MORSEDISTILL: SPATIO-TEMPORAL MORSE-GUIDED DISTILLATION FOR EFFICIENT WILDFIRE FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Wildfires have become one of the most widespread natural hazards worldwide, posing severe threats to communities and ecosystems. Recent series of devastating wildfires have further underscored a critical need for more holistic, accurate, and efficient prediction of *when and where* new fire events may occur. Emerging generative AI tools, particularly Transformer-based diffusion models, can capture the complex spatio-temporal dependencies exhibited by wildfires, but their architectural complexity and iterative denoising make inference computationally expensive. Knowledge distillation offers a path toward more efficient models, yet standard approaches apply uniform supervision across all events and ignore their distinct structural roles. We propose *MorseDistill*, a Morse-guided knowledge distillation framework that distills a Transformer-based teacher into a lightweight MLP student. Inspired by the concepts from low-dimensional topology, our central idea is to use the latent topological structure of the event sequence to guide knowledge transfer. Specifically, we construct a graph abstraction for spatio-temporal event representation and identify critical events through discrete Morse functions and their gradient pairings. These critical events then shape the distillation loss, concentrating teacher supervision on the most essential spatial and temporal structures. Extensive experiments on 10 wildfire benchmarks demonstrate that MorseDistill achieves superior predictive performance at substantially lower computational costs. We validate MorseDistill with respect to the state-of-the-art baselines on a diverse set of wildfire datasets, demonstrating MorseDistill's superior predictive capabilities of up to 40.6% in relative gains under reduced computational costs of up to 96.2%.

1 INTRODUCTION

Wildfires have emerged as a pressing global challenge, transcending geographical and disciplinary boundaries, disrupting human lives and well-being, causing severe economic losses, and permanently altering ecosystems worldwide. In Canada, for example, the August 2026 Bald Range wildfire forced more than 20,000 people to evacuate (Fallico, 2026). Effects of the Canadian wildfires extended across North America, degrading air quality and prompting health alerts not only in Ontario but also across the Upper Midwest, the Great Lakes region, and the East Coast of the United States (Green, 2026a). The economic toll of these disasters is equally vast. For instance, the 2026 Canadian wildfires caused an estimated $225 million in insured losses, while the January 2025 Los Angeles wildfires resulted in estimated total losses of $250 billion (Kishore, 2026; Green, 2026b). These recent events and their escalating societal impacts emphasize the critical need for more integrated approaches to proactive wildfire risk management. This, in turn, requires systematic, data-informed, accurate, and efficient predictive risk analytics that address the critical questions of *when and where* new fire events may emerge. To be effective and reliable, such analytics must both account for the complex, interdependent, multi-scale nature of wildfire drivers, including local winds, topography, and human activity, and also adapt promptly to newly arriving observations. As such, both reliability and computational efficiency emerge as twin pillars of proactive wildfire risk management (Jain et al., 2020).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Emerging advances within deep learning (DL) and AI tools provide formidable tools for addressing these challenges. Because wildfire observations form irregular event sequences across space and time, spatio-temporal point processes (STPPs) offer a natural framework for modeling wildfire occurrence (Pereira et al., 2013; Koh et al., 2023; Xu et al., 2023). However, statistical treatment of STPPs is often associated with the restrictive assumptions that are often unverifiable in practice. In turn, DL offers a more flexible modeling alternative, and the early approaches in this direction include neural STPPs, which capture dependencies among historical events (Chen et al., 2021; Zhou et al., 2022), as well as more recent diffusion-based models that are capable to characterizing the joint distribution of future event times and locations (Yuan et al., 2023; Liu et al., 2026; Yang et al., 2026). Despite their predictive power, the emerging techniques that combine Transformer-based encoders with iterative denoising incur substantial computational costs, which hinders their operational use when forecasts must be updated frequently across multiple affected areas. Existing scalable approaches, such as ones based on Kalman and particle filtering (Rochoux et al., 2014; Zhou et al., 2021), offer a more efficient alternative, but their limited modeling capacity makes it difficult to capture complex spatio-temporal dependencies among wildfire drivers at both local and global scales. This trade-off motivates for design of knowledge distillation, or the transfer of predictive behavior and representation of an expressive teacher model to a more compact, lightweight student model. Indeed, knowledge distillation provides a practical way to leverage a teacher's predictions and intermediate representations to guide student training, while requiring only the student at inference time (Hinton et al., 2015). As such, the ultimate objective can be formulated as development of a compact model that *aims to retain the predictive capabilities of an expressive wildfire forecaster, while reducing computational cost.*

Recent advances in time-series distillation encompass the transfer of both temporal and frequency patterns across architectures, the alignment of predictive representations, and the adaptive integration of guidance from teacher models (Floratos et al., 2022; Liu et al., 2025). Together, these efforts demonstrate the potential to improve compact forecasters through teacher supervision without incurring an inference cost to the teacher. These advances motivate distillation for wildfire forecasting, but leave open *how to tailor the transfer to irregular event sequences modeled by STPPs*. In this setting, both event times and locations vary, and their joint arrangement determines the local structure of the observation history. Matching teacher outputs alone does not explicitly specify *which aspects of this structure the student should retain*. Moreover, treating all historical events equally during feature alignment may allow densely sampled regions to dominate the supervision, emphasizing less the smaller, distinct clusters of activity. This in turn may negatively impact wildfire predictive analytics at a local scale. Here we aim to take a more holistic approach, asking *which* events and spatio-temporal relationships should guide distillation so that a compact student *preserves information* needed to predict subsequent fire activity.

To address this fundamental question, we leverage the ideas from low-dimensional topology. In particular, we focus on the local *latent* structure of wildfire activity, in which scattered fire events form distinct patterns across space and time. An event's contribution to this structure depends *not only on how densely its region is sampled, but also on its relationship to neighboring observations*. We make this distinction *explicit* through insights from low-dimensional topology and, in particular, discrete Morse theory. Discrete Morse theory's strength lies in its ability to simplify a complex structure, while retaining its latent topological structural foundation. Informally, it pairs locally redundant elements so that they can be collapsed, leaving a smaller collection of critical elements that summarize structurally distinct parts of the data. This ability makes discrete Morse theory an attractive machinery to characterize wildfire spread. While discrete Morse theory has been successfully integrated into DL in conjunction with the machinery of persistent homology (Hu et al., 2022; Horn et al., 2022; Yan et al., 2022; Verma et al., 2024), these approaches use topology primarily as a representation, constraint, or generative-modeling signal. Instead, we use a decomposition lent to us by discrete Morse to organize knowledge transfer, where critical events and their induced basins determine where additional teacher supervision is imposed during distillation. Specifically, we construct a graph connecting nearby fire detections in space and time and define an ordering based on local event concentration. A discrete Morse pairing then partitions the history into basins, each rooted at a critical event that represents a local concentration maximum. This construction provides a structural criterion for allocating distillation supervision, where each basin contributes a representative event and allows locally distinct wildfire activity to receive additional emphasis even when it contains relatively few observations.

Building on this idea, we propose **MorseDistill**, a novel framework that leverages a spatio-temporal Morse function and transfers knowledge from a pretrained teacher model to a smaller student. Both models encode the complete event history and the Morse-selected critical subsequence, then fuse their representations at the corresponding event positions. We integrate prediction and feature distillation over the full history with dedicated supervision on the critical subsequence. The full-history objectives retain broad contextual information, while the critical-event objectives emphasize representative local activity and relationships among its representatives. As such, Morse selection determines where additional guidance is to be applied, while keeping the complete observation history available to the student. Our studies demonstrate that MorseDistill consistently achieves competitive performance with a compact student model across diverse wildfire datasets, underscoring the effectiveness of the Morse-guided distillation in combining predictive accuracy with computational efficiency.

The contributions of this paper are:

(1) we introduce MorseDistill, a new knowledge distillation framework that uses the spatio-temporal latent topological structure of irregular wildfire observations to guide knowledge transfer from an expressive diffusion teacher to a compact student; (2) we bring the concepts from low-dimensional topology to knowledge distillation and formulate a mechanism for knowledge distillation through discrete Morse structure, using gradient basins and their critical representatives to determine where additional prediction- and feature-level teacher supervision is to be applied; and (3) we validate MorseDistill with respect to the state-of-the-art baselines on a diverse set of wildfire datasets, demonstrating MorseDistill's superior predictive capabilities of up to 40.6% in relative gains under reduced computational costs of up to 96.2%.

2 Related Work

Knowledge distillation for time-series forecasting. Knowledge distillation transfers knowledge from a computationally expensive teacher model to a lightweight student model, enabling the student to retain much of the teacher's predictive capability while reducing model size and inference cost (Gou et al., 2021). Recent studies have also explored knowledge distillation for time-series forecasting. TimeKD (Liu et al., 2025) employs a calibrated language model as a privileged teacher and transfers future representations to a lightweight forecasting model through feature and correlation distillation. TimeDistill (Ni et al., 2026) introduces a cross-architecture distillation framework that transfers knowledge from Transformer- and CNN-based teachers to a lightweight model with multi-scale temporal patterns and multi-period frequency characteristics. **Spatio-temporal point processes.** From a nonparametric perspective, Xu et al. (2026) employs Gaussian processes to flexibly characterize background intensities and spatio-temporal event interactions in STPPs, while exploiting Kronecker structure to improve scalability. From a generative perspective, Lüdke et al. (2025) introduces Point Set Diffusion to directly model distributions over spatial and spatio-temporal point sets, whereas Liu et al. (2026) incorporates topological signature into a diffusion framework for spatio-temporal event prediction. Another research direction focuses on recovering continuous dynamics from irregular event observations. For example, Iakovlev & Lähdesmäki (2025) jointly model an underlying spatio-temporal dynamical system and the probabilistic locations and times of its observations. For the comprehensive overview of DL for STPPs see Mukherjee et al. (2025).

3 Preliminaries

We represent a wildfire STTP as a sequence of events observed at irregular times and locations. Let $\mathcal{H}_n = \{(t_i, s_i)\}_{i=1}^{n}$ denote the history available at time t_n, where $t_1 \leq \cdots \leq t_n$ and $s_i \in \mathbb{R}^{d_s}$ is the spatial location of event i. Here, n is the number of historical events and d_s is the spatial dimension. Given $\mathcal{H}_n$, our goal is to predict *when and where the next event may occur*. We express the target as $\boldsymbol{x}^0 = (\tau_{n+1}, s_{n+1})$, $\tau_{n+1} = t_{n+1} - t_n$, where τ_{n+1} is the waiting time to the next event. Because the diffusion process itself does not constrain its outputs to $\tau_{n+1} \geq 0$, we normalize τ to $[0, 1]$ prior to diffusion and clip the predicted $\boldsymbol{x}^0$ to $[-1, 1]$ at every denoising step, so the unnormalized $\hat{\tau}_{n+1}$ is guaranteed to lie in $[0, 1]$. The forecasting model learns the conditional distribution $p(\tau_{n+1}, s_{n+1} \mid \mathcal{H}_n)$. A predicted waiting time is converted to an occurrence time by $\hat{t}_{n+1} = t_n + \hat{\tau}_{n+1}$.

