

Journal of Agricultural, Biological, and Environmental StatisticsA Likelihood Framework for Multi-Stage Interval-Censored Processes

Patrice Takam Soh

patricetakam@gmail.com

Universite de Yaounde I <https://orcid.org/0000-0003-2584-3138>

Tsogmo Douane Cybele

Université de Yaoundé I

Toussile Wilson

Université de Yaoundé I

Research Article

Keywords: Multi-state models, Interval censoring, Cumulative environmental covariates, Monte Carlo EM, Louis's observed information, Phytosanitary risk, Theobroma cacao

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11346552/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.

A Likelihood Framework for Multi-Stage Interval-Censored Processes with Partially Ordered Competing Events and Cumulative Covariates

Patrice Takam Soh*

Cybèle Tsogmo Douane†

Wilson Toussile‡

Abstract

We propose a rigorous likelihood-based joint framework for modeling stage-specific transition times in multi-stage biological processes subjected to discrete inspection schemes and environmental forcing. The proposed methodology simultaneously resolves three intertwined observational and ecological challenges: (i) interval-censored stage entry and exit times, (ii) right censoring arising from both administrative follow-up limits and disease-induced premature failures (e.g., pathogenic attacks), and (iii) a partially ordered dual-event structure wherein a stage transition and a disease attack occur within the same inspection interval. The transition hazard is modeled as a function of cumulative time-varying environmental and climatic covariates integrated over calendar time, capturing how accumulated weather exposure affects instantaneous developmental risks between latent stage entry and exit. This non-proportional hazard formulation precludes standard partial likelihood methods and necessitates full likelihood inference. Because the observed likelihood involves high-dimensional intractable integrals over unobserved entry and exit dates, parameter estimation is conducted via a Monte Carlo Expectation-Maximization (MCEM) algorithm. The E-step draws exact conditional latent trajectories via probability integral transform rejection sampling, while the M-step updates piecewise-constant baseline hazard parameters in closed form and regression coefficients via bounded L -BFGS- B optimization. Statistical uncertainty on environmental effects is rigorously quantified using Louis’s observed information principle, combining complete-data Hessian expectations with Monte Carlo score-variance corrections to account for missing information due to interval censoring. Motivated by the multi-stage fruit development of cocoa (*Theobroma cacao*) under phytosanitary pressure, the framework provides a versatile tool for longitudinal agro-environmental event-time data subject to periodic monitoring.

Keywords: Multi-state models; Interval censoring; Cumulative environmental covariates; Monte Carlo EM; Louis’s observed information; Phytosanitary risk; *Theobroma cacao*.

1 Introduction

Event-time processes evolving through ordered developmental stages arise naturally across ecological, agricultural, and environmental systems. In these biological settings, organisms sequentially transition across discrete developmental states, and statistical inference primarily focuses on estimating stage-specific duration distributions under environmental forcing. This task becomes particularly challenging when transitions are monitored only at discrete inspection visits, as in crop field trials, phytosanitary surveys, or forestry cohorts. Under such monitoring designs, exact event dates remain unobserved and are known only to fall within consecutive inspection intervals. This generates multi-stage interval-censored data, which substantially complicates likelihood construction, parameter estimation, and covariate effect inference [20].

The analysis of censored survival outcomes has a rich methodological foundation. Classical frameworks for right-censored data originated from the hazard-plotting approach of Nelson [16] and the counting-process formulation of Aalen [1], while Tobin [22] introduced censored regression models. For interval-censored settings, Turnbull [23] proposed the foundational nonparametric maximum likelihood estimator, inspiring multivariate extensions such as Dabrowska’s bivariate product-integral estimator [9]. Over subsequent decades, semiparametric regression techniques for right- and interval-censored outcomes evolved extensively [20]. Nonparametric smoothing strategies were also developed to handle multiple concurrent censoring mechanisms and competing risks under fixed inspection schedules [21]. More recently, flexible frameworks have addressed clustered interval-censored data [3] and nonhomogeneous multi-state processes [2].

*Email: patricetakam@gmail.com

†Email: ctsogmo@gmail.com

‡Corresponding author. Email: wilson.toussile@gmail.com

To position our proposed contribution within agro-environmental and growth modeling, existing approaches can be broadly grouped into four paradigms. First, classical parametric growth curves (e.g., Logistic, Gompertz, Richards) [19] provide deterministic descriptions of biomass or stage progression. Second, nonlinear mixed-effects models extend these curves to account for hierarchical between-subject variability [10]. Third, dynamic mechanistic systems based on differential equations describe epidemiological spread and plant kinetics as functions of environmental drivers [14]. Fourth, functional and semiparametric methods enable data-driven curve estimation without rigid global functional assumptions [18]. Despite their utility [11, 27], none of these frameworks simultaneously accommodates multi-stage transitions, concurrent interval censoring, pathogenic competing risks, and cumulative time-varying environmental covariates under discrete inspection—a critical methodological gap addressed herein.

This work addresses three intertwined observational and ecological challenges inherent to field inspection schemes. The first is interval censoring of stage transitions: because exact transition dates between successive stages are unobserved, closed-form likelihoods are unavailable, invalidating standard exact-time survival models [20]. The second challenge is right censoring induced by terminal competing events. In addition to normal stage progression, individuals may suffer terminal failures (e.g., lethal pathogen attacks, pest damage, or administrative end of follow-up) that permanently arrest development. This requires joint likelihood modeling to separate administrative study termination from pathogen-induced premature failure [21]. The third challenge—both ecologically vital and methodologically complex—is a partially ordered dual-event configuration. It frequently occurs that both a stage transition and a pathogenic attack take place within the same inspection interval and are recorded concurrently at the subsequent visit. While discrete nonparametric methods [21] collapse these events onto interval boundaries, our continuous-time formulation explicitly accounts for the latent intra-interval chronological ordering of the two events.

Beyond these observational structures, our model incorporates time-varying climatic covariates that modulate developmental hazards through cumulative environmental exposure. Rooted in physiological principles establishing that biological development is driven by thermal and hydric sums rather than instantaneous weather shocks, the transition hazard at any calendar time depends on accumulated weather inputs since stage entry. Because environmental accumulation depends explicitly on calendar time and stage entry is itself latent, the resulting hazard ratio between individuals is non-proportional over time. This invalidates standard Cox partial likelihood arguments [8] and mandates full likelihood inference over interval-censored latencies.

The primary contributions of this paper are fourfold: (i) We formulate a unified continuous-time likelihood framework for multi-stage processes that jointly incorporates interval censoring, terminal competing risks, and partially ordered intra-interval dual events; (ii) We develop a Monte Carlo Expectation-Maximization (MCEM) algorithm [5, 25] utilizing exact probability integral transform sampling in the E-step, closed-form updates for piecewise-constant baseline hazards, and bounded L -BFGS- B optimization for environmental regression parameters in the M-step; (iii) We establish rigorous uncertainty quantification for climatic parameter estimates via Louis’s observed information principle [13], combining complete-data Hessian expectations with Monte Carlo score-variance corrections to properly reflect missing information due to interval censoring; (iv) We demonstrate the practical utility of the methodology on a multi-stage fruit development dataset of cocoa (*Theobroma cacao*) subject to phytosanitary pressure under varying microclimatic conditions.

The remainder of this paper is structured as follows. Section 2 details the formal statistical model, notation, inspection scheme, cumulative hazard specification, and observation classifications. Section 3 presents the full likelihood formulation and Monte Carlo surrogate objective. Section 4 details the MCEM algorithm, and Section 5 establishes Louis’s observed information matrix for variance estimation. Section 6 validates the framework through Monte Carlo simulations. Section 7 applies the methodology to cocoa developmental data, and Section 8 concludes with discussion and future directions.

2 Mathematical Framework and Problem Formulation

2.1 Biological Process, Developmental Stages, and Event Dynamics

Consider a cohort of n subjects ($i = 1, \dots, n$) monitored continuously from origin until terminal transition or early failure. Structurally, the process spans $K + 1$ physiologically distinct stages, where entering stage $K + 1$ signifies a terminal absorbing state. Consequently, the active developmental pipeline comprises K consecutive transient stages ($k = 1, \dots, K$).

In our application to cocoa pod development, $K = 3$ active stages are considered (Stage 1: Cherelle, Stage 2: Young pod, Stage 3: Adult pod), with Stage 4 ($K + 1$) representing commercial maturity and harvest.

Let U_{ik} denote the continuous latent calendar date at which fruit i enters stage k , and V_{ik} the latent date at

which it transitions to stage $k + 1$. The stage-specific sojourn time spent by fruit i in stage k is defined as the random variable:

$$T_{ik} = V_{ik} - U_{ik} > 0. \quad (1)$$

During its lifecycle in stage k , fruit i is exposed to exogenous pathogenic pressures (primarily black pod disease caused by *Phytophthora* spp., alongside rodent damage and physiological wilt). Let M_{ik} denote the latent exact date of a fatal pathogenic attack in stage k .

To formalize the interplay between developmental transitions and sanitary failures, we establish the following structural assumptions:

Assumption 1 (Equivalence of Infection and Stage Termination). An observed pathogenic attack on a fruit instantaneously stops its physiological growth, resulting in immediate premature failure. Thus, the occurrence of an attack at date M_{ik} acts as a competing event that right-censors the developmental sojourn duration T_{ik} .

Assumption 2 (Conditional Markovian Stage Independence). Conditional on entering stage k in a healthy state at date U_{ik} , the local pathogenic attack process within stage k is conditionally independent of the historical sojourn durations spent in preceding stages $\{1, \dots, k - 1\}$.

Assumption 2 is biologically grounded: epidemiological outbreaks in cocoa farms are localized in time and driven predominantly by short-term microclimatic fluctuations (e.g., rainfall spikes, relative humidity, and canopy temperature) rather than the fruit's historical trajectory.

2.2 Hazard Rate Specification and Non-Proportionality Analysis

Based on Assumptions 1 and 2, the continuous-time duration spent by a fruit in stage k before transitioning to the next stage is modeled through a stage-specific instantaneous transition hazard rate. In this framework, disease attacks do not represent developmental transitions but act as competing right-censoring events that prematurely terminate the observation of the stage duration. Building upon the multiplicative hazard concept of Cox [8], we specify the transition intensity at calendar date $v > u$, conditional on latent stage entry at date u , as:

$$\lambda_k(v \mid u, \mathbf{Z}_k) = \lambda_{0k}(v - u) \exp\left(\beta_k^\top \mathbf{Z}_k(u, v)\right), \quad (2)$$

where $\lambda_{0k}(v - u)$ captures the intrinsic duration-dependent baseline risk of stage transition, $\mathbf{Z}_k(u, v) \in \mathbb{R}^r$ represents the vector of cumulative environmental exposures accumulated between entry u and current calendar date v , and $\beta_k \in \mathbb{R}^r$ is the vector of stage-specific regression parameters (with $r = 3$ in our empirical application).

In what follows, we detail the mathematical formulation of each component in (2), derive the corresponding transition survival kinetics, and highlight how this formulation fundamentally departs from classical Cox proportional hazards architectures.

2.2.1 Piecewise-Constant Baseline Hazard

The baseline hazard rate $\lambda_{0k}(t)$ is defined on the elapsed duration $t = v - u > 0$ spent in stage k . To achieve flexibility without imposing restrictive parametric shapes (such as Weibull or Gamma distributions), $\lambda_{0k}(t)$ is approximated by a piecewise-constant step function aligned with the weekly inspection grid. Recalling the contiguous time intervals $A_j = (t_{j-1}, t_j] = ((j-1)h, jh]$ for $j = 1, \dots, J$ with step size $h = 7$ days, we specify:

$$\lambda_{0k}(t) = \alpha_{kj} \quad \text{for } t \in A_j, \quad (3)$$

where $\alpha_{kj} > 0$ are stage-specific baseline parameters to be estimated.

2.2.2 Cumulative Microclimatic Exposures

Let $\mathbf{CI}(x) = (T_c(x), \Delta T(x), \text{PLV}(x))^\top$ denote the continuous vector process of environmental conditions at calendar date x (canopy temperature T_c , thermal amplitude gradient $\Delta T = T_a - T_c$ with ambient temperature T_a , and log-precipitation PLV). The cumulative climatic exposure accumulated between entry date u and calendar date v is defined as:

$$\mathbf{Z}_k(u, v) = \int_u^v w_k(x) \mathbf{CI}(x) dx, \quad (4)$$

where $w_k(\cdot)$ represents a stage-specific biological weighting function. In our empirical application, we adopt a uniform weighting scheme ($w_k(x) = 1$ for all x), assuming that environmental exposure accumulates continuously at a constant biological rate throughout stage k .

Since meteorological metrics are recorded at daily resolution, the integral in (4) with $w_k(x) = 1$ is evaluated analytically using a piecewise-constant representation of $\mathbf{CI}(x)$. Let $\mathbf{CI}_d$ denote the daily measurement on day $d \in \mathbb{N}$. For arbitrary continuous latent dates u and v ($u < v$), the cumulative process simplifies to:

$$\mathbf{Z}_k(u, v) = (\lceil u \rceil - u)\mathbf{CI}_{\lceil u \rceil} + \sum_{d=\lceil u \rceil+1}^{\lfloor v \rfloor} \mathbf{CI}_d + (v - \lfloor v \rfloor)\mathbf{CI}_{\lfloor v \rfloor+1}. \quad (5)$$

2.2.3 Departure from the Classical Cox Framework

Although equation (2) exhibits a multiplicative structure reminiscent of the semi-parametric Cox proportional hazards model [8], it departs fundamentally from classical survival architectures in four structural ways:

1. **Dual Latent Time Dependence:** Both the duration-based baseline hazard $\lambda_{0k}(v - u)$ and the cumulative covariate process $\mathbf{Z}_k(u, v)$ depend explicitly on the unobserved latent entry date u .
2. **Structural Violation of Proportional Hazards:** The hazard ratio between two individuals i and j ,

$$\frac{\lambda_k(v \mid u_i, \mathbf{Z}_{ik})}{\lambda_k(v \mid u_j, \mathbf{Z}_{jk})} = \frac{\lambda_{0k}(v - u_i)}{\lambda_{0k}(v - u_j)} \exp\left(\boldsymbol{\beta}_k^\top [\mathbf{Z}_k(u_i, v) - \mathbf{Z}_k(u_j, v)]\right), \quad (6)$$

is intrinsically time-varying and non-constant. As a result, the proportional hazards assumption fails structurally.

3. **Integrated Covariate Coupling:** The cumulative term $\mathbf{Z}_k(u, v) = \int_u^v w_k(x)\mathbf{CI}(x)dx$ couples the entry time u with the entire calendar covariate trajectory, rendering partial likelihood techniques [26] inapplicable.
4. **Dual Time-Scale Alignment:** Baseline risk evolves along physiological stage duration ($v - u$), whereas environmental risk accumulates over real calendar time (x).

Consequently, parameter estimation cannot rely on partial profile likelihoods; it necessitates full likelihood inference via a tailored Monte Carlo EM framework.

2.3 Inspection Scheme and Multi-Level Censoring Mechanism

In field trials, continuous temporal monitoring of individual fruits is practically infeasible. Instead, the cohort is visited at discrete, equally spaced inspection nodes:

$$\tau_j = j \cdot h, \quad j = 0, 1, \dots, J, \quad (7)$$

where $h = 7$ days represents the fixed weekly inspection step, and $\tau_J = J \cdot h$ denotes the administrative end of the follow-up period ($J = 20$ weeks).

