive mechanistic structural uncertainty.

Prior distributions are prespecified and screened using prior-predictive simulation. The prior distributions used for the inferred parameters are summarized in Table 1. The current implementation does not autonomously generate or modify priors.

Table 1: Prior distributions used in the Bayesian SEIR models.

Parameter	Meaning	Prior
R_0	Apparent reproduction number	TruncNormal(1.4, 0.4; [1.01, 3.5])
ρ	Hospitalization/reporting fraction	LogNormal(log 0.01, 1)
I_0	Initial infectious count	LogNormal(log 5, 1)
b	Background hospitalization rate	HalfNormal(0.5)
ϕ	NB concentration	Exponential(1/8)

Posterior inference is performed with the No-U-Turn Sampler (NUTS) [14] implemented in NumPyro [15]. Four chains are run with 1500 warm-up iterations and 1500 retained draws per chain, using a target acceptance probability of 0.95. The mechanistic ODE solver is evaluated within the probabilistic program so that parameter uncertainty propagates to posterior predictive trajectories. Sampling quality is assessed using ArviZ [16], including rank-normalized $\hat{R}$ [17], effective sample size, and divergent transitions. Random seeds, sampler settings, model specifications, and preprocessing choices are stored with each workflow run.

4.2 Automated Adequacy Assessment and Layered Uncertainty

APMM converts outputs from validated scientific modules into deterministic workflow-routing decisions. No LLM or autonomous scientific agent determines whether a model passes or fails in the present implementation. Posterior sampling quality and predictive adequacy are evaluated separately: sampling quality is assessed using rank-normalized $\widehat{R}$ and divergent transitions, whereas predictive adequacy is evaluated jointly using the continuous ranked probability score (CRPS) and empirical coverage of the nominal 80% predictive interval. The criteria in Table 2 serve different purposes and should not be interpreted as having identical statistical status. In particular, empirical coverage of 1.00 for a nominal 80% predictive interval is not treated as evidence of exact calibration because complete coverage may also reflect conservative predictive uncertainty. Similarly, stacking weights are interpreted only as relative predictive support within the restricted candidate set. The two-revision limit is a configurable engineering safeguard against indefinite workflow cycling rather than a universal statistical threshold. If the revision budget is exhausted without an adequate model state, automatic progression is terminated and the case is escalated for human review.

Table 2: Principal evidence used for APMM workflow decisions.

Stage	Evidence	Interpretation
Sampling	$\widehat{R} \leq 1.05$; 0 divergences	Sampling quality
Forecast	CRPS; 80% coverage	Accuracy/calibration
Comparison	Leave-One-Out Cross-Validation (LOO) + stacking	Relative model support
Control	Maximum 2 revisions	Human escalation

APMM organizes uncertainty into four traceable components. *Parametric uncertainty* is represented by posterior variation in R_0 , ρ , I_0 , ϕ , and b where applicable. *Observational uncertainty* arises from mapping latent infections to reported hospitalizations, including reporting scale, overdispersion, and the possible background hospitalization floor. *Structural uncertainty* is conditional on the two prespecified observation-model variants and is therefore interpreted narrowly as observation-model sensitivity rather than comprehensive mechanistic uncertainty. *Scenario uncertainty* is represented by explicit counterfactual changes in apparent transmissibility and is not assigned posterior probabilities. Keeping these components separate preserves the origin and interpretation of each uncertainty statement.

4.3 Model Comparison, Stress Testing, and Forecast Evaluation

The two observation-model specifications are compared using approximate leave-one-out cross-validation (LOO) [22] and Bayesian predictive stacking [23]. Because both candidates share the same mechanistic SEIR core, a large stacking weight is interpreted only as evidence favoring one observation specification over the other and not as validation of broad mechanistic structural uncertainty. No universal stacking-weight cutoff is used.

