When to Trust, How to Distill: Multi-Foundation Model Guidance for Lightweight, Robust Scientific Time Series Forecasting

Rupasree Dey
rupasree.dey@colostate.edu
Department of Computer Science
Colorado State University
Fort Collins, Colorado, USA

Abdul Matin
abdul.matin@colostate.edu
Department of Computer Science
Colorado State University
Fort Collins, Colorado, USA

Nathan Orwick
nathan.orwick@colostate.edu
Department of Computer Science
Colorado State University
Fort Collins, Colorado, USA

Yao Zhang
yao.zhang@colostate.edu
Department of Soil and Crop Sciences
Colorado State University
Fort Collins, Colorado, USA

Shrideep Pallickara
shrideep.pallickara@colostate.edu
Department of Computer Science
Colorado State University
Fort Collins, Colorado, USA

Sangmi Lee Pallickara
sangmi.pallickara@colostate.edu
Department of Computer Science
Colorado State University
Fort Collins, Colorado, USA

Abstract

The deployment of Time-Series Foundation Models (TSFMs) in physical sciences is hindered by a critical trade-off: while these models encode rich, universal temporal dynamics, they suffer from severe distributional misalignment when applied zero-shot to specific scientific domains, and their computational cost prohibits deployment in edge-computing sensor networks. We address a fundamental challenge: How can we extract latent structural knowledge from misaligned foundation models (FM) to train lightweight, specialized forecasters? We propose **G**ated **U**ncertainty-**A**ware **R**outing for **D**istillation (**GUARD**), a novel framework that reframes multi-teacher distillation as an instance-wise decision process with two adaptive mechanisms: (1) a Contextual Router that dynamically selects the most relevant teacher based on local input statistics, exploiting complementarity across diverse foundation models; and (2) an Uncertainty-Gated Temperature mechanism that acts as a "circuit-breaker," automatically attenuating distillation strength when teacher confidence diverges from domain reality. We evaluate our proposed lightweight framework on four climate-critical domains: meteorology, ecosystem carbon flux, soil moisture, and energy grids. Our method significantly reduces RMSE relative to a fixed-weight multi-teacher distillation baseline, successfully distilling knowledge from pretrained FMs (teachers) even when they exhibit suboptimal zero-shot accuracy due to distribution shift between the original and target data domains. We demonstrate that these domain-misaligned teachers can still serve as critical correctives, outperforming the globally superior FMs on 28.5% of the hardest instances. Ultimately, this enables high-precision scientific forecasting suitable for resource-constrained edge deployment.

Conference acronym 'XX, Jeju, Korea

ACM ISBN 978-1-4503-XXXX-X/2026/06
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts

• **Computing methodologies** → **Machine learning**; *Knowledge representation and reasoning*; *Neural networks*.

Keywords

Scientific AI, Knowledge Distillation, Foundation Models, Multi-teacher Learning, Time Series Forecasting

ACM Reference Format:
Rupasree Dey, Abdul Matin, Nathan Orwick, Yao Zhang, Shrideep Pallickara, and Sangmi Lee Pallickara. 2026. When to Trust, How to Distill: Multi-Foundation Model Guidance for Lightweight, Robust Scientific Time Series Forecasting. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Accurate and resilient time-series forecasting is critical to understand natural phenomena and to monitor the performance of long-lasting infrastructures such as meteorological systems, ecosystem dynamics, and energy grid operations. However, capturing complex temporal trends and patterns using conventional modeling strategies presents significant challenges due to inherent label scarcity, high-stakes data acquisition costs, and the complexity of ancillary factors influencing system behavior [7, 13, 30].

Recent machine learning advances in time-series foundation models (TSFMs) have demonstrated strong potential with their generalizability across diverse domains [1, 3, 23, 27]. TSFMs are large-scale models trained on voluminous and heterogeneous temporal corpora. While these models capture rich, universal temporal dynamics (e.g., seasonality and trends), their direct deployment in scientific workflows via zero-shot inference is often infeasible [21]. In particular, applying TSFMs to new or specialized scientific phenomena presents significant challenges in maintaining acceptable performance without task or domain specific model refinement [25].

Fine-tuning billion-parameter models is computationally prohibitive for most scientific labs [25]. Several approaches have been proposed to develop efficient, robust, and domain-adaptive methods that transfer the latent temporal reasoning of FMs into lightweight, task-specific forecasters. Knowledge distillation [12], which extracts learned representations and insights from a high-capacity teacher

model into a compact student model. However, conventional distillation strategies typically treat the teacher model's outputs as target labels, an assumption that becomes problematic when the student model operates on data whose scope, distribution, or modality is misaligned with that used to train the teacher. Such data misalignment is common in environmental monitoring, where new observations are continuously assimilated and system dynamics are strongly influenced by ancillary conditions. These factors widen the gap between the knowledge encoded in the teacher and what the student can effectively learn. As a result, the student models often struggle to capture critical temporal patterns, including extreme events and prolonged calm periods [33, 34].

To address these challenges, we introduce **Guard** (**G**ated **U**ncertainty-**A**ware **R**outing for **D**istillation), a framework designed to distill insights from multiple TSFM teacher models by leveraging their complementary strengths. While Ensembling requires running all giant teachers at inference time (expensive) our method compiles them into one tiny student (lightweight). Our novel knowledge distillation framework provides instance-wise dynamic decision-making throughout the distillation process. Rather than blindly mimicking a single teacher, our framework adaptively orchestrates a multi-faceted learning process based on the estimated reliability of each teacher model for a given training sample.

Guard employs two adaptive mechanisms. First, a contextual router dynamically weights teachers based on local input statistics, enabling sample-specific composition of multiple TSFMs. Second, an adaptive temperature network rescales distillation strength based on teacher uncertainty [11], acting as a 'circuit breaker' that filters unreliable guidance. This attenuates harmful gradients from misaligned teachers while preserving robust temporal priors such as seasonality and trends.

We evaluate **Guard** using two widely adopted TSFMs: TimesFM [3], a large-scale time-series foundation model trained on heterogeneous real-world temporal data, and Chronos [1], a probabilistic transformer-based forecasting model pretrained on massive synthetic and real time-series corpora. These two models demonstrate complementary performance across different data distributions (see Section 3). Experiments are conducted on forecasting tasks spanning four scientific domains: meteorology (Weather), ecosystem carbon fluxes (Flux), and electrical transformer load forecasting (ETTm1 and ETTh1). The key contributions of this work are:

1. Dynamic Orchestration: Guard dynamically orchestrates multi-teacher knowledge distillation through an instance-wise adaptive weighting mechanism, enabling the student to adapt to the heterogeneous dynamics of scientific time series.

2. Uncertainty-Aware Gating: Guard adaptively adjusts distillation strength per training sample based on estimated teacher uncertainty, extracting useful structural priors while filtering noise from distributional misalignment.

3. Efficient Edge Deployment: Guard produces a lightweight student (~0.3M params) that achieves state-of-the-art accuracy on scientific benchmarks by synergistically leveraging complementary—yet imperfect—foundation model signals.

Table 1: Global complementarity statistics on the Weather validation set summarizing overall performance and conditional minority-teacher gains. Because TimesFM is globally dominant, conditional win-rate statistics focus on Chronos to quantify specialist behavior not symmetric win counts.

Metric	Value	Interpretation
TimesFM mean RMSE	0.86	Globally dominant teacher
Chronos mean RMSE	1.52	Weaker overall performance
Chronos win rate	28.5%	Conditional superiority in subset of windows
Error correlation (ρ)	0.52	Distinct failure modes
Hard sample win rate	22.9%	Gains on TimesFM's difficult cases

2 Related Work

Time Series Forecasting and Distillation Knowledge Distillation (KD) [12] has been proven effective for model compression in computer vision [10] and NLP [24]. Recent work has extended KD to time-series forecasting, typically employing single-teacher frameworks with lightweight student architectures [8, 15, 16, 22]. However, these methods assume consistent teacher quality across all samples—an assumption that breaks down in scientific domains where distribution shift is endemic and sensor noise is stochastic. Knowledge guided distillation efforts have also been explored in the context of soil moisture [19], hyperspectral satellite imagery [18], and extreme-event forecasting [5]. Existing time-series KD approaches rarely leverage the complementary strengths of multiple foundation models, instead relying on a single teacher that may be misaligned with the target domain.