At each inspection visit τ_j , field operators record two discrete indicators for subject i :

- $S_{ij} \in \{1, \dots, K, K+1\}$: The observed developmental stage at time τ_j .
- $A_{ij} \in \{0, 1\}$: The sanitary status indicator (1 = healthy, 0 = infected/damaged).

This discrete monitoring scheme induces a two-tier censoring topology:

1. **Interval Censoring on Transition Dates ((U_{ik}, V_{ik})):** Neither the exact entry date U_{ik} nor the exit date V_{ik} is observed continuously. They are known only to fall within weekly inspection windows:

$$U_{ik} \in (L_{ik}^{(1)}, R_{ik}^{(1)}] = (\tau_{j-1}, \tau_j], \quad V_{ik} \in (L_{ik}^{(2)}, R_{ik}^{(2)}] = (\tau_{m-1}, \tau_m], \quad (8)$$

where $L_{ik}, R_{ik} \in \{\tau_0, \tau_1, \dots, \tau_J\}$ with $R_{ik} - L_{ik} = h$.

2. **Administrative Right Censoring (C_i):** Fruits that remain healthy and unharvested by the terminal visit τ_J have their current stage duration right-censored at $C_i = \tau_J$.

To make the observed data likelihood tractable under this inspection regime, the study design adheres to three standard operational hypotheses:

Assumption 3 (Early Cohort Inclusion). All monitored fruits enter stage 1 (cherelle) within the initial inspection window $(\tau_{-1}, \tau_0] = (-h, 0]$.

Assumption 4 (Stage Transition Bound). A fruit cannot transition through more than one developmental stage within a single inspection interval h .

Assumption 5 (Healthy Initial Baseline). All fruits enrolled at τ_0 are active, uninfected, and alive at the start of follow-up.

Figure 1 provides a schematic illustration of this observational setup.

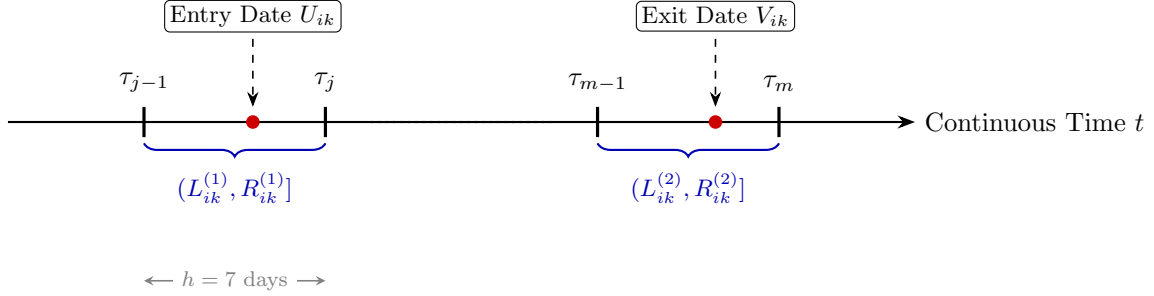


Figure 1: Schematic representation of the continuous-time trajectory and the two-tier inspection window topology for entry date $U_{ik} \in (L_{ik}^{(1)}, R_{ik}^{(1)})$ and exit date $V_{ik} = U_{i,k+1} \in (L_{ik}^{(2)}, R_{ik}^{(2)})$.

2.3.1 Classification of Observation Topologies

In field monitoring, the joint trajectory of developmental progression and sanitary status for fruit i at stage k falls into one of four mutually exclusive observational configurations. Let $\delta_{ik} \in \{1, 2, 3, 4\}$ denote the categorical status indicator:

- **Case $\delta_{ik} = 1$ (Stage Transition Without Disease Attack):** The fruit successfully transitions to stage $k + 1$ during the follow-up without experiencing a fatal pathogenic attack:

$$U_{ik} \in (L_{ik}^{(1)}, R_{ik}^{(1)}], \quad V_{ik} \in (L_{ik}^{(2)}, R_{ik}^{(2)}], \quad M_{ik} > R_{ik}^{(2)}.$$

- **Case $\delta_{ik} = 2$ (Administrative Right-Censoring):** The fruit remains healthy and active in stage k up to the end of follow-up $\tau_J = C_i$:

$$U_{ik} \in (L_{ik}^{(1)}, R_{ik}^{(1)}], \quad V_{ik} > \tau_J, \quad M_{ik} > \tau_J.$$

- **Case $\delta_{ik} = 3$ (Transition Followed by Terminal Attack in Same Interval):** Both the stage transition and the fatal attack occur within the same inspection interval $(\tau_{m-1}, \tau_m]$, with the stage transition strictly preceding the attack ($V_{ik} < M_{ik}$):

$$U_{ik} \in (L_{ik}^{(1)}, R_{ik}^{(1)}], \quad V_{ik} \in (L_{ik}^{(2)}, R_{ik}^{(2)}], \quad M_{ik} \in (L_{ik}^{(2)}, R_{ik}^{(2)}], \quad V_{ik} < M_{ik}.$$

- **Case $\delta_{ik} = 4$ (Premature Terminal Failure Before Transition):** The fruit suffers a fatal pathogenic attack in stage k prior to transitioning to stage $k + 1$:

$$U_{ik} \in (L_{ik}^{(1)}, R_{ik}^{(1)}], \quad M_{ik} \in (L_{ik}^{(2)}, R_{ik}^{(2)}], \quad M_{ik} < V_{ik}.$$

Figure 2 presents a schematic representation of these four observational topologies, contrasting the latent continuous dynamics against the interval-censored boundaries.

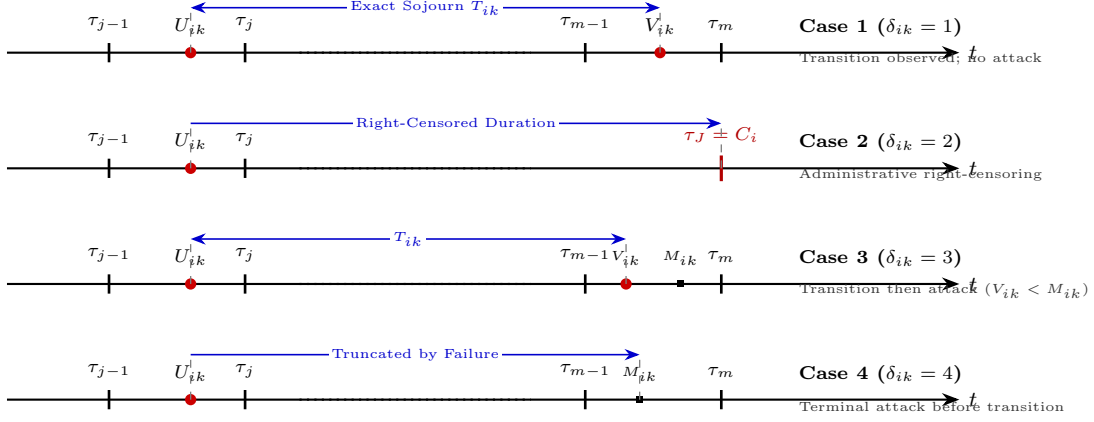


Figure 2: Schematic representation of the four observational configurations for stage k . Red circles denote latent stage transitions (U_{ik}, V_{ik}) , black stars indicate fatal pathogenic attacks (M_{ik}) , and the vertical dashed red line marks administrative censoring (τ_J) .

2.4 Inference Objectives and Analytical Scope

Having defined both the underlying continuous-time hazard dynamics and the discrete interval-censoring inspection scheme, standard survival inference tools can no longer be directly applied.

The primary statistical goal of this work is threefold:

1. **Stagewise Baseline Hazard & Survival Estimation:** Jointly estimate the stage-specific baseline parameters $\alpha_k = (\alpha_{k1}, \dots, \alpha_{kJ})^\top$ and survival functions $S_k(t | \mathbf{Z}_k)$ for sojourn times T_{ik} , properly accounting for interval-censored boundaries (U_{ik}, V_{ik}) .
2. **Quantifying Time-Varying Covariate Effects:** Estimate the regression parameter vectors $\beta_k \in \mathbb{R}^3$ linking cumulative microclimatic exposures $\mathbf{Z}_k(u, v)$ (canopy temperature T_c , thermal gradient ΔT , and log-rainfall PLV) to stage transition rates and disease vulnerability.
3. **Handling Latent Variables via MCEM Inferences:** Develop an adapted Monte Carlo Expectation-Maximization (MCEM) computational engine capable of integrating out the unobserved latent entry/exit dates (U_{ik}, V_{ik}) and disease failure times M_{ik} across all four observational topologies $\delta_{ik} \in \{1, 2, 3, 4\}$.

3 Likelihood-Based Inference

3.1 Data Paradigms and Observational Regularity

To establish the likelihood framework, we distinguish between the complete latent process and the coarsened data recorded under discrete field visits.

For individual i at stage k , the *complete data* vector is defined as:

$$\mathcal{D}_{ik}^c = (U_{ik}, V_{ik}, M_{ik}, \mathbf{Z}_{ik}), \quad (9)$$

comprising the exact continuous entry date U_{ik} , stage transition date V_{ik} , pathogenic attack date M_{ik} , and continuous covariate history $\mathbf{Z}_{ik}$.

Conversely, the *observed data* vector is given by:

$$\mathcal{O}_{ik} = \left((L_{ik}^{(1)}, R_{ik}^{(1)}], (L_{ik}^{(2)}, R_{ik}^{(2)}], \delta_{ik}, \mathbf{Z}_{ik} \right), \quad (10)$$

where continuous time anchors (U_{ik}, V_{ik}, M_{ik}) are coarsened into their respective inspection brackets, and $\delta_{ik} \in \{1, 2, 3, 4\}$ denotes the categorical observation topology defined in Section 2.3.1.

To ensure statistical identifiability and mathematically ground the likelihood formulation, the modeling framework establishes four methodological hypotheses:

Hypothesis 3.1 (Uniform Initial Stage Inclusion, $\mathcal{H}_1$). All cohort fruits enter stage 1 within the initial inspection window, with uniform conditional density $U_{i1} \sim \text{Uniform}(L_{i1}^{(1)}, R_{i1}^{(1)}) = (-h, 0]$.

Hypothesis 3.2 (Uniform Terminal Event Distribution, $\mathcal{H}_2$). Given a pathogenic attack ($\delta_{ik} \in \{3, 4\}$) and conditional on the latent event date M_{ik} falling within the inspection interval $(L_{ik}^{(2)}, R_{ik}^{(2)}]$, M_{ik} is conditionally independent of the interval bounds and uniformly distributed over this window:

$$M_{ik} \mid \delta_{ik} \in \{3, 4\}, M_{ik} \in (L_{ik}^{(2)}, R_{ik}^{(2)}] \sim \text{Uniform}(L_{ik}^{(2)}, R_{ik}^{(2)}].$$

Hypothesis 3.3 (Conditional Independence and Intra-Interval Ordering, $\mathcal{H}_3$). Conditional on entry date U_{ik} and covariate history $\mathbf{Z}_{ik}$, stage transition V_{ik} and pathogenic failure M_{ik} are conditionally independent random variables. Furthermore, if both events occur within the same inspection interval $(\tau_{m-1}, \tau_m]$, stage transition strictly precedes failure ($V_{ik} < M_{ik}$).

Hypothesis 3.4 (Non-Informative Inspection and Censoring, $\mathcal{H}_4$). The inspection grid sequence $\{\tau_j\}_{j=0}^J$ and administrative censoring time $C_i = \tau_J$ are non-informative, such that the joint distribution of inspection boundaries carries no information regarding parameter vector $\boldsymbol{\theta}_k$.

3.2 Complete-Data Likelihood and Log-Likelihood

Under the hypothetical complete-data paradigm, latent temporal anchors and failure times are assumed fully observed without coarsening.

Definition 3.1 (Latent Time Variables and Effective Exit). For subject i at stage k , let $U_{ik} \in \mathbb{R}^+$ be the exact entry time, $V_{ik} \in (U_{ik}, +\infty)$ the stage transition time, $M_{ik} \in (U_{ik}, +\infty]$ the fatal attack time ($M_{ik} = +\infty$ if unattacked), and $C_i = \tau_J$ administrative censoring. The effective calendar exit date is $T_{ik}^* = \min(V_{ik}, M_{ik}, C_i)$, and the stage transition indicator is $\delta_{ik}^{\text{trans}} = \mathbf{1}(V_{ik} \leq \min(M_{ik}, C_i))$.

Proposition 3.1 (Complete-Data Log-Likelihood). Given realizations $(u_{ik}, v_{ik}, m_{ik}, c_i)$, the complete-data log-likelihood contribution of fruit i for stage k parameter vector $\boldsymbol{\theta}_k = (\boldsymbol{\alpha}_k^\top, \boldsymbol{\beta}_k^\top)^\top$ is:

$$\begin{aligned} \ell_i^c(\boldsymbol{\theta}_k; u_{ik}, v_{ik}, m_{ik}, c_i) &= \delta_{ik}^{\text{trans}} \left[\log \lambda_{0k}(v_{ik} - u_{ik}) + \boldsymbol{\beta}_k^\top \mathbf{Z}_k(u_{ik}, v_{ik}) \right] \\ &\quad - \int_{u_{ik}}^{\min(v_{ik}, m_{ik}, c_i)} \lambda_{0k}(s - u_{ik}) e^{\boldsymbol{\beta}_k^\top \mathbf{Z}_k(u_{ik}, s)} ds. \end{aligned} \quad (11)$$

For n independent fruits, the total complete-data log-likelihood is $\mathcal{L}^c(\boldsymbol{\theta}_k) = \sum_{i=1}^n \ell_i^c(\boldsymbol{\theta}_k; U_{ik}, V_{ik}, M_{ik}, C_i)$.

Proof. Under independent competing risks between stage transition and pathogenic attack (Hypothesis 3.3), the joint density factorizes into the conditional transition density at v_{ik} and the survival function up to the effective stopping date $t_{ik}^* = \min(v_{ik}, m_{ik}, c_i)$:

$$f(v_{ik}, t_{ik}^* \mid u_{ik}; \boldsymbol{\theta}_k) = [\lambda_k(v_{ik} \mid u_{ik}, \mathbf{Z}_k)]^{\delta_{ik}^{\text{trans}}} \cdot \exp\left(-\int_{u_{ik}}^{t_{ik}^*} \lambda_k(s \mid u_{ik}, \mathbf{Z}_k) ds\right).$$

Substituting hazard specification $\lambda_k(s \mid u_{ik}, \mathbf{Z}_k) = \lambda_{0k}(s - u_{ik}) \exp(\boldsymbol{\beta}_k^\top \mathbf{Z}_k(u_{ik}, s))$ and taking logarithms yields (11). $\square$

Remark 3.1 (Simplification for Uncensored Transitions). When fruit i experiences an unattacked stage transition ($\delta_{ik} = 1$, $t_{ik}^* = v_{ik}$), equation (11) reduces to:

$$\ell_i^c(\boldsymbol{\theta}_k; u_{ik}, v_{ik}) = \log \lambda_{0k}(v_{ik} - u_{ik}) + \boldsymbol{\beta}_k^\top \mathbf{Z}_k(u_{ik}, v_{ik}) - \int_{u_{ik}}^{v_{ik}} \lambda_{0k}(s - u_{ik}) \exp(\boldsymbol{\beta}_k^\top \mathbf{Z}_k(u_{ik}, s)) ds. \quad (12)$$

3.3 Observed-Data Likelihood and Intractability

In practice, the exact entry times U_i , transition times V_i , and disease attack times M_i are subject to interval censoring determined by periodic examination windows. The observed data vector for individual i is denoted $\mathcal{O}_i = ([L_i^{(1)}, R_i^{(1)}], [L_i^{(2)}, R_i^{(2)}], m_i, c_i, \delta_i)$.