Global sensitivity analysis is performed with variance-based Sobol indices [24] using SALib [25]. First-order indices S_1 measure direct contributions to output variability, while total-effect indices S_T include interactions. The principal inputs are R_0 , ρ , and I_0 , and the analysis is interpreted as a stress-testing and consistency check rather than as evidence of novel SEIR dynamics. Scenario stress tests additionally scale apparent transmissibility over 0.8–1.5 times its fitted value. These curves are counterfactual trajectories and are not interpreted as probability-weighted forecasts.

Probabilistic forecast performance is assessed through rolling-origin evaluation within the same FluSurv-NET 2023–24 season. At each origin, the model is refitted using only information available at that time and generates forecasts for horizons $h = 1, \dots, 4$. Forecast accuracy is summarized using the continuous ranked probability score (CRPS) [5]; lower values indicate better probabilistic performance.

Predictive adequacy is evaluated using the continuous ranked probability score (CRPS) and empirical coverage of the nominal 80% predictive interval. These quantities are interpreted jointly rather than treating coverage as an isolated acceptance criterion. In particular, the increase in 80% empirical coverage from 0.12 to 1.00 after adding the background hospitalization component is interpreted as correction of severe post-peak undercoverage rather than proof of exact nominal calibration. Because complete empirical coverage can also arise from a conservative predictive distribution, the result is reported cautiously. The evaluation retains the CDC FluSurv-NET 2023–24 proof-of-concept while providing repeated within-season out-of-sample checks through rolling-origin evaluation. It does not establish cross-season, cross-region, or cross-pathogen generalizability.

5 Results

The CDC FluSurv-NET 2023–24 experiment is used as a proof-of-concept evaluation of whether APM can identify predictive inadequacy and alter workflow execution accordingly. The results are not intended to establish cross-season forecasting superiority.

5.1 Model Fit, Predictive Failure, and Revision

The deterministic SEIR core reproduced the broad timing and shape of the seasonal hospitalization wave, with $R^2 \approx 0.84$, and predicted the peak within approximately two weeks of the observed peak. However, it underestimated peak magnitude (approximately 6.2 vs. 8.9 hospitalizations per 100,000) and decayed more rapidly toward zero in the post-peak period. Preliminary unconstrained fitting also drove the latent-period parameter toward its admissible boundary, supporting the decision to fix the epidemiological periods because of weak practical identifiability.

The initial Bayesian observation specification, $\mu_t = \rho \Delta C_t$, showed satisfactory posterior sampling behavior but poor predictive uncertainty. The maximum rank-normalized $\widehat{R}$ was approximately 1.01 with no divergent transitions, yet empirical coverage of the nominal 80% predictive interval in the eight-week post-peak window was only 0.12. Thus, acceptable convergence did not imply adequate forecasting performance.

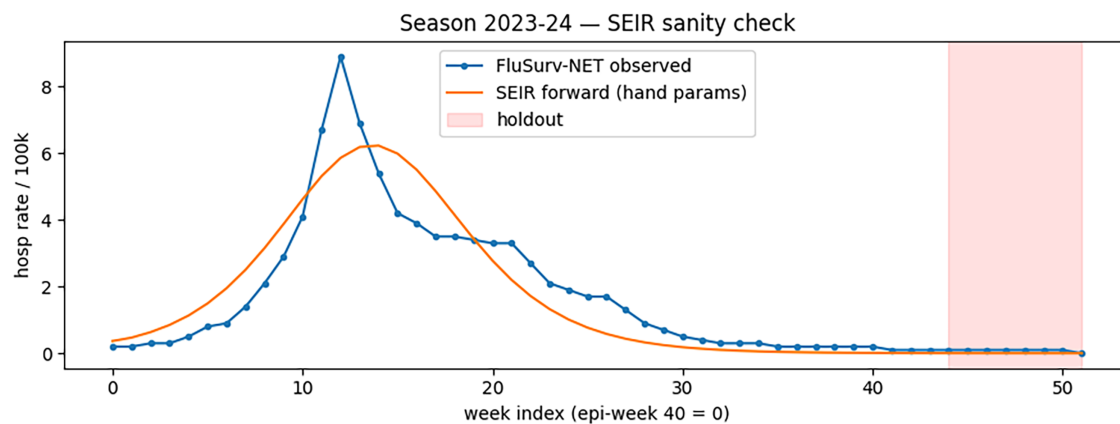
The deficiency arose because predicted hospitalizations collapsed toward zero as the SEIR wave decayed, whereas the observed FluSurv-NET series retained a small non-zero hospitalization floor. The workflow therefore evaluated the prespecified alternative observation model $\mu_t = b + \rho \Delta C_t$. After recalibration, nominal 80% empirical coverage increased to 1.00 and the held-out CRPS was approximately 0.10. This result is interpreted as removal of severe undercoverage, not as evidence of exact 80% calibration: complete empirical coverage may also indicate conservative predictive intervals. Table 3 summarizes the principal quantitative results and Fig. 2 shows the fitted deterministic SEIR forward model vs. FluSurv-NET 2023–24.