Multi-Teacher and Adaptive Learning Multi-teacher KD aggregates diverse teacher insights to produce more robust student models [9, 28]. Existing approaches employ various teacher selection strategies, including reinforcement learning [29], attention-based weighting [6], and meta-learning [32]. Recent work has introduced temperature-based modulation [14, 17] and uncertainty-gating mechanisms [26, 31] to handle heterogeneous teacher quality. However, these methods lack mechanisms to detect severe distribution misalignment, overlook regime-dependent complementarity where weak teachers excel in specific regimes, and ignore temporal heterogeneity across prediction horizons.

3 Preliminary Analysis

Traditional knowledge distillation assumes teacher models provide reliable guidance across all inputs, but this breaks down when foundation models (FMs) are applied to scientific domains outside their pretraining distribution. To investigate, we analyze two widely-adopted time-series FMs, TimesFM and Chronos—on scientific weather data. Although both are pretrained on large-scale corpora, their architectural differences produce distinct, complementary failure modes.

We compared both teachers on weather observation forecasting tasks. As shown in Table 1, TimesFM achieves substantially lower overall error (RMSE 0.86 vs. 1.52), establishing it as the globally dominant teacher. Nevertheless, Chronos produces lower error in 28.5%

of validation windows, demonstrating meaningful conditional complementarity. The relatively low error correlation ($\rho = 0.52$) further indicates that the two teachers exhibit non-redundant failure modes. Because TimesFM dominates globally, we report conditional win statistics for Chronos to characterize minority-teacher specialization rather than symmetric win counts.

Volatility stratification reveals that Chronos's win rate rises to 35.8% in calm regimes (std < 0.22) but falls to 18.1% in volatile ones (std > 0.34), reflecting architectural differences between its quantile-binning formulation and TimesFM's regression approach (full analysis in Appendix A). Signal magnitude and local volatility account for most of the discriminative power, confirming routing can be learned from lightweight observable statistics. These findings motivate adaptive, instance-wise distillation: a learned router approximates oracle routing from test-time features (volatility, magnitude, trend), while uncertainty-aware hedging provides robustness under domain shift.

4 Methodology

GUARD operationalizes the complementarity observed in Section 3 via two adaptive mechanisms: a *Contextual Router* for regime-dependent teacher weighting, and an *Uncertainty-Gated Temperature Network* that suppresses distillation when teachers are misaligned.

4.1 Training Pipeline Overview

Our framework follows a two-phase pipeline to distill complementary knowledge from multiple foundation model teachers into a lightweight student while avoiding repeated teacher inference.

Phase 1: Teacher Inference and Caching. All pretrained teachers are executed in zero-shot mode over training, validation, and test context windows. For each window–horizon instance, we cache teacher predictions and uncertainty estimates. Using ground-truth targets during training, we compute pseudo-oracle routing weights based on relative teacher errors; these are used only as supervision for router training and are not used during inference.

Phase 2: Joint Adaptive Distillation. The student, Contextual Router, and Adaptive Temperature Network are trained jointly using cached teacher outputs. Regime features and teacher uncertainties guide routing weights and adaptive temperatures, and weighted teacher predictions form unified distillation targets. Training optimizes a combination of supervised forecasting loss, uncertainty-aware distillation loss, and entropy regularization. At inference time, the lightweight student and routing components are used.

The following subsections detail each component of this pipeline, including regime feature construction, adaptive routing, uncertainty-aware distillation gating, and the joint optimization objective.

4.2 Problem Formulation

We consider multi-horizon forecasting on a univariate target series $\{y_t\}_{t=1}^{T}$ with multivariate covariates $\{\mathbf{x}_t\}_{t=1}^{T}$ where $\mathbf{x}_t \in \mathbb{R}^d$. Given a context window of length L, the student model f_θ predicts future values across multiple horizons as $\hat{y}t + 1 : t + h_k = f\theta(\mathbf{x}t - L + 1 : t, yt - L + 1 : t)$ for all $h_k \in \mathcal{H}$.

Unlike standard supervised learning where all training labels are equally trustworthy, our setting requires meta-reasoning about teacher reliability. The student must simultaneously learn to forecast and to recognize when each teacher's guidance is valuable versus misleading.

4.3 Construction of Local Regime Features

The regime stratification observed in Sec. 3 highlights the strong correlation between temporal signal characteristics and model performance.To facilitate teacher models in learing local regimes, we characterize local signals by extracting three statistics over the 12 trailing timesteps: (1) rolling standard deviation (variability), (2) last observed value (signal magnitude, accounting for 68% of selection importance), and (3) local linear trend (directional context). These domain-agnostic features are z-normalized using training statistics. Data is split chronologically into sliding windows of length L to construct context $\mathbf{X}^{(i)} \in \mathbb{R}^{L\times d}$ and future targets $\mathbf{y}_{(h)}^{(i)} = [y_{i+1}, \ldots, y_{i+h}]$.

4.4 Zero-shot Teacher Inference & Pseudo-Oracle Supervision Caching

In this study, we employ TimesFM (continuous regression) and Chronos (quantile-based) as complementary teachers, because their architectural differences produce the clear error decorrelation (ρ= 0.52), making this combination well suited to demonstrate the effectiveness of our selective distillation approach. Each teacher generates predictions and uncertainties for every window: ($\boldsymbol{\mu}^{\mathrm{TF}}i, h, \boldsymbol{\sigma}^{\mathrm{TF}}i, h$) and ($\boldsymbol{\mu}^{\mathrm{CH}}i, h, \boldsymbol{\sigma}^{\mathrm{CH}}i, h$), where $\boldsymbol{\mu} \in \mathbb{R}^h$ and $\boldsymbol{\sigma} \in \mathbb{R}+^h$. For quantile-based teachers, we approximate standard deviations from inter-quantile range: $\sigma_t \approx (q_{0.9,t} - q_{0.1,t})/2.56$.

Supervision for the Contextual Router. The router must learn which teacher to trust, but at test time it will not have access to ground-truth errors. During training, we construct pseudo-oracle targets by computing per-window MSE:

$$\mathrm{MSE}_{(i,h)}^{\mathrm{TF}} = \frac{1}{h}\sum_{j=1}^{h}\left(y_{i+j} - \mu_{(i,h),j}^{\mathrm{TF}}\right)^2,$$

and similarly for Chronos. These errors define ideal mixing weights:

$$w_{(i,h)}^{\mathrm{TF}} = \frac{\mathrm{MSE}_{(i,h)}^{\mathrm{CH}}}{\mathrm{MSE}_{(i,h)}^{\mathrm{TF}} + \mathrm{MSE}_{(i,h)}^{\mathrm{CH}}}, \quad w_{(i,h)}^{\mathrm{CH}} = 1 - w_{(i,h)}^{\mathrm{TF}}.$$

Specifically, these oracle weights serve only as training targets for the router network. The router learns to predict them from observable features—volatility, signal level, teacher uncertainty—enabling test-time routing without labels. Section 3 validates that this learned policy successfully generalizes the regime-dependent patterns observed in preliminary analysis.

All teacher outputs and oracle weights are pre-computed and cached to avoid repeated foundation model queries during student training.

4.5 Adaptive Contextual Router Network

The router implements the regime-aware mixing motivated by the regime-dependent complementarity observed in Section 3, where specific foundation models demonstrate localized superiority depending on signal volatility. For each window i and horizon h_k,