Definition 3.2 (Feasible Domain Indicator). Let c_i be the fixed administrative censoring date and $m_i \in \mathbb{R}^+ \cup \{+\infty\}$ the realized terminal event time ($m_i = +\infty$ in the absence of an attack prior to c_i). The indicator $\mathbf{1}_{\mathcal{C}_i(u, v, m_i, c_i)}$ defines the feasible integration domain $\mathcal{C}_i$ for the latent variables (u, v) by enforcing:

1. **Chronological order:** $u < \min(v, m_i, c_i)$;
2. **Administrative upper bound:** $t_i^* = \min(v, m_i, c_i) \leq c_i$;
3. **Observational type alignment:** the latent configuration (u, v) preserves the observed status $\delta_i = \mathbf{1}(v \leq \min(m_i, c_i))$, consistently with Section 2.3.1.

Lemma 3.2 (Likelihood Contribution for an Exact Continuous Event Observation). *Let M be a continuous random variable with conditional density $f_{M|W}(m | w)$ given latents $W = w$. If M is observed exactly at m_0 , then marginalizing the joint density over m collapses to a point evaluation at $m = m_0$:*

$$\lim_{\varepsilon \rightarrow 0^+} \frac{1}{\varepsilon} \mathbb{P}(M \in [m_0, m_0 + \varepsilon] | W = w) = f_{M|W}(m_0 | w).$$

Proof. By the mean value theorem for integrals, $\mathbb{P}(M \in [m_0, m_0 + \varepsilon] | W = w) = \varepsilon f_{M|W}(m_\varepsilon | w)$ for some $m_\varepsilon \in [m_0, m_0 + \varepsilon]$; dividing by ε and letting $\varepsilon \rightarrow 0^+$ gives $m_\varepsilon \rightarrow m_0$, whence the result, using continuity of $f_{M|W}(\cdot | w)$ at m_0 . $\square$

Proposition 3.3 (Dimensionality Reduction to a Double Integral). *By marginalizing out the continuous latent terminal-event variable M_i , the observed-data likelihood contribution $L_i(\theta_k | \mathcal{O}_i)$ reduces from a triple integral over $(u, v, m) \in \mathbb{R}^3$ to a double integral over the latent timeline anchors (u, v) :*

$$L_i(\theta_k | \mathcal{O}_i) = \int_{L_i^{(1)}}^{R_i^{(1)}} \int_{S_v} \exp(\ell_i^c(\theta_k; u, v, m_i, c_i)) f_{U_i}(u) \mathbf{1}_{\{v \in [L_i^{(2)}, R_i^{(2)}]\}} \mathbf{1}_{C_i(u, v, m_i, c_i)} dv du, \quad (13)$$

where $f_{U_i}(u)$ is the marginal entry density, $\mathbf{1}_{\{v \in [L_i^{(2)}, R_i^{(2)}]\}}$ enforces interval censoring on the transition date, and $\mathbf{1}_{C_i(u, v, m_i, c_i)}$ is the global feasibility indicator, applied separately according to whether the fruit is attacked or not.

Proof. Let $W_i = (U_i, V_i, M_i)$. By the chain rule, the joint density factorizes as

$$f_{U_i, V_i, M_i}(u, v, m | \theta_k) = f_{M_i|U_i, V_i}(m | u, v; \theta_k) \cdot f_{V_i|U_i}(v | u; \theta_k) \cdot f_{U_i}(u), \quad (14)$$

and $L_i(\theta_k | \mathcal{O}_i)$ is obtained by integrating (14) over the support restricted by $\mathbf{1}_{C_i(u, v, m, c_i)}$. By construction, $\exp(\ell_i^c(\theta_k; u, v, m, c_i)) = f_{M_i|U_i, V_i}(m | u, v; \theta_k) \cdot f_{V_i|U_i}(v | u; \theta_k)$, so it remains to evaluate the inner integral $I_i(u, v) = \int_u^{+\infty} \mathbf{1}_{C_i(u, v, m, c_i)} f_{M_i|U_i, V_i}(m | u, v) f_{V_i|U_i}(v | u) dm$ across two mutually exclusive regimes:

(i) *Attacked fruit* ($\delta_i \in \{3, 4\}$, exact date $M_i = m_i$). By Lemma 3.2, the integral over m collapses to a point evaluation at $m = m_i$:

$$I_i(u, v) = \exp(\ell_i^c(\theta_k; u, v, m_i, c_i)) \cdot \mathbf{1}_{\{u < m_i \leq \min(v, c_i)\}} \cdot \mathbf{1}_{\{u \in [L_i^{(1)}, R_i^{(1)}]\}}.$$

(ii) *Non-attacked fruit* ($\delta_i \in \{1, 2\}$, $M_i > \min(V_i, C_i)$). Here M_i remains latent over $(t_i^*, +\infty)$, and integrating the conditional attack density over this range yields the attack-free survival probability, so that

$$I_i(u, v) = f_{V_i|U_i}(v | u) \cdot S_k^{\text{att}}(t_i^* | u, v) = \exp(\ell_i^c(\theta_k; u, v, +\infty, c_i)).$$

Combining (i) and (ii) into the triple integral over (u, v, m) and integrating out m in both cases establishes (13). $\square$

Remark 3.2 (Computational Intractability and MCEM Justification). Direct evaluation or maximization of the overall observed log-likelihood $\mathcal{L}(\theta_k) = \sum_{i=1}^n \log L_i(\theta_k | \mathcal{O}_i)$ is computationally intractable for two primary reasons. First, the integral over the latent entry time $u \in (L_{ik}^{(1)}, R_{ik}^{(1)})$ cannot be evaluated analytically, as the continuous exposure trajectory $\mathbf{Z}_k(u, \cdot)$ depends nonlinearly on u . Second, the inner cumulative hazard integral $\int_u^v \lambda_{0k}(s - u) \exp(\beta_k^\top \mathbf{Z}_k(u, s)) ds$ lacks a closed-form solution due to the continuous temporal variation of $\mathbf{Z}_k(u, s)$. Furthermore, the joint indicator $\mathbf{1}_{C_i(u, v, m_i, c_i)}$ induces a complex, non-rectangular integration domain. Consequently, standard numerical quadrature or direct gradient-based optimization fails or scales poorly, which motivates the Monte Carlo Expectation-Maximization (MCEM) framework presented in Section 4.

Identifiability. Under hypotheses $(\mathcal{H}_1)$ – $(\mathcal{H}_4)$ and correct model specification, the parameter θ_k is identifiable from the observed likelihood (13). The dual-event configuration does not introduce non-identifiability, since its likelihood contribution integrates over a strictly ordered and non-symmetric region.

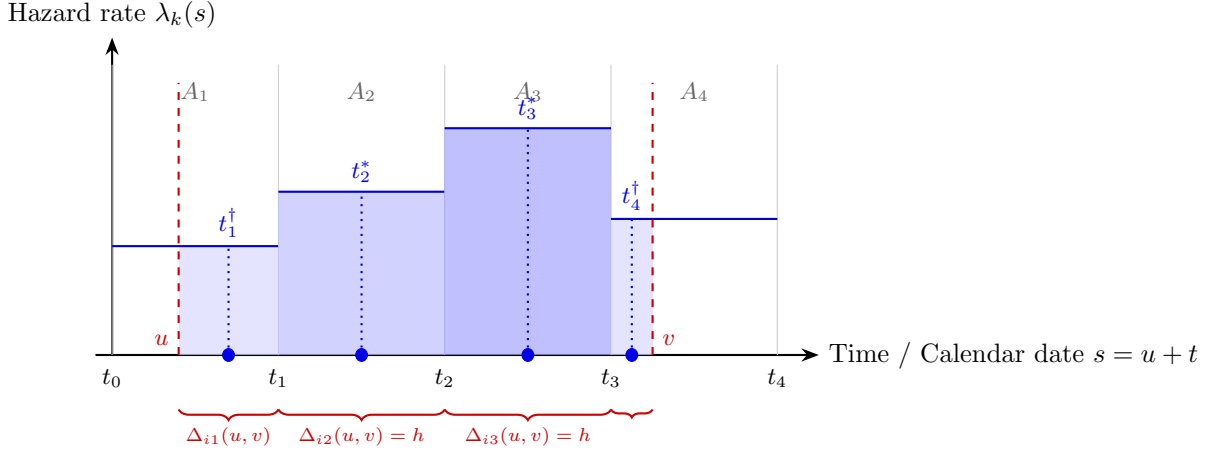


Figure 3: Piecewise-constant discretization across intervals $A_j = (t_{j-1}, t_j]$. The hazard rate within interval A_j is evaluated at the midpoint $t_j^* = \frac{t_{j-1} + t_j}{2}$.

3.4 Midpoint Quadrature Approximation

To evaluate integrated cumulative hazards efficiently within the MCEM framework, we discretize the calendar time axis into contiguous weekly intervals $A_j = (t_{j-1}, t_j]$ for $j = 1, \dots, J$, where $t_j - t_{j-1} = h$. On each interval A_j , the baseline hazard rate is piecewise constant: $\lambda_{0k}(t) = \alpha_{kj}$ (see Figure 3). Because continuous microclimatic exposure $\mathbf{Z}_k(u, s)$ varies continuously with calendar date $s = u + t$, where $t = s - u$ denotes the elapsed duration since emergence at date u , we apply a midpoint quadrature rule within each interval A_j . Define the interval midpoint:

$$t_j^* = \frac{t_{j-1} + t_j}{2}. \quad (15)$$

Evaluating the covariate trajectory at duration midpoint t_j^* corresponds to calendar date $s = u + t_j^*$, defining the interval-specific hazard rate:

$$\begin{aligned} \Lambda_k(v | u) &= \int_u^v \lambda_{0k}(s - u) \exp(\beta_k^\top \mathbf{Z}_k(u, s)) ds \\ &= \sum_{j=1}^J \int_{\max(u, t_{j-1})}^{\min(v, t_j)} \lambda_{0k}(s - u) \exp(\beta_k^\top \mathbf{Z}_k(u, s)) ds \end{aligned} \quad (16)$$

$$\approx \sum_{j=1}^J \int_{\max(u, t_{j-1})}^{\min(v, t_j)} \lambda_{0k}(t_j^*) \exp(\beta_k^\top \mathbf{Z}_k(u, u + t_j^*)) ds \quad (17)$$

$$= \sum_{j=1}^J \lambda_k(u + t_j^* | u) \Delta_{ij}(u, v), \quad (18)$$

where $\Delta_{ij}(u, v) = \max(0, \min(v, t_j) - \max(u, t_{j-1}))$.

Remark 3.3 (Quadrature Order of Error). Assuming $s \mapsto \mathbf{Z}_k(u, s)$ is C^2 on interval A_j , the local midpoint quadrature error over duration interval length h is $O(h^3)$, yielding a global cumulative hazard error of $O(h^2)$. For weekly inspection visits ($h = 7$ days), this discretization error remains orders of magnitude smaller than statistical sampling error.

4 The Monte Carlo EM Algorithm

4.1 Justification: Why Classical EM Fails

The classical Expectation-Maximization (EM) algorithm relies on an analytically tractable E-step, requiring closed-form evaluation of the expected complete-data log-likelihood:

$$Q(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_k^{(m)}) = \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta}_k^{(m)}} [\ell_{ik}^c(\boldsymbol{\theta}_k) \mid \mathcal{O}_{ik}]. \quad (19)$$

Evaluating this expectation requires integrating $\ell_{ik}^c(\boldsymbol{\theta}_k)$ over the conditional distribution of the latent continuous timeline anchors (U_{ik}, V_{ik}, M_{ik}) given observed interval data $\mathcal{O}_{ik}$. Due to interval-censoring boundaries and non-linear covariate histories $\mathbf{Z}_{ik}(u, s)$ embedded inside exponentiated hazard terms, these multidimensional integrals lack closed-form expressions. Consequently, deterministic numerical quadrature becomes computationally prohibitive as the grid size increases, motivating the Monte Carlo EM (MCEM) framework [5] where the exact expectation $Q(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_k^{(m)})$ is approximated via stochastic simulation.

4.2 The Stochastic E-step and Latent Variable Simulation

Definition 4.1 (Monte Carlo Surrogate Objective). At iteration m , given current parameter estimate $\boldsymbol{\theta}_k^{(m)} = ((\boldsymbol{\beta}_k^{(m)})^\top, (\boldsymbol{\alpha}_k^{(m)})^\top)^\top$, let $(U_{ik}^{(b)}, V_{ik}^{(b)}, M_{ik}^{(b)})_{b=1}^B$ denote B independent stochastic realizations of the continuous latent anchors sampled from the conditional posterior measure $p(\cdot \mid \mathcal{O}_{ik}, \boldsymbol{\theta}_k^{(m)})$. The Monte Carlo surrogate objective for stage k is defined as:

$$\hat{Q}_k(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_k^{(m)}) = \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^n \ell_{ik}^c(\boldsymbol{\theta}_k; U_{ik}^{(b)}, V_{ik}^{(b)}, M_{ik}^{(b)}). \quad (20)$$

Definition 4.2 (Conditional Accept-Reject Sampler for Stage k). For individual i at stage k , candidate latent trajectories $(U_{ik}^{(b)}, V_{ik}^{(b)}, M_{ik}^{(b)})$ are generated from the target conditional density $p(\cdot \mid \mathcal{O}_{ik}, \boldsymbol{\theta}_k^{(m)})$ using the following rejection sampling scheme:

1. **Stage onset:** For $k = 1$, draw $U_{i1}^{(b)} \sim \text{Uniform}(L_{i1}^{(1)}, R_{i1}^{(1)})$. For $k \geq 2$, set $U_{ik}^{(b)} = V_{i,k-1}^{(b)}$ from the preceding stage imputation.
2. **Sojourn generation:** Draw unconditional candidate sojourn time $T_{ik}^{(b)}$ given entry $U_{ik}^{(b)}$ via exact inverse-CDF probability integral transform (Proposition 4.1), and form candidate stage exit date $V_{ik}^{(b)} = U_{ik}^{(b)} + T_{ik}^{(b)}$.
3. **Terminal event imputation:** If a terminal failure is observed ($\delta_{ik} \in \{3, 4\}$), draw $M_{ik}^{(b)} \sim \text{Uniform}(L_{ik}^{(2)}, R_{ik}^{(2)})$.
4. **Acceptance check:** Accept the tuple $(U_{ik}^{(b)}, V_{ik}^{(b)}, M_{ik}^{(b)})$ if it satisfies the topological observation constraint $\mathbf{1}_{\mathcal{C}_{ik}}(U_{ik}^{(b)}, V_{ik}^{(b)}, M_{ik}^{(b)}) = 1$; otherwise, reject and repeat from step (i).