5.2 Model Comparison and Stress-Test Results

Predictive stacking assigned approximately all relative weight to the SEIR observation specification containing the background hospitalization component. Because both candidates share the same SEIR dynamical core, this result is interpreted only as strong relative support for the background observation term within the restricted two-model set; it is not evidence of comprehensive mechanistic structural validation.

Table 3: Principal diagnostic results for the CDC FluSurv-NET 2023–24 proof-of-concept.

Diagnostic	Value	Interpretation
Maximum $\widehat{R}$	1.01	No evident convergence failure
Divergences	0	Satisfactory NUTS behavior
Held-out CRPS	0.10	Descriptive; scale dependent
80% coverage, no b	0.12	Severe undercoverage
80% coverage, with b	1.00	Undercoverage removed; conservative
Rolling 4-week CRPS	0.135	Within-season forecast summary
Stacking weight, with b	≈ 1.00	Relative support in two-model set

**Figure 2:** Deterministic SEIR forward model vs. FluSurv-NET 2023–24. The model captures wave timing and shape but underestimates the sharp peak and is more symmetric than the data. Shaded region is the held-out tail used for calibration scoring.

The Sobol analysis of peak hospitalization burden produced first-order indices of approximately 0.50 for R_0 , 0.35 for ρ , and approximately zero for I_0 , with corresponding total-effect indices of approximately 0.63, 0.50, and zero. These results are treated as a consistency and robustness check rather than as a novel epidemiological finding. In particular, the small contribution of I_0 to peak magnitude is compatible with standard compartmental dynamics, while the effect of ρ reflects its direct role in scaling latent infections to observed hospitalizations.

Scenario stress testing showed that increasing apparent transmissibility over the prespecified 0.8–1.5 range produced progressively earlier and larger epidemic trajectories. These curves are interpreted as counterfactual “what-if” scenarios rather than probability-weighted forecasts.

Fig. 3 summarizes the predictive uncertainty, Sobol sensitivity, and transmission-scenario outputs. The predictive panel shows that adding the background component prevents the post-peak uncertainty band from collapsing to zero, while the sensitivity and scenario panels illustrate the distinct roles of parameter and counterfactual uncertainty.

5.3 Summary of the Proof-of-Concept

The principal finding is that APM distinguished sampling adequacy from predictive adequacy: the initial model converged but exhibited severe post-peak undercoverage, and this deficiency changed the workflow routing state. The prespecified background observation component corrected that specific failure,

although its 1.00 empirical coverage is interpreted cautiously because high coverage alone does not imply exact calibration.