4 METHODOLOGY

Fig. 1 illustrates an overview of MorseDistill. Our framework transfers knowledge from a pretrained teacher model to a lightweight student model through three components. (i) **Spatio-Temporal Morse Construction** identifies critical events on a graph of the observed history, providing a structural criterion for selecting where to emphasize teacher supervision. (ii) **Full and Critical Event Encoding** represents the complete history and its critical subsequence through separate encoders in both models, then fuses their features at corresponding event positions. (iii) **Morse-Guided Knowledge Distillation** matches teacher predictions and features across the full sequence, with additional supervision on Morse-selected critical events.

4.1 SPATIO-TEMPORAL MORSE FUNCTION

Fire histories are highly redundant, as many detections occur within regions that are already active, while comparatively few indicate changes in the spatial or temporal organization of activity. For example, detection at the core of a newly dense cluster, or the initial return after a quiet interval marks a change in the arrangement of activity. We seek a deterministic set of criteria that distinguishes the structurally informative events using only the observed history, without introducing additional learned components. To achieve this goal, we construct such a criterion using discrete Morse theory (Forman, 2002). Borrowed from low-dimensional topology, discrete Morse theory equips the constructed fire-history graph with a real-valued function whose induced gradient field identifies vertex–edge pairs that are structurally redundant and can be collapsed without changing the graph's homotopy type. The cells that cannot be eliminated through this reduction are referred to as *critical cells*. In our setting, the discrete Morse function provides the notion of criticality a direct physical interpretation, as our noncritical detections are locally represented by neighboring wildfire activity, whereas critical detections mark locally distinct features of the spatio-temporal event structure. By choosing a Morse function that reflects local spatio-temporal concentration, the critical vertices act as representatives of distinct regions of fire activity rather than selecting the most frequent or globally largest events. We then use the critical vertices as a *topology-aware compression of each history*, retaining observations that are structurally nonredundant with respect to the geometry encoded by the event graph.

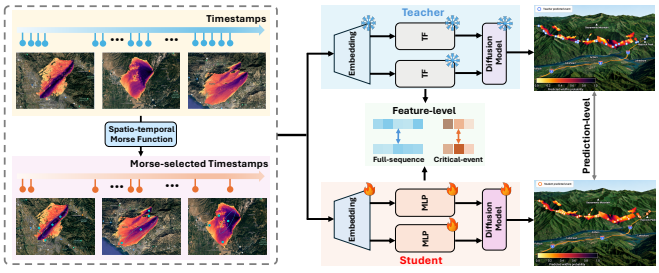


Figure 1: Overview of MorseDistill.

Event graph. Let $H_n = \{(t_i, s_i)\}_{i=1}^n$ be a fixed history of n fire detections, where t_i is the observation time and s_i its spatial coordinate. We independently min-max normalize time and each spatial coordinate to $[0, 1]$, and write $\tilde{x}_i = (\tilde{t}_i, \tilde{s}_i)$. We measure distances between events jointly in space and time, denoting:

$$d_w(\tilde{x}_i, \tilde{x}_j) = \sqrt{w^2(\tilde{t}_i - \tilde{t}_j)^2 + |\tilde{s}_i - \tilde{s}_j|_2^2}, \tag{1}$$

where $w \geq 0$ controls the relative contribution of temporal separation (exchange rate between elapsed time and distance). Note that $w = 0$ recovers a purely spatial graph construction. We then define the undirected *event graph* $G_n = (V, E)$, such that $V = \{1, \ldots, n\}$ and an undirected edge $\{i, j\}$ whenever j is among the k nearest neighbors of i under d_w (or vice versa) with $k = \lceil \sqrt{n} \rceil$. Symmetrization keeps G_n undirected, and $k = \lceil \sqrt{n} \rceil$ retains sparsity in the graph structure while retaining its cycle structure.

Spatio-temporal discrete Morse construction. We now use discrete Morse theory to obtain a topology-aware reduction of the event graph $G_n = (V, E)$. We construct the discrete gradient vector field directly from the ordering induced by a spatio-temporal density score.

For each event $i \in V$, we define its local spatio-temporal concentration by:

$$\rho(i) = \sum_{j \in N(i)} \exp\left(-d_w(\tilde{x}_i, \tilde{x}_j)^2 / (2\sigma^2)\right) \tag{2}$$

where $N(i)$ is the neighborhood of i in G_n and $\sigma > 0$ is the kernel bandwidth. Large values of $\rho(i)$ correspond to events that lie in locally concentrated regions of the spatio-temporal event graph. Discrete Morse theory requires only an ordering of the cells compatible with the gradient construction, so we can define an injective vertex ordering deterministically. Let $a_i = \big(-\rho(i),\, i\big)$, ordered lexicographically, and let $g(i) = \mathrm{rank}_{\mathrm{lex}}\{a_1, \ldots, a_n\}(a_i) \in \{1, \ldots, n\}$. So, smaller $g(i)$ corresponds to larger local concentration, while the event index is used to only resolve exact ties. This deterministic tie-breaking replaces random jitter and makes the Morse construction reproducible. For every vertex $v \in V$ that has a neighbor of strictly smaller score, we define its descending parent:

$$p(v) = \arg \min_{u \in N(v):\, g(u) < g(v)} g(u). \tag{3}$$

Since g is injective, $p(v)$ is unique whenever it exists. We pair such a vertex with the incident edge $e_v = \{v, p(v)\}$. The resulting collection of vertex-edge pairs is:

$$V_g = \left\{ \big(v, e_v\big) : \exists u \in N(v) \text{ with } g(u) < g(v) \right\}. \tag{4}$$

We now turn to the theoretical guarantees on validity and characterization of the designed spatio-temporal discrete Morse construction.

Proposition 4.1 (Validity of the spatio-temporal gradient field) *The collection V_g defined in Eq. 4 is an acyclic discrete vector field on G_n, and is the gradient vector field of a discrete Morse function on G_n. See proof in Appx. A.*

Proposition 4.2 (Characterization of critical events) *Let g be the injective ordering defined in Eq. 4.1, and let $\mathcal{V}_g$ be the gradient field defined in Eq. 4. Then*

$$\mathrm{Crit}_0(\mathcal{V}_g) = \{v \in V : g(v) < g(u) \textit{ for all } u \in N(v)\}. \tag{5}$$

If no adjacent vertices have equal concentration, this is equivalent to

$$\mathrm{Crit}_0(\mathcal{V}_g) = \{v \in V : \rho(v) > \rho(u) \textit{ for all } u \in N(v)\}. \tag{6}$$

Critical vertices are strict local minima of the Morse ordering g, and, in the absence of adjacent density ties, strict graph-local maxima of the spatio-temporal concentration ρ. Proof is in Appx. A.

We define the critical-event index set by $I_C = \mathrm{Crit}_0(\mathcal{V}_g)$, where $m = |I_C|$, and order the selected indices chronologically before passing them to the forecasting model. Each selected event retains its original timestamp t_i and spatial coordinate s_i. As such, the resulting Morse construction changes only which events receive additional emphasis, not their spatio-temporal coordinates. Propositions 4.1 and 4.2 establish that critical-event selection follows a valid discrete Morse gradient and identifies local maxima of spatio-temporal concentration, with deterministic resolution of density ties. Every noncritical event follows a unique descending path to a critical root, so the selected events represent the gradient basins of the observed history. So, we have a structural criterion for assigning additional teacher supervision to locally distinct activity. We note that the above concerns the constructed event graph, but does not establish that noncritical events are predictively redundant or can be discarded without affecting forecasting accuracy. Accordingly, MorseDistill retains full-history supervision and uses critical events for additional, targeted distillation.

Morse Basins & critical-event representatives. Each critical event is the unique root of a gradient basin, consisting of the events whose descending paths terminate at that root. So, I_C contains one representative per basin, including comparatively small regions of activity when they form distinct basins under the chosen graph and density ordering. Contracting the paired-edge forest and retaining all unpaired edges (including any resulting self-loops and parallel edges) produces a quotient graph with the same homotopy type as G_n. This applies to the complete quotient graph, rather than solely to the selected event sequence, and MorseDistill uses the basin representatives to allocate additional distillation supervision. Formal definitions and proofs are provided in Appx.A.

4.2 Full and Critical Event Representations

Forecasting wildfire activity requires representing individual events within the spatial and temporal context of the observed history. The full sequence retains this context, while the Morse-selected

critical sequence provides a focused view of its structure. We first construct embeddings for both views using a common temporal encoding and spatial projection. Each critical event retains its original timestamp and location, so the two views describe the same history at different levels of selection. These embeddings allow the teacher and student to process the complete event record and its critical subsequence within a consistent representation space.

Full-sequence embedding. Given a historical event sequence $\{(t_i, s_i)\}_{i=1}^n$, we represent each event through its occurrence time and spatial location. That is, we encode the timestamps directly using sinusoidal functions at multiple frequencies, i.e., $[\boldsymbol{z}_i^t]_j = \cos(t_i/\xi^{(j-1)/d})$ for odd j and $[\boldsymbol{z}_i^t]_j = \sin(t_i/\xi^{(j-1)/d})$ for even j, where $j = 1, \ldots, d$, d is the temporal embedding dimension, and $\xi = 10{,}000$ controls the frequency range. Spatial coordinates are mapped to latent features through a learnable nonlinear projection, $\boldsymbol{z}_i^s = \text{MLP}(s_i)$. Keeping the original event order yields two aligned embedding sequences, i.e., $\boldsymbol{z}_F^t = (\boldsymbol{z}_1^t, \ldots, \boldsymbol{z}_n^t)$, $\boldsymbol{z}_F^s = (\boldsymbol{z}_1^s, \ldots, \boldsymbol{z}_n^s)$.

Morse-guided critical embedding. We construct a complementary representation of the critical events selected from the full event sequence by the spatio-temporal Morse function. Let $I_C = \{c_1, \ldots, c_m\} \subseteq \{1, \ldots, n\}$ denote their indices, ordered chronologically, where m is the number of selected events. We apply the same temporal encoding and spatial projection used for the full sequence to obtain both spatial and temporal embedded features, i.e., $\boldsymbol{z}_C^t = (\boldsymbol{z}_{c_1}^t, \ldots, \boldsymbol{z}_{c_m}^t)$, $\boldsymbol{z}_C^s = (\boldsymbol{z}_{c_1}^s, \ldots, \boldsymbol{z}_{c_m}^s)$. Note that each critical event retains its original timestamp and spatial coordinates, preserving its position within the full spatio-temporal sequence.

4.3 Teacher and Student Models

A wildfire history unfolds across space and time, with each event situated within a larger pattern. The full sequence preserves this broader context, while the Morse-selected events offer a focused view of its structure. Our teacher-student framework brings these two perspectives together: a pretrained Transformer teacher provides contextual representations, and a lightweight MLP student learns to approximate them. Both models encode the full and critical sequences separately, then return the critical-event features to their original positions in the full history.