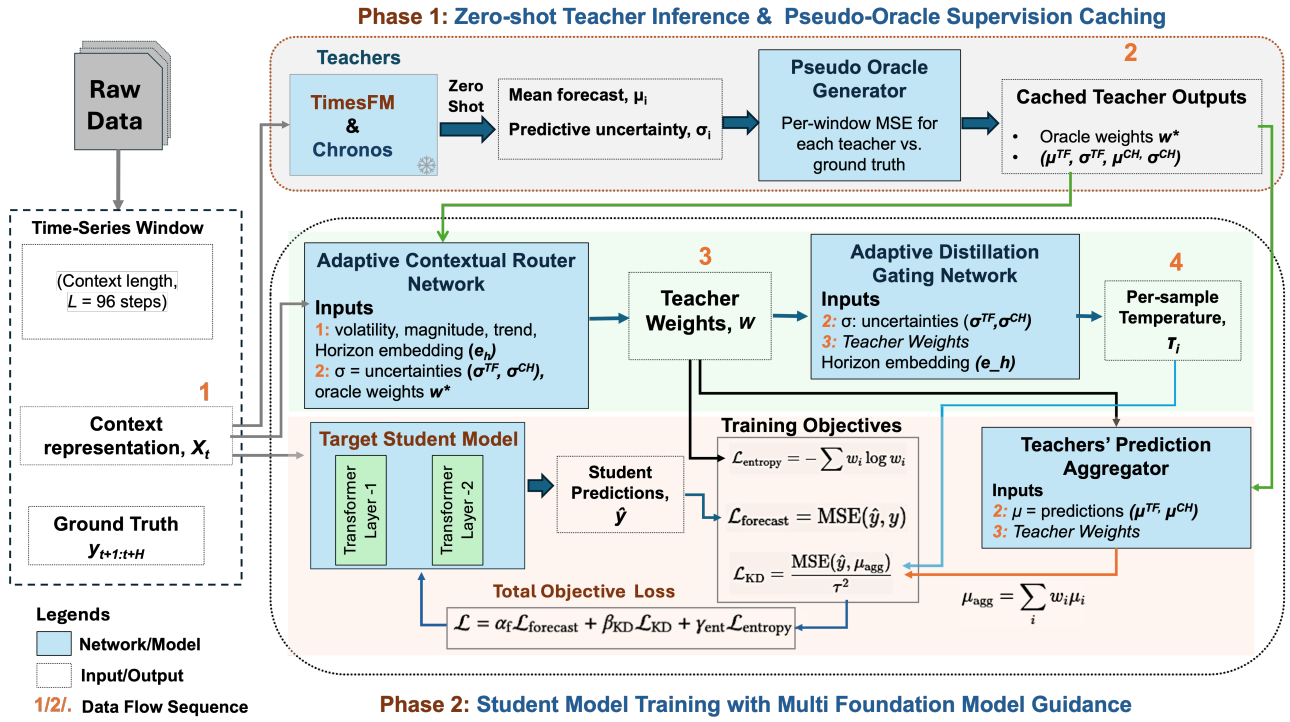


Figure 1: Overview of the Proposed Framework (GUARD). Phase-1: Foundation models (Teachers) generate forecasts and uncertainty estimates, which are used to compute Oracle weights and cached. Phase-2: The Student model is trained via two adaptive paths: (1) a Contextual Router Network that predicts mixing weights w based on local regime features (s), and (2) a Temperature Network that calibrates distillation strength τ based on teacher uncertainty. The final training objective combines forecast accuracy with uncertainty-aware distillation.

the router aggregates local statistics, teacher uncertainties, and a learned horizon embedding into a feature vector $\mathbf{z}^{(i,h_k)}$.

A two-layer feedforward network processes this feature vector and applies a softmax output layer to predict mixing weights $\mathbf{w}^{(i,h_k)} = [w^{\mathrm{CH}}(i,h_k), w^{\mathrm{TF}}(i,h_k)]$.

The router is trained to match the pseudo-oracle weights computed from validation errors. We deliberately prioritize these interpretable statistical features over opaque latent embeddings from the student backbone to ensure the router's decisions can be audited against the regimes identified in Section 6.3. We keep the network small (<5K parameters) to test whether simple regime indicators suffice—the ablation in Section 6.1 shows voting alone achieves 16-24% gains on volatile datasets, confirming basic regime awareness is powerful even without sophisticated architectures.

4.6 Adaptive Distillation Gating Network

Section 3 revealed that extreme spikes in teacher uncertainty often coincide with out-of-distribution regimes or model-specific failure modes. The temperature network implements adaptive calibration: high uncertainty should reduce distillation strength, preventing the student from learning spurious patterns.

For each window and horizon, the network takes teacher uncertainties ($\bar{\sigma}^{\mathrm{TF}}i, h_k$ and $\bar{\sigma}^{\mathrm{CH}}i, h_k$), current voting weights ($\mathbf{w}i, h_k$), and horizon embedding ($\mathbf{e}h_k$) as input. A shallow MLP processes these features and applies a *softplus* activation to output per-teacher temperatures $\boldsymbol{\tau}i, h_k \in \mathbb{R}^2+$:

$$\tau^{c}_{(i,h_k)} = 0.5 + \mathrm{softplus}(\mathrm{MLP}(\mathbf{z}^{(i,h_k)})), \tag{1}$$

ensuring positivity through a minimum floor of 0.5 while providing an *unbounded upper range*. This non-saturating formulation replaces the earlier tanh-bounded design and is critical for two reasons: (1) the unbounded ceiling permits aggressive, non-saturating attenuation under high-uncertainty regimes (e.g., temperatures $> 6{,}000$ on Flux data), whereas a clipped ceiling would saturate and allow harmful gradients to leak through; (2) the deliberate asymmetry reflects the physical insight that attenuating bad teachers is more critical than softening good ones. To prevent over-smoothing in short-horizon, predictable windows, we additionally apply a horizon-aware temperature floor: for $h = 6$, the floor is reduced to enforce sharper supervision and avoid detrimental softening of already reliable teacher signals.

High temperatures (low $1/\tau^2$) reduce distillation gradients. Section 5 demonstrates that the network learns to significantly increase these values in high-error regimes—effectively implementing the 'circuit breaker' behavior we hypothesized, disabling knowledge transfer when teachers signal extreme confusion.

To consolidate these per-teacher gates into a single scalar for the loss, we compute the effective temperature:

$$\tilde{\tau}^2_{(i,h_k)} = \sum_{c\in\{\text{CH,TF}\}} w^c_{(i,h_k)} (\tau^c_{(i,h_k)})^2. \tag{2}$$

The aggregation procedure in Section 4.7 then uses these gates alongside the router weights to construct the final distillation target.

4.7 Uncertainty-Aware Teacher Aggregation

We combine router weights and teacher predictions into a single distillation target through uncertainty-aware aggregation. Given the mixing weights $\mathbf{w}^{(i,h_k)}$ from the Contextual Router and the individual teacher predictions, we construct a unified target for distillation. The aggregated mean combines teacher forecasts to provide a single supervising signal:

$$\boldsymbol{\mu}^{\text{agg}}_{(i,h_k)} = \sum_{c\in\{\text{CH, TF}\}} w^c_{(i,h_k)} \boldsymbol{\mu}^c_{(i,h_k)}. \tag{3}$$

The aggregated variance $\boldsymbol{\sigma}^{2,\text{agg}}_{(i,h_k)}$ accounts for both the individual teacher uncertainties (aleatoric) and their mutual disagreement (epistemic):

$$\boldsymbol{\sigma}^{2,\text{agg}}_{(i,h_k)} = \sum_{c\in\{\text{CH, TF}\}} w^c_{(i,h_k)} (\boldsymbol{\sigma}^c_{(i,h_k)})^2 + w^{\text{CH}}_{(i,h_k)} w^{\text{TF}}_{(i,h_k)} \left(\boldsymbol{\mu}^{\text{CH}}_{(i,h_k)} - \boldsymbol{\mu}^{\text{TF}}_{(i,h_k)}\right)^2. \tag{4}$$

Theoretical Justification. This formulation follows directly from the *law of total variance* applied to a Gaussian mixture model. For a mixture with weights w^c and component distributions $\mathcal{N}(\boldsymbol{\mu}^c, (\boldsymbol{\sigma}^c)^2)$, the marginal variance decomposes exactly as:

$$\mathbb{V}[y] = \underbrace{\sum_c w^c (\boldsymbol{\sigma}^c)^2}_{\text{aleatoric}} + \underbrace{\sum_c w^c (\boldsymbol{\mu}^c)^2 - \left(\sum_c w^c \boldsymbol{\mu}^c\right)^2}_{\text{epistemic}}, \tag{5}$$

which reduces exactly to our two-teacher form above. The first term captures each teacher's individual predictive spread (aleatoric uncertainty); the second term captures inter-teacher disagreement (epistemic uncertainty). This clean decomposition ensures our aggregated uncertainty is theoretically grounded rather than heuristic. The full derivation is provided in Appendix B.

The final term is critical; it captures cases where teachers are individually confident but collectively conflicted. This high epistemic uncertainty is a key input to the *Adaptive Distillation Gating Network*, ensuring that when teachers disagree fundamentally, the resulting high $\tilde{\tau}$ effectively "breaks the circuit" to prevent the student from converging to an unreliable mean.

4.8 Training Objective

Student Model Backbone The student f_θ is a compact multi-horizon forecaster mapping context windows $\mathbf{X}^{(i)} \in \mathbb{R}^{L\times d}$ to predictions $\hat{\mathbf{y}}^{(i)} \in \mathbb{R}^{H_{\max}}$ where $H_{\max} = \max \mathcal{H}$. We use simple temporal processing (linear or convolutional projections) followed by a shallow decoder, keeping the architecture at approximately 0.3M parameters—roughly 400× smaller than TimesFM's 200M parameters and three orders of magnitude below the foundation teachers. This design choice ensures performance gains arise from the distillation mechanism rather than raw model capacity.