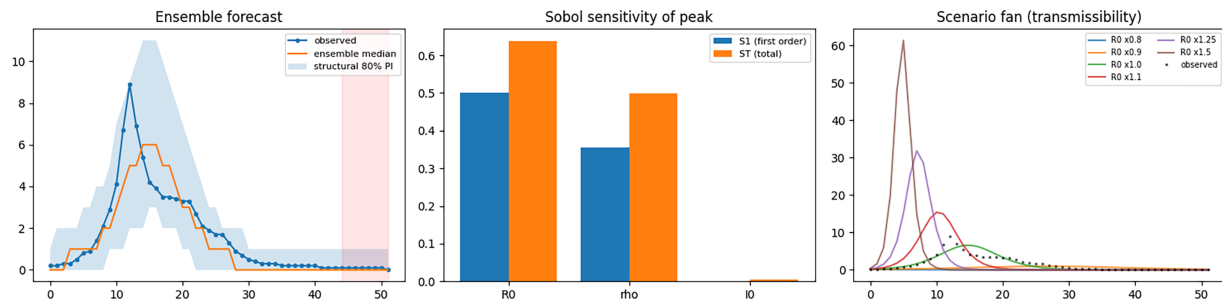


Figure 3: Layered uncertainty and stress-test outputs for the revised observation specification. **(Left)** ensemble forecast over the observed weekly hospitalization rate with the 80% predictive interval; the red region indicates the held-out tail. **(Center)** first-order (S_1) and total-effect (S_T) Sobol indices for R_0 , ρ , and I_0 . **(Right)** counterfactual epidemic trajectories obtained by scaling apparent transmissibility from 0.8 to 1.5, with observed values overlaid.

The rolling-origin CRPS provides additional within-season out-of-sample evidence beyond the original eight-week tail. Nevertheless, all forecasts originate from the same FluSurv-NET 2023–24 season, and the model-comparison exercise involves only two closely related observation specifications. The results therefore support the feasibility of failure-aware workflow routing under the demonstrated case study, not cross-season generalizability or comprehensive structural validation.

6 Discussion

The FluSurv-NET 2023–24 case study demonstrates the central purpose of APMM: to make quantitative diagnostic evidence operational within an automated Bayesian workflow. The SEIR observation specification showed satisfactory posterior sampling behavior and reasonable in-sample agreement, yet produced severe post-peak undercoverage. Rather than treating numerical convergence as sufficient for reporting, APMM used the predictive deficiency to trigger evaluation of the prespecified background-hospitalization specification. Thus, the contribution of *self-auditing* lies not in autonomous scientific reasoning, but in ensuring that predefined diagnostics can alter workflow execution, that failed states remain visible, and that unresolved deficiencies can be escalated to human review.

The change in nominal 80% empirical coverage from approximately 0.12 to 1.00 illustrates both the value and the limits of this mechanism. The revised observation model removed the severe undercoverage caused by the original specification's tendency to collapse toward zero in the post-peak tail. However, coverage of 1.00 is not interpreted as exact 80% calibration because excessively broad predictive intervals may also produce complete coverage. Forecast adequacy is therefore assessed using complementary evidence from CRPS and empirical 80% coverage rather than treating coverage as an isolated criterion. The rolling-origin evaluation further strengthens the original eight-week hold-out analysis by providing repeated within-season out-of-sample checks, although it does not establish performance beyond the observed 2023–24 season.

The layered uncertainty analysis should likewise be interpreted according to the scope of the experiment. Parametric uncertainty reflects posterior variation in inferred quantities, observational uncertainty reflects the mapping from latent infections to reported hospitalizations, and scenario uncertainty represents explicit counterfactual changes in apparent transmissibility rather than probability-weighted forecasts. Structural uncertainty is more limited: both candidate specifications share the same SEIR dynamics and differ primarily through the inclusion of the background hospitalization component. Consequently, the stacking weight of

approximately one for the revised specification indicates strong relative predictive support only within this restricted two-model set and should not be interpreted as comprehensive mechanistic structural validation. Similarly, the Sobol results are treated as consistency and robustness checks of the workflow rather than as novel epidemiological findings.

Several limitations therefore remain. First, the empirical analysis is restricted to a single CDC FluSurv-NET influenza season; rolling-origin evaluation improves within-season validation but does not establish cross-season, cross-region, or cross-pathogen generalizability. Second, the candidate model library is narrow and does not represent alternative mechanisms such as waning immunity, age structure, time-varying transmission, or multi-wave dynamics. Third, several workflow settings, including the revision budget and prior-predictive plausibility rules, are configurable engineering choices rather than universal statistical criteria. Finally, the current implementation is deterministic and does not employ an LLM for model-selection or routing decisions. Numerical quantities are produced only by validated scientific modules, which supports provenance and reproducibility but means that APMM should be understood as an automated self-auditing workflow rather than an autonomous scientific agent.