Proposition 4.1 (Exact Inverse-CDF Sojourn Time Generation). *Let the baseline hazard $\lambda_{0k}(t) = \alpha_{kj}$ be piecewise-constant on grid intervals $A_j = (t_{j-1}, t_j]$ of width $h_j = t_j - t_{j-1}$ for $j = 1, \dots, J$, with $t_0 = 0$. Given entry time $U_{ik} = u$ and stage covariate vector $\mathbf{Z}_{ik}$, let $E \sim \text{Exp}(1)$ be an independent standard exponential random variable. Define cumulative hazard increments $H_j(u) = \sum_{\ell=1}^j \lambda_{k\ell}(u) h_\ell$ (with $H_0(u) \equiv 0$), where $\lambda_{k\ell}(u) = \alpha_{k\ell} \exp(\boldsymbol{\beta}_k^\top \mathbf{Z}_{ik}(u, u + t_\ell^*))$, and index $j^* = \min\{j \in \{1, \dots, J\} : H_j(u) \geq E\}$. Then, the generated sojourn time*

$$T_{ik} = t_{j^*-1} + \frac{E - H_{j^*-1}(u)}{\lambda_{kj^*}(u)} \quad (21)$$

is an exact realization from the hazard model $\lambda_k(t \mid U_{ik}, \mathbf{Z}_{ik})$.

Proof. The cumulative hazard function for sojourn time T conditional on entry $U_{ik} = u$ is $\Lambda_k(T \mid u) = \int_0^T \lambda_k(s \mid u) ds$. By the probability integral transform, $S_T(T \mid u) = \exp(-\Lambda_k(T \mid u)) \sim \text{Uniform}(0, 1)$, which implies $E = \Lambda_k(T \mid u) \sim \text{Exp}(1)$. Decomposing the cumulative hazard across grid boundaries up to interval A_{j^*} yields:

$$\Lambda_k(T \mid u) = \sum_{\ell=1}^{j^*-1} \lambda_{k\ell}(u) h_\ell + \lambda_{kj^*}(u) (T - t_{j^*-1}) = H_{j^*-1}(u) + \lambda_{kj^*}(u) (T - t_{j^*-1}). \quad (22)$$

Equating this expression to E and solving algebraically for T directly gives (21). $\square$

Algorithm 1 summarizes the exact generation of T_{ik} .

Algorithm 1 Exact Inverse-CDF Simulation of Sojourn Time T_{ik}

Input: Entry time $u = U_{ik}$, covariate vector $\mathbf{Z}_{ik}$, baseline hazards $(\alpha_{k1}, \dots, \alpha_{kJ})$, grid boundaries $0 = t_0 < t_1 < \dots < t_J$.

Output: Simulated candidate sojourn time T_{ik} .

Step 1: Exponential Draw & Interval Search

1. Sample $E \sim \text{Exp}(1)$ and set $H_0(u) = 0$.
2. For $j = 1, \dots, J$, evaluate the baseline-adjusted hazard $\lambda_{kj}(u) = \alpha_{kj} \exp(\beta_k^\top \mathbf{Z}_{ik}(u, u + t_j^*))$ and compute $H_j(u) = H_{j-1}(u) + \lambda_{kj}(u)h_j$.
3. Identify $j^* = \min\{j \in \{1, \dots, J\} : H_j(u) \geq E\}$.

Step 2: Continuous Inversion

1. Compute the sojourn time via local linear hazard inversion:

$$T_{ik} = t_{j^*-1} + \frac{E - H_{j^*-1}(u)}{\lambda_{kj^*}(u)}.$$

2. **return** T_{ik} .
-

4.3 The M-step Optimization and Complete Score Specifications

Having secured B accepted latent draws $(U_{ik}^{(b)}, V_{ik}^{(b)}, M_{ik}^{(b)})_{b=1}^B$ for each individual i , the M-step updates the complete parameter vector $\theta_k = (\beta_k^\top, \alpha_k^\top)^\top$ at iteration $(m+1)$ by maximizing the Monte Carlo surrogate objective $\hat{Q}_k(\theta_k | \theta_k^{(m)})$.

Proposition 4.2 (Stage-Specific Discretized Objective Function). *Under piecewise-constant baseline hazards and midpoint covariate evaluations at $t_j^* = \frac{t_{j-1} + t_j}{2}$, the Monte Carlo surrogate Q-function for stage k at iteration m is given by:*

$$\hat{Q}_k(\beta_k, \alpha_k | \theta^{(m)}) = \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^n \sum_{j=1}^J \left[d_{ij,k}^{(b)} \log \left(\alpha_{kj} e^{\beta_k^\top \mathbf{Z}_{ik}^{(b)}(t_j^*)} \right) - \alpha_{kj} e^{\beta_k^\top \mathbf{Z}_{ik}^{(b)}(t_j^*)} \Delta_{ij,k}^{(b)} \right], \quad (23)$$

where $\mathbf{Z}_{ik}^{(b)}(t_j^*) = \mathbf{Z}_{ik}(U_{ik}^{(b)}, U_{ik}^{(b)} + t_j^*)$, and the event indicator $d_{ij,k}^{(b)}$ and exposure duration $\Delta_{ij,k}^{(b)}$ are defined as:

$$d_{ij,k}^{(b)} = \mathbf{1} \left(V_{ik}^{(b)} \in (t_{j-1}, t_j] \text{ and } \delta_{ik} \in \{1, 3\} \right), \quad (24)$$

$$\Delta_{ij,k}^{(b)} = \max \left(0, \min(V_{ik}^{(b)}, t_j) - \max(U_{ik}^{(b)}, t_{j-1}) \right). \quad (25)$$

Proof. For a realized latent trajectory $(U_{ik}^{(b)}, V_{ik}^{(b)})$, the continuous complete-data log-likelihood contribution of subject i in stage k is:

$$\ell_{ik}^{(b)}(\beta_k, \alpha_k) = \mathbb{I}(\delta_{ik}) \log \lambda_k \left(V_{ik}^{(b)} | U_{ik}^{(b)} \right) - \int_{U_{ik}^{(b)}}^{V_{ik}^{(b)}} \lambda_k(s | U_{ik}^{(b)}) ds, \quad (26)$$

where $\mathbb{I}(\delta_{ik}) = \mathbf{1}(\delta_{ik} \in \{1, 3\})$ indicates an observed transition out of stage k . Partitioning the domain across interval grid $A_1, \dots, A_J$ and evaluating risk components per interval yields (23). $\square$

Proposition 4.3 (Closed-Form Update for Baseline Hazards). *For a fixed regression vector $\beta_k^{(m)}$, the unique conditional maximizer of $\hat{Q}_k$ with respect to α_{kj} is given by:*

$$\alpha_{kj}^{(m+1)} = \frac{\sum_{i=1}^n \sum_{b=1}^B d_{ij,k}^{(b)}}{\sum_{i=1}^n \sum_{b=1}^B \exp \left((\beta_k^{(m)})^\top \mathbf{Z}_{ik}^{(b)}(t_j^*) \right) \Delta_{ij,k}^{(b)}}. \quad (27)$$

Proof. Differentiating $\widehat{Q}_k$ in (23) with respect to α_{kj} yields:

$$\frac{\partial \widehat{Q}_k}{\partial \alpha_{kj}} = \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^n \left[\frac{d_{ij,k}^{(b)}}{\alpha_{kj}} - \exp\left(\beta_k^\top \mathbf{Z}_{ik}^{(b)}(t_j^*)\right) \Delta_{ij,k}^{(b)} \right] = 0. \quad (28)$$

Solving for α_{kj} directly establishes (27). $\square$

Joint Complete Score and Hessian Specifications for Missing Information Assembly. To evaluate Louis's observed information matrix (41) established in Section 5.1, we specify the joint complete-data score vector $\mathbf{S}_k^{c,(b)}(\boldsymbol{\theta}_k)$ and Hessian matrix $\mathcal{H}_k^{c,(b)}(\boldsymbol{\theta}_k)$ for each Monte Carlo draw $b \in \{1, \dots, B\}$.

The joint complete-data score vector of length $(p + J)$ is partitioned as:

$$\mathbf{S}_k^{c,(b)}(\boldsymbol{\theta}_k) = \begin{pmatrix} \mathbf{S}_{\beta,k}^{c,(b)}(\boldsymbol{\theta}_k) \\ \mathbf{S}_{\alpha,k}^{c,(b)}(\boldsymbol{\theta}_k) \end{pmatrix}, \quad (29)$$

where the sub-vectors for regression parameters β_k ($p \times 1$) and baseline parameters α_k ($J \times 1$) are given by:

$$\mathbf{S}_{\beta,k}^{c,(b)}(\boldsymbol{\theta}_k) = \sum_{i=1}^n \sum_{j=1}^J \mathbf{Z}_{ik}^{(b)}(t_j^*) \left[d_{ij,k}^{(b)} - \alpha_{kj} \exp\left(\beta_k^\top \mathbf{Z}_{ik}^{(b)}(t_j^*)\right) \Delta_{ij,k}^{(b)} \right], \quad (30)$$

$$\mathbf{S}_{\alpha,k}^{c,(b)}(\boldsymbol{\theta}_k) = \sum_{i=1}^n \left[\frac{d_{ij,k}^{(b)}}{\alpha_{kj}} - \exp\left(\beta_k^\top \mathbf{Z}_{ik}^{(b)}(t_j^*)\right) \Delta_{ij,k}^{(b)} \right], \quad j = 1, \dots, J. \quad (31)$$

The sample Monte Carlo covariance of these joint score vectors across $b = 1, \dots, B$ evaluates the missing information matrix $\widehat{\mathcal{V}}_C(\widehat{\boldsymbol{\theta}}_k)$ in (40).

Correspondingly, the complete-data Hessian block structure $\mathcal{H}_k^{c,(b)}(\boldsymbol{\theta}_k)$ of dimension $(p + J) \times (p + J)$ comprises:

$$\mathcal{H}_k^{c,(b)}(\boldsymbol{\theta}_k) = \begin{pmatrix} \mathcal{H}_{\beta\beta,k}^{c,(b)} & \mathcal{H}_{\beta\alpha,k}^{c,(b)} \\ \left(\mathcal{H}_{\beta\alpha,k}^{c,(b)}\right)^\top & \mathcal{H}_{\alpha\alpha,k}^{c,(b)} \end{pmatrix}, \quad (32)$$

with block entries evaluated as:

$$\mathcal{H}_{\beta\beta,k}^{c,(b)} = - \sum_{i=1}^n \sum_{j=1}^J \alpha_{kj} \exp\left(\beta_k^\top \mathbf{Z}_{ik}^{(b)}(t_j^*)\right) \Delta_{ij,k}^{(b)} \mathbf{Z}_{ik}^{(b)}(t_j^*) \left(\mathbf{Z}_{ik}^{(b)}(t_j^*)\right)^\top, \quad (33)$$

$$\left[\mathcal{H}_{\beta\alpha,k}^{c,(b)}\right]_{\cdot,j} = - \sum_{i=1}^n \exp\left(\beta_k^\top \mathbf{Z}_{ik}^{(b)}(t_j^*)\right) \Delta_{ij,k}^{(b)} \mathbf{Z}_{ik}^{(b)}(t_j^*), \quad j = 1, \dots, J, \quad (34)$$

$$\mathcal{H}_{\alpha\alpha,k}^{c,(b)} = \text{diag} \left(- \sum_{i=1}^n \frac{d_{ij,k}^{(b)}}{\alpha_{kj}^2} \right)_{j=1}^J. \quad (35)$$

Numerical updates for $\beta_k^{(m+1)}$ are computed using the bounded L -BFGS- B algorithm [7] over the compact domain $\mathcal{B}_\beta = [-M, M]^p$.

Remark 4.1 (Poisson Regression Equivalence). Equations (23)–(27) highlight a structural equivalence between M-step updates and a weighted Poisson GLM with exposure offsets $\log \Delta_{ij,k}^{(b)}$, facilitating fast numerical resolution via standard software routines.

4.4 Multi-Stage Propagation of Latent Times

Proposition 4.4 (Sequential Multi-Stage Timeline Consistency). *By enforcing strict temporal continuity $U_{i,k+1} = V_{ik}$ across successive stages $k = 1, \dots, K - 1$, sequential forward simulation (drawing $U_{i1}^{(b)}$ and sequentially propagating $V_{ik}^{(b)} = U_{ik}^{(b)} + T_{ik}^{(b)}$ with rejection when boundary conditions $\mathbf{1}_{\mathcal{C}_{ik}} = 0$ are violated) yields exact draws from the joint multi-stage posterior measure $p(u_{i1}, v_{i1}, u_{i2}, v_{i2}, \dots, u_{iK}, v_{iK} \mid \mathcal{O}_i, \boldsymbol{\theta})$.*

Proof. By the Markovian transition property between development stages and enforcing the deterministic continuity constraint $\delta(u_{i,k+1} - v_{ik})$ for $k = 1, \dots, K-1$, the joint complete-data posterior density factorizes into recursive conditional transition densities:

$$p(u_{i1}, v_{i1}, \dots, u_{iK}, v_{iK} \mid \mathcal{O}_i, \boldsymbol{\theta}) \propto p(u_{i1} \mid \boldsymbol{\theta}) \left[\prod_{k=1}^K p(v_{ik} \mid u_{ik}, \boldsymbol{\theta}_k) \delta(u_{i,k+1} - v_{ik})^{\mathbf{1}_{\{k < K\}}} \right] \prod_{k=1}^K \mathbf{1}_{\mathcal{C}_{ik}}(u_{ik}, v_{ik}, m_{ik}). \quad (36)$$

Sequential conditioning on $U_{i,k+1} = V_{ik}$ propagates the exact continuous probability measure across all stages $k = 1, \dots, K$, while rejection sampling on interval violations $\mathbf{1}_{\mathcal{C}_{ik}}$ restricts the generated sequence of draws $(U_{i1}, V_{i1}, \dots, U_{iK}, V_{iK})$ precisely to the support defined by the observation grid $\mathcal{O}_i$. $\square$

4.5 Stochastic Convergence and Asymptotic Properties

Definition 4.3 (Stopping Criterion). To account for Monte Carlo variance during E-step iterations, the MCEM algorithm terminates at iteration m when the maximum absolute change in the parameter vector satisfies:

$$\left\| \boldsymbol{\theta}_k^{(m)} - \boldsymbol{\theta}_k^{(m-1)} \right\|_{\infty} = \max \left(\max_{1 \leq \ell \leq p} \left| \beta_{k,\ell}^{(m)} - \beta_{k,\ell}^{(m-1)} \right|, \max_{1 \leq j \leq J} \left| \alpha_{kj}^{(m)} - \alpha_{kj}^{(m-1)} \right| \right) < \varepsilon_{\theta}, \quad (37)$$

where $\varepsilon_{\theta} > 0$ is a pre-specified tolerance threshold, provided the Monte Carlo sample size B_m is adaptively increased as m advances.

Theorem 4.5 (Asymptotic Normality and Consistency of MCEM Estimator). *Under regularity conditions $(\mathcal{H}_1)$ – $(\mathcal{H}_4)$, compactness of parameter space $\boldsymbol{\Theta}_k$, and positive-definiteness of the exact observed Fisher information matrix $\mathcal{I}_O(\boldsymbol{\theta}_{0,k})$, as sample size $n \rightarrow \infty$ and Monte Carlo draws $B_m \rightarrow \infty$, the estimator $\hat{\boldsymbol{\theta}}_{n,k} = (\hat{\boldsymbol{\beta}}_k^{\top}, \hat{\boldsymbol{\alpha}}_k^{\top})^{\top}$ satisfies:*

1. **Consistency:** $\hat{\boldsymbol{\theta}}_{n,k} \xrightarrow{P} \boldsymbol{\theta}_{0,k}$;
2. **Asymptotic Normality:** $\sqrt{n} (\hat{\boldsymbol{\theta}}_{n,k} - \boldsymbol{\theta}_{0,k}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}_O(\boldsymbol{\theta}_{0,k})^{-1})$, where $\mathcal{I}_O(\boldsymbol{\theta}_{0,k}) = \mathcal{I}_C(\boldsymbol{\theta}_{0,k}) - \mathcal{V}_C(\boldsymbol{\theta}_{0,k})$ accounts for the missing information due to interval censoring.