Teacher model. Let $q \in \{F, C\}$ index the full and critical event sequences, respectively, and let $\boldsymbol{Z}_q = (\boldsymbol{z}_q^t, \boldsymbol{z}_q^s)$ denote their temporal and spatial embeddings. The teacher encodes each sequence using a Transformer:

$$\boldsymbol{H}_F^T = \text{Transformer}_F(\boldsymbol{Z}_F), \qquad \boldsymbol{H}_C^T = \text{Transformer}_C(\boldsymbol{Z}_C), \tag{7}$$

where $\boldsymbol{Z}_F$ and $\boldsymbol{Z}_C$ denote the embedded full and critical sequences, respectively. The full-sequence representation captures the observed event history, while the critical-sequence representation captures relationships among the Morse-selected events. The pretrained teacher remains frozen during distillation. We fuse the two representations by mapping the critical-event features back to their original positions in the full sequence: $\boldsymbol{H}^T = \boldsymbol{H}_F^T + \boldsymbol{P}_C \boldsymbol{H}_C^T$, where $\boldsymbol{P}_C \in \{0, 1\}^{n \times m}$ is an index mapping matrix with $[\boldsymbol{P}_C]_{c_j, j} = 1$ and all other entries equal to zero.

Student model. The student uses two independently parameterized MLP encoders to process the full and critical event sequences:

$$\boldsymbol{H}_F^S = \text{MLP}_F(\boldsymbol{Z}_F), \qquad \boldsymbol{H}_C^S = \text{MLP}_C(\boldsymbol{Z}_C), \tag{8}$$

where MLP_F and MLP_C have separate trainable parameters and produce features with the same hidden dimension. Following the above fusion scheme, we map the critical-event features back to their original positions and add them to the full-sequence representation: $\boldsymbol{H}^S = \boldsymbol{H}_F^S + \boldsymbol{P}_C \boldsymbol{H}_C^S$.

4.4 Conditional Spatio-Temporal Diffusion

The teacher and student representations condition their respective diffusion models to predict the next event. Let $\boldsymbol{x}^0 = (\tau_{n+1}, s_{n+1})$ denote the target event, where $\tau_{n+1} = t_{n+1} - t_n$ is the inter-event time and s_{n+1} is its spatial location. During training, the forward process corrupts the target with Gaussian noise as: $\boldsymbol{x}^k = \sqrt{\bar{\alpha}_k}\boldsymbol{x}^0 + \sqrt{1 - \bar{\alpha}_k}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $\bar{\alpha}_k = \prod_{\ell=1}^k (1 - \beta_\ell)$, and $\beta_\ell \in (0, 1)$ defines the noise schedule. For each model $r \in \{T, S\}$, a conditional denoiser uses the historical representation $\boldsymbol{H}^r$ to reverse this process as $p_{\phi_r}\left(\boldsymbol{x}^{k-1} \mid \boldsymbol{x}^k, \boldsymbol{H}^r\right) =$

$\mathcal{N}\big(\boldsymbol{x}^{k-1}; \boldsymbol{\mu}_{\phi_r}(\boldsymbol{x}^k, k, \boldsymbol{H}^r), \sigma_k^2 \boldsymbol{I}\big)$, where ϕ_r denotes the denoiser parameters and σ_k^2 is the reverse-process variance. At inference, denoising starts from Gaussian noise and produces a predicted event $\hat{\boldsymbol{x}}^r = (\hat{\tau}^r_{n+1}, \hat{s}^r_{n+1})$, with predicted occurrence time $\hat{t}^r_{n+1} = t_n + \hat{\tau}^r_{n+1}$.

4.5 Morse-inspired Knowledge Distillation

Wildfire event histories contain observations with different degrees of redundancy and structural relevance. Repeated detections within an active region may provide overlapping information, while changes in the spatial and temporal arrangement of events may offer additional clues about evolving fire activity. When transferring knowledge to a student model, matching all events with equal weight may devote much of the supervision to frequently observed patterns, leaving less emphasis on a small set of structurally distinct events. This raises a question: *What should a lightweight student learn from a high-capacity teacher, and where should that supervision be concentrated?* We answer this question by distilling knowledge through four complementary objectives grouped into two categories: (1) prediction distillation over the full sequence and Morse-selected critical events, and (2) feature distillation over the corresponding full and critical representations.

Full-sequence and critical-event prediction distillation. We first supervise what the student predicts. Given the same historical sequence, let $\hat{\boldsymbol{x}}^T$ and $\hat{\boldsymbol{x}}^S$ denote the next-event predictions of the teacher and student. We align their predicted inter-event times and spatial locations through: $\mathcal{L}_{\text{pred}} = \frac{1}{d_x}||\hat{\boldsymbol{x}}^S - \hat{\boldsymbol{x}}^T||_2^2$, where d_x is the number of scalar coordinates in an event (here $d_x = 3$). Teacher and student predictions are generated from the same history with shared sampling randomness, reducing variation due to independently sampled diffusion noise. This objective transfers the teacher's forecasting behavior to the student. We further use the Morse structure to determine which events receive additional prediction supervision, which is defined as follows: $\mathcal{L}_{\text{M-pred}} = \frac{1}{|I_C|\, d_x} \sum_{i \in I_C} ||\hat{\boldsymbol{x}}_i^S - \hat{\boldsymbol{x}}_i^T||_2^2$, where $\hat{\boldsymbol{x}}_i^S, \hat{\boldsymbol{x}}_i^T$ denote the student and teacher predictions of the critical event i, respectively.

Full-sequence and critical-event feature distillation. Similar predictions do not necessarily imply similar representations of the event history. To guide the student's encoding of that history, we also match the latent features of corresponding events: $\mathcal{L}_{\text{feat}} = \frac{1}{n d_h}||\boldsymbol{H}_F^S - \boldsymbol{H}_F^T||_F^2$, where d_h is the common feature dimension. By aligning features across all historical events, this objective guides the student to learn the teacher's context-dependent representations, complementing supervision from the final prediction. To place additional emphasis on structurally selected events, we complement full-sequence supervision with a dedicated loss on the Morse-selected critical sequence. Specifically, the student learns to match the features produced by the teacher when encoding these events as: $\mathcal{L}_{\text{M-feat}} = \frac{1}{m d_h}||\boldsymbol{H}_C^S - \boldsymbol{H}_C^T||_F^2$, where $\boldsymbol{H}_C^S$ and $\boldsymbol{H}_C^T$ contain the student and teacher features of the same m critical events in chronological order. The teacher encodes the critical sequence separately, so its features capture relationships among the selected events. This objective emphasizes the critical events while the full-sequence loss continues to supervise all historical observations.

From critical events to basin-level alignment. The Morse gradient partitions the event history into basins rooted at critical events. We examine when matching these roots also controls representation discrepancies within their basins. Fix a nonempty event graph $G_n = (V, E)$ and the acyclic gradient matching $\mathcal{V}_g$ induced by the ordering in Eq. 4.1. Its critical vertices are $I_C = \{c_1, \ldots, c_m\}$. Let γ_i be the descending path from event i to its critical root $c(i)$, and write $B_j = \{i : c(i) = c_j\}$ and $b_j = |B_j|$. For fixed positive edge lengths ℓ_e, we have:

$$D_g(i) = \sum_{e \in \gamma_i} \ell_e, \qquad R_g^2 = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{b_j} \sum_{i \in B_j} D_g(i)^2. \tag{9}$$

Hence, R_g measures the root-mean-square descending distance under equal weighting of the basins, critical roots have $D_g(c_j) = 0$, and unit edge lengths recover descending-path step counts. For $r \in \{S, T\}$, let $\boldsymbol{h}^r_{F,i}$ and $\boldsymbol{h}^r_{C,j}$ denote the feature vectors in rows i and j of $\boldsymbol{H}^r_F$ and $\boldsymbol{H}^r_C$, respectively:

$$\nu_g^r = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{b_j} \sum_{i \in B_j} \delta_{\boldsymbol{h}^r_{F,i}}, \qquad \nu^r_{g,\eta} = \nu^r_g * \mathcal{N}(\boldsymbol{0}, \eta^2 \boldsymbol{I}_{d_h}), \tag{10}$$

where $\delta_{\boldsymbol{h}}$ is a point mass, $*$ denotes convolution, and $\boldsymbol{I}_{d_h}$ is the identity matrix. The distribution ν_g^r assigns equal mass to each basin. The bandwidth $\eta > 0$ is used only for this distributional analysis.

Proposition 4.3 (Feature alignment over Morse basins) *Under the construction above, suppose that, for each $r \in \{S, T\}$, feature variation along paired edges and cross-view discrepancy at critical roots satisfy: $||\boldsymbol{h}^r_{F,i} - \boldsymbol{h}^r_{F,p(i)}||_2 \le L_r \ell_{\{i,p(i)\}}$, $\frac{1}{m}\sum_{j=1}^{m} ||\boldsymbol{h}^r_{F,c_j} - \boldsymbol{h}^r_{C,j}||_2^2 \le \delta_r^2$, where constants $L_r, \delta_r \ge 0$. Here, $p(i)$ is the descending parent of i, L_r bounds feature variation per unit path length, and δ_r bounds the root-mean-square discrepancy between the two representations of the same critical events. Then, $\forall \eta > 0$, $D_{KL}\big(\nu^S_{g,\eta} \,\big\|\, \nu^T_{g,\eta}\big) \le \frac{\left(\sqrt{d_h \mathcal{L}_{M\text{-}feat}} + (L_S + L_T) R_g + \delta_S + \delta_T\right)^2}{2\eta^2}$, where D_{KL} denotes the Kullback–Leibler divergence. (Proof is in Appx. A.6.)*

Proposition A.12 provides an upper bound on divergence between feature distributions corresponding to two different Morse pairings as a function on discrepancy between the respective Morse basins. Coupled with stability of Morse (Gerber et al., 2013), it implies stability of the resulting feature distributions with respect to minor perturbations.

Training objective. The student learns from both observed outcomes and teacher supervision as $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} + \lambda_{\text{M-pred}}\mathcal{L}_{\text{M-pred}} + \lambda_{\text{M-feat}}\mathcal{L}_{\text{M-feat}}$, where $\mathcal{L}_{\text{task}}$ is the task-specific loss, i.e., supervised diffusion loss on observed next events, and $\lambda_{\text{pred}}, \lambda_{\text{feat}}, \lambda_{\text{M-pred}}, \lambda_{\text{M-feat}}$ are hyperparameters that control contributions of distillation objectives.

	eMAS		GK2A		MASTER		CAN		CA	
Model	Spatial↓	Temporal↓	Spatial↓	Temporal↓	Spatial↓	Temporal↓	Spatial↓	Temporal↓	Spatial↓	Temporal↓
DeepSTPP	1.1550±0.5060	3.1521±0.0043	0.2698±0.0411	11.3095±0.0118	6.9058±0.0271	5.5440±0.0550	0.7560±0.0129	0.6305±0.0053	2.8299±0.0099	0.0984±0.0010
AutoSTPP	5.6562±0.9502	3.1497±0.0211	0.2859±0.0167	10.0344±0.2910	7.4287±0.4858	5.7621±0.1593	1.3593±0.5238	0.6343±0.0050	2.1590±0.0140	0.0901±0.0015
STNPP	6.0892±0.8858	2.6642±0.0270	0.2347±0.0083	9.9660±0.0760	8.8121±1.5331	4.6379±0.0600	6.3500±0.1116	0.6696±0.0010	2.8593±0.0211	0.1010±0.0110
DSTPP	0.4426±0.2097	5.1174±0.2107	0.2305±0.0136	10.0302±0.8737	***6.6397±0.1490**	5.7939±0.2328	1.4497±0.4123	0.5843±0.0018	3.1310±0.0501	4.0242±1.0273
TEN-DM	0.3564±0.0129	3.7440±0.4039	0.1962±0.0231	9.3528±0.3208	6.9349±0.0229	4.6819±0.1420	0.8380±0.0577	0.6017±0.0144	3.0837±0.0113	0.1307±0.0053
Guard	0.7788±0.1198	7.2316±0.8153	0.2016±0.0181	9.9357±0.3902	7.2603±0.0253	5.5422±0.2756	1.8678±0.1047	0.8182±0.2902	3.0584±0.0328	0.1278±0.0032
TimeDistill	0.2217±0.0348	3.2168±0.0208	0.1938±0.0170	9.6851±1.0510	7.0928±0.0055	5.8099±0.0723	0.7664±0.0170	0.6363±0.0167	2.0566±0.0119	0.0964±0.0170
MorseDistill (Ours)	**0.1806**±0.0184	****2.5125**±0.0222	**0.1769**±0.0165	**9.3061**±0.0606	6.8536±0.0018	****4.4160**±0.0330	**0.7396**±0.0337	**0.5775**±0.0051	****1.2407**±0.0166	***0.0852**±0.0010