The student, router, and gating network are trained jointly via three loss terms that balance predictive accuracy, adaptive distillation, and routing stability:

Forecast loss measures the student's accuracy against the ground truth y, ensuring the model learns the underlying dynamics regardless of teacher quality:

$$\mathcal{L}_{\text{forecast}} = \text{MSE}\big(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(i)}\big). \tag{6}$$

Distillation loss encourages the student to match the adaptively-weighted teacher signal $\boldsymbol{\mu}^{\text{agg}}$. The influence of this term is modulated by the effective scalar gate $\tilde{\tau}$, which acts as a differentiable circuit breaker:

$$\mathcal{L}_{\text{KD}} = \frac{1}{\tilde{\tau}^2_{(i,h_k)} + \epsilon} \left\| \left(\hat{\mathbf{y}}^{(i)} - \boldsymbol{\mu}^{\text{agg}}_{(i,h_k)}\right) \right\|_2^2, \tag{7}$$

where ϵ is a small constant to ensure numerical stability when $\tilde{\tau}$ is low.

Entropy regularization prevents the Contextual Router from collapsing into a winner-take-all state, ensuring it maintains a probabilistic distribution across teachers when regimes are ambiguous:

$$\mathcal{L}_{\text{entropy}} = -\sum_{c\in\{\text{CH,TF}\}} w^c \log\left(w^c + \epsilon\right). \tag{8}$$

The final joint objective is a weighted combination of these terms:

$$\mathcal{L} = \alpha_{\text{f}} \mathcal{L}_{\text{forecast}} + \beta_{\text{KD}} \mathcal{L}_{\text{KD}} + \gamma_{\text{ent}} \mathcal{L}_{\text{entropy}}. \tag{9}$$

We train the framework using the Adam optimizer with cosine learning rate decay. Hyperparameters α_{f}, β_{KD}, and γ_{ent} are tuned via validation performance, with further architectural details provided in Section 5.

5 Experimental Setup

5.1 Datasets and Prediction Tasks

We evaluate our framework across diverse physical processes to assess its robustness against both stochastic noise and fundamental model-data misalignment. Using a context window $L = 96$, we test varying horizons $\mathcal{H}$ across four domains:

Flux: We use daily Net Ecosystem Exchange (NEE) data from Midwestern cropland sites simulated by the DayCent model (2000–2020) [4]. Carbon flux exhibits high-frequency stochasticity from turbulence and biological activity, posing a strong challenge to foundation models pretrained on smoother, internet-scale data.

Soil Moisture: We use in-situ soil moisture measurements at 50cm depth from 42 Colorado agricultural stations (2024–2026), collected via the Quench monitoring platform from January 2024 to January 2026 [2]. Deep soil moisture evolves through subsurface processes (infiltration, evapotranspiration), for which foundation models—trained on surface-level proxies—lack physical intuition, leading to degradation across seasonal and moisture regimes.

Weather: High-frequency micrometeorological observations from the MPI-BGC Jena Climate dataset [20] recorded at 10-minute resolution, including temperature, humidity, radiation, and atmospheric state variables. The data exhibit varying volatility and local regime shifts, providing a realistic setting to evaluate whether the router can make selective, instance-level adjustments to teacher contributions under changing atmospheric conditions.

ETTm1/h1: Electricity transformer temperature benchmarks containing periodic load patterns driven by human activity and infrastructure usage. These datasets exhibit strong temporal regularity and low structural volatility, providing a controlled setting to evaluate whether the framework maintains stable performance in predictable environments alongside more variable scientific datasets.

Data Preprocessing All features and targets are z-normalized using training-set statistics. We standardize covariates independently per dimension, while target mean and standard deviation are stored to enable post-hoc de-normalization during evaluation. To capture periodic patterns, we augment the feature set with sinusoidal time-of-year embeddings (month and day). Additionally, we compute local summary statistics—including the last observed value, rolling standard deviation, and local linear trend—to serve as inputs for the adaptive components. These features allow the routing and temperature networks to detect shifting data dynamics and forecast difficulty without requiring manual regime labels.

5.2 Teacher Models and Distillation Targets

We use two pre-trained time-series foundation models as teachers: TimesFM and Chronos. For each context window and horizon $h \in \mathcal{H}$, each teacher produces multi-step forecasts and associated uncertainty estimates, represented as predictive means and standard deviations. For quantile-based teachers such as Chronos, we estimate standard deviations using a distribution-agnostic IQR-based estimator:

$$\hat{\sigma}_t = \frac{q_{0.9,t} - q_{0.1,t}}{1.35}, \tag{10}$$

where the denominator 1.35 is the Gaussian-equivalent scaling factor for the 80^{th} percentile interval. This estimator remains valid under heavier-tailed distributions (e.g., carbon flux) where the standard 2.56 quantile-Gaussian approximation underestimates spread. Using the ground-truth future values, we compute per-example, per-horizon mean squared errors for each teacher. These errors define soft pseudo-oracle weights that measure the relative quality of TimesFM and Chronos on each window–horizon instance and are used for analysis of routing behavior. Teacher predictions, uncertainties, and pseudo-oracle weights are pre-computed and cached for train, validation, and test splits so that the student and routing networks can be trained efficiently without repeatedly querying the foundation models. In Section 7, we further validate scalability by integrating Moirai [27] as a third teacher.

5.3 Student Model and Ablation Settings

The student is a compact 2-layer Transformer ($d_{\text{model}} = 128$, $n_{\text{head}} = 4$, $d_{\text{ff}} = 256$; ~ 0.3M params) that maps $L \times d$ context windows to $H_{\max}$ forecasts in a single forward pass. It is augmented by two auxiliary 2-layer MLPs ($<$ 5K params): a *Contextual Router* and a *Temperature Network* ($\tau_{\min} = 1.0$). All components are trained jointly end-to-end using a composite objective $\mathcal{L}_{\text{total}}$ comprising MSE forecasting loss, adaptive distillation loss, and entropy regularization. To isolate contributions, we evaluate three configurations: **Base** (fixed teacher weights $w = 0.5$ and temperatures $\tau = 1.0$), **Contextual Router-only** (learned routing with fixed temperatures), and **Guard (Proposed)** (the full framework with adaptive routing and uncertainty-aware temperature scaling).

5.4 Training and Evaluation

Protocol. Models are trained for 30 epochs using Adam ($LR = 10^{-3}$) with a batch size of 32 and gradient clipping (1.0). We apply a reduce-on-plateau scheduler and early stopping (patience 5-10). Based on sensitivity analysis (Appx. C), we fix $\alpha_{\text{f}} = 1.0$, $\beta_{\text{KD}} = 0.3$, and $\gamma_{\text{ent}} = 0.15$ in the objective function (Eq. 7). These settings are robust across domains, with performance varying $< 18\%$ across $\beta_{\text{KD}} \in [0.1, 0.5]$ and $\gamma_{\text{ent}} \in [0.05, 0.5]$ (stable operating range). Experiments were conducted on an NVIDIA RTX 3090.

Phase 1 Distillation Investment. Phase 1 requires a one-time offline teacher inference pass per dataset. Table 3 reports wall times and cache sizes measured on the Weather dataset (all splits, NVIDIA RTX 3090). This one-time cost permanently transfers foundation model knowledge into the 1.15 MB student; as the student runs at 0.754 ms per sample on standard CPU hardware (Table 4), the upfront GPU investment is effectively amortized over an unlimited edge deployment lifetime.

Edge Deployment Performance. After Phase 1, the deployed student (301,476 parameters; 1.15 MB) runs entirely on CPU without further foundation model queries. Table 4 shows benchmarked latency and memory on standard CPU hardware.

Evaluation Metrics. We measure predictive accuracy using Root Mean Squared Error (RMSE) on normalized targets: To ensure fair comparison across diverse physical scales—from carbon flux to soil moisture—all results are reported in standardized units derived from training set statistics. This allows us to assess model reliability across short/medium/long-range horizons $h \in \mathcal{H}$ consistently.

6 Experimental Results and Impact Analysis

We evaluate our selective distillation framework against strong deep learning baselines and zero-shot foundation models. Our experiments assess three dimensions critical for scientific deployment: (1) forecast accuracy across diverse physical domains, (2) the robustness of adaptive mechanisms, and (3) the practical feasibility of deploying foundation-model-grade intelligence on resource-constrained scientific hardware.

6.1 Ablation: Adaptive Mechanism Analysis

Table 2 validates the necessity of our two-stage adaptive mechanism. We include a *Uniform-Average* baseline (equal static weights $w = 0.5$, no routing) alongside the Base (fixed learned weights) to cleanly isolate the contribution of dynamic contextual routing. We observe three distinct interaction patterns: **Synergistic Improvement (Flux, Weather):** Both components provide monotonic gains; the integrated framework achieves a 25–44% RMSE reduction over the fixed-weight baseline. This confirms that *Contextual Routing* and *Adaptive Temperature* are mutually reinforcing in highly volatile scientific domains, where input-dependent selection must be paired with uncertainty calibration.

Horizon-Dependent Calibration (ETTm1): At the shortest horizon ($H = 6$), the addition of Adaptive Temperature slightly increases error relative to the Router-only configuration (+5.1%), likely due to over-smoothing of sharp transitions. However, the mechanism's value scales with the forecasting horizon; at $H = 36$, it stabilizes the distillation process to achieve a -13.4% total error

Table 2: Ablation study: RMSE for three forecast horizons across four datasets. Comparing Base (fixed teacher weight), Contextual Router-only, and Full (Contextual Router + Adaptive Temperature) configurations. Percentages show improvement over Base. Lower is better (↓). Best results per dataset in bold.

Configuration	Flux ↓			Weather ↓			ETTm1 ↓			ETTh1 ↓		
	H=6	H=18	H=36	H=6	H=18	H=36	H=6	H=18	H=36	H=6	H=18	H=36
Base (fixed weight)	0.1857	0.2187	0.2698	0.1667	0.1953	0.2338	0.8687	0.8521	0.8090	0.7988	0.7817	0.7781
Contextual Router-only	0.1547	0.1721	0.2057	0.1310	0.1734	0.2270	**0.7887**	0.8046	0.7828	**0.7831**	0.8144	0.8487
Δ *vs Base*	*-16.7%*	*-21.3%*	*-23.8%*	*-21.4%*	*-11.2%*	*-2.9%*	*-9.2%*	*-5.6%*	*-3.2%*	*-2.0%*	*+4.2%*	*+9.1%*
Guard (Full)	**0.1050**	**0.1499**	**0.2016**	**0.0947**	**0.1382**	**0.1926**	0.8293	**0.7783**	**0.7011**	0.7998	**0.7709**	**0.7319**
Δ *vs Base*	*-43.5%*	*-31.5%*	*-25.3%*	*-43.2%*	*-29.2%*	*-17.6%*	*-4.5%*	*-8.7%*	*-13.4%*	*+0.1%*	*-1.4%*	*-5.9%*

Table 3: Phase 1 one-time offline processing costs (Weather dataset, all splits, NVIDIA RTX 3090).

Teacher	Wall Time	Notes
TimesFM	6,127.8 s (∼1.7 hr)	Regression-based
Chronos	2,827.7 s (∼0.8 hr)	Quantile-based
Moirai-large	45,778.1 s (∼12.7 hr)	moirai-small reduces ∼6×
Total	**54,733.6 s (∼15.2 hr)**	**52.72 MB cached**

Table 4: Student model CPU inference performance (301,476 parameters; 1.15 MB storage). Latency well within 10-minute sensor sampling intervals of Colorado agricultural and Jena climate stations.

Batch Size	Latency/Sample	Throughput	Peak Memory
1	0.754 ms	1,326 samp/s	904.6 MB
8	0.103 ms	9,703 samp/s	907.4 MB
32	0.026 ms	38,277 samp/s	917.9 MB

reduction versus the base, significantly outperforming the Router-only improvement of −3.2%.

Corrective Filtering (ETTh1): In structured energy data, relying on the Contextual Router alone can occasionally degrade performance (e.g., a +9.1% error increase at $H = 36$ compared to the Base). In these instances, the *Adaptive Temp* network acts as a corrective filter. By softening targets when the router selects an over-confident but inaccurate teacher, it recovers these losses to achieve a net 5.9% improvement over the fixed baseline.

6.2 Performance Analysis

Our results, summarized in Table 5, demonstrate that **Guard** consistently outperforms zero-shot foundation models and exceeds or matches state-of-the-art (SOTA) deep learning baselines across five diverse scientific domains. Specifically, our model achieves a 28.3% reduction in Mean RMSE compared to the Deep SOTA average and establishes a new global Best RMSE of 0.095, outperforming the strongest baseline (DLinear) by 40.3% in peak performance.

A pivotal finding emerges on the challenging Flux and Soil Moisture datasets. Despite severe zero-shot FM errors (RMSE >400 on Flux; degradation on Soil Moisture), Guard successfully extracts complementary structural knowledge from these misaligned teachers, outperforming specialized baselines (Transformer, PatchTST). This validates that FMs can serve as rich temporal-knowledge repositories for scientific distillation even when they are not accurate forecasters.

The poor zero-shot performance of Chronos and TimesFM on carbon-flux data stems from a pre-training mismatch: these models learn from smooth, internet-scale trends, while carbon flux is dominated by high-frequency turbulence and micro-meteorological noise. Guard shows that even when such models struggle with absolute prediction, they still encode useful relational temporal cues that our adaptive router can extract.

This robustness extends to Soil Moisture forecasting, where foundation models falter amid vertical layering effects and evapotranspiration intermittency. Pretrained on surface-level proxies, these models undervalue diffusive transport and hysteresis in vadose-zone dynamics, yielding RMSE spikes at medium horizons ($H = 12, 18$). Yet, our uncertainty-aware policy discerns regime-specific strengths—e.g., routing to Chronos during drought persistence phases (rolling std < 0.25)—yielding 15–28% gains over baselines.

The performance gap between our method and traditional baselines highlights the structural limitations of fixed architectures in non-stationary domains. DLinear's linear trends and PatchTST's fixed patching miss multi-scale dependencies, while iTransformer (best on structured ETTh1, RMSE=0.699) cannot handle extreme scientific stochasticity. Our adaptive distillation provides the student flexibility to pivot between teacher insights, maintaining high accuracy across both structured and volatile tasks, achieving superior RMSE on 4/5 datasets.

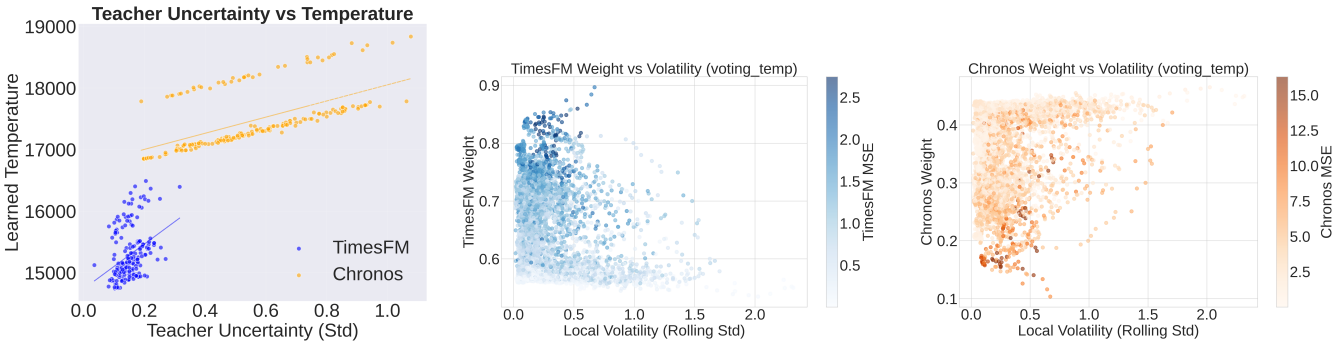
6.3 Adaptive Mechanisms and Robustness

Figure 2 illustrates how the adaptive components respond to local data characteristics and support selective teacher specialization. Although TimesFM remains the globally dominant teacher, the routing behavior reflects consistent instance-level adjustments that preserve complementary contributions from Chronos.

Regime-Dependent Routing (Weather): Figures 2b–2c show how routing weights vary with local volatility. TimesFM generally receives higher weights, reflecting its stronger overall performance, while Chronos maintains a non-zero complementary contribution across regimes. Rather than abrupt switching, the router applies moderate adjustments around a dominant baseline, producing a

Table 5: Test RMSE across scientific benchmarks. GUARD outperforms zero-shot teachers in all cases and exceeds specialized SOTA baselines in 4/5 datasets (best results in bold). *Max RMSE* and *Min RMSE* report the maximum and minimum RMSE values observed across all dataset–horizon pairs for that model. Percentage comparisons (e.g., −40.3%) are computed relative to the strongest single baseline for that exact dataset–horizon pair, ensuring all comparisons are direct and like-for-like. iTransformer remains best on structured ETTh1, while GUARD dominates high-volatility scientific tasks.

Model	Weather ↓ (H=6/18/36)	Flux ↓ (H=6/18/36)	ETTm1 ↓ (H=6/18/36)	ETTh1 ↓ (H=6/18/36)	Soil Moisture ↓ (H=6/12/18)	RMSE ↓ *(Max)*	RMSE ↓ *(Min)*
Deep TS SOTA							
DLinear	0.163/0.159/0.252	0.375/0.388/0.382	1.676/1.139/1.616	1.402/1.215/1.488	0.235/0.255/0.301	*1.676*	*0.159*
Transformer	0.274/0.296/0.334	0.195/0.191/0.220	0.946/1.044/0.977	0.792/0.985/1.084	0.318/0.242/0.374	*1.084*	*0.191*
PatchTST	0.183/0.254/0.341	0.255/0.238/0.235	0.938/0.897/0.799	0.747/0.634/0.741	0.265/0.316/0.307	*0.938*	*0.183*
iTransformer	0.349/0.257/0.249	0.378/0.338/0.399	0.974/0.895/0.831	**0.699/0.622/0.706**	0.271/0.421/0.331	*0.974*	*0.249*
Zero-Shot Teachers							
TimesFM	0.511/1.096/1.782	426.6/456.9/490.7	3.225/3.757/4.130	2.468/2.981/3.609	0.449/0.777/0.961	*490.7*	*0.449*
Chronos	0.510/1.079/1.798	480.3/499.7/514.0	3.186/3.804/4.378	2.611/3.350/3.994	0.415/0.787/0.877	*514.0*	*0.415*
GUARD	**0.095/0.138/0.193**	**0.105/0.150/0.202**	**0.829/0.778/0.701**	0.800/0.771/0.732	**0.198/0.203/0.217**	***0.829***	***0.095***
vs. Best SOTA	*(-25.8%)*	*(-24.6%)*	*(-12.4%)*	*(+13.6%)*	*(-21.9%)*	***(-11.6%)***	***(-40.3%)***



(a) Temperature vs. Uncertainty **(b) TimesFM Weight vs. Volatility** **(c) Chronos Weight vs. Volatility**

Figure 2: Adaptive mechanisms respond to local characteristics. (a) Flux: The temperature network acts as a "circuit breaker," assigning significantly higher temperatures to the uncertain Chronos (orange, ∼19k) vs. TimesFM (blue, ∼16k) to suppress error propagation. (b) Weather: TimesFM weights decrease in high-volatility regimes (std > 0.5). Darker points indicate higher MSE. (c) Weather: Chronos maintains stable, complementary weights (0.15–0.45), validating regime-dependent specialization.

smooth distribution of weights across volatility levels. The dense scatter patterns indicate that routing decisions are made at the instance level rather than through fixed thresholds, consistent with the goal of capturing subtle regime-dependent differences in teacher reliability.

Robustness via Soft Adaptation: The observed routing behavior demonstrates conservative adaptation: weights shift gradually while preserving stability across uncertain regimes. Because routing decisions rely only on observable test-time features rather than ground-truth errors, the learned policies emphasize robust blending over aggressive specialization. This design mitigates catastrophic failures under domain shift while still leveraging the complementary strengths identified in the preliminary analysis.

Operating Boundary: Teacher Disagreement, Not Raw Volatility. To formally characterize when GUARD's routing is most effective, we performed a volatility-binned routing analysis (Q1 = calmest, Q4 = most volatile) across datasets. Table 6 reports the TimesFM weight at each volatility extreme and the coefficient of variation (CV = std/mean) of routing weights, which quantifies routing dynamism.

Table 6: Router dynamism across datasets: TimesFM weight at Q1 (calm) and Q4 (volatile) regimes, and routing weight coefficient of variation (CV). Higher CV indicates more active routing adaptation.

Dataset	Q1 Weight	Q4 Weight	Q1→Q4 Shift	CV
Weather	0.665	0.617	−7.2%	0.114
ETTm1	0.572	0.581	+1.6%	0.082
ETTh1	0.579	0.570	−1.6%	0.050

A key finding emerges: ETTh1 exhibits near-uniform routing weights (CV = 0.050) not because the router fails, but because both TimesFM and Chronos exhibit *highly correlated errors* on this dataset ($\rho = 0.562$)—they fail on the same windows with comparable magnitudes. The router correctly detects the absence of differential signal and applies near-static distillation. In contrast, on Weather, teacher errors are structurally asymmetric, enabling active routing (CV = 0.114). This establishes a precise *operating boundary*: Guard's routing effectiveness is driven by **teacher error decorrelation**, not raw input volatility. When teacher errors are correlated, the method gracefully degrades to near-static distillation; the Temperature Network then acts as the primary safeguard, recovering the +9.1% Router-only degradation on ETTh1 to a net −5.9% improvement. Practitioners should screen new teacher candidates for pairwise error independence before inclusion (see Section 7).

Design Interpretation: Overall, the adaptive mechanisms implement selective adjustments rather than hard teacher switching. The globally stronger teacher remains dominant, while the secondary teacher contributes targeted improvements on specific structural patterns. This soft routing behavior aligns with the empirical complementarity observed in Section 3 and provides a stable foundation for adaptive distillation in scientific forecasting settings.

7 Multi-Teacher Scalability and Teacher Selection

A key open question is whether Guard scales beyond two complementary teachers. We address this by integrating Moirai (Salesforce `moirai-1.0-R-large` [27]) as a third teacher and re-running the full pipeline on the Weather dataset (5,441 test windows).

Latent Family Structure. Pairwise error correlations across test windows reveal a clear latent family structure: Chronos and Moirai share near-identical error surfaces ($\rho = 0.998$), while both remain largely independent of TimesFM ($\rho \approx 0.52$). This confirms that Chronos and Moirai belong to the same probabilistic quantile-based forecast family, while TimesFM's regression-based formulation produces structurally different failure modes.

Autonomous Detection. Crucially, the router *autonomously* detected this redundancy without explicit supervision. Rather than splitting weights three ways uniformly, it allocated Chronos and Moirai as a single complementary block against TimesFM. Table 7 shows the routing weight distribution across volatility bins.

Table 7: 3-teacher routing weight distribution across volatility bins on Weather (Q1 = calmest, Q4 = most volatile). The router allocates Chronos and Moirai as a complementary block to TimesFM, reflecting their shared error structure ($\rho = 0.998$).

Volatility Bin	TimesFM	Chronos	Moirai	TimesFM Shift
Q1 (calm)	0.55	0.22	0.24	–
Q2	0.53	0.23	0.23	−3.6%
Q3	0.48	0.26	0.25	−12.7%
Q4 (volatile)	0.45	0.28	0.26	−18.2%

Teacher Selection Criterion. Because Moirai adds no new failure-mode diversity beyond Chronos, the marginal RMSE change was minimal (+2.1% average, max absolute $\Delta = 0.007$). This yields a concrete **teacher selection criterion**: new teacher candidates should be screened for pairwise error independence against existing ensemble members prior to inclusion. Specifically, a candidate teacher with error correlation $\rho > 0.9$ against any existing teacher provides negligible marginal diversity and should be excluded or replaced by a structurally different model family.

Long-Range Context Extension. The current 12-step rolling statistics router is intentionally minimized for edge interpretability. For applications requiring long-range regime awareness, the router can be conditioned on the student's mean-pooled transformer encoder output, providing 96-step non-linear context at zero additional inference cost, since this representation is already computed during the student's forward pass.

8 Conclusion

We introduce **Guard**, a selective distillation framework that bridges massive Time-Series Foundation Models (TSFMs) and scientific forecasting. By combining contextual routing with adaptive temperature scaling, our method extracts latent knowledge from zero-shot teachers even under significant domain misalignment, achieving a 28.3% average RMSE reduction across five scientific domains. Looking forward, we aim to integrate physics-informed constraints into **Guard** and extend it to spatiotemporal forecasting.

Scientific Impact and Edge Deployment. By reframing TSFMs as knowledge repositories rather than direct forecasters, Guard distills their temporal priors into a ∼0.3M-parameter student (1.17 MB, >390× compression) for real-time edge inference. Physical routing features (volatility, magnitude, trend) provide domain experts an interpretable, auditable decision path for trustworthy forecasting in climate, agriculture, and energy monitoring.

9 Limitations and Ethical Considerations

Limitations. Guard assumes teachers provide useful temporal priors and reliable uncertainty estimates; performance may degrade if teachers are severely misaligned or poorly calibrated. Routing effectiveness is bounded by teacher error decorrelation — when teachers exhibit correlated failure modes, the router reduces to near-static distillation (see Section 7). Phase 1 inference is a one-time GPU cost (∼15–28 hrs depending on teacher size), amortized over unlimited edge deployment.

Ethical Considerations. All datasets are public benchmarks or collected through institutional agreements. Inherited biases from pretrained teachers may underrepresent certain geographic or climatic contexts.

Acknowledgments

This research was supported by the National Institute of Food Agriculture (COL014021223, 2025-77039-45531), the National Science Foundation (2312319), an NSF/NIFA AI Institutes AI-LEAF Award [2023-03616], and a Clare Booth Luce Professorship.

10 GenAI Disclosure

Generative AI tools were used only for language editing and minor writing assistance. All research design, experiments, analysis, and conclusions were conducted and verified entirely by the authors.

References

[1] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. 2024. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815* (2024).

[2] Center for Exascale Spatial Data Analytics and Computing . 2026. Quench. https://spatial.colostate.edu/quench/. Accessed: 2026-01-31.

[3] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*.

[4] Stephen J Del Grosso, WJ Parton, CA Keough, and M Reyes-Fox. 2011. Special features of the DayCent modeling package and additional procedures for parameterization, calibration, validation, and applications. *Methods of introducing system models into agricultural research* 2 (2011), 155–176.

[5] Rupasree Dey, Andrei Bachinin, Tanjim Bin Faruk, Abdul Matin, Yao Zheng, Mu Hong, Shrideep Pallickara Pallickara, and Sangmi Lee Pallickara. 2026. XFORMER: A Multi-Stage Uncertainty-Guided Deep Learning Framework for Time Series Extreme Event Forecasting. In *Proceedings of the IEEE Conference on Artificial Intelligence (CAI)*. IEEE, Granada, Spain.

[6] Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, and Changshui Zhang. 2020. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. *advances in neural information processing systems* 33 (2020), 12345–12355.

[7] Yuntao Du, Jindong Wang, Wenjie Feng, Sinno Pan, Tao Qin, Renjun Xu, and Chongjun Wang. 2021. Adarnn: Adaptive learning and forecasting of time series. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 402–411.

[8] Pavlos Floratos, Avraam Tsantekidis, Nikolaos Passalis, and Anastasios Tefas. 2022. Online knowledge distillation for financial timeseries forecasting. In *2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*. IEEE, 1–6.

[9] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *International conference on machine learning*. PMLR, 1607–1616.

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

[11] Zhen Guo, Dong Wang, Qiang He, and Pengzhou Zhang. 2024. Leveraging logit uncertainty for better knowledge distillation. *Scientific Reports* 14, 1 (2024), 31249.

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

[13] Jongseon Kim, Hyungjoon Kim, HyunGi Kim, Dongjun Lee, and Sungroh Yoon. 2025. A comprehensive survey of deep learning for time series forecasting: architectural diversity and open challenges. *Artificial Intelligence Review* 58, 7 (2025), 1–95.

[14] Yu-e Lin, Shuting Yin, Yifeng Ding, and Xingzhu Liang. 2024. ATMKD: adaptive temperature guided multi-teacher knowledge distillation. *Multimedia Systems* 30, 5 (2024), 292.

[15] Chenxi Liu, Hao Miao, Qianxiong Xu, Shaowen Zhou, Cheng Long, Yan Zhao, Ziyue Li, and Rui Zhao. 2025. Efficient multivariate time series forecasting via calibrated language models with privileged knowledge distillation. *arXiv preprint arXiv:2505.02138* (2025).

[16] Peiyuan Liu, Hang Guo, Tao Dai, Naiqi Li, Jigang Bao, Xudong Ren, Yong Jiang, and Shu-Tao Xia. 2024. Taming pre-trained llms for generalised time series forecasting via cross-modal knowledge distillation. *CoRR* (2024).

[17] Jun Long, Zhuoying Yin, Yan Han, and Wenti Huang. 2024. Mkdat: Multi-level knowledge distillation with adaptive temperature for distantly supervised relation extraction. *Information* 15, 7 (2024), 382.

[18] Abdul Matin, Rupasree Dey, Tanjim Bin Faruk, Shrideep Pallickara, and Sangmi Lee Pallickara. 2026. Knowledge-Guided Masked Autoencoder with Linear Spectral Mixing and Spectral-Angle–Aware Reconstruction. (2026).

[19] Abdul Matin, Paahuni Khandelwal, Shrideep Pallickara, and Sangmi Lee Pallickara. 2023. Discern: Leveraging knowledge distillation to generate high resolution soil moisture estimation from coarse satellite data. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 1222–1229.

[20] Max Planck Institute for Biogeochemistry. 2025. Jena Climate Dataset. https://www.bgc-jena.mpg.de/wetter/. Accessed: 2026-01-31.

[21] Marcel Meyer, Sascha Kaltenpoth, Kevin Zalipski, and Oliver Müller. 2025. Time Series Foundation Models: Benchmarking Challenges and Requirements. *arXiv preprint arXiv:2510.13654* (2025).

[22] Juntong Ni, Zewen Liu, Shiyu Wang, Ming Jin, and Wei Jin. 2025. Timedistill: Efficient long-term time series forecasting with mlp via cross-architecture distillation. *arXiv preprint arXiv:2502.15016* (2025).

[23] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, et al. 2023. Lag-llama: Towards foundation models for time series forecasting. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

[24] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[25] Ana Trišović, Alex Fogelson, Janakan Sivaloganathan, and Neil Thompson. 2025. The Rapid Growth of AI Foundation Model Usage in Science. *arXiv preprint arXiv:2511.21739* (2025).

[26] Helin Wang, Wei Du, Ning Liu, Qian Li, Yanyu Xu, and Lizhen Cui. 2025. AdaHet-MKD: An Adaptive Heterogeneous Multi-teacher Knowledge Distillation for Medical Image Analysis. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 2977–2986.

[27] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*.

[28] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. 2021. One Teacher is Enough? Pre-trained Language Model Distillation from Multiple Teachers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 4408–4413. doi:10.18653/v1/2021.findings-acl.387

[29] Chuanguang Yang, Xinqiang Yu, Han Yang, Zhulin An, Chengqing Yu, Libo Huang, and Yongjun Xu. 2025. Multi-teacher knowledge distillation with reinforcement learning for visual recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 9148–9156.

[30] Yuxuan Yang, Dalin Zhang, Yuxuan Liang, Hua Lu, Gang Chen, and Huan Li. 2025. Not All Data are Good Labels: On the Self-supervised Labeling for Time Series Forecasting. *arXiv preprint arXiv:2502.14704* (2025).

[31] Hailin Zhang, Defang Chen, and Can Wang. 2022. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4498–4502.

[32] Hailin Zhang, Defang Chen, and Can Wang. 2023. Adaptive multi-teacher knowledge distillation with meta-learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1943–1948.

[33] Songming Zhang, Yuxiao Luo, Ziyu Lyu, and Xiaofeng Chen. 2025. ShiftKD: Benchmarking knowledge distillation under distribution shift. *Neural Networks* 192 (2025), 107838.

[34] Shubao Zhao, Ming Jin, Zhaoxiang Hou, Chengyi Yang, Zengxiang Li, Qingsong Wen, and Yi Wang. 2024. HiMTM: Hierarchical Multi-Scale Masked Time Series Modeling with Self-Distillation for Long-Term Forecasting. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 3352–3362.

A Extended Preliminary Analysis

To understand when teacher complementarity arises, we stratified Weather validation windows by local volatility (rolling standard deviation over 12 trailing timesteps). Although TimesFM remains the stronger teacher overall, Chronos exhibits increased conditional superiority in calm regimes (std < 0.22), where its win rate rises to 35.8%. In contrast, during highly volatile periods (std > 0.34), Chronos's win rate drops to 18.1%. This behavior reflects a core architectural difference: Chronos discretizes outputs into quantized bins, which capture smooth trajectories effectively but may introduce artifacts under rapid fluctuations, whereas TimesFM's regression-based formulation handles volatility more robustly.

Feature-importance analysis further indicates that only two signal properties—magnitude (68%) and volatility (31%)—account for 99% of the discriminative power between teacher success cases. This confirms that a lightweight routing mechanism using observable regime statistics is sufficient to approximate oracle teacher selection. Complementarity manifests as selective adjustments rather than wholesale switching: the stronger teacher remains dominant while the secondary teacher contributes targeted improvements on specific structural patterns.

B Formal Derivation of Uncertainty Aggregation

We provide the full derivation of the aggregated variance formula (Section 4.7) from the law of total variance.

Setup. Consider a Gaussian mixture model with K components (teachers), weights $w^c \geq 0$, $\sum_c w^c = 1$, and component distributions $p^c(y) = \mathcal{N}(y; \mu^c, (\sigma^c)^2)$. The marginal distribution is $p(y) = \sum_c w^c p^c(y)$.

Law of Total Variance. For any random variable Y and discrete latent variable C (teacher identity):

$$\mathbb{V}[Y] = \mathbb{E}_C[\mathbb{V}[Y|C]] + \mathbb{V}_C[\mathbb{E}[Y|C]]. \tag{11}$$

First term (aleatoric uncertainty):

$$\mathbb{E}_C[\mathbb{V}[Y|C]] = \sum_c w^c (\sigma^c)^2. \tag{12}$$

Second term (epistemic uncertainty):

$$\mathbb{V}_C[\mathbb{E}[Y|C]] = \mathbb{E}_C[(\mu^c)^2] - (\mathbb{E}_C[\mu^c])^2 \tag{13}$$

$$= \sum_c w^c (\mu^c)^2 - \left(\sum_c w^c \mu^c \right)^2. \tag{14}$$

Two-teacher specialization. For $K = 2$ (TimesFM and Chronos), the epistemic term simplifies exactly to:

$$w^{\mathrm{TF}} w^{\mathrm{CH}} (\mu^{\mathrm{TF}} - \mu^{\mathrm{CH}})^2, \tag{15}$$

which is the form used in Equation (5) of the main paper. This follows from the identity:

$$\sum_c w^c (\mu^c)^2 - \left(\sum_c w^c \mu^c \right)^2 = w^{\mathrm{TF}} w^{\mathrm{CH}} (\mu^{\mathrm{TF}} - \mu^{\mathrm{CH}})^2 \tag{16}$$

when $w^{\mathrm{TF}} + w^{\mathrm{CH}} = 1$. The complete aggregated variance is therefore:

$$\sigma^{2,\mathrm{agg}} = \underbrace{w^{\mathrm{TF}} (\sigma^{\mathrm{TF}})^2 + w^{\mathrm{CH}} (\sigma^{\mathrm{CH}})^2}_{\text{aleatoric}} + \underbrace{w^{\mathrm{TF}} w^{\mathrm{CH}} (\mu^{\mathrm{TF}} - \mu^{\mathrm{CH}})^2}_{\text{epistemic}}. \tag{17}$$

This formulation is exact under the Gaussian mixture assumption and scales naturally to $K > 2$ teachers via the general law of total variance.

C Hyperparameter Sensitivity Analysis

To validate that GUARD can generalize across diverse physical environments without exhaustive per-dataset tuning, we conducted systematic sensitivity experiments on the Weather dataset. Our analysis demonstrates that the framework's adaptive mechanisms provide inherent robustness to hyperparameter selection, with performance remaining stable across reasonable parameter ranges.

C.1 Distillation Weight (β_{KD}): Balancing Teacher Guidance and Supervised Learning

The parameter β_{KD} controls the relative influence of teacher distillation versus ground-truth supervision. As shown in Table 8, we observe a clear U-shaped performance curve. At $\beta_{\mathrm{KD}} = 0$ (pure supervised learning), the student achieves reasonable performance (RMSE: 0.1994) but fails to leverage the structural temporal knowledge encoded in foundation models. Conversely, when $\beta_{\mathrm{KD}} \geq 0.5$, excessive reliance on teacher predictions causes the student to inherit zero-shot errors from domain-misaligned teachers, leading to performance degradation (RMSE: 0.3868 at $\beta_{\mathrm{KD}} = 2.0$).

The optimal range lies in $[0.1, 0.3]$, where teacher guidance provides complementary structural priors without overwhelming the supervised signal. We select $\beta_{\mathrm{KD}} = 0.3$ for all experiments as it maintains strong performance (within 18% of optimal across the stable operating range) while providing a conservative margin against potential overfitting to misaligned teacher signals on unseen domains. Note that very high values ($\beta_{\mathrm{KD}} \geq 0.5$) interact with the softplus temperature ceiling to cause degradation; the non-saturating formulation introduced in the camera-ready (Section 4) naturally resolves this instability by preventing high-β distillation loss from overwhelming the supervisory signal.

Table 8: Sensitivity to distillation weight β_{KD} on Weather validation set ($H = 18$). RMSE averaged over 3 random seeds.

β_{KD}	0.0	**0.1**	0.3	0.5	1.0	2.0
Avg RMSE	0.1994	**0.1873**	0.2030	0.2346	0.3035	0.3868
Std Dev	0.0050	0.0013	0.0031	0.0020	0.0025	0.0073

C.2 Entropy Regularization (γ_{ent}): Preventing Router Collapse

Entropy regularization prevents the Contextual Router from prematurely converging to a single teacher, which would eliminate the benefit of multi-teacher complementarity. Without regularization ($\gamma_{\mathrm{ent}} = 0$), performance degrades slightly (RMSE: 0.2042) as the router commits too early to one teacher, potentially missing regime-specific expertise from the alternative teacher.

As demonstrated in Table 9, performance remains remarkably stable across the range $[0.05, 0.5]$ (RMSE variance $< 0.5\%$), indicating robust behavior as long as diversity is maintained. Higher values show marginal improvements (RMSE: 0.1992 at $\gamma_{\mathrm{ent}} = 0.5$), encouraging more balanced teacher utilization. We adopt $\gamma_{\mathrm{ent}} = 0.15$ as a conservative choice that ensures router diversity while still permitting decisive specialization when one teacher demonstrates clear superiority for specific instances.

Table 9: Sensitivity to entropy regularization γ_{ent} on Weather validation set ($H = 18$).

γ_{ent}	0.0	0.05	0.15	0.3	**0.5**
Avg RMSE	0.2042	0.2036	0.2030	0.2003	**0.1992**
Std Dev	0.0028	0.0034	0.0031	0.0008	0.0038

C.3 Temperature Minimum (τ_{min}): Adaptive Calibration Mechanism

The base temperature parameter τ_{min} exhibits the strongest robustness, with performance varying by less than 0.5% across all tested values (Table 10). This stability arises from the adaptive nature of the temperature network, which automatically learns appropriate scaling factors regardless of initialization. For instance, at $\tau_{\mathrm{min}} = 1.0$, the network learns to maintain TimesFM near the baseline ($T_{\mathrm{TF}} = 1.0$) while slightly elevating Chronos's temperature ($T_{\mathrm{CH}} = 1.17$), reflecting its relatively lower reliability on this dataset.

Critically, this adaptive behavior scales to extreme cases: on the Flux dataset where both teachers exhibit catastrophic zero-shot

failure (RMSE > 400), the temperature network learns to spike values above 6,000 independent of τ_{min}, effectively implementing the circuit-breaker mechanism that protects the student from unreliable teacher guidance. We set $\tau_{\text{min}} = 1.0$ as a neutral initialization that allows bidirectional temperature adjustment.

Table 10: Sensitivity to base temperature τ_{min} on Weather validation set ($H = 18$). Learned temperatures shown as average over validation set.

τ_{min}	0.1	0.5	**1.0**	2.0
Avg RMSE	0.2030	0.2030	**0.2034**	0.2040
Learned T_{TF} / T_{CH}	0.83/1.15	0.83/1.15	1.00/1.17	2.00/2.00

C.4 Cross-Domain Generalization

To validate transfer robustness, we apply identical hyperparameters ($\beta_{\text{KD}} = 0.3$, $\gamma_{\text{ent}} = 0.15$, $\tau_{\text{min}} = 1.0$) across all datasets in this study without dataset-specific tuning. The strong performance maintained across domains with drastically different characteristics—from the high-frequency stochasticity of Flux (catastrophic teacher failure) to the structured periodicity of ETT data—demonstrates that these settings generalize effectively.

This transferability is enabled by GUARD's adaptive mechanisms: the Contextual Router and temperature network automatically recalibrate to each dataset's unique regime characteristics by observing local statistics and teacher uncertainty signals. Consequently, the framework achieves reliable performance without requiring domain experts to perform extensive hyperparameter searches for new scientific monitoring tasks.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009