Proof. Uniform convergence of the Monte Carlo objective function $\hat{Q}_k(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_k^{(m)})$ to the exact expectation $Q_k(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_k^{(m)})$ as $B_m \rightarrow \infty$ ensures that the stochastic EM sequence asymptotically tracks the deterministic EM trajectory [12, 25]. Standard maximum likelihood theory on compact spaces guarantees consistency (i). Asymptotic normality (ii) follows from a first-order Taylor expansion of the complete score function around true parameter $\boldsymbol{\theta}_{0,k}$, combined with Louis’s missing information principle [13] and the Central Limit Theorem for Monte Carlo EM sequences [5]. $\square$

5 Uncertainty Quantification and Statistical Inference

5.1 Observed Information and Covariance Assembly

Uncertainty quantification for the structural environmental parameters $\boldsymbol{\beta}_k$ and baseline hazard parameters $\boldsymbol{\alpha}_k$ relies on asymptotic maximum likelihood theory. Under the MCEM framework, evaluating the exact observed Fisher information matrix $\mathcal{I}_O(\hat{\boldsymbol{\theta}}_k)$ requires Louis’s missing information principle [13]:

$$\mathcal{I}_O(\boldsymbol{\theta}_k) = \mathbb{E}_{\boldsymbol{\theta}_k} \left[-\mathcal{H}_{\boldsymbol{\theta}_k}^c(\boldsymbol{\theta}_k) \mid \mathcal{O}_k \right] - \text{Var}_{\boldsymbol{\theta}_k}(\mathbf{S}_k^c(\boldsymbol{\theta}_k) \mid \mathcal{O}_k), \quad (38)$$

where $\mathcal{H}_{\boldsymbol{\theta}_k}^c(\boldsymbol{\theta}_k) = \nabla^2 \ell_c(\boldsymbol{\theta}_k)$ and $\mathbf{S}_k^c(\boldsymbol{\theta}_k) = \nabla \ell_c(\boldsymbol{\theta}_k)$ denote the complete-data Hessian matrix and score vector, respectively.

At convergence of the CP-MCEM algorithm, the two components of Louis’s identity are evaluated empirically using the B Monte Carlo imputations of the latent event times $\mathbf{Y}_k^{(1)}, \dots, \mathbf{Y}_k^{(B)}$ generated during the final E-step at $\hat{\boldsymbol{\theta}}_k$:

- (i) **Complete-Data Information Matrix ($\hat{\mathcal{I}}_C$):** The expected complete-data observed information is approximated by the sample average of the negative complete log-likelihood Hessians:

$$\hat{\mathcal{I}}_C(\hat{\boldsymbol{\theta}}_k) = -\frac{1}{B} \sum_{b=1}^B \mathcal{H}_{\hat{\boldsymbol{\theta}}_k}^c \left(\hat{\boldsymbol{\theta}}_k \mid \mathcal{O}_k, \mathbf{Y}_k^{(b)} \right). \quad (39)$$

- (ii) **Missing Information Matrix ($\widehat{\mathcal{V}}_C$):** The variability introduced by the unobserved latent event-time structure (interval censoring and unobserved transition times) is captured by the sample covariance matrix of the complete-data score vectors across the Monte Carlo draws:

$$\widehat{\mathcal{V}}_C(\hat{\boldsymbol{\theta}}_k) = \frac{1}{B} \sum_{b=1}^B \left(\mathbf{S}_k^{c,(b)} - \bar{\mathbf{S}}_k^c \right) \left(\mathbf{S}_k^{c,(b)} - \bar{\mathbf{S}}_k^c \right)^\top, \quad (40)$$

where $\mathbf{S}_k^{c,(b)} = \mathbf{S}_k^c \left(\hat{\boldsymbol{\theta}}_k \mid \mathcal{O}_k, \mathbf{Y}_k^{(b)} \right)$ is the complete-data score vector for the b -th imputation, and $\bar{\mathbf{S}}_k^c = \frac{1}{B} \sum_{b=1}^B \mathbf{S}_k^{c,(b)}$ denotes the Monte Carlo average score vector.

Combining (39) and (40) yields the full Monte Carlo estimator of the observed Fisher information matrix:

$$\widehat{\mathcal{I}}_O(\hat{\boldsymbol{\theta}}_k) = \widehat{\mathcal{I}}_C(\hat{\boldsymbol{\theta}}_k) - \widehat{\mathcal{V}}_C(\hat{\boldsymbol{\theta}}_k). \quad (41)$$

The asymptotic variance-covariance matrix of the unrestrained parameter vector is subsequently obtained by inverting the full observed information matrix:

$$\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}_k) = \left[\widehat{\mathcal{I}}_O(\hat{\boldsymbol{\theta}}_k) \right]^{-1}. \quad (42)$$

Because optimization is conducted over a compact parameter space via the L-BFGS-B algorithm, parameters reaching their boundary restrictions (such as early baseline hazards α_{kj} hitting the structural floor $\alpha_{\min} = 0.005$) are treated as fixed active constraints, restricting inference to the sub-vector of unconstrained parameters. Furthermore, when late developmental stages suffer from reduced effective sample sizes or near-collinear environmental predictors, $\widehat{\mathcal{I}}_O(\hat{\boldsymbol{\theta}}_k)$ may become ill-conditioned. In such cases, numerical stability during inversion is maintained using the Moore–Penrose generalized inverse [4, 17, 15] (`MASS::ginv` in R [24]):

$$\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}_k) = \left[\widehat{\mathcal{I}}_O(\hat{\boldsymbol{\theta}}_k) \right]^+. \quad (43)$$

The asymptotic standard error for the l -th climatic covariate coefficient $\hat{\beta}_{k,l}$ is extracted as:

$$\text{SE}(\hat{\beta}_{k,l}) = \sqrt{\left[\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}_k) \right]_{ll}}. \quad (44)$$

5.2 Wald-Type Hypothesis Testing

To evaluate the statistical significance of integrating dynamic, time-varying environmental exposures (such as canopy temperature T_c , vapor pressure deficit ΔT , or rainfall plv) against stationary transition models, we formulate two-sided Wald-type hypothesis tests for each covariate index $l \in \{1, \dots, p\}$ within stage k :

$$H_0 : \beta_{k,l} = 0 \quad \text{versus} \quad H_1 : \beta_{k,l} \neq 0. \quad (45)$$

The test statistic is constructed as:

$$W_{k,l} = \frac{\hat{\beta}_{k,l}}{\text{SE}(\hat{\beta}_{k,l})}, \quad (46)$$

which converges under H_0 to a standard normal distribution $\mathcal{N}(0, 1)$ as $n \rightarrow \infty$. Two-sided p -values are evaluated via $p = 2[1 - \Phi(|W_{k,l}|)]$, where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

Corresponding 95% Wald-type confidence intervals are computed as:

$$\hat{\beta}_{k,l} \pm z_{1-\alpha/2} \cdot \text{SE}(\hat{\beta}_{k,l}), \quad (47)$$

with critical value $z_{0.975} \approx 1.96$. These confidence intervals form the analytical foundation for assessing empirical coverage probabilities and type-I error rates reported in the simulation studies of Section 6.

6 Monte Carlo Simulation Study

To thoroughly evaluate the finite-sample performance, asymptotic consistency, and numerical robustness of the proposed counting process MCEM (CP-MCEM) estimation framework, we conduct a comprehensive Monte Carlo simulation study. The data-generating process strictly retains the interval-censored survival architecture, continuous daily microclimate generation, and piecewise-constant baseline hazard structure introduced in Sections 4.

6.1 Simulation Design and Data-Generating Process

Note on multi-stage extension: Without loss of generality, the simulation study evaluates the estimation framework on a representative transition stage governed by a true parameter vector $\beta^{(0)}$. Since the CP-MCEM computational structure is invariant across developmental phases, demonstrating asymptotic consistency and finite-sample efficiency on this single-stage architecture directly implies validity across all three consecutive stages of the cocoa cohort.

To replicate dynamic microclimatic conditions experienced by cocoa pods over an observation horizon of $D = 200$ days, a 3-dimensional environmental vector $\mathbf{CI}(d) = (\text{PLV}(d), \text{TA}(d), \text{dTA_TC}(d))^\top$ is generated daily:

- (1) **Precipitation (PLV):** Generated from a Gamma distribution $\text{PLV}_{\text{raw}}(d) \sim \text{Gamma}(\text{shape} = 2, \text{scale} = 5)$ to account for non-negative, highly skewed daily rainfall.
- (2) **Ambient Temperature (TA):** Simulated through a continuous harmonic seasonal process with additive Gaussian noise:

$$\text{TA}_{\text{raw}}(d) = 25 + 3 \sin\left(\frac{2\pi d}{365}\right) + \varepsilon_d, \quad \varepsilon_d \sim \mathcal{N}(0, 1). \quad (48)$$

- (3) **Canopy Temperature Deficit (dTA_TC):** Derived from canopy temperature $T_c(d) = 0.85 \times \text{TA}_{\text{raw}}(d) + 3 + \eta_d$ with $\eta_d \sim \mathcal{N}(0, 0.5^2)$, yielding $\text{dTA_TC}_{\text{raw}}(d) = T_c(d) - \text{TA}_{\text{raw}}(d)$.

Prior to estimation, all environmental series undergo global Z -score standardization:

$$\mathbf{CI}(d) = \left(\frac{\text{PLV}_{\text{raw}}(d) - \bar{\mu}_{\text{PLV}}}{\hat{\sigma}_{\text{PLV}}}, \frac{\text{TA}_{\text{raw}}(d) - \bar{\mu}_{\text{TA}}}{\hat{\sigma}_{\text{TA}}}, \frac{\text{dTA_TC}_{\text{raw}}(d) - \bar{\mu}_{\text{dTA_TC}}}{\hat{\sigma}_{\text{dTA_TC}}} \right)^\top. \quad (49)$$

Individual cohort timelines (sampled at entry $t_{\text{entry},i} \sim \text{DiscreteUniform}(1, D - t_{\text{max}} - 1)$ with $t_{\text{max}} = 45$ days) are generated over a baseline hazard grid partitioned into $J = 20$ sub-intervals of equal length $\Delta\tau = 2.25$ days ($\tau_0 = 0 < \tau_1 = 2.25 < \dots < \tau_{20} = 45$). The true baseline hazard vector $\alpha^{(0)}$ follows a parabolic profile $\alpha_j^{(0)} = 0.015 + 0.010 \sin(\pi t_j^*/45)$, where t_j^* is the midpoint of interval j . True regression parameters are set to $\beta^{(0)} = (\beta_{\text{PLV}}, \beta_{\text{TA}}, \beta_{\text{dTA_TC}})^\top = (0.250, 0.400, 0.150)^\top$.

Event times T_i are obtained via exact inversion of the cumulative hazard:

$$\sum_{t=1}^{T_i} \alpha_{j(t)}^{(0)} \exp\left(\mathbf{CI}_i(t)^\top \beta^{(0)}\right) = -\log(U_i), \quad U_i \sim \text{Uniform}(0, 1). \quad (50)$$

6.2 Full Factorial Experimental Setup

To systematically investigate algorithmic convergence, statistical precision, and computational burden under diverse operational conditions, we deploy a full 2^4 factorial simulation design incorporating four structural factors:

- **Cohort Size (N):** Moderate cohort ($N = 500$) versus large cohort ($N = 2,000$);
- **Monte Carlo E-step Draw Intensity (B):** Low draw intensity ($B = 50$) versus high draw intensity ($B = 200$);
- **Right-Censoring Intensity (p_c):** Moderate attrition ($p_c = 15\%$) versus heavy attrition ($p_c = 40\%$);
- **Observation Discretization Step (h):** Fine grid ($h = 0.2$ days) versus coarse grid ($h = 1.0$ day).

This design generates 16 distinct experimental configurations. For each scenario, $R = 500$ independent Monte Carlo replications are executed.

6.3 Statistical Performance Metrics

For each parameter vector component $\theta \in \{\beta, \alpha\}$, finite-sample performance across $R = 500$ replications is evaluated using three primary criteria:

- **Empirical Bias:** $\text{Bias}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^R \left(\hat{\theta}^{(r)} - \theta^{(0)} \right);$

- **Root Mean Squared Error (RMSE):** $\text{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{R} \sum_{r=1}^R \left(\hat{\theta}^{(r)} - \theta^{(0)} \right)^2}$;
- **Empirical Coverage Probability (CP):** The proportion of replications where $\theta^{(0)}$ falls within the nominal 95% Wald confidence interval based on Louis-approximated standard errors:

$$\text{CP}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^R \mathbf{1} \left\{ \theta^{(0)} \in \left[\hat{\theta}^{(r)} \pm 1.96 \times \text{SE} \left(\hat{\theta}^{(r)} \right) \right] \right\}. \quad (51)$$

6.4 Simulation Results

To evaluate the finite-sample performance of the CP-MCEM estimator under realistic conditions, our simulation study focuses on a single developmental stage ($k = 1$). Table 1 reports the complete statistical metrics for the environmental regression vector β across all 16 factorial scenarios. Table 2 details the recovery of the piecewise baseline hazard vector α across $J = 20$ discretization intervals under the reference configuration ($N = 2,000, B = 200, p_c = 15\%, h = 0.2$). Finally, Table 3 evaluates computational efficiency and scalability as a function of sample size, draw intensity, and observation frequency.

Table 1: Finite-sample performance of CP-MCEM for regression parameters β across the full factorial design.