	AZ		MT		OK		OR		TX	
Model	Spatial↓	Temporal↓	Spatial↓	Temporal↓	Spatial↓	Temporal↓	Spatial↓	Temporal↓	Spatial↓	Temporal↓
DeepSTPP	1.5097±0.0080	0.3855±0.0053	3.1197±0.3002	0.5214±0.0244	1.2224±0.0740	2.5987±0.0077	1.6569±0.2666	1.6433±0.0026	2.2055±0.0323	0.8625±0.0071
AutoSTPP	1.5034±0.0065	0.3867±0.0043	3.0886±0.1759	0.5102±0.0305	1.1893±0.0246	2.5889±0.0473	1.6471±0.1616	1.6333±0.0058	2.1075±0.0194	0.9631±0.0697
STNPP	1.6090±0.0043	0.4180±0.0100	3.4414±0.0690	0.5527±0.0146	1.3612±0.0099	2.3268±0.0206	1.6767±0.0133	1.2769±0.0051	2.6647±0.0364	0.9303±0.0080
DSTPP	1.7711±0.0156	14.6783±0.7431	3.6359±0.0127	26.4391±5.6145	1.4256±0.0400	31.4209±4.4240	1.8996±0.0182	26.5491±2.8291	2.7081±0.0957	44.6834±6.6879
TEN-DM	1.9148±0.1751	0.4492±0.0098	3.6745±0.0363	0.6957±0.0158	1.4756±0.0073	2.8151±0.1921	1.9463±0.0153	1.8219±0.2069	2.8322±0.0255	2.2232±0.2808
Guard	1.8008±0.0118	0.5668±0.0405	3.6468±0.0215	0.9967±0.2233	1.4934±0.0168	4.0630±0.0757	1.9321±0.0162	3.2768±0.4639	2.8589±0.0657	4.1305±0.8794
TimeDistill	1.5133±0.0004	0.3761±0.0066	3.1316±0.1455	****0.3940±0.0140**	1.2241±0.0148	2.5895±0.0280	1.6687±0.0159	1.6022±0.0062	2.2470±0.0098	0.9237±0.0126
MorseDistill (Ours)	****0.8926**±0.0063	**0.3652**±0.0079	**2.9960**±0.1720	0.4901±0.0145	****0.8381**±0.0118	**2.3260**±0.0186	***1.0272**±0.0143	**1.2757**±0.0014	****1.4509**±0.0080	****0.8343**±0.0052

Table 1: Comparison of spatio-temporal event prediction performance on 10 wildfire datasets.

	eMAS		CA		AZ		OR		TX	
Model	Spatial↓	Temporal↓	Spatial↓	Temporal↓	Spatial↓	Temporal↓	Spatial↓	Temporal↓	Spatial↓	Temporal↓
MorseDistill (ours)	**0.1806**±0.0184	**2.5125**±0.0222	**1.2407**±0.0166	**0.0852**±0.0010	**0.8926**±0.0063	**0.3652**±0.0079	**1.0272**±0.0143	**1.2757**±0.0014	**1.4509**±0.0080	**0.8343**±0.0052
Teacher	0.2070±0.0361	2.5785±0.0917	1.5957±0.0554	0.0932±0.0028	1.0091±0.0125	0.3679±0.0086	1.2747±0.1533	1.2892±0.0091	1.4598±0.0174	0.8625±0.0313
Student	0.8651±0.1836	5.3613±0.9063	1.9662±0.0361	0.0996±0.0009	1.5994±0.0292	0.4097±0.0095	1.5370±0.0892	1.8959±0.1137	1.9231±0.0423	1.9621±0.1231
W/o Prediction level	0.2435±0.0308	2.6027±0.0592	1.2838±0.0733	0.0874±0.0013	0.9872±0.0402	0.3697±0.0103	1.2144±0.1476	1.3001±0.0117	1.4817±0.0226	0.8666±0.0168
W/o Feature level	0.2654±0.0307	2.5498±0.0208	1.3203±0.0692	0.0890±0.0018	1.0186±0.0423	0.3745±0.0094	1.2419±0.1395	1.2927±0.0209	1.4944±0.0179	0.8707±0.0534
W/o Morse-guided prediction level	0.2611±0.0518	2.5961±0.0836	1.3166±0.0760	0.0875±0.0013	1.0013±0.0421	0.3716±0.0105	1.2148±0.1366	1.2835±0.0075	1.4803±0.0204	0.8651±0.0352
W/o Morse-guided feature level	0.2339±0.0141	2.5717±0.0759	1.3327±0.0951	0.0890±0.0013	1.0221±0.0339	0.3756±0.0136	1.2297±0.1330	1.2866±0.0096	1.5062±0.0244	0.8535±0.0155

Table 2: Ablation studies.

5 Evaluation Protocol

Datasets. We conduct experiments on 10 wildfire datasets, including eMAS, MASTER, GK2A, Wildfire-Canada (CAN), Wildfire-California (CA), Wildfire-Arizona (AZ), Wildfire-Montana (MT), Wildfire-Oklahoma (OK), Wildfire-Oregon (OR), and Wildfire-Texas (TX). Detailed dataset descriptions and statistics on sequence lengths (and data splits) are provided in Appx. B.1.

Baselines and evaluation metrics. We compare MorseDistill against diverse baselines, including DeepSTPP (Zhou et al., 2022), AutoSTPP (Zhou & Yu, 2023), STNPP (Dong et al., 2026),

DSTPP (Yuan et al., 2023), TEN-DM (Liu et al., 2026), TimeDistill (Ni et al., 2026), and Guard (Dey et al., 2026). We evaluate each method over 3 independent runs with different seeds and report the mean and standard deviation. Next-event prediction is assessed in both space and time: spatial error is the mean Euclidean distance between predicted and observed locations, and temporal error is the root-mean-square error (RMSE) between predicted and observed inter-event intervals. We conduct a systematic search over all hyperparameters, and the resulting model configurations are summarized in Table 3, Appx. B.2. All experiments are conducted on a server with 4 NVIDIA A5000 GPUs. The best and second-best results are highlighted in **bold** and underlined. *, ** denote significant, statistically significant results. Codes are available at `https://anonymous.4open.science/r/MorseDistill-05CD/`.

Forecasting performance. As shown in Table 1, MorseDistill consistently outperforms most baselines across datasets. Compared with the strongest competing baseline for each dataset, MorseDistill achieves an average relative spatial gain of 20.78% across all 10 datasets. For instance, for spatial prediction, MorseDistill achieves relative gains of 40.63%, 39.67%, and 37.64% on AZ, CA, and OR, respectively, compared with the runner-up on each dataset. It also achieves spatial gains of 31.16% on TX and 29.53% on OK; for temporal prediction, MorseDistill reduces error by 5.69% on eMAS, 5.44% on CA, and 4.78% on MASTER. Compared with knowledge distillation baselines, MorseDistill outperforms Guard on both metrics across all 10 datasets, with average spatial and temporal error reductions of 42.27% and 42.47%, respectively. It also surpasses TimeDistill on spatial prediction across all 10 datasets and temporal prediction on 9 out of 10 datasets, with average relative gains of 22.46% and 8.94%, respectively, across all 10 datasets. We analyze hyperparameter sensitivity in Tables 4 and 5 in Appx. B.3. Table 6 in Appx. B.4 also demonstrates the computational efficiency of MorseDistill. MorseDistill achieves the lowest temporal error on MASTER and the lowest spatial error on MT, but trails the strongest baseline in the complementary dimension, i.e., on MASTER, airborne acquisition geometry may influence local event density, weakening its correspondence with the next detection location; on MT, a mixture of clustered detections and longer quiet intervals could make temporal RMSE sensitive to errors on a few long gaps.

Ablation studies. We validate the effectiveness of each component in MorseDistill through the following ablation studies: (1) *W/o Prediction level* removes the prediction-level distillation objective; (2) *W/o Feature level* removes feature-level distillation; (3) *W/o Morse-guided prediction level* removes Morse-guided critical-event selection; and (4) *W/o Morse-guided feature level* removes Morse guidance from feature-level distillation. We additionally compare against teacher and student models. Table 2 presents the ablation results. MorseDistill consistently outperforms both the teacher and the standalone student across all datasets and metrics. Compared with the student, it reduces spatial and temporal errors by 43.59% and 33.73% on average, respectively. Compared with the teacher, the corresponding reductions are 13.31% and 3.24%. Both prediction- and feature-level distillation contribute to these improvements. The full model achieves 11.25% and 14.11% lower spatial error than the variants without prediction-level and feature-level distillation, respectively. Morse guidance provides additional benefits: compared with the variants without Morse-guided prediction-level and feature-level distillation, the full model reduces spatial error by 12.98% and 12.50%, respectively. of prediction matching, feature alignment, and Morse-guided supervision.

6 Conclusion, Limitation, and Future Directions

By leveraging the concepts from low-dimensional topology, we have introduced MorseDistill, a new knowledge distillation framework that identifies topologically critical events through discrete Morse theory and then integrates targeted supervision on the most representative local activity with prediction and feature alignment over the complete history. We have illustrated the predictive capabilities of MorseDistill on a diverse set of wildfire datasets. While the MorseDistill forecasts are found to be highly competitive, overall, highlighting the potential of the knowledge-distillation solutions based on the discrete Morse theory, the current approach has some limitations. It does not explicitly account for additional exogenous fire drivers such as weather and terrain information. Also, the distillation framework would benefit from formal statistical inference linking the sparsity of observations with the distillation performance. In the future we will advance the MorseDistill ideas toward simultaneous knowledge distillation across multiple fire drivers, investigate MorseDistill transfer capabilities across regions and sensing modalities.

REFERENCES

Carol J Bruegge, G Thomas Arnold, Jeffrey Czapla-Myers, RoseAnne Dominguez, Mark C Helmlinger, David R Thompson, Jeannette Van den Bosch, and Brian N Wenny. Vicarious calibration of eMAS, AirMSPI, and AVIRIS sensors during FIREX-AQ. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12):10286–10297, 2021.

Ricky T. Q. Chen, Brandon Amos, and Maximilian Nickel. Neural spatio-temporal point processes. In *International Conference on Learning Representations*, 2021.

Rupasree Dey, Abdul Matin, Nathan Orwick, Yao Zhang, Shrideep Pallickara, and Sangmi Lee Pallickara. When to trust, how to distill: Multi-foundation model guidance for lightweight, robust scientific time series forecasting. In *ACM SIGKDD*, 2026.

Tamal K. Dey, Anil N. Hirani, Bala Krishnamoorthy, and Gavin Smith. Edge contractions and simplicial homology, 2013. URL https://arxiv.org/abs/1304.0664.

Zheng Dong, Jorge Mateu, and Yao Xie. Spatiotemporal-network point processes for modeling crime events with landmarks. *The Annals of Applied Statistics*, 20(2):1646–1670, 2026.

Adriana Fallico. Why did the 'absolutely shocking' b.c. wildfire intensify so fast? *Global News*, August 10, 2026.

Pavlos Floratos, Avraam Tsantekidis, Nikolaos Passalis, and Anastasios Tefas. Online knowledge distillation for financial timeseries forecasting. In *2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 2022.

Robin Forman. A user's guide to discrete morse theory. *Séminaire Lotharingien de Combinatoire [electronic only]*, 48:B48c, 35 p., electronic only–B48c, 35 p., electronic only, 2002. URL http://eudml.org/doc/123837.

Samuel Gerber, Oliver Rübel, Peer-Timo Bremer, Valerio Pascucci, and Ross T. Whitaker. Morse-smale regression. *Journal of Computational and Graphical Statistics*, 22(1):193–214, 2013. ISSN 10618600. URL http://www.jstor.org/stable/43304823.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021.

Fiona Green. Understanding the Canadian wildfires: Widespread air quality impacts with smoke stretching across the U.S. *Clarity*, September 3, 2026a. URL https://www.clarity.io/blog/understanding-the-canadian-wildfires-widespread-air-quality-impacts-with-smoke-stretching-across-the-u-s.

Fiona Green. The invisible costs of wildfires: What is the true financial impact? *Clarity*, September 3, 2026b. URL https://www.clarity.io/blog/the-invisible-costs-of-wildfires-what-is-the-true-financial-impact.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Max Horn, Edward De Brouwer, Michael Moor, Yves Moreau, Bastian Rieck, and Karsten Borgwardt. Topological graph neural networks. In *ICLR*, 2022.

Xiaoling Hu, Dimitris Samaras, and Chao Chen. Learning probabilistic topological representations using discrete morse theory, 2022. URL https://arxiv.org/abs/2206.01742.

Valerii Iakovlev and Harri Lähdesmäki. Learning spatiotemporal dynamical systems from point process observations. In *International Conference on Learning Representations*, 2025.

Piyush Jain, Sean CP Coogan, Sriram Ganapathi Subramanian, Mark Crowley, Steve Taylor, and Mike D Flannigan. A review of machine learning applications in wildfire science and management. *Environmental Reviews*, 28(4):478–505, 2020.

Michael D King, W Paul Menzel, Patrick S Grant, Jeffrey S Myers, G Thomas Arnold, Steven E Platnick, Liam E Gumley, Si-Chee Tsay, Christopher C Moeller, Michael Fitzgerald, et al. Airborne scanning spectrometer for remote sensing of cloud, aerosol, water vapor, and surface properties. *Journal of Atmospheric and Oceanic Technology*, 13(4):777–794, 1996.

Kamal Kishore. The invisible costs of wildfire disasters in 2025. *The United Nations Office for Disaster Risk Reduction*, January 13, 2026. URL https://www.undrr.org/news/invisible-costs-wildfire-disasters-2025.

Jonathan Koh, François Pimont, Jean-Luc Dupuy, and Thomas Opitz. Spatiotemporal wildfire modeling through point processes with moderate and extreme marks. *The Annals of Applied Statistics*, 17(1):560–582, 2023.

Chenxi Liu, Hao Miao, Qianxiong Xu, Shaowen Zhou, Cheng Long, Yan Zhao, Ziyue Li, and Rui Zhao. Efficient multivariate time series forecasting via calibrated language models with privileged knowledge distillation. In *IEEE 41st International Conference on Data Engineering (ICDE)*, 2025.

Yuxin Liu, Kaiming Wang, Chenguang Yang, Yulia Gel, and Yuzhou Chen. TEN-DM: Topology-enhanced diffusion model for spatio-temporal event prediction. In *The Fourteenth International Conference on Learning Representations*, 2026.

David Lüdke, Enric Rabasseda Raventós, Marcel Kollovieh, and Stephan Günnemann. Unlocking point processes through point set diffusion. In *International Conference on Learning Representations*, 2025.

Sumantrak Mukherjee, Mouad Elhamdi, George Mohler, David Antony Selby, Yao Xie, Sebastian Josef Vollmer, and Gerrit Großmann. Neural spatiotemporal point processes: Trends and challenges. *Transactions on Machine Learning Research*, 2025.

Juntong Ni, Zewen Liu, Shiyu Wang, Ming Jin, and Wei Jin. Timedistill: Efficient long-term time series forecasting with mlp via cross-architecture distillation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1113–1124, 2026.

Paula Pereira, Kamil Feridun Turkman, Maria Antónia Amaral Turkman, Ana Sá, and José MC Pereira. Quantification of annual wildfire risk; a spatio-temporal point process approach. *Statistica*, 73(1):55–68, 2013.

Mélanie C Rochoux, Sophie Ricci, Didier Lucor, Bénédicte Cuenot, and Arnaud Trouvé. Towards predictive data-driven simulations of wildfire spread–part i: Reduced-cost ensemble Kalman filter based on a polynomial chaos surrogate model for parameter estimation. *Natural Hazards and Earth System Sciences*, 14(11):2951–2973, 2014.

Yogesh Verma, Amauri H Souza, and Vikas Garg. Topological neural networks go persistent, equivariant, and continuous. In *ICML Workshop on AI for Science: Scaling in AI for Scientific Discovery*, 2024.

Chen Xu, Yao Xie, Daniel A Zuniga Vazquez, Rui Yao, and Feng Qiu. Spatio-temporal wildfire prediction using multi-modal data. *IEEE Journal on Selected Areas in Information Theory*, 4: 302–313, 2023.

Zhitong Xu, Qiwei Yuan, Yinghao Chen, Yan Sun, Bin Shen, and Shandian Zhe. Kronecker-structured nonparametric spatiotemporal point processes. In *Uncertainty in Artificial Intelligence*, 2026.

Zuoyu Yan, Tengfei Ma, Liangcai Gao, Zhi Tang, Yusu Wang, and Chao Chen. Neural approximation of graph topological features. In *Advances in neural information processing systems*, 2022.

Yiyuan Yang, Ming Jin, Haomin Wen, Chaoli Zhang, Yuxuan Liang, Lintao Ma, Yi Wang, Chenghao Liu, Bin Yang, Zenglin Xu, Shirui Pan, and Qingsong Wen. A survey on diffusion models for time series and spatio-temporal data. *ACM Computing Surveys*, 58(8):1–39, 2026.

Yuan Yuan, Jingtao Ding, Chenyang Shao, Depeng Jin, and Yong Li. Spatio-temporal diffusion point processes. In *ACM SIGKDD*, 2023.

Tengjiao Zhou, Long Ding, Jie Ji, and Shengfeng Luo. VWETKF for wildfire propagation prediction employing observations with missing values and/or spatial variations of error variance. *Proceedings of the Combustion Institute*, 38(3):5091–5099, 2021.

Zihao Zhou and Rose Yu. Automatic integration for spatiotemporal neural point processes. In *Advances in Neural Information Processing Systems*, 2023.

Zihao Zhou, Xingyi Yang, Ryan Rossi, Handong Zhao, and Rose Yu. In *Learning for Dynamics and Control Conference*, 2022.

A Discrete Morse Theory

Proof of Proposition 4.1.

Proof: Every vertex appears in at most one pair by construction. Moreover, edge $\{u, v\}$ cannot be paired with both endpoints: if $g(u) < g(v)$, then the edge can only be selected by v, since u cannot descend to v. Hence, it remains to establish acyclicity.

Along any nontrivial V_g-path

$$v_0, e_0, v_1, e_1, \ldots,$$

the second endpoint of each paired edge satisfies

$$v_{r+1} = p(v_r), \qquad g(v_{r+1}) < g(v_r).$$

Hence, g decreases strictly along every gradient path. A closed V_g-path is then impossible. By the discrete Morse realization theorem Forman (2002), an acyclic discrete vector field is induced by a discrete Morse function, which concludes the proof. ■

Gradient basins The paired edges

$$E_{\text{pair}} = \{\{v, p(v)\} : v \notin \text{Crit}_0(V_g)\} \tag{11}$$

form a forest. Iterating the parent map from any vertex produces a strictly decreasing sequence of g-values and hence terminates at a unique critical vertex. For a critical vertex c, define its gradient basin by

$$B(c) = \left\{ v \in V : p^{(r)}(v) = c \text{ for some } r \geq 0 \right\}. \tag{12}$$

Sets $\{B(c) : c \in \text{Crit}_0(V_g)\}$ partition V, and each critical event is the root of one combinatorial spatio-temporal basin induced by the density ordering.

Critical edges and topology An edge that does not belong to a vertex–edge gradient pair is a critical 1-cell:

$$E_{\text{crit}} = E \setminus E_{\text{pair}} = \text{Crit}_1(V_g). \tag{13}$$

Let

$$c = \beta_0(G_n)$$

denote the number of connected components of G_n. Since exactly one paired edge is associated with each noncritical vertex,

$$|E_{\text{pair}}| = |V| - m.$$

Consequently,

$$\begin{aligned} |E_{\text{crit}}| &= |E| - |V| + m \\ &= \beta_1(G_n) + m - c. \end{aligned} \tag{14}$$

Eq. 14 is the graph-level Morse relation. In particular, a critical edge should not in general be interpreted as an independent cycle. When several critical vertices lie in the same connected component, some critical edges are required simply to connect distinct gradient basins.

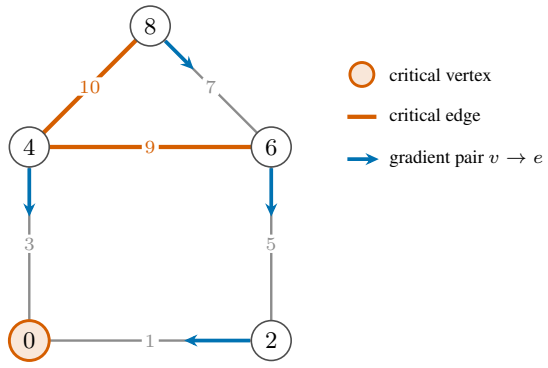


Figure 2: A discrete Morse function on a graph. Values are shown on vertices and edge midpoints. Arrows denote gradient pairs (v, e) with $f(v) \geq f(e)$; highlighted cells are critical. The numbers of critical vertices and edges equal the Betti numbers $\beta_0 = 1$ and $\beta_1 = 2$.

A useful equivalent description is obtained by contracting every component of the gradient forest (V, E_{pair}) to its critical root. The resulting Morse quotient graph

$$\mathcal{M}_g = G_n / E_{\text{pair}} \tag{15}$$

has m vertices and $|E_{\text{crit}}|$ edges and is homotopy equivalent to G_n. Each critical edge therefore becomes either (i) an edge between two distinct critical-event basins, or (ii) a loop at one critical basin when both endpoints descend to the same root. The latter directly records an intra-basin cycle, while collections of inter-basin critical edges encode the remaining connectivity and cycle structure.

The critical-event sequence used below corresponds to the vertices of $\mathcal{M}_g$. So, the Morse-selected representation does not arbitrarily subsample the history, instead it selects the roots of the gradient forest determined by local spatio-temporal concentration.

We now recall Forman's definition of Morse function on graphs (Forman, 2002).

Definition A.1 (Discrete Morse function on a graph) *A function*

$$f : V \sqcup E \longrightarrow \mathbb{R}$$

is a discrete Morse function if the following conditions hold:

1. *for every vertex $v \in V$,*

$$\#\{e \in E : v \in e,\ f(e) \leq f(v)\} \leq 1; \tag{16}$$

2. *for every edge $e = \{u, v\} \in E$,*

$$\#\{w \in \{u, v\} : f(w) \geq f(e)\} \leq 1. \tag{17}$$

A vertex is critical when the set in Eq. 16 is empty, and an edge is critical when the set in Eq. 17 is empty. If a vertex v and incident edge e satisfy $f(e) \leq f(v)$, then (v, e) forms a gradient pair.

A.1 Discrete vector fields and gradient paths

It is often more convenient to construct the gradient vector field directly rather than first constructing numerical values for f.

Definition A.2 (Discrete vector field) *A discrete vector field V on a graph G is a collection of vertex–edge pairs (v, e), $v \in e$, such that every cell of G occurs in at most one pair.*

Thus V is a partial matching on the Hasse diagram of the cell complex.

Definition A.3 (V-path) *A V-path is a sequence*

$$v_0, e_0, v_1, e_1, \ldots, e_{r-1}, v_r \tag{18}$$

such that $(v_i, e_i) \in V$ and v_{i+1} is the endpoint of e_i distinct from v_i.

A V-path is closed if $v_r = v_0$. It is nontrivial if $r > 0$.

Theorem A.4 (Forman's acyclicity criterion) *A discrete vector field V is the gradient vector field of a discrete Morse function if and only if V contains no nontrivial closed V-paths.*

Theorem A.4 allows us to work directly with an acyclic matching. Once acyclicity has been established, a real-valued discrete Morse function inducing the matching is guaranteed to exist.

A.2 The density-induced gradient field

Let $g : V \longrightarrow \mathbb{R}$ be injective. In our applications, g is exactly the deterministic rank ordering defined in Eq. (4), obtained by lexicographically ordering $a_v = (-\rho(v), v)$. In particular, we see that $g(V) = \{1, \ldots, n\}$.

For every vertex v possessing a lower-scoring neighbor, define

$$p(v) = \arg \min_{u \in N(v):\, g(u) < g(v)} g(u). \tag{19}$$

We then define

$$V_g = \{(v, \{v, p(v)\}) : p(v) \text{ exists}\}. \tag{20}$$

Proposition A.5 *The collection V_g is a discrete vector field.*

Proof: Every vertex chooses at most one edge. Suppose an edge $e = \{u, v\}$ were selected by both of its endpoints. Selection by u would require $g(v) < g(u)$, whereas selection by v would require $g(u) < g(v)$, a contradiction. Hence every edge also belongs to at most one pair. ■

Proposition A.6 *The discrete vector field V_g contains no nontrivial closed V_g-path. Consequently, V_g is the gradient vector field of a discrete Morse function.*

Proof: Consider a V_g-path $v_0, e_0, v_1, e_1, \ldots$. By construction, $v_{i+1} = p(v_i)$, and therefore $g(v_{i+1}) < g(v_i)$. Hence $g(v_0) > g(v_1) > g(v_2) > \cdots$. A finite strictly decreasing sequence cannot return to its starting value, so a nontrivial closed V_g-path cannot exist. The result then follows from Theorem A.4. ■

A.3 Critical vertices and gradient basins

Proposition A.7 *A vertex v is critical with respect to V_g if and only if*

$$g(v) < g(u) \qquad \textit{for every } u \in N(v). \tag{21}$$

Proof: A vertex is paired precisely when it possesses a neighbor of lower g-value. Therefore it is unpaired precisely when no such neighbor exists, which is equivalent to Eq. 21. ■

Because g is the lexicographic rank induced by $(-\rho(v), v)$, whenever adjacent vertices have unequal concentrations,

$$g(v) < g(u) \quad \Longleftrightarrow \quad \rho(v) > \rho(u).$$

Therefore, in the absence of adjacent density ties, Proposition A.7 implies that the critical vertices are exactly the graph-local maxima of the spatio-temporal concentration ρ.

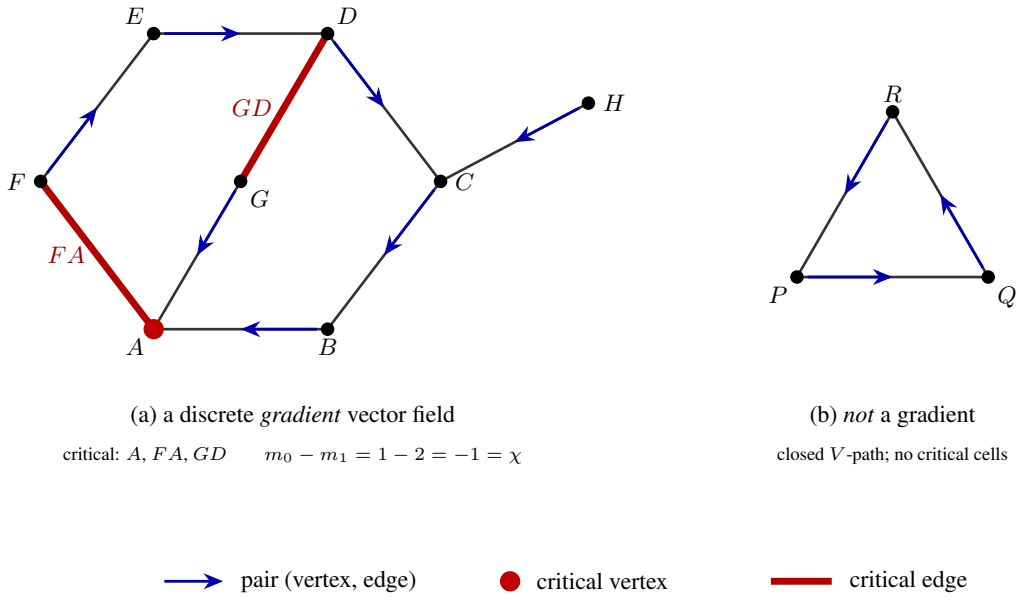
For every noncritical vertex v, repeated application of the parent map gives $v,\ p(v),\ p^2(v), \ldots$. Because g strictly decreases, this sequence terminates at a critical vertex.

Definition A.8 (Gradient basin) *For a critical vertex c, its gradient basin is*

$$B(c) = \left\{ v \in V : p^{(r)}(v) = c \textit{ for some } r \geq 0 \right\}. \tag{22}$$

Proposition A.9 *The gradient basins form a partition of V, and each component of the paired-edge subgraph (V, E_{pair}) contains exactly one critical vertex.*

Proof: Each vertex has at most one descending parent, so its descending sequence is unique. Strict decrease of g prevents cycles and guarantees termination at a critical vertex. Thus every vertex belongs to exactly one basin. Moreover, the paired-edge subgraph contains no cycle. Hence it is a forest, and each of its components is rooted at the unique critical vertex reached by its vertices. ■



(a) a discrete *gradient* vector field

critical: A, FA, GD $\quad m_0 - m_1 = 1 - 2 = -1 = \chi$

(b) *not* a gradient

closed V-path; no critical cells

A.4 Critical edges and the graph Morse relation

Let $m_0 = |\mathrm{Crit}_0(V_g)|$ and $m_1 = |\mathrm{Crit}_1(V_g)|$;, Since each noncritical vertex is paired with exactly one edge,

$$|E_{\text{pair}}| = |V| - m_0. \tag{23}$$

Every remaining edge is critical, so

$$m_1 = |E| - |E_{\text{pair}}| = |E| - |V| + m_0. \tag{24}$$

Therefore, we find that

$$m_1 = \beta_1(G) + m_0 - \beta_0(G). \tag{25}$$

Equivalently,

$$m_0 - m_1 = \beta_0(G) - \beta_1(G) = \chi(G). \tag{26}$$

In particular,

$$m_0 \geq \beta_0(G), \qquad m_1 \geq \beta_1(G), \tag{27}$$

which are the Morse inequalities specialized to graphs.

Remark A.10 *If G is connected, then $\beta_0(G) = 1$, and*

$$m_1 = \beta_1(G) + m_0 - 1. \tag{28}$$

Therefore, whenever more than one critical vertex occurs in a connected event graph, not every critical edge can be interpreted as an independent cycle. Exactly $m_0 - 1$ additional critical edges are needed, in a graph-theoretic sense, to connect the critical-vertex basins before the remaining cycle rank $\beta_1(G)$ is obtained.

A.5 THE MORSE QUOTIENT GRAPH

The paired edges form a forest $F_g = (V, E_{\text{pair}})$. Contract each connected component of this forest to its critical root. We denote the resulting quotient graph by $\mathcal{M}_g = G/E_{\text{pair}}$. Multiple edges and self-loops are retained in $\mathcal{M}_g$, since they carry information about the topology of the original graph.

Proposition A.11 (Morse quotient) *The quotient graph $\mathcal{M}_g$ is homotopy equivalent to G, where vertices are in one-to-one correspondence with the critical vertices of V_g, and edges are in one-to-one correspondence with the critical edges.*

Proof: Each connected component of E_{pair} is a tree and therefore contractible. Contracting a forest in a graph preserves homotopy type (Dey et al., 2013). After contraction, every basin is represented by its critical root, while precisely the unpaired edges remain. ■

If we let $e = \{u, v\}$ be a critical edge and $r(u)$ and $r(v)$ denote the critical roots reached from its endpoints, in the Morse quotient, e becomes $\{r(u), r(v)\}$. Further, if $r(u) = r(v)$, the critical edge becomes a self-loop and records an intra-basin cycle. If $r(u) \neq r(v)$, it records adjacency between two distinct gradient basins. Such an inter-basin edge may either contribute to connectivity or participate in a cycle, depending on the remainder of the quotient graph.

A.6 FEATURE ALIGNMENT OVER MORSE BASINS

Proposition A.12 (Feature alignment over Morse basins) *Under the construction above, suppose that, for each $r \in \{S, T\}$, feature variation along paired edges and cross-view discrepancy at critical roots satisfy:*

$$||\boldsymbol{h}^r_{F,i} - \boldsymbol{h}^r_{F,p(i)}||_2 \leq L_r \ell_{\{i,p(i)\}}, \qquad \frac{1}{m}\sum_{j=1}^{m} ||\boldsymbol{h}^r_{F,c_j} - \boldsymbol{h}^r_{C,j}||_2^2 \leq \delta_r^2,$$

where constants $L_r, \delta_r \geq 0$. Here, $p(i)$ is the descending parent of i, L_r bounds feature variation per unit path length, and δ_r bounds the root-mean-square discrepancy between the two representations of the same critical events. Then, for every $\eta > 0$,

$$D_{KL}\big(\nu^S_{g,\eta} \,\big\|\, \nu^T_{g,\eta}\big) \leq \frac{\big(\sqrt{d_h \mathcal{L}_{\textit{M-feat}}} + (L_S + L_T)R_g + \delta_S + \delta_T\big)^2}{2\eta^2},$$

where D_{KL} denotes the Kullback–Leibler divergence.

Proof: Because the event graph is nonempty and the gradient matching is acyclic, the descending paths partition the vertices into nonempty basins $B_1, \ldots, B_m$, each rooted at a critical event c_j. For every $i \in B_j$, define:

$$\alpha_i = \frac{1}{m b_j}, \qquad b_j = |B_j|.$$

Then $\sum_{i \in V} \alpha_i = 1$, and each basin has total weight $1/m$.

We first bound the discrepancy between the full-sequence features. For $r \in \{S, T\}$ and $i \in B_j$, telescoping along the descending path γ_i and applying the edgewise variation assumption give:

$$\begin{aligned} ||\boldsymbol{h}^r_{F,i} - \boldsymbol{h}^r_{F,c_j}||_2 &\leq \sum_{(u,p(u)) \in \gamma_i} ||\boldsymbol{h}^r_{F,u} - \boldsymbol{h}^r_{F,p(u)}||_2 \\ &\leq L_r \sum_{e \in \gamma_i} \ell_e = L_r D_g(i). \end{aligned}$$

For $i = c_j$, the same inequality holds because both sides are zero. We have:

$$e^r_j = ||\boldsymbol{h}^r_{F,c_j} - \boldsymbol{h}^r_{C,j}||_2, \qquad a_j = ||\boldsymbol{h}^S_{C,j} - \boldsymbol{h}^T_{C,j}||_2.$$

Inserting the full-sequence and critical-sequence representations of the root between $\boldsymbol{h}^S_{F,i}$ and $\boldsymbol{h}^T_{F,i}$ yields

$$||\boldsymbol{h}^S_{F,i} - \boldsymbol{h}^T_{F,i}||_2 \leq (L_S + L_T) D_g(i) + e^S_j + a_j + e^T_j.$$

Applying Minkowski's inequality with weights $\{\alpha_i\}$, and using $\sum_{i\in B_j}\alpha_i = 1/m$ yields:

$$\begin{aligned}\Delta &= \left(\sum_{i\in V}\alpha_i||\boldsymbol{h}^S_{F,i} - \boldsymbol{h}^T_{F,i}||_2^2\right)^{1/2} \\ &\le \big(L_S + L_T\big)\left(\sum_{i\in V}\alpha_i D_g(i)^2\right)^{1/2} + \left(\frac{1}{m}\sum_{j=1}^{m}(e^S_j)^2\right)^{1/2} \\ &\quad + \left(\frac{1}{m}\sum_{j=1}^{m}a_j^2\right)^{1/2} + \left(\frac{1}{m}\sum_{j=1}^{m}(e^T_j)^2\right)^{1/2} \\ &\le (L_S + L_T)R_g + \delta_S + \sqrt{d_h\mathcal{L}_{\text{M-feat}}} + \delta_T.\end{aligned}$$

The last inequality follows from the assumptions and the identity:

$$\frac{1}{m}\sum_{j=1}^{m}a_j^2 = \frac{1}{m}||\boldsymbol{H}^S_C - \boldsymbol{H}^T_C||_F^2 = d_h\mathcal{L}_{\text{M-feat}}.$$

We next relate this feature discrepancy to the KL divergence. For any $\eta > 0$, distributions can be written as Gaussian mixtures with identical component weights:

$$\nu^r_{g,\eta} = \sum_{i\in V}\alpha_i\mathcal{N}\big(\boldsymbol{h}^r_{F,i}, \eta^2 I_{d_h}\big)\,, \qquad r \in \{S, T\}.$$

By the log-sum inequality, the KL divergence between these mixtures is bounded by the weighted sum of the componentwise divergences. Since the paired Gaussian components have the same covariance,

$$\begin{aligned}D_{\text{KL}}\big(\nu^S_{g,\eta}\,\big\|\,\nu^T_{g,\eta}\big) &\le \sum_{i\in V}\alpha_i D_{\text{KL}}\big(\mathcal{N}(\boldsymbol{h}^S_{F,i}, \eta^2 I_{d_h})\,\big\|\,\mathcal{N}(\boldsymbol{h}^T_{F,i}, \eta^2 I_{d_h})\big) \\ &= \frac{1}{2\eta^2}\sum_{i\in V}\alpha_i||\boldsymbol{h}^S_{F,i} - \boldsymbol{h}^T_{F,i}||_2^2 = \frac{\Delta^2}{2\eta^2}.\end{aligned}$$

Substituting the bound on Δ, we have:

$$D_{\text{KL}}\big(\nu^S_{g,\eta}\,\big\|\,\nu^T_{g,\eta}\big) \le \frac{\big(\sqrt{d_h\mathcal{L}_{\text{M-feat}}} + (L_S + L_T)R_g + \delta_S + \delta_T\big)^2}{2\eta^2}.$$

■

B Experiment Details

Computational complexity. Let $G_n = (V, E)$ be the symmetrized k-nearest-neighbor event graph where $n = |V|$. Given G_n, computing all concentration scores $\rho(i)$ requires $O(|E|)$ time, constructing the injective ordering g requires $O(n \log n)$ time, and determining all descending parents requires $O(|E|)$ time. Hence, the Morse construction has time complexity $O(n \log n + |E|)$ and memory complexity $O(n + |E|)$, excluding the implementation-dependent cost of constructing the k-nearest-neighbor graph.

B.1 Datasets

We conduct experiments on 10 wildfire datasets, including: eMAS contains measurements from NASA's MODIS Airborne Simulator (King et al., 1996; Bruegge et al., 2021); MASTER is collected by the MODIS/ASTER Airborne Simulator aboard NASA's DC-8 aircraft; GK2A consists of observations over the Korean Peninsula acquired by the Advanced Meteorological Imager (AMI) onboard Geo-KOMPSAT-2A; Wildfire-Canada (CAN) contains active-fire hotspot observations from the June 2023 Canadian wildfire season, compiled from NASA's Fire Information for

Resource Management System (FIRMS); the remaining 6 datasets (April 2014 - February 2025), i.e., Wildfire-California (CA), Wildfire-Arizona (AZ), Wildfire-Montana (MT), Wildfire-Oklahoma (OK), Wildfire-Oregon (OR), and Wildfire-Texas (TX) contain state-level wildfire observations curated from the NOAA National Centers for Environmental Information (NCEI). **eMAS:** The dataset is partitioned into 214 training sequences, 46 validation sequences, and 47 test sequences; Sequence lengths are fixed at 50. **GK2A:** The dataset is partitioned into 11 training sequences, 2 validation sequences, and 3 test sequences; Sequence lengths are fixed at 10. **MASTER:** The dataset is partitioned into 177 training sequences, 38 validation sequences, and 39 test sequences; Sequence lengths are fixed at 50. **CAN:** The dataset is partitioned into 219 training sequences, 47 validation sequences, and 47 test sequences; Sequence lengths range from 5 to 50. **CA:** The dataset is partitioned into 3,367 training sequences, 753 validation sequences, and 753 test sequences; Sequence lengths range from 5 to 93. **AZ:** The dataset is partitioned into 525 training sequences, 112 validation sequences, and 113 test sequences; Sequence lengths range from 13 to 43. **MT:** The dataset is partitioned into 385 training sequences, 83 validation sequences, and 83 test sequences; Sequence lengths range from 35 to 43. **OK:** The dataset is partitioned into 308 training sequences, 66 validation sequences, and 67 test sequences; Sequence lengths range from 25 to 43. **OR:** The dataset is partitioned into 295 training sequences, 63 validation sequences, and 64 test sequences; Sequence lengths range from 17 to 43. **TX:** The dataset is partitioned into 244 training sequences, 52 validation sequences, and 53 test sequences; Sequence lengths range from 21 to 43.

B.2 Implementation Details

Table 3 summarizes the hyperparameter settings for MorseDistill across ten wildfire datasets.

Hyperparameter	**eMAS**	**GK2A**	**MASTER**	**CAN**	**CA**	**AZ**	**MT**	**OK**	**OR**	**TX**
Dropout	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
MLP Hidden Dim. $\times$ Layers	256×2	256×2	256×2	256×2	256×2	256×2	256×2	256×2	256×2	256×2
History Window	16	16	16	16	16	16	16	16	16	16
Batch Size	64	64	64	64	64	64	64	64	64	64
Learning Rate (MLP Encoder)	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3
Learning Rate (Diffusion/Morse)	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
Diffusion Timesteps	500	500	500	500	500	500	500	500	500	500
Sampling Steps	500	500	500	500	500	500	500	500	500	500
Noise Schedule	cosine	cosine	cosine	cosine	cosine	cosine	cosine	cosine	cosine	cosine
λ_{pred}	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
λ_{feat}	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
$\lambda_{\text{M-pred}}$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
$\lambda_{\text{M-feat}}$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Training Epochs	100	100	100	100	100	100	100	100	100	100

Table 3: Hyperparameter settings for MorseDistill.

B.3 Hyperparameter Studies

We examine how the distillation loss weights affect forecasting performance on eMAS, CA, AZ, OR, and TX. Each experiment varies one weight while keeping the others at their default values. Table 4 reports the resulting spatial and temporal errors. Prediction-level distillation generally performs best with a small weight: increasing either λ_{pred} or $\lambda_{\text{M-pred}}$ consistently increases spatial error across all five datasets. Full-sequence feature alignment instead generally benefits from $\lambda_{\text{feat}} = 5.0$, although TX achieves its lowest spatial error at 0.1. For critical-event feature alignment, $\lambda_{\text{M-feat}} = 0.1$ provides the best or tied-best results throughout the evaluated settings.

B.4 Computation Studies

Table 6 compares model size, memory usage, and training time per epoch (in seconds) across all wildfire datasets. MorseDistill uses 0.98M parameters, reducing model size by 38.0% relative to DSTPP and 19.7% relative to TimeDistill. Compared with DSTPP, it reduces memory usage by 11.7-73.2% and achieves a 7.9-26.2$\times$ speedup per training epoch. On CA, for example, memory usage decreases from 558.7 to 150.0 MiB, while training time drops from 46.64 to 1.78 seconds per epoch. Together with the forecasting results, these measurements demonstrate that MorseDistill combines strong predictive performance with a compact model and lower training costs.

Loss Weight	Value	eMAS		CA		AZ		OR		TX	
		Spatial ↓	Temporal ↓	Spatial ↓	Temporal ↓	Spatial ↓	Temporal ↓	Spatial ↓	Temporal ↓	Spatial ↓	Temporal ↓
λ_{pred} (prediction level)	0.1	**0.1806**$_{\pm 0.0184}$	**2.5125**$_{\pm 0.0222}$	**1.2407**$_{\pm 0.0166}$	**0.0852**$_{\pm 0.0010}$	**0.8926**$_{\pm 0.0063}$	**0.3652**$_{\pm 0.0079}$	**1.0272**$_{\pm 0.0143}$	**1.2757**$_{\pm 0.0014}$	**1.4509**$_{\pm 0.0080}$	**0.8343**$_{\pm 0.0052}$
	0.5	0.2440$_{\pm 0.0298}$	2.5647$_{\pm 0.0647}$	1.3022$_{\pm 0.0544}$	0.0854$_{\pm 0.0017}$	0.9829$_{\pm 0.0408}$	0.3681$_{\pm 0.0088}$	1.1735$_{\pm 0.1347}$	1.2785$_{\pm 0.0022}$	1.4719$_{\pm 0.0213}$	0.8423$_{\pm 0.0061}$
	2.0	0.2456$_{\pm 0.0058}$	2.5791$_{\pm 0.0794}$	1.3363$_{\pm 0.0834}$	0.0854$_{\pm 0.0010}$	1.0259$_{\pm 0.0333}$	0.3736$_{\pm 0.0111}$	1.2432$_{\pm 0.1441}$	1.2880$_{\pm 0.0064}$	1.5041$_{\pm 0.0152}$	0.8511$_{\pm 0.0188}$
	5.0	0.2797$_{\pm 0.0350}$	2.5908$_{\pm 0.0997}$	1.3551$_{\pm 0.1027}$	0.0864$_{\pm 0.0020}$	1.0499$_{\pm 0.0221}$	0.3766$_{\pm 0.0119}$	1.2788$_{\pm 0.1437}$	1.2933$_{\pm 0.0084}$	1.5267$_{\pm 0.0187}$	0.8929$_{\pm 0.0872}$
λ_{feat} (feature level)	0.1	0.2231$_{\pm 0.0266}$	2.5831$_{\pm 0.0637}$	1.2696$_{\pm 0.0371}$	0.0904$_{\pm 0.0026}$	0.9682$_{\pm 0.0589}$	0.3689$_{\pm 0.0086}$	1.1448$_{\pm 0.1361}$	1.2967$_{\pm 0.0160}$	**1.4484**$_{\pm 0.0244}$	0.8546$_{\pm 0.0146}$
	0.5	0.2434$_{\pm 0.0369}$	2.6017$_{\pm 0.1012}$	1.2719$_{\pm 0.0367}$	0.0893$_{\pm 0.0026}$	0.9670$_{\pm 0.0539}$	0.3667$_{\pm 0.0081}$	1.1413$_{\pm 0.1390}$	1.2852$_{\pm 0.0114}$	1.4520$_{\pm 0.0289}$	0.8445$_{\pm 0.0160}$
	2.0	0.2565$_{\pm 0.0699}$	2.5566$_{\pm 0.0517}$	1.2889$_{\pm 0.0727}$	0.0884$_{\pm 0.0009}$	0.9585$_{\pm 0.0414}$	0.3659$_{\pm 0.0084}$	1.1425$_{\pm 0.1376}$	1.2793$_{\pm 0.0041}$	1.4634$_{\pm 0.0228}$	0.8558$_{\pm 0.0201}$
	5.0	**0.1806**$_{\pm 0.0184}$	**2.5125**$_{\pm 0.0222}$	**1.2407**$_{\pm 0.0166}$	**0.0852**$_{\pm 0.0010}$	**0.8926**$_{\pm 0.0063}$	**0.3652**$_{\pm 0.0079}$	**1.0272**$_{\pm 0.0143}$	**1.2757**$_{\pm 0.0014}$	1.4509$_{\pm 0.0080}$	**0.8343**$_{\pm 0.0052}$
$\lambda_{\text{M-pred}}$ (Morse-guided prediction level)	0.1	**0.1806**$_{\pm 0.0184}$	**2.5125**$_{\pm 0.0222}$	**1.2407**$_{\pm 0.0166}$	**0.0852**$_{\pm 0.0010}$	**0.8926**$_{\pm 0.0063}$	**0.3652**$_{\pm 0.0079}$	**1.0272**$_{\pm 0.0143}$	**1.2757**$_{\pm 0.0014}$	**1.4509**$_{\pm 0.0080}$	**0.8343**$_{\pm 0.0052}$
	0.5	0.2221$_{\pm 0.0300}$	2.5772$_{\pm 0.0517}$	1.2880$_{\pm 0.0768}$	0.0855$_{\pm 0.0007}$	0.9731$_{\pm 0.0392}$	0.3672$_{\pm 0.0086}$	1.1693$_{\pm 0.1329}$	1.2788$_{\pm 0.0021}$	1.4956$_{\pm 0.0138}$	0.8554$_{\pm 0.0167}$
	2.0	0.2461$_{\pm 0.0313}$	2.5991$_{\pm 0.0735}$	1.3286$_{\pm 0.0714}$	0.0873$_{\pm 0.0019}$	1.0134$_{\pm 0.0411}$	0.3720$_{\pm 0.0120}$	1.2359$_{\pm 0.1412}$	1.2858$_{\pm 0.0010}$	1.5171$_{\pm 0.0212}$	0.8465$_{\pm 0.0128}$
	5.0	0.2907$_{\pm 0.0205}$	2.5936$_{\pm 0.0593}$	1.3518$_{\pm 0.0952}$	0.0890$_{\pm 0.0021}$	1.0361$_{\pm 0.0304}$	0.3792$_{\pm 0.0186}$	1.2680$_{\pm 0.1333}$	1.2880$_{\pm 0.0049}$	1.5279$_{\pm 0.0182}$	0.8584$_{\pm 0.0277}$
$\lambda_{\text{M-feat}}$ (Morse-guided feature level)	0.1	**0.1806**$_{\pm 0.0184}$	**2.5125**$_{\pm 0.0222}$	**1.2407**$_{\pm 0.0166}$	**0.0852**$_{\pm 0.0010}$	**0.8926**$_{\pm 0.0063}$	**0.3652**$_{\pm 0.0079}$	**1.0272**$_{\pm 0.0143}$	**1.2757**$_{\pm 0.0014}$	**1.4509**$_{\pm 0.0080}$	**0.8343**$_{\pm 0.0052}$
	0.5	0.2084$_{\pm 0.0372}$	2.5681$_{\pm 0.0428}$	1.2746$_{\pm 0.0369}$	0.0853$_{\pm 0.0015}$	0.9590$_{\pm 0.0496}$	0.3658$_{\pm 0.0080}$	1.1390$_{\pm 0.1430}$	1.2786$_{\pm 0.0022}$	1.4551$_{\pm 0.0309}$	0.8369$_{\pm 0.0004}$
	2.0	0.2143$_{\pm 0.0417}$	2.5781$_{\pm 0.0568}$	1.2735$_{\pm 0.0506}$	0.0857$_{\pm 0.0010}$	0.9568$_{\pm 0.0511}$	0.3654$_{\pm 0.0076}$	1.1527$_{\pm 0.1468}$	1.2861$_{\pm 0.0147}$	1.4695$_{\pm 0.0324}$	0.8560$_{\pm 0.0315}$
	5.0	0.2110$_{\pm 0.0380}$	2.5590$_{\pm 0.0318}$	1.2721$_{\pm 0.0484}$	0.0857$_{\pm 0.0018}$	0.9545$_{\pm 0.0480}$	0.3652$_{\pm 0.0075}$	1.1617$_{\pm 0.1492}$	1.2846$_{\pm 0.0123}$	1.4513$_{\pm 0.0275}$	0.8385$_{\pm 0.0005}$

Table 4: Hyperparameter sensitivity analysis of the loss weights. Each block varies a single weight while the remaining weights are held at the default configuration.

k	eMAS		CA		AZ		OR		TX	
	Spatial ↓	Temporal ↓	Spatial ↓	Temporal ↓	Spatial ↓	Temporal ↓	Spatial ↓	Temporal ↓	Spatial ↓	Temporal ↓
3	0.1835$_{\pm 0.0187}$	2.5326$_{\pm 0.0688}$	**1.0484**$_{\pm 0.0475}$	0.1053$_{\pm 0.0034}$	1.2245$_{\pm 0.0639}$	**0.3344**$_{\pm 0.0068}$	1.2972$_{\pm 0.0909}$	1.2852$_{\pm 0.0139}$	1.5841$_{\pm 0.0246}$	0.8425$_{\pm 0.0125}$
4	0.2019$_{\pm 0.0254}$	2.5235$_{\pm 0.0289}$	1.2340$_{\pm 0.0961}$	0.0953$_{\pm 0.0010}$	1.2027$_{\pm 0.0840}$	0.3580$_{\pm 0.0065}$	1.2190$_{\pm 0.1006}$	1.2766$_{\pm 0.0024}$	1.5200$_{\pm 0.0435}$	0.8725$_{\pm 0.0231}$
5	**0.1806**$_{\pm 0.0184}$	**2.5125**$_{\pm 0.0222}$	1.2407$_{\pm 0.0166}$	0.0852$_{\pm 0.0010}$	**0.8926**$_{\pm 0.0063}$	0.3652$_{\pm 0.0079}$	**1.0272**$_{\pm 0.0143}$	**1.2757**$_{\pm 0.0014}$	**1.4509**$_{\pm 0.0080}$	**0.8343**$_{\pm 0.0052}$
6	0.2568$_{\pm 0.0530}$	2.5994$_{\pm 0.0692}$	1.2356$_{\pm 0.0268}$	**0.0843**$_{\pm 0.0009}$	0.9981$_{\pm 0.0163}$	0.3743$_{\pm 0.0055}$	1.1778$_{\pm 0.0100}$	1.2825$_{\pm 0.0037}$	1.6263$_{\pm 0.0213}$	0.8450$_{\pm 0.0019}$

Table 5: Sensitivity analysis of the k-NN neighbor count k used to build the spatio-temporal Morse graph.

Method	Metric	eMAS	GK2A	MASTER	CAN	CA	AZ	MT	OK	OR	TX
MorseDistill	Param. (M)	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
	Mem. (MiB)	124.7	50.7	126.0	102.6	150.0	119.0	116.3	117.5	116.1	115.6
	Time (s/epoch)	0.24	0.11	0.22	0.18	1.78	0.37	0.37	0.26	0.25	0.22
DSTPP	Param. (M)	1.58	1.58	1.58	1.58	1.58	1.58	1.58	1.58	1.58	1.58
	Mem. (MiB)	335.7	57.4	335.3	295.8	558.7	287.1	288.1	286.6	286.6	286.3
	Time (s/epoch)	3.61	0.87	2.79	3.43	46.64	8.07	6.29	4.47	4.51	3.63
TimeDistill	Param. (M)	1.22	1.22	1.22	1.22	1.22	1.22	1.22	1.22	1.22	1.22
	Mem. (MiB)	218.6	63.9	187.2	203.9	312.4	169.8	196.5	177.5	209.1	188.1
	Time (s/epoch)	1.17	0.31	0.68	0.94	8.53	1.26	1.64	0.73	1.19	0.87

Table 6: Computational efficiency comparison including parameters, memory usage, and training time.

AI USE STATEMENT

Large language models (LLMs) were only used for rephrasing sentences and correcting grammar, with no further application beyond this purpose.

ETHICS STATEMENT

This work aims to support accurate and computationally efficient wildfire forecasting, with potential benefits for disaster preparedness and resource allocation. We do not foresee immediate negative societal impacts arising from this work.

REPRODUCIBILITY STATEMENT

In the main text, we have formally defined the model architecture with equations. Appendix B also presents hyperparameter sensitivity analyses and computational efficiency comparisons. We provide an openly accessible, anonymized code repository to facilitate reproduction of the main experimental results.