Future work should therefore prioritize validation across additional influenza seasons and broader surveillance settings, expansion to structurally distinct mechanistic models, and systematic assessment of calibration and sharpness across heterogeneous epidemic conditions. LLM-assisted components may later be introduced for tasks such as summarizing diagnostics or proposing revisions from a validated model library, provided that numerical inference, scoring, and sensitivity analysis remain grounded in auditable computational tools. Overall, the present study supports APMM as a proof-of-concept for failure-aware mechanistic forecasting in which model inadequacy can trigger revision rather than being silently carried into the final report.

7 Conclusion

We presented APMM, a deterministic self-auditing Bayesian workflow in which convergence, probabilistic forecast evaluation, model comparison, stress testing, revision, and human escalation are connected through explicit routing rules. In the CDC FluSurv-NET 2023–24 proof-of-concept, the initial SEIR observation specification converged but produced severe post-peak undercoverage (0.12 for a nominal 80% interval). Adding a prespecified background hospitalization component removed this undercoverage, although the resulting 1.00 coverage is interpreted as potential conservatism rather than exact calibration. Model comparison is restricted to the two observation specifications, and sensitivity results are treated as consistency checks rather than new epidemiological findings. The study therefore establishes the feasibility and auditability of the workflow, not cross-season forecasting superiority. Broader validation and richer model libraries remain priorities for future work.

Acknowledgement: The authors acknowledge the Arkansas High Performance Computing Center for providing computational resources that supported this research.

Funding Statement: This material is based upon work partially supported by the National Science Foundation under Award No. OIA-1946391.

Author Contributions: Conceptualization, Emmanuel Addai; methodology, Emmanuel Addai and Xiaowei Xu; formal analysis, Emmanuel Addai, Zi Wang, Devika Ambekar and Xiaowei Xu; software, Emmanuel Addai, Zi Wang, Devika Ambekar and Xiaowei Xu; visualization, Emmanuel Addai and Xiaowei Xu; writing—original draft preparation, Emmanuel Addai; writing—review and editing, Emmanuel Addai, Zi Wang, Devika Ambekar and Xiaowei Xu; supervision, Xiaowei Xu; funding acquisition, Xiaowei Xu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this study are publicly available from the U.S. Centers for Disease Control and Prevention (CDC) Influenza Hospitalization Surveillance Network (FluSurv-NET) through the FluView Interactive platform. The FluSurv-NET hospitalization data can be accessed and downloaded at: <https://gis.cdc.gov/GRASP/Fluview/FluHospRates.html>. Additional information about FluSurv-NET and its surveillance methodology is available at: <https://www.cdc.gov/fluview/overview/influenza-hospitalization-surveillance.html>.

Ethics Approval: Not applicable. This study used publicly available data and involved no direct recruitment, interaction, or intervention with human participants.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rumack A, Tibshirani RJ, Rosenfeld R. Recalibrating probabilistic forecasts of epidemics. *PLoS Comput Biol*. 2022;18(12):e1010771. doi:10.1371/journal.pcbi.1010771.
2. Kamarthi H, Kong L, Rodríguez A, Zhang C, Prakash BA. When in doubt: neural non-parametric uncertainty quantification for epidemic forecasting. *Adv Neural Inf Process Syst*. 2021;34:19796–807.
3. Sung C-L, Hung Y. Efficient calibration for imperfect epidemic models with applications to the analysis of COVID-19. arXiv:2009.12523. 2020. doi:10.48550/arXiv.2009.12523.
4. Liyanage YR, Saucedo O, Tuncer N, Chowell G. A tutorial on structural identifiability of epidemic models using StructuralIdentifiability.jl. arXiv:2505.10517. 2025. doi:10.48550/arXiv.2505.10517.
5. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc*. 2007;102(477):359–78. doi:10.1198/016214506000001437.
6. Lu C, Lu C, Lange RT, Foerster J, Clune J, Ha D. The AI scientist: towards fully automated open-ended scientific discovery. arXiv:2408.06292. 2024. doi:10.48550/arXiv.2408.06292.
7. Gottweis J, Weng WH, Daryin A, Tu T, Sirkovic P, Myaskovsky A, et al. Accelerating scientific discovery with co-scientist. *Nature*. 2026;655(8122):487–96. doi:10.1038/s41586-026-10644-y.
8. Schmidgall S, Su Y, Wang Z, Sun X, Wu J, Yu X, et al. Agent laboratory: using LLM agents as research assistants. In: Findings of the Association for Computational Linguistics: EMNLP 2025. Stroudsburg, PA, USA: Association for Computational Linguistics; 2025. doi:10.18653/v1/2025.findings-emnlp.320.
9. Wu Q, Bansal G, Zhang J, Wu Y, Li B, Zhu E, et al. AutoGen: enabling next-generation LLM applications via multi-agent conversation. arXiv:2308.08155. 2023. doi:10.48550/arXiv.2308.08155.
10. LangChain. LangGraph: workflows and agents. LangGraph Documentation. 2026 [cited 2026 Jan 1]. Available from: <https://docs.langchain.com/oss/python/langgraph/overview>.
11. Xu H, Zhu Z, Ma D, Zhang S, Fan S, Chen L, et al. Reducing tool hallucination via reliability alignment. arXiv:2412.04141. 2024. doi:10.48550/arXiv.2412.04141.
12. Lin X, Ning Y, Zhang J, Dong Y, Liu Y, Wu Y, et al. LLM-based agents suffer from hallucinations: a survey of taxonomy, methods, and directions. arXiv:2509.18970. 2025. doi:10.48550/arXiv.2509.18970.
13. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. *Proc R Soc Lond Ser A Contain Pap A Math Phys Character*. 1927;115(772):700–21. doi:10.1098/rspa.1927.0118.
14. Hoffman MD, Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J Mach Learn Res*. 2014;15(47):1593–623.
15. Phan D, Pradhan N, Jankowiak M. Composable effects for flexible and accelerated probabilistic programming in NumPyro. arXiv:1912.11554. 2019. doi:10.48550/arXiv.1912.11554.
16. Kumar R, Carroll C, Hartikainen A, Martin O. ArviZ a unified library for exploratory analysis of Bayesian models in Python. *J Open Source Softw*. 2019;4(33):1143. doi:10.21105/joss.01143.
17. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC. Rank-normalization, folding, and localization: an improved $\widehat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Anal*. 2021;16(2):667–718. doi:10.1214/20-ba1221.

18. Tuncer N, Le TT. Structural and practical identifiability analysis of outbreak models. *Math Biosci.* 2018;299(1):1–18. doi:10.1016/j.mbs.2018.02.004.
19. Gallo L, Frasca M, Latora V, Russo G. Lack of practical identifiability may hamper reliable predictions in COVID-19 epidemic models. *Sci Adv.* 2022;8(3):eabg5234. doi:10.1126/sciadv.abg5234.
20. Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc Natl Acad Sci U S A.* 2019;116(8):3146–54. doi:10.1073/pnas.1812594116.
21. Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc Natl Acad Sci U S A.* 2022;119(15):e2113561119. doi:10.1073/pnas.2113561119.
22. Vehtari A, Ojanen J. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statist Surv.* 2012;6:142–228. doi:10.1214/12-ss102.
23. Yao Y, Vehtari A, Simpson D, Gelman A. Using stacking to average Bayesian predictive distributions. *Bayesian Anal.* 2018;13(3):917–1007. doi:10.1214/17-BA1091.
24. Sobol' IM. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math Comput Simul.* 2001;55(1–3):271–80. doi:10.1016/s0378-4754(00)00270-6.
25. Herman J, Usher W. SALib: an open-source Python library for sensitivity analysis. *J Open Source Softw.* 2017;2(9):97. doi:10.21105/joss.00097.