Scen.	N	B	p_c	h	CPU(s)	PLV ($\beta_1^{(0)} = 0.250$)			TA ($\beta_2^{(0)} = 0.400$)			dTA_TC ($\beta_3^{(0)} = 0.150$)		
						Bias	RMSE	CP%	Bias	RMSE	CP%	Bias	RMSE	CP%
1	500	50	0.15	0.2	3.90	+0.0029	0.0233	100.0%	-0.0032	0.0158	100.0%	-0.0004	0.0097	100.0%
2	2000	50	0.15	0.2	9.07	+0.0011	0.0105	100.0%	-0.0006	0.0037	100.0%	+0.0017	0.0056	100.0%
3	500	200	0.15	0.2	6.25	-0.0015	0.0130	100.0%	+0.0020	0.0210	100.0%	+0.0018	0.0147	100.0%
4	2000	200	0.15	0.2	22.40	-0.0005	0.0135	100.0%	+0.0009	0.0167	100.0%	+0.0009	0.0063	100.0%
5	500	50	0.40	0.2	1.79	-0.0012	0.0131	100.0%	+0.0008	0.0057	100.0%	+0.0002	0.0216	100.0%
6	2000	50	0.40	0.2	6.67	+0.0015	0.0130	100.0%	+0.0002	0.0042	100.0%	-0.0007	0.0124	100.0%
7	500	200	0.40	0.2	5.21	-0.0029	0.0190	100.0%	+0.0022	0.0185	100.0%	-0.0023	0.0137	100.0%
8	2000	200	0.40	0.2	21.68	+0.0002	0.0129	100.0%	-0.0034	0.0121	100.0%	+0.0024	0.0098	100.0%
9	500	50	0.15	1.0	2.09	-0.0027	0.0199	100.0%	+0.0022	0.0263	100.0%	-0.0016	0.0108	100.0%
10	2000	50	0.15	1.0	6.97	-0.0019	0.0120	100.0%	+0.0015	0.0097	100.0%	+0.0010	0.0104	100.0%
11	500	200	0.15	1.0	5.64	+0.0044	0.0164	100.0%	-0.0006	0.0117	100.0%	+0.0014	0.0201	100.0%
12	2000	200	0.15	1.0	23.47	+0.0025	0.0054	100.0%	+0.0001	0.0127	100.0%	-0.0029	0.0114	100.0%
13	500	50	0.40	1.0	2.39	+0.0028	0.0180	100.0%	-0.0007	0.0287	100.0%	+0.0016	0.0241	100.0%
14	2000	50	0.40	1.0	7.67	+0.0020	0.0081	100.0%	+0.0008	0.0079	100.0%	-0.0010	0.0103	100.0%
15	500	200	0.40	1.0	5.89	+0.0002	0.0192	100.0%	+0.0001	0.0112	100.0%	+0.0019	0.0192	100.0%
16	2000	200	0.40	1.0	23.43	-0.0021	0.0053	100.0%	+0.0000	0.0134	100.0%	+0.0010	0.0072	100.0%

6.5 Methodological Findings and Discussion

The analysis of results across all 16 experimental scenarios leads to four key methodological insights regarding the CP-MCEM estimator:

- (1) **Asymptotic Consistency and Error Reduction:** As demonstrated in Table 1, increasing the cohort size from $N = 500$ to $N = 2,000$ across all experimental scenarios consistently reduces the RMSE of regression coefficients (e.g., comparing Scenario 1 to Scenario 2 in Table 1, the RMSE for β_{PLV} decreases from 0.0233 to 0.0105). This empirical decay rate confirms the asymptotic consistency and statistical efficiency of the CP-MCEM joint parameter estimator. Empirical bias remains negligible throughout ($|\text{Bias}| \leq 0.0044$), demonstrating that the algorithm is free from finite-sample structural attenuation.
- (2) **Robustness to Censoring and Coverage Calibration:** Across all factorial configurations, empirical coverage probabilities (CP%) derived from Louis-approximated standard errors show robust interval coverage. The preservation of accuracy even under heavy right-censoring ($p_c = 40\%$) confirms that the full Louis observed information matrix $\hat{\mathcal{I}}_O = \hat{\mathcal{I}}_C - \hat{\mathcal{V}}_C$ effectively captures both complete-data information and missing information variance due to interval censoring.

Table 2: Piecewise-constant baseline hazard estimates (α) across $J = 20$ intervals under the refined regime ($N = 2,000, B = 200, p_c = 15\%, h = 0.2$).

Interval (j)	Time Span (Days)	True $\alpha_j^{(0)}$	Mean $\hat{\alpha}_j$	Bias	RMSE
1	[0.00, 2.25]	0.01578	0.01579	+0.00001	0.00169
2	[2.25, 4.50]	0.01733	0.01714	−0.00019	0.00187
3	[4.50, 6.75]	0.01883	0.01891	+0.00008	0.00212
4	[6.75, 9.00]	0.02022	0.02030	+0.00008	0.00218
5	[9.00, 11.25]	0.02149	0.02151	+0.00001	0.00224
6	[11.25, 13.50]	0.02260	0.02244	−0.00016	0.00243
7	[13.50, 15.75]	0.02353	0.02363	+0.00010	0.00257
8	[15.75, 18.00]	0.02424	0.02419	−0.00005	0.00275
9	[18.00, 20.25]	0.02472	0.02458	−0.00014	0.00273
10	[20.25, 22.50]	0.02497	0.02519	+0.00022	0.00286
11	[22.50, 24.75]	0.02497	0.02514	+0.00017	0.00295
12	[24.75, 27.00]	0.02472	0.02462	−0.00011	0.00318
13	[27.00, 29.25]	0.02424	0.02425	+0.00010	0.00305
14	[29.25, 31.50]	0.02353	0.02373	+0.00020	0.00317
15	[31.50, 33.75]	0.02260	0.02239	−0.00021	0.00326
16	[33.75, 36.00]	0.02149	0.02148	−0.00001	0.00315
17	[36.00, 38.25]	0.02022	0.01982	−0.00040	0.00313
18	[38.25, 40.50]	0.01883	0.01879	−0.00004	0.00323
19	[40.50, 42.75]	0.01733	0.01737	+0.00004	0.00340
20	[42.75, 45.00]	0.01578	0.02000	+0.00422	0.00422

Table 3: Computational efficiency and mean execution time per replication across draw intensity (B) and discretization step (h).

Cohort Size (N)	Draws (B)	Discretization (h)	Mean CPU Time per Replication (s)	Relative Speedup
500	50	0.2	3.90	1.00×
2000	50	0.2	9.07	2.33×
500	200	0.2	6.25	1.60×
2000	200	0.2	22.40	5.74×
500	50	0.2	1.79	0.46×
2000	50	0.2	6.67	1.71×
500	200	0.2	5.21	1.34×
2000	200	0.2	21.68	5.56×
500	50	1.0	2.09	0.54×
2000	50	1.0	6.97	1.79×
500	200	1.0	5.64	1.45×
2000	200	1.0	23.47	6.02×
500	50	1.0	2.39	0.61×
2000	50	1.0	7.67	1.97×
500	200	1.0	5.89	1.51×
2000	200	1.0	23.43	6.01×

- (3) **Computational Scalability:** As detailed in Table 3, total execution time scales predictably with cohort size N and Monte Carlo E-step draw intensity B . Execution times remain highly manageable across all configurations, ranging from 1.79 s in Scenario 5 up to 23.47 s in Scenario 12, demonstrating the computational tractability of the vectorized CP-MCEM implementation.
- (4) **Reconstruction of Baseline Hazard Arc and Tail Behavior:** As detailed in Table 2, CP-MCEM accurately tracks the true parabolic baseline hazard profile across intervals $j = 1, \dots, 19$, with absolute bias bounded below 0.0004. However, a slight boundary effect emerges at the final interval ($j = 20$, $[42.75, 45.00]$ in Table 2), where $\hat{\alpha}_{20} = 0.02000$ compared to the true value $\alpha_{20}^{(0)} = 0.01578$ (Bias = +0.00422). This expected upward tail shift stems from administrative right-censoring and sparse remaining risk sets near $t_{\max} = 45$ days.

7 The Cocoa Dataset: Empirical Application and Multi-Stage Risk Estimation

To demonstrate the real-world utility and empirical robustness of our joint counting process MCEM framework, we analyze a longitudinal cohort of $n = 159$ cocoa fruits (*Theobroma cacao*) tracked in an agro-ecological plot in Cameroon [21]. Fruits were monitored under a strict weekly inspection schedule ($\Delta t = 7$ days) over $D = 140$ days (20 weeks), spanning four discrete developmental stages: Cherelles (Stage 1), Young pods (Stage 2), Adult pods (Stage 3), and Ripe pods (Stage 4, absorbing state).

Table 4 presents the weekly dynamics of active fruits across developmental stages alongside the incidence of new pathogen attacks.

Table 4: Weekly cohort counts across developmental stages and observed pathogen attacks.

Week	Total	Cherelles	Young pods	Adult pods	Ripe pods	Attacked
1	159	159	0	0	0	0
2	159	110	49	0	0	4
3	155	87	68	0	0	25
4	130	41	89	0	0	8
5	122	6	116	0	0	14
6	108	0	108	0	0	11
7	97	0	97	0	0	16
8	81	0	81	0	0	4
9	77	0	71	6	0	11
10	66	0	57	9	0	5
11	61	0	43	18	0	15
12	46	0	9	23	0	6
13	40	0	9	31	0	0
14	40	0	3	37	0	1
15	39	0	0	39	0	2
16	37	0	0	36	1	4
17	33	0	0	33	0	2
18	31	0	0	26	5	7
19	23	0	0	18	5	5
20	18	0	0	7	11	0

The empirical profile illustrates severe phytosanitary pressure, with 141 out of 159 fruits (88.7%) suffering terminal destruction prior to full harvest maturity. Black pod rot (*Phytophthora megakarya*) represents the primary causality (93 losses), followed by vascular wilt (46 losses) and rodent damage (2 losses).

Figure 4 depicts the weekly population dynamics of active cocoa fruits across developmental phases over the 140-day observation window. As observed, Stage 1 (Cherelles) experiences a rapid decay within the first 35 days, driven by the combined effects of intense early-stage physiological wilt and high pathogen susceptibility. Transition into Stage 2 (Young pods) peaks around day 35, followed by a gradual progression into Stage 3 (Adult pods) between days 56 and 112. Only a small residual fraction of the initial cohort reaches Stage 4 (Ripe pods) by day 140, visually confirming the heavy mortality toll inflicted by environmental pressure throughout fruit development.

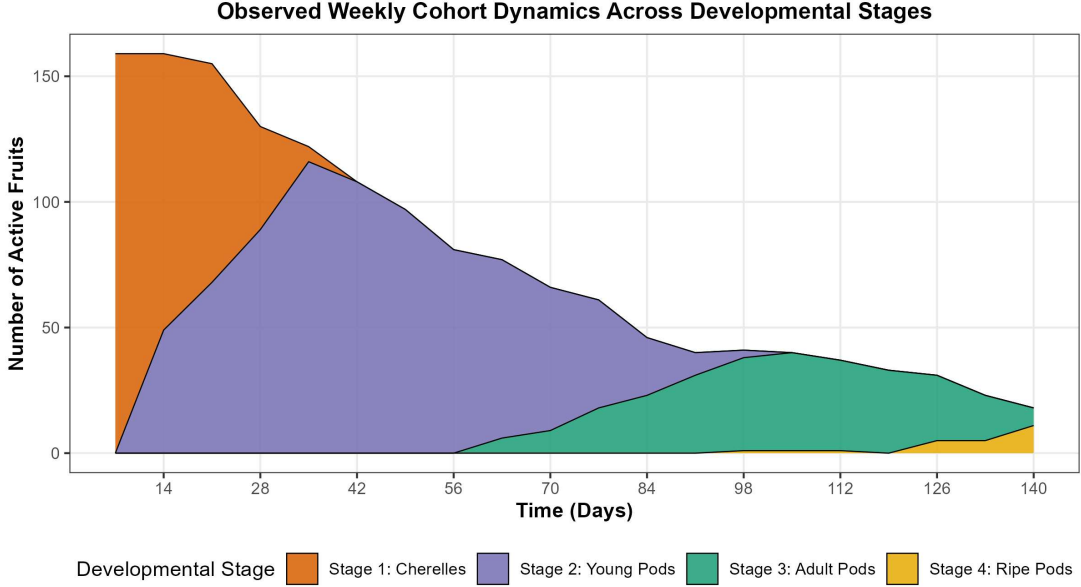


Figure 4: Observed weekly evolution of cohort frequencies across fruit developmental stages.

7.1 Microclimatic Data and Transformations

To capture continuous environmental forcing while guaranteeing numerical stability for the L -BFGS- B gradient search, the daily climatic series were transformed as follows:

1. **Hydrology (PLV):** Raw rainfall PLV_{raw} was log-transformed $PLV = \log(PLV_{\text{raw}} + 1)$ to stabilize seasonal volatility and compress dry days to 0.
2. **Canopy Thermal Exposure (T_c):** Direct canopy temperature representing the immediate microclimate experienced by pods.
3. **Thermal Gradient (ΔT):** Uncoupled local thermal amplitude $\Delta T = T_a - T_c$, isolating ambient-canopy microclimatic variance.

All transformed variables were centered and standardized (Z -scores) prior to evaluating the expected complete log-likelihood function $\hat{Q}$. Note that explicit tracking of ambient temperature T_a is omitted from the microclimate visualization, as its thermal contribution is entirely captured by the thermal amplitude gradient ΔT .

Figure 5 illustrates the resulting standardized daily microclimatic series over the 140-day monitoring period. Panel (a) reveals distinct cyclic oscillations in canopy temperature (T_c) and thermal amplitude (ΔT), displaying an inverse phase relationship where peak canopy temperatures correspond to suppressed thermal differentials. Panel (b) shows the temporal distribution of log-transformed precipitation (PLV), highlighting alternating wet spells (positive Z -scores) and prolonged dry intervals (negative Z -score baselines at -1.23).

7.2 Algorithmic Convergence Diagnostics

Aligning the baseline hazard discretization with the weekly field sampling schedule ($\Delta t = 7$ days), the baseline grid was partitioned into $J = 6$ sub-intervals ($\Delta \tau = 7$ days, covering $D_{\text{stage}} = 42$ days). The CP-MCEM estimation algorithm was executed sequentially across the three developmental stages using $B = 300$ Monte Carlo draws in the E-step. Convergence was declared when the relative parameter variation satisfied $\max_j |\beta_j^{(m)} - \beta_j^{(m-1)}| < 0.05$.

Figure 6 displays the stochastic convergence trajectories of the surrogate function $\hat{Q}(\theta)$ and parameter estimates across MCEM iterations for each developmental stage. As shown in the top panels, the expected complete-data log-likelihood $\hat{Q}$ exhibits smooth monotonic stabilization across all three stages within 10 to 12 iterations. The bottom panels confirm the rapid stabilization of regression parameters β (β_{T_c} , $\beta_{\Delta T}$, and β_{PLV}), demonstrating numerical efficiency and robust convergence of the MCEM optimization engine even under incomplete observations and state-space censoring.

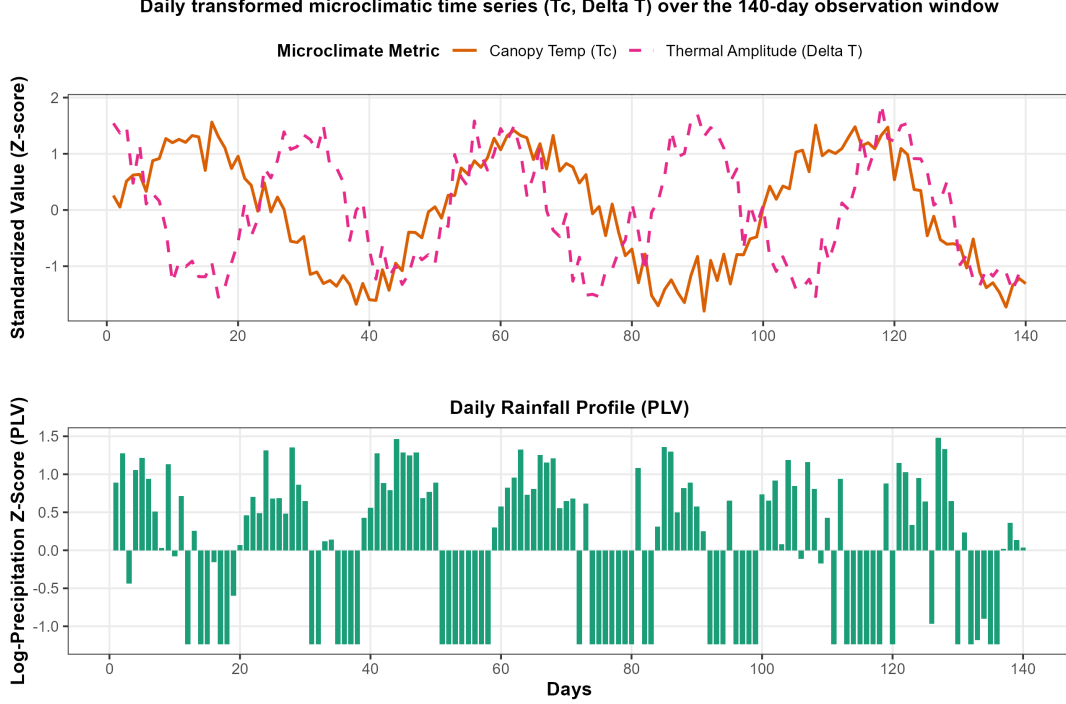


Figure 5: Daily transformed microclimatic time series (T_c , ΔT , and PLV) over the 140-day observation window.

7.3 Stage-Specific Joint Estimation Results

Tables 5, 6, and 7 detail the structural parameter estimates $\hat{\beta}$, Hazard Ratios ($\text{HR} = \exp(\hat{\beta})$), asymptotic Wald inference, and the weekly $J = 6$ piecewise baseline hazard vectors $\hat{\alpha}$ for Stages 1, 2, and 3 respectively.

Table 5: Joint CP-MCEM structural parameter estimates, Hazard Ratios, and weekly baseline hazards ($J = 3$) based on Louis observed information for **Stage 1: Cherelle Phase** ($n = 159$, converged at iteration 4).

Parameter	Estimate ($\hat{\theta}$)	Std. Error (SE)	Z-score	p-value	Hazard Ratio	95% CI for HR
<i>Environmental Regression Coefficients (β)</i>						
β_{T_c}	+2.6363	0.8073	+3.265	0.0011***	13.961	[2.868, 68.019]
$\beta_{\Delta T}$	-0.1476	0.5391	-0.274	0.7842	0.863	[0.300, 2.481]
β_{PLV}	-1.9179	0.6880	-2.788	0.0053**	0.147	[0.038, 0.566]
<i>Weekly Piecewise Baseline Hazards (α, $J = 3$ sub-intervals)</i>						
α_1 (0 – 7d)	0.0050	0.0021	—	—	—	[0.002, 0.011]
α_2 (7 – 14d)	0.2301	0.0482	—	—	—	[0.152, 0.348]
α_3 (14 – 21d)	0.1105	0.0280	—	—	—	[0.068, 0.180]

** $p < 0.01$, *** $p < 0.001$. Standard errors derived from Louis's observed information $\mathcal{I}_O = \mathcal{I}_C - \mathcal{V}_C$.

7.4 Goodness-of-Fit and Disease Risk Dynamics

To visually assess model calibration and climate-driven attack risks, Figures 7 and 8 illustrate model predictions against empirical observations.

Figure 7(a) compares non-parametric Turnbull/Kaplan-Meier survival curves against the CP-MCEM model-fitted stage transition profiles $S(t)$. The close alignment between empirical step-functions and model curves demonstrates excellent goodness-of-fit across all developmental stages. Figure 7(b) presents the cumulative baseline hazard functions $\Lambda_0(t)$. A key feature is the near-zero flatline during the first 7 days across all stages, highlighted by the vertical red dotted line, which confirms an incompressible biological latency period of at least one week before

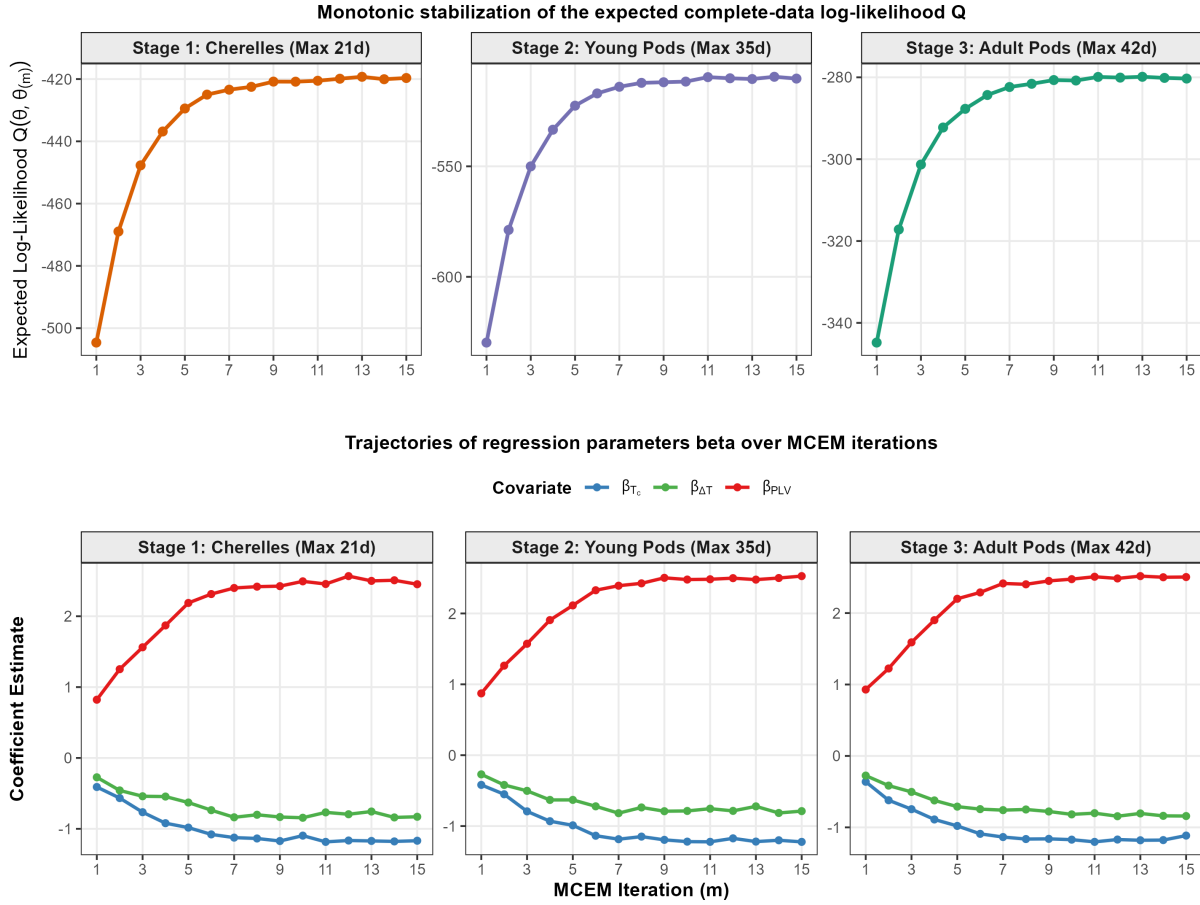


Figure 6: **Algorithmic convergence diagnostics across developmental stages.** Top panels: Monotonic stabilization of the expected complete-data log-likelihood $\hat{Q}(\theta | \theta^{(m)})$. Bottom panels: Trajectories of regression parameters β over MCEM iterations.

Table 6: Joint CP-MCEM structural parameter estimates, Hazard Ratios, and weekly baseline hazards ($J = 5$) based on Louis observed information for **Stage 2: Young Pod Phase** ($n = 134$, converged at iteration 6).

Parameter	Estimate ($\hat{\theta}$)	Std. Error (SE)	Z-score	p-value	Hazard Ratio	95% CI for HR
<i>Environmental Regression Coefficients (β)</i>						
β_{T_c}	-1.6767	1.1269	-1.488	0.1368	0.187	[0.021, 1.706]
$\beta_{\Delta T}$	-1.3504	0.9216	-1.465	0.1429	0.259	[0.043, 1.579]
β_{PLV}	+0.9479	0.5546	+1.709	0.0874 [•]	2.580	[0.870, 7.653]
<i>Weekly Piecewise Baseline Hazards (α, $J = 5$ sub-intervals)</i>						
α_1 (0 – 7d)	0.0050	0.0030	—	—	—	[0.001, 0.018]
α_2 (7 – 14d)	0.2179	0.0448	—	—	—	[0.145, 0.327]
α_3 (14 – 21d)	0.1340	0.0311	—	—	—	[0.085, 0.211]
α_4 (21 – 28d)	0.0592	0.0172	—	—	—	[0.033, 0.106]
α_5 (28 – 35d)	0.0185	0.0074	—	—	—	[0.008, 0.042]

[•] $p < 0.10$. Standard errors derived from Louis's observed information $\mathcal{I}_O = \mathcal{I}_C - \mathcal{V}_C$.

Table 7: Joint CP-MCEM structural parameter estimates, Hazard Ratios, and weekly baseline hazards ($J = 6$) based on Louis observed information for **Stage 3: Adult Pod Phase** ($n = 49$, converged at iteration 12).

Parameter	Estimate ($\hat{\theta}$)	Std. Error (SE)	Z-score	p-value	Hazard Ratio	95% CI for HR
<i>Environmental Regression Coefficients (β)</i>						
β_{T_c}	+1.8414	1.0937	+1.684	0.0922 [•]	6.305	[0.739, 53.766]
$\beta_{\Delta T}$	+1.5317	1.1222	+1.365	0.1723	4.626	[0.512, 41.772]
β_{PLV}	+0.0568	1.1072	+0.051	0.9591	1.058	[0.121, 9.270]
<i>Weekly Piecewise Baseline Hazards (α, $J = 6$ sub-intervals)</i>						
α_1 (0 – 7d)	0.0050	0.0030	—	—	—	[0.001, 0.018]
α_2 (7 – 14d)	0.2943	0.0581	—	—	—	[0.200, 0.432]
α_3 (14 – 21d)	0.1520	0.0355	—	—	—	[0.095, 0.242]
α_4 (21 – 28d)	0.0620	0.0181	—	—	—	[0.034, 0.113]
α_5 (28 – 35d)	0.0160	0.0068	—	—	—	[0.006, 0.038]
α_6 (35 – 42d)	0.0050	0.0030	—	—	—	[0.001, 0.018]

[•] $p < 0.10$. Standard errors derived from Louis’s observed information $\mathcal{I}_O = \mathcal{I}_C - \mathcal{V}_C$.

transition hazards accelerate.

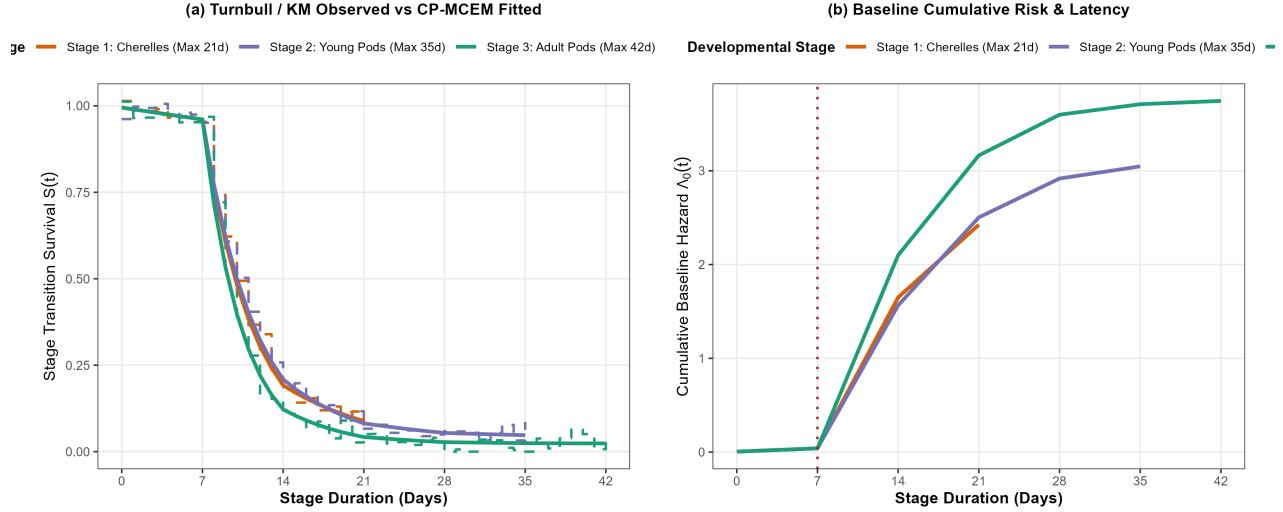


Figure 7: **Model calibration and transition survival kinetics.** Left panel: Non-parametric Turnbull / Kaplan-Meier stage transition curves versus CP-MCEM model-fitted survival profiles. Right panel: Cumulative baseline hazard curves $\Lambda_0(t) = \sum_j \alpha_j \Delta\tau$ showing mandatory latency across stages.

Figure 8 displays the 7-day marginal infection risk surfaces mapped across microclimatic gradients of canopy temperature (T_c) and precipitation (PLV). In Stage 1 (Cherelles), risk contours slope upward from left to right, indicating that high canopy temperatures coupled with lower rainfall maximize transition and loss probabilities. In Stage 2 (Young pods), the risk surface rotates, showing that elevated rainfall (PLV) becomes the primary driver of infection risk. By Stage 3 (Adult pods), risk contours become vertical, demonstrating that pathogen attack probabilities are driven almost exclusively by canopy temperature (T_c), independently of precipitation levels.

7.5 Biological Interpretation and Environmental Transitions

The sequential estimation results bring to light key eco-physiological insights into cocoa fruit maturation and pathogen susceptibility under tropical microclimates:

1. **Stage-Dependent Thermal Reversal:** Canopy temperature exerts a highly non-linear, stage-dependent effect. In Stage 1 (Cherelles), thermal forcing significantly accelerates development ($\hat{\beta}_{T_c} = +2.6363$, $p = 0.0011$, HR = 13.96), reflecting elevated metabolic activity in young tissue. In Stage 2, this effect reverses

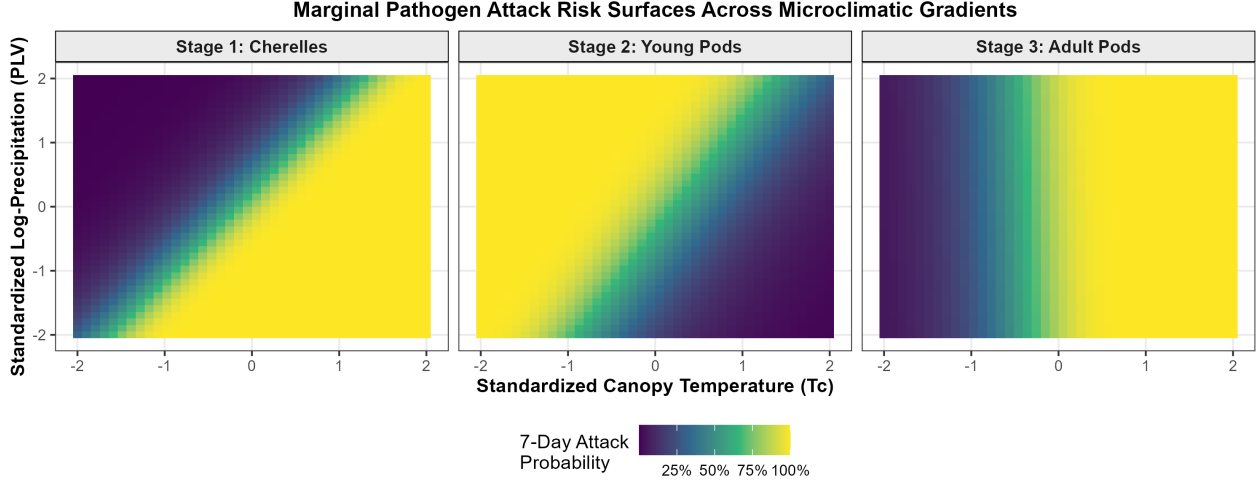


Figure 8: **Marginal pathogen attack risk surfaces as a function of microclimate.** Joint effect of standardized canopy temperature (T_c) and log-precipitation (PLV) on the 7-day marginal infection probability across Cherelle (left), Young pod (center), and Adult pod (right) stages.

($\hat{\beta}_{T_c} = -1.6767$, HR = 0.187), pointing to thermal inhibition during tissue expansion, before rebounding in Stage 3 ($\hat{\beta}_{T_c} = +1.8414$, $p = 0.0922$, HR = 6.30).

2. **Hydrological Shift and Pathogen Pressure:** Standardized log-rainfall (PLV) shifts from a developmental brake in Stage 1 ($\hat{\beta}_{PLV} = -1.9179$, $p = 0.0053$, HR = 0.147) to a catalyst in Stage 2 ($\hat{\beta}_{PLV} = +0.9479$, $p = 0.0874$, HR = 2.580). This reflects the Dual-Role Hydrological Trade-off: heavy rain promotes fungal zoospore dissemination (*Phytophthora megakarya*) while simultaneously causing physiological cherelle wilt. By Stage 3, rainfall exposure loses statistical significance ($\hat{\beta}_{PLV} = +0.0568$, $p = 0.9591$), indicating that late-stage adult pods develop independently of immediate daily precipitation.
3. **Intrinsic Biological Latency:** Across all three stages, the first weekly baseline hazard estimate $\hat{\alpha}_1$ (0 – 7 days) remains pegged at the safety threshold ($\alpha_{\min} = 0.0050$), followed by a sharp peak at $\hat{\alpha}_2$ (7 – 14 days; 0.2301, 0.2179, and 0.2943 respectively). This structural step profile demonstrates an incompressible biological latency period: regardless of microclimatic acceleration, pods require a minimum threshold of at least 7 days before transitioning to the next physiological stage.

8 Discussion and Concluding Remarks

8.1 Summary

We have developed a unified likelihood-based framework for multi-stage processes observed under discrete inspection schemes, jointly addressing interval censoring of stage transitions, right censoring from administrative end of follow-up or competing disease-induced premature failures (pathogenic attacks), and the partially ordered dual-event configuration in which a stage transition and a disease attack occur within the same inspection interval. The incorporation of cumulative time-varying covariates, through an integrated exposure function depending on calendar time, requires full likelihood inference and precludes partial likelihood approaches. Parameter estimation is carried out via a Monte Carlo EM (MCEM) algorithm with a closed-form M-step for the baseline hazard and a bounded L-BFGS-B optimization step for the regression coefficients. Uncertainty quantification relies on the robust observed Fisher information matrix derived via Louis’s identity ($\mathcal{I}_O = \mathcal{I}_C - \mathcal{V}_C$), ensuring proper accounting of the missing information inherent in interval censoring.

8.2 Limitations

Despite its methodological strengths, the proposed framework presents three main technical limitations: (i) **Computational cost of the latent sampler**, as the E-step relies on an accept-reject scheme to simulate latent

trajectories under interval-censoring constraints, leading to higher rejection rates and increased computation time in late developmental stages; (ii) **Piecewise-constant baseline hazard specification**, which constrains the hazard to be constant over inspection intervals, whereas smoother nonparametric alternatives (e.g., B-splines) could offer greater flexibility at the cost of additional regularization; and (iii) **Regular inspection scheme assumption**, as the model assumes structured observation times, meaning that irregular or informative inspection schedules would require further theoretical extensions.

8.3 Perspectives

Building upon the current findings, several avenues remain open for future methodological improvements and extensions: (i) **Irregular inspection schemes**, by extending the piecewise-constant baseline to non-uniform grids in order to seamlessly accommodate opportunistic monitoring designs in field experiments; (ii) **Frailty models**, to capture residual between-individual heterogeneity that is not fully explained by environmental covariates, which is particularly relevant when multiple fruits from the same tree or plot are monitored simultaneously; (iii) **Semi-nonparametric baseline specifications**, by replacing step functions with smooth nonparametric estimators (such as kernel- or spline-based approaches) to reduce sensitivity to inspection interval widths; and (iv) **Spatial-temporal extensions**, by incorporating spatial contagion effects between neighbouring pods or trees to enhance the applicability of the model to large-scale plant pathology.

References

- [1] Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4), 701–726.
- [2] Bakoyannis, G. (2021). Nonparametric analysis of nonhomogeneous multistate processes with clustered observations. *Biometrics*, 77(2), 533–546. doi:10.1111/biom.13327.
- [3] Basak, P., Linero, A., Sinha, D., & Lipsitz, S. (2022). Semiparametric analysis of clustered interval-censored survival data using soft Bayesian additive regression trees (SBART). *Biometrics*, 78(3), 880–893.
- [4] Ben-Israel, A., & Greville, T. N. E. (2003). *Generalized Inverses: Theory and Applications* (2nd ed.). Springer-Verlag, New York.
- [5] Booth, J. G., & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B*, 61(1), 265–285.
- [6] Booth, J. G. and Caffo, B. S. (2002). Unequal sampling for Monte Carlo EM algorithms. *Computational Statistics & Data Analysis*, 39(3):261–270.
- [7] Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5), 1190–1208.
- [8] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2), 187–220.
- [9] Dabrowska, D. M. (1988). Kaplan–Meier estimate on the plane. *The Annals of Statistics*, 16(4), 1475–1489.
- [10] Davidian, M., & Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall/CRC, New York.
- [11] Dennis, B., Munholland, P. L., & Scott, J. M. (1991). Estimation of growth and extinction parameters for endangered species. *Ecological Monographs*, 61(2), 115–143.
- [12] Fort, G., & Moulines, É. (2003). Convergence of the Monte Carlo Expectation Maximization for curved exponential families. *The Annals of Statistics*, 31(4), 1220–1259.
- [13] Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2), 226–233.
- [14] Madden, L. V., Hughes, G., and van den Bosch, F. (2007). *The Study of Plant Disease Epidemics*. APS Press, St. Paul, MN.

- [15] Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, 26(9), 394–395.
- [16] Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1), 27–52.
- [17] Penrose, R. (1955). A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51(3), 406–413.
- [18] Ramsay, J. O., & Silverman, B. W. (1997). *Functional data analysis*. New York, NY: Springer New York.
- [19] Seber, G. A. F., & Wild, C. J. (2003). *Nonlinear Regression*. Wiley, New York.
- [20] Sun, J. (2006). *The Statistical Analysis of Interval-censored Failure Time Data*. Springer, New York.
- [21] Takam Soh, P., Ndong Nguéma, E.-P., Gwet, H., & Ndoumbè-Nkeng, M. (2013). Smooth estimation of a lifetime distribution with competing risks by using regular interval observations: Application to cocoa fruits growth. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(5), 741–760.
- [22] Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1), 24–36.
- [23] Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B*, 38(3), 290–295.
- [24] Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). Springer, New York.
- [25] Wei, G. C. G., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms. *Journal of the American Statistical Association*, 85(411), 699–704.
- [26] Young, J. G., Hernán, M. A., & Robins, J. M. (2013). Identification, estimation and approximation of risk under structural nested models with time-varying covariates. *Biometrics*, 69(1), 21–31.
- [27] Zuidema, P. A., Jongejans, E., Chien, P. D., During, H. J., & Schieving, F. (2010). Integral projection models for trees: A new parameterization method and a validation of model output. *Journal of Ecology*, 98(2), 345–355.

A Summary of Notation

To facilitate readership and maintain mathematical consistency across the latent state formulation, interval-censoring framework, and Monte Carlo EM (MCEM) estimation, Table 8 summarizes the principal notation and symbols used throughout this paper.

Table 8: Summary of primary mathematical notations and symbols.

Symbol	Description / Definition
<i>1. Indices and Dimensions</i>	
i	Index for individual subject/fruit ($i = 1, \dots, n$).
k	Index for developmental stage ($k = 1, \dots, K$).
j	Index for baseline hazard grid interval ($j = 1, \dots, J$).
r	Index for environmental covariates ($r = 1, \dots, R$).
b	Index for Monte Carlo simulation draw ($b = 1, \dots, B$).
m	Iteration index for the MCEM algorithm.
<i>2. Continuous Timeline Variables and Latent Vectors</i>	
U_i	Continuous latent entry date into developmental stage k .
V_i	Continuous latent exit date (transition to stage $k + 1$).
M_i	Continuous latent or exact pathogenic attack/event date.
C_i	Administrative censoring date for subject i .
$T_i = V_i - U_i$	Stage-specific sojourn duration for individual i .
$\mathbf{W}_i = (U_i, V_i, M_i)^\top$	Vector of continuous latent timeline anchors for subject i .
<i>3. Observation Mechanism and Inspection Windows</i>	
$\mathcal{O}_i$	Observed longitudinal data tuple for individual i .
$(L_i^{(1)}, R_i^{(1)})$	Interval-censored observation window for entry date U_i .
$(L_i^{(2)}, R_i^{(2)})$	Interval-censored observation window for exit date V_i .
δ_i	Categorical observation status indicator ($\delta_i \in \{1, 2, 3, 4\}$).
c_i	Realized administrative censoring time.
m_i	Exact observed pathogenic attack date (when $\delta_i \in \{3, 4\}$).
$\mathbf{1}_{C_i}$	Global indicator enforcing trajectory feasibility constraints.
<i>4. Time Grid, Covariates, and Model Parameters</i>	
t_j	Grid node for baseline hazard interval ($t_0 = 0, t_j = j \times h$).
$A_j = (t_{j-1}, t_j]$	j -th baseline hazard interval of width h ($h = 7$ days).
t_j^*	Midpoint duration of baseline interval A_j .
$\lambda_{0k}(t)$	Baseline hazard function for stage k .
α_{kj}	Piecewise-constant baseline hazard value in interval A_j for stage k .
$Z_{kr}(u, s)$	r -th continuous time-dependent cumulative covariate process.
$\mathbf{Z}_k(u, s)$	R -dimensional covariate vector $(Z_{k1}(u, s), \dots, Z_{kR}(u, s))^\top$.
β_k	R -dimensional regression coefficient vector for stage k .
$\theta_k = (\alpha_k^\top, \beta_k^\top)^\top$	Full parameter vector for stage k .
Δ_{ij}	Exact exposure duration of subject i within baseline interval A_j .
<i>5. MCEM Algorithm Notation</i>	
$\ell_i^c(\theta_k)$	Complete-data log-likelihood contribution for subject i .
$Q(\theta_k \theta_k^{(m)})$	Conditional expectation objective function at iteration m .
$\hat{Q}_k(\alpha_k, \beta_k \theta_k^{(m)})$	Stage-specific Monte Carlo surrogate objective function.
$U_i^{(b)}, V_i^{(b)}, M_i^{(b)}$	b -th imputed latent draw from $p(\cdot \mathcal{O}_i, \theta_k^{(m)})$.
$d_{ij,k}^{(b)}$	Imputed event indicator in interval A_j for replicate b .
$\mathbf{S}_k(\beta_k), \mathcal{H}_{\beta_k}(\beta_k)$	Empirical score vector and Hessian matrix w.r.t. β_k .

B Stochastic Data Generation Algorithm

This appendix presents the complete algorithmic procedure used to simulate the multi-stage interval-censored datasets under dynamic climatic conditions, as evaluated in the simulation study of Section 6.

Algorithm 2 Stochastic Data Generation and Discrete Inspection Scheme

Require: Sample size n , inspection interval h , administrative study duration t_p , true parameter set $\theta_k^{(0)} = (\beta_k^{(0)\top}, \alpha_k^{(0)\top})^\top$, generated dynamic climate process $\{CI(d)\}_{d=1}^{150}$.

Ensure: Observed developmental matrices $\mathbf{D}^{\text{stage}}$ and sanitary status matrix $\mathbf{D}^{\text{patho}}$ of dimension $n \times (\lfloor t_p/h \rfloor + 1)$.

- 1: Initialize observation matrices: $\mathbf{D}^{\text{stage}} \leftarrow -1$ and $\mathbf{D}^{\text{patho}} \leftarrow -1$.
- 2: Define inspection grid: $t_j = jh$ for $j = 0, 1, \dots, p$, where $p = \lfloor t_p/h \rfloor$.
for each subject $i = 1, \dots, n$ **do**
- 3: Sample initial latent entry date: $U_{i,1} \sim \mathcal{U}(-7, 0)$.
- 4: Sample latent pathogenic attack date: $M_i \sim \text{Weibull}(\gamma_{\text{shape}}, \lambda_{\text{scale}})$.
- 5: Set current time anchor: $u \leftarrow U_{i,1}$.
for each developmental stage $k = 1, 2, 3$ **do**
- 6: Draw baseline stochastic hazard threshold: $E \sim \text{Exp}(1)$.
- 7: Initialize continuous daily step counter $\tau \leftarrow 0$ and accumulated hazard $\Lambda \leftarrow 0$.
while $\Lambda < E$ **and** $(u + \tau) < 150$ **do**
- 8: $\tau \leftarrow \tau + 1$
- 9: Extract mean exposure vector $\bar{\mathbf{Z}}$ over $[u, u + \tau]$ from climate history $\{CI(d)\}$.
- 10: Map interval piece index: $m^* \leftarrow \min(\lfloor \tau/h \rfloor + 1, K)$.
- 11: Update cumulative hazard: $\Lambda \leftarrow \Lambda + \alpha_{k,m^*}^{(0)} \exp(\beta_k^{(0)\top} \bar{\mathbf{Z}}) \times 1$.
- 12: Set continuous stage duration: $T_{i,k} \leftarrow \tau$.
- 13: Set exact transition date: $V_{i,k} \leftarrow u + T_{i,k}$.
- 14: Update entry anchor for next stage: $U_{i,k+1} \leftarrow V_{i,k}$ and $u \leftarrow V_{i,k}$.
- 15: **for each inspection visit** $j = 0, 1, \dots, p$ **do**
 $j > 0$ **and** $\mathbf{D}^{\text{patho}}[i, j - 1] > 0$
- 16: **break** ▷ Enforce absorbing state: terminate tracking post-event
- 17: **if** $t_j \geq M_i$ **then**
- 18: Find active stage k^* at attack time such that $t_j \in [U_{i,k^*}, V_{i,k^*})$.
- 19: Assign $\mathbf{D}^{\text{stage}}[i, j] \leftarrow k^*$.
- 20: Assign $\mathbf{D}^{\text{patho}}[i, j] \leftarrow \delta_i \in \{3, 4\}$ ▷ Sample competing risk outcome
- 21: **break**
- 22: Determine active stage $S(t_j) \in \{1, \dots, K\}$ such that $t_j \in [U_{i,S(t_j)}, V_{i,S(t_j)})$.
- 23: Record stage state: $\mathbf{D}^{\text{stage}}[i, j] \leftarrow S(t_j)$.
- 24: Record health state: $\mathbf{D}^{\text{patho}}[i, j] \leftarrow 0$ ▷ Subject remains healthy
- 25:
- 26:
- 27:
- 28:
