DSSER: DYNAMIC SPARSE SPATIOTEMPORAL EVIDENCE RESTORATION FOR DRIFT-ROBUST MULTIVARIATE TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multivariate time series forecasting predicts future states from multiple interdependent historical series, yet real-world observations are often incomplete. Existing approaches typically address missingness by reconstructing unavailable values or learning representations robust under incomplete inputs. However, we observe that forecasts from incomplete observations drift away from those obtained with complete histories as missingness conditions vary. This drift arises from missingness-induced changes in the availability of historical evidence and the reliability of the remaining evidence. To mitigate such drift, we propose Dynamic Sparse Spatiotemporal Evidence Restoration (DSSER). The framework selectively leverages reliable evidence available under incomplete observations to improve forecasting stability. Reliability-Guided Evidence Routing (RGER) further regulates the contribution of restored evidence through learned routing confidence, introducing it as a bounded residual correction to the base forecast. To maintain reliability across training and inference, DSSER further introduces Cross-Stage Uncertainty Calibration (CSUC). Teacher predictive uncertainty and branch missingness jointly regulate teacher–student supervision during training, while learned routing confidence regulates restored evidence utilization at inference. Across three public traffic and air-quality benchmarks, DSSER consistently outperforms existing methods under random missingness, achieving the best results across all 36 dataset-rate-metric combinations. Under unseen-rate and dynamic-missing shifts, DSSER maintains robust performance, reducing MAE by 2.5%–3.6% on average relative to the best-performing non-DSSER baselines.

1 INTRODUCTION

Multivariate time series (MTS) forecasting plays a crucial role in various real-world applications, including traffic management, environmental monitoring, and urban sensing, where future states are predicted from historical sensor observations (Chen et al., 2025; Naghashi et al., 2025; Taga et al., 2025). In practice, real-world time series are often incomplete due to sensor failures, communication interruptions, and unreliable measurements (Al-Fakih et al., 2025; Sun et al., 2026). Such missing observations not only reduce available historical information but also disrupt the temporal and spatial cues underlying forecasting (Deng et al., 2024; Liu et al., 2026a). Recent advances in MTS forecasting have achieved remarkable progress by modeling complex temporal and spatial dependencies (Chen et al., 2024a). However, many existing methods assume sufficiently observed historical inputs (Qiu et al., 2025). Varying missingness alters both the composition and reliability of the spatiotemporal evidence available for prediction.

Existing approaches to MTS forecasting with incomplete observations mainly follow two paradigms, including two-stage imputation-then-forecast methods (Gupta et al., 2025) and end-to-end forecasting methods (Peng et al., 2025). Imputation-based methods reconstruct missing values before forecasting but may propagate reconstruction errors, while improved reconstruction quality does not necessarily yield better forecasting utility (Choi & Lee, 2024). In contrast, end-to-end methods directly incorporate incomplete observations into forecasting models and improve robustness under varying missing conditions. These methods mainly focus on recovering missing information or adapting forecasting models to incomplete inputs. However, when observation patterns change, the

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.

reliability of available spatiotemporal evidence may also vary, indicating that robustness depends on how information is recovered and whether the recovered evidence remains trustworthy.

As missingness patterns change, evidence that is informative under mild missingness may become unreliable under severe or shifted missingness (Kim et al., 2025; Liu et al., 2026b). This degradation causes forecasts from incomplete observations to deviate from those under complete histories, which we refer to as forecast drift under missingness. Figure 1 illustrates this phenomenon on PEMS04. Mitigating such drift requires more than recovering missing values. It requires reliability-aware regulation of restored forecasting evidence.

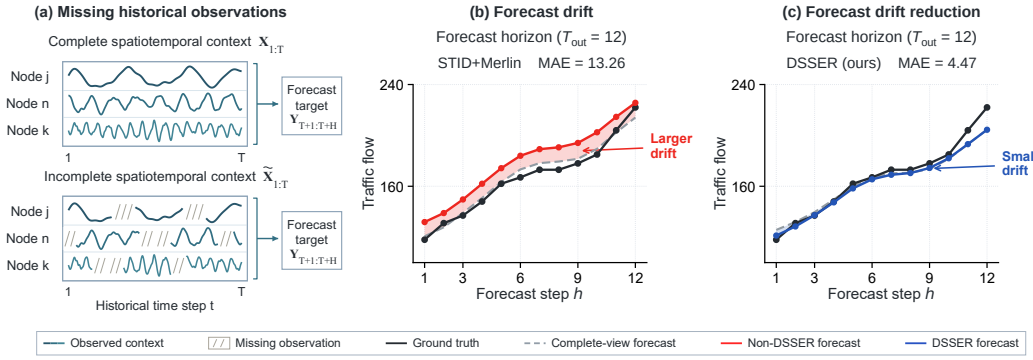


Figure 1: Missingness-induced forecast drift caused by changes in spatiotemporal evidence availability and reliability on PEMS04. (a) Missing observations reduce the available spatiotemporal evidence for forecasting. (b) Forecasting from incomplete histories can lead to deviations from the fully observed trajectory, resulting in forecast drift under missingness. (c) DSSER mitigates such drift by restoring sparse predictive evidence and regulating its contribution through reliability-guided evidence routing and bounded residual correction.

To mitigate forecast drift under missing observations, we propose Dynamic Sparse Spatiotemporal Evidence Restoration (DSSER), which combines a stable forecasting path with sparse evidence restoration. DSSER employs a lightweight forecasting backbone to produce a stable base forecast from incomplete histories, while sparse spatiotemporal evidence is restored from available observations, missingness cues, and temporal context. Reliability-Guided Evidence Routing (RGER) then regulates the contribution of restored evidence through learned routing confidence, introducing it as a bounded residual correction to the base forecast. During training, DSSER adopts a teacher-guided supervision framework, where a shared teacher receives the original historical input before masking and supervises student branches constructed from the same history under different missing rates (Hossain et al., 2025). Although the teacher accesses complete historical evidence, its predictions may still exhibit sample-dependent uncertainty (Hu et al., 2025a; Song et al., 2025), causing supervision reliability to vary across samples. To maintain reliability across training and inference, DSSER further introduces Cross-Stage Uncertainty Calibration (CSUC). CSUC combines teacher uncertainty with branch missingness to calibrate teacher–student supervision during training, while learned routing confidence regulates restored evidence utilization during inference. Our main contributions are summarized as follows:

- We characterize forecast drift under missingness in incomplete MTS forecasting, where changes in evidence availability and reliability cause forecasts from incomplete histories to deviate from their complete-view counterparts.
- We propose DSSER, which integrates dynamic sparse spatiotemporal evidence restoration with RGER to preserve a stable base forecasting path and introduce restored evidence as a bounded residual correction to the base forecast.
- We introduce CSUC, which adapts teacher–student supervision according to teacher uncertainty and branch missingness during training and leverages RGER for reliability-aware evidence utilization during inference.

2 Related Work

Forecasting from incomplete observations has attracted increasing attention, with recent studies incorporating missingness into model design. Marisca et al. (2024) and GraFITi (Yalavarthi et al., 2024) explore missing-aware representations for incomplete and irregular time series, while recent works further investigate probabilistic forecasting and representation learning under irregular observations and extensive missingness (Yalavarthi et al., 2025; Hu et al., 2025b). For regularly sampled MTSF, S4M (Peng et al., 2025) and GinAR+ (Yu et al., 2025a) integrate missing-data handling into forecasting, whereas Merlin (Yu et al., 2025b) improves robustness under varying missing rates through representation alignment. However, these methods mainly adapt model structures or representations, without explicitly regulating the reliability of information contributed under changing missingness conditions.

Another line of research addresses incomplete forecasting through missing information recovery. Recent studies develop uncertainty-aware and generative recovery methods for time series. CIB (Choi & Lee, 2024) preserves temporal dependencies through an information bottleneck framework while modeling uncertainty in imputed values, whereas BayOTIDE (Fang et al., 2024) performs Bayesian multivariate imputation with uncertainty quantification. Other works explore diffusion-based imputation (Chen et al., 2024b) and optimal transport approaches for distribution alignment under non-stationarity (Wang et al., 2025). Le Morvan & Varoquaux (2025) show that accurate imputations may provide limited benefits for downstream prediction. However, improved reconstruction quality does not necessarily guarantee reliable forecasting, since recovered values may contain unreliable evidence under changing missingness patterns. DSSER therefore restores forecast-relevant evidence rather than replacing missing inputs with reconstructed sequences.

Reliability-aware learning provides another perspective by estimating and weighting uncertain information. In knowledge distillation, Tang et al. (2024) introduce uncertainty-guided knowledge transfer to selectively exploit teacher information under domain shifts. Jain et al. (2024) further study predictive uncertainty through instance-dependent weighting, showing that uncertainty estimation can facilitate adaptive model optimization. For time series forecasting, Liu et al. (2025) propose an instance-aware post-hoc revision strategy that identifies unreliable forecasts and refines predictions using contextual information. However, existing approaches mainly consider reliability during supervision or inference separately, overlooking reliability shifts under changing missingness. We propose DSSER, which couples dynamic sparse spatiotemporal evidence restoration with RGER to address changes in evidence availability and reliability, selectively introducing restored evidence as a bounded residual correction to a stable base forecast.

3 Method

3.1 Problem Formulation

Let $\mathbf{X}_{1:T} = [X_{n,t,c}] \in \mathbb{R}^{N\times T\times C}$ denote a multivariate historical time series, where N, T, and C represent the numbers of nodes, historical time steps, and input features, respectively. The indices $n \in \{1,\dots,N\}$, $t \in \{1,\dots,T\}$, and $c \in \{1,\dots,C\}$ correspond to node, time, and feature dimensions. We designate the first input channel ($c = 1$) as the forecasting target and use $x_{n,t} = X_{n,t,1}$. Its observation status is represented by a binary mask $\mathbf{M}_{1:T} = [m_{n,t}] \in \{0,1\}^{N\times T}$, where $m_{n,t} = 1$ indicates an observed value and $m_{n,t} = 0$ indicates a missing value. The incomplete historical input is defined as $\widetilde{\mathbf{X}}_{1:T} = [\widetilde{X}_{n,t,c}]$, where $\widetilde{X}_{n,t,c} = m_{n,t}X_{n,t,c}$ for $c = 1$, while $\widetilde{X}_{n,t,c} = X_{n,t,c}$ for $c > 1$. Thus, the mask is applied to historical observations at the sensor-time level, while temporal covariates remain unchanged. Given the incomplete history and its missingness pattern, the objective is to predict $\mathbf{Y}_{T+1:T+H} = [y_{n,h}] \in \mathbb{R}^{N\times H}$, where $h \in \{1,\dots,H\}$ denotes the forecasting horizon and $y_{n,h}$ represents the ground-truth value of node n at horizon h. We study forecasting under varying missing rates and missing patterns, where missingness changes the available spatiotemporal evidence and makes both supervision and inference-time correction reliability-dependent.

3.2 Missingness-Aware Base Forecast

DSSER addresses incomplete forecasting through a stable forecasting path and a missingness-aware evidence path. The stable path produces a base forecast from incomplete histories, while the evi-

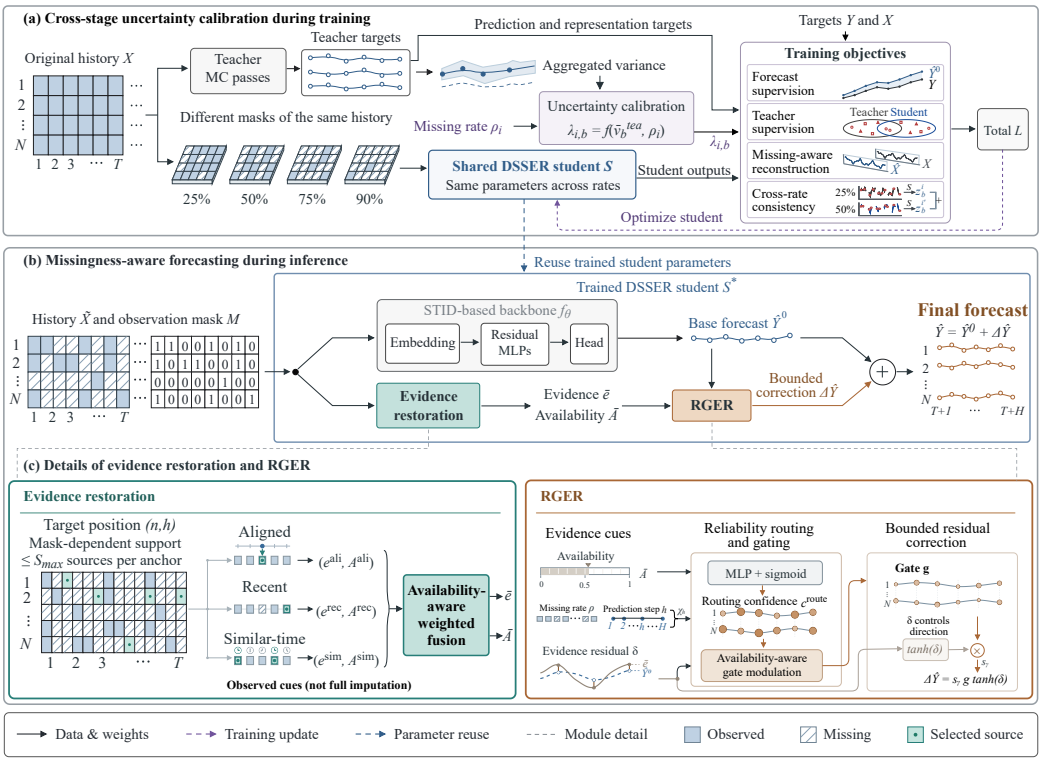


Figure 2: Overview of DSSER. (a) Cross-stage uncertainty calibration during training. (b) Missingness-aware forecasting during inference. (c) Dynamic sparse spatiotemporal evidence restoration and RGER.

dence restoration module restores sparse spatiotemporal evidence based on the current observation pattern. RGER then regulates the contribution of restored evidence to the base forecast through a bounded residual correction. Figure 2 provides an overview of DSSER. We adopt STID (Shao et al., 2022) as a standard forecasting substrate to provide the base forecast. To support forecasting from incomplete histories, the student additionally takes the observation mask as an explicit input and produces a historical reconstruction alongside the forecast:

$$(\mathbf{Z}, \hat{\mathbf{Y}}^0, \hat{\mathbf{X}}) = f_\theta(\widetilde{\mathbf{X}}, \mathbf{M}), \tag{1}$$

where f_θ is the student forecaster with parameters θ, $\mathbf{Z}$ is its hidden state, $\hat{\mathbf{Y}}^0 \in \mathbb{R}^{N\times H}$ is the base forecast, and $\hat{\mathbf{X}}$ is a historical reconstruction used only during training.

3.3 Spatiotemporal Evidence Restoration and RGER

Unlike impute-then-forecast pipelines, DSSER does not replace the forecasting input with reconstructed sequences. Instead, it maintains a base forecasting path and introduces restored evidence through RGER, avoiding the assumption that restored evidence is uniformly reliable under severe missingness. For each prediction position (n, h), DSSER constructs a tri-anchor evidence pool indexed by $\mathcal{K} = \{\mathrm{ali}, \mathrm{rec}, \mathrm{sim}\}$, corresponding to aligned, recent, and similar-time evidence, respectively. These three evidence types capture complementary temporal contexts, with e^{ali}, e^{rec}, and e^{sim} being derived from temporally aligned positions, recent available contexts, and analogous historical periods, respectively. Each evidence source is further associated with an availability score A reflecting its reliability under missing observations. For each anchor $k \in \mathcal{K}$, $\mathcal{S}^k_{n,h} \subseteq \{1,\dots,N\} \times \{1,\dots,T\}$ denotes the candidate source set for target (n, h), and $q^k_{n,h,j,t} \geq 0$ denotes the normalized weight assigned to source position $(j, t) \in \mathcal{S}^k_{n,h}$, with $\sum_{(j,t)\in\mathcal{S}^k_{n,h}} q^k_{n,h,j,t} = 1$, where n and h denote the target node and forecast horizon, while j and t denote the historical source node and time step. To keep evidence retrieval sparse, each anchor

retains at most $S_{\max}$ source positions, i.e., $|\mathcal{S}^k_{n,h}| \le S_{\max} \ll NT$. The effective observed support is computed as $a^k_{n,h} = \sum_{(j,t)\in\mathcal{S}^k_{n,h}} q^k_{n,h,j,t} m_{j,t}$, where $m_{j,t}$ denotes the observation mask. The restored evidence is obtained as:

$$e^k_{n,h} = \frac{\sum_{(j,t)\in\mathcal{S}^k_{n,h}} q^k_{n,h,j,t} m_{j,t} x_{j,t}}{a^k_{n,h} + \epsilon}. \tag{2}$$

where $x_{j,t}$ denotes the observed feature, $a^k_{n,h}$ the effective support, and $e^k_{n,h}$ the anchor-specific evidence. The corresponding availability is defined as $A^k_{n,h} = 1 - \exp(-a^k_{n,h})$. Lower support reduces availability and weakens the corresponding contribution during fusion. We then perform availability-aware anchor fusion using learned priors β_k, where the anchor mixture weight is computed as $\pi^k_{n,h} = \frac{\exp(\beta_k)(A^k_{n,h}+\epsilon)}{\sum_{k'\in\mathcal{K}} \exp(\beta_{k'})(A^{k'}_{n,h}+\epsilon)}$. The aggregated evidence and availability are given by:

$$(\bar{e}_{n,h}, \bar{A}_{n,h}) = \sum_{k\in\mathcal{K}} \pi^k_{n,h} (e^k_{n,h}, A^k_{n,h}), \tag{3}$$

where β_k denotes the learned anchor prior, and $\bar{e}_{n,h}$ and $\bar{A}_{n,h}$ denote the aggregated evidence and availability, respectively. Rather than replacing the base forecast, RGER introduces the restored evidence as a reliability-conditioned residual. With $s_y > 0$ denoting the training-set target scale, the normalized evidence residual is:

$$\delta_{n,h} = \frac{\bar{e}_{n,h} - \hat{\mathbf{Y}}^0_{n,h}}{s_y + \epsilon}. \tag{4}$$

RGER conditions the routing decision on residual disagreement, evidence availability, and the current missingness-and-horizon context. We compute the window-level missing rate as $\rho = 1 - \frac{1}{NT}\sum_{\nu=1}^{N}\sum_{t=1}^{T} m_{\nu,t}$. Missingness level and forecast position are encoded into a context signal $\chi_h = \rho + \gamma[\sin(\frac{2\pi(h-1)}{H}) + \cos(\frac{2\pi(h-1)}{H})]$, where γ controls the scaling of forecast-position encoding. The router maps the residual disagreement $|\delta_{n,h}|$, evidence availability $\bar{A}_{n,h}$, and forecast-step-aware context χ_h into a shared routing state:

$$\mathbf{r}_{n,h} = \mathbf{W}_r[|\delta_{n,h}|; \bar{A}_{n,h}; \chi_h] + \mathbf{b}_r, \tag{5}$$

where $\mathbf{r}_{n,h}$ is the routing state, mapped by a multilayer perceptron to the evidence-routing logit $\ell_{n,h} = \mathbf{w}_2^\top \tanh(\mathbf{W}_1\mathbf{r}_{n,h} + \mathbf{b}_1) + b_2$. The final gate combines learned routing confidence, evidence availability, and missingness activation:

$$g_{n,h} = s_{\max}\bar{A}_{n,h}\sigma(\eta(\rho - \rho_0))\, c^{\text{route}}_{n,h}, \qquad 0 \le g_{n,h} \le s_{\max}, \tag{6}$$

where $c^{\text{route}}_{n,h} = \sigma(\ell_{n,h})$ denotes the learned evidence-routing confidence, $\sigma(\cdot)$ is the logistic sigmoid, $s_{\max} > 0$ bounds the gate magnitude, ρ_0 is the missing-rate activation threshold, and $\eta > 0$ controls the transition sharpness. The bounded forecast correction is computed as $\Delta\hat{Y}_{n,h} = s_y g_{n,h} \tanh(\delta_{n,h})$, where $\tanh(\delta_{n,h}) \in [-1, 1]$ preserves the correction direction and $g_{n,h}$ controls the correction magnitude. The final prediction is obtained by adding the bounded residual correction to the base forecast, i.e., $\hat{\mathbf{Y}}_{n,h} = \hat{\mathbf{Y}}^0_{n,h} + \Delta\hat{Y}_{n,h}$.

3.4 Cross-Stage Uncertainty Calibration

CSUC couples uncertainty-aware supervision during training with reliability-guided evidence utilization during inference. We construct multiple student branches from the same history under different missing rates, supervised by a teacher receiving the original history. Let $i \in \mathcal{I}$ index student branches with missing rate ρ_i, and b index samples in a batch of size B. Following calibration-aware knowledge distillation (KD) (Fan et al., 2024), DSSER models the sample-wise reliability of teacher supervision, since teacher predictions may remain uncertain despite access to complete histories. DSSER estimates teacher uncertainty from R Monte Carlo forward passes to calibrate supervision and reconstruction. During inference, the routing confidence $c^{\text{route}}_{n,h}$ produced by RGER regulates the contribution of restored evidence. Let $\hat{\mathbf{Y}}^{\text{tea},(r)}_b$ denote the r-th stochastic teacher prediction for sample b. The predictive mean and variance are:

$$(\boldsymbol{\mu}^{\text{tea}}_b, \mathbf{v}^{\text{tea}}_b) = \left(\frac{1}{R}\sum_{r=1}^{R} \hat{\mathbf{Y}}^{\text{tea},(r)}_b, \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\mathbf{Y}}^{\text{tea},(r)}_b - \boldsymbol{\mu}^{\text{tea}}_b\right)^2\right). \tag{7}$$

DSSER aggregates predictive variance over nodes and horizons to obtain the sample-level teacher uncertainty $\bar{v}_b^{\mathrm{tea}}$:

$$\bar{v}_b^{\mathrm{tea}} = \frac{1}{NH} \sum_{n=1}^{N} \sum_{h=1}^{H} v_{b,n,h}^{\mathrm{tea}}, \tag{8}$$

where $v_{b,n,h}^{\mathrm{tea}}$ denotes the teacher predictive variance at node n and horizon h. Similar to uncertainty-aware weighting approaches (Jain et al., 2024), we estimate sample-level reliability from predictive uncertainty. The aggregated uncertainty $\bar{v}_b^{\mathrm{tea}}$ and branch missing rate ρ_i determine the calibration weight $\lambda_{i,b} = \sigma(\kappa_0 - \kappa_1 \log(\bar{v}_b^{\mathrm{tea}} + \epsilon) - \kappa_2 \rho_i)$, where κ_0, κ_1, and κ_2 control calibration offset, uncertainty sensitivity, and missing-rate sensitivity, respectively, and $0 \le \lambda_{i,b} \le 1$. Lower teacher uncertainty and missing rates lead to stronger teacher supervision. Let $\mathcal{L}_{\mathrm{MIT}}^{i,b}$ and $\mathcal{L}_{\mathrm{ORT}}^{i,b}$ denote reconstruction losses on missing-input-target and observed-reconstruction-target entries for branch i and sample b, respectively. The reconstruction loss for branch i is adaptively allocated according to the calibration weight as $\mathcal{L}_{\mathrm{rec}}^{i} = \frac{1}{B} \sum_{b=1}^{B} [(1-\lambda_{i,b}) \mathcal{L}_{\mathrm{MIT}}^{i,b} + \lambda_{i,b} \mathcal{L}_{\mathrm{ORT}}^{i,b}]$. A smaller calibration weight shifts reconstruction emphasis toward missing entries. The calibration weight $\lambda_{i,b}$ also regulates teacher–student knowledge distillation. For branch i and sample b, we define representation- and prediction-level knowledge distillation (KD) losses using mean squared error (MSE), denoted as $\mathcal{D}_{\mathrm{rep}}^{i,b} = \mathrm{MSE}(\mathbf{Z}_{i,b}^{\mathrm{stu}}, \mathbf{Z}_b^{\mathrm{tea}})$ and $\mathcal{D}_{\mathrm{pred}}^{i,b} = \mathrm{MSE}(\hat{\mathbf{Y}}_{i,b}^{\mathrm{stu}}, \boldsymbol{\mu}_b^{\mathrm{tea}})$, respectively, where $\mathbf{Z}_{i,b}^{\mathrm{stu}}$, $\mathbf{Z}_b^{\mathrm{tea}}$, and $\hat{\mathbf{Y}}_{i,b}^{\mathrm{stu}}$ denote the student representation, teacher representation, and student prediction, respectively. The sample-wise KD discrepancy is then defined as $\mathcal{D}_{\mathrm{KD}}^{i,b} = \omega_{\mathrm{rep}} \mathcal{D}_{\mathrm{rep}}^{i,b} + \omega_{\mathrm{pred}} \mathcal{D}_{\mathrm{pred}}^{i,b}$, where ω_{rep} and ω_{pred} balance the two distillation terms. We further preserve the teacher node-relation structure through correlation distillation. The calibrated correlation loss is defined as:

$$\mathcal{L}_{\mathrm{corr}} = \sum_{i \in \mathcal{I}} \frac{1}{B} \sum_{b=1}^{B} \lambda_{i,b} \ell_{\mathrm{corr}}^{i,b}, \tag{9}$$

where $\ell_{\mathrm{corr}}^{i,b} = N^{-2} \sum_{n=1}^{N} \sum_{n'=1}^{N} \mathrm{SmoothL1}(R_{b,n,n'}^{i,\mathrm{stu}}, R_{b,n,n'}^{\mathrm{tea}})$, and $\mathbf{R}_b^{i,\mathrm{stu}}$ and $\mathbf{R}_b^{\mathrm{tea}}$ denote the node-similarity matrices of the L_2-normalized student and MC-mean teacher representations, respectively, and $\mathrm{SmoothL1}$ is the standard Smooth-L_1 discrepancy. We further introduce cross-rate contrastive regularization across student branches. For two branches i and i', the same historical sample b under different missing rates forms a positive pair, while samples $b' \neq b$ from branch i' serve as negatives for anchor sample b. Negative contributions are reweighted by teacher representation similarity, and each anchor sample is weighted according to teacher uncertainty. The contrastive regularizer is:

$$\mathcal{L}_{\mathrm{CL}} = \frac{1}{|\mathcal{I}|(|\mathcal{I}|-1)B} \sum_{\substack{i,i' \in \mathcal{I} \\ i' \neq i}} \sum_{b=1}^{B} \alpha_b \left[\log \left(\sum_{\substack{b'=1 \\ b' \neq b}}^{B} w_{bb'} \exp\left(s_{bb'}^{ii'} \right) + \epsilon \right) - s_{bb}^{ii'} \right], \tag{10}$$

where $s_{bb'}^{ii'} = (\mathbf{z}_b^i)^\top \mathbf{z}_{b'}^{i'} / \tau_c$ denotes the cross-branch similarity, $w_{bb'} = \sigma((\psi_0 - (\bar{\mathbf{z}}_b^{\mathrm{tea}})^\top \bar{\mathbf{z}}_{b'}^{\mathrm{tea}})/\tau_w)$ reweights negative pairs according to teacher similarity, and $\alpha_b = \exp(-\zeta u_b)$ provides uncertainty-aware sample weighting. Here, $\mathbf{z}_b^i$ and $\bar{\mathbf{z}}_b^{\mathrm{tea}}$ denote normalized student and teacher contrastive representations. The parameter τ_c is the contrastive temperature, ψ_0 and τ_w control teacher-similarity-based negative reweighting, u_b denotes the sample-level MC variability of the teacher contrastive representation, and ζ controls variability-based sample weighting. Combining these training terms, the overall objective is:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\mathrm{pred}} &+ \lambda_{\mathrm{kd}} \sum_{i \in \mathcal{I}} \frac{1}{B} \sum_{b=1}^{B} \lambda_{i,b} \mathcal{D}_{\mathrm{KD}}^{i,b} \\ &+ \lambda_{\mathrm{rec}} \sum_{i \in \mathcal{I}} \mathcal{L}_{\mathrm{rec}}^{i} + \lambda_{\mathrm{corr}} \mathcal{L}_{\mathrm{corr}} + \lambda_{\mathrm{cl}} \bar{\lambda} \mathcal{L}_{\mathrm{CL}}, \end{aligned} \tag{11}$$

where $\bar{\lambda} = (|\mathcal{I}|B)^{-1} \sum_{i \in \mathcal{I}} \sum_{b=1}^{B} \lambda_{i,b}$ is the average calibration weight over branches and samples, and λ_{kd}, λ_{rec}, λ_{corr}, and λ_{cl} are non-negative loss weights. $\mathcal{L}_{\mathrm{pred}}$ denotes the supervised forecasting loss between the student base forecast $\hat{\mathbf{Y}}^0$ and the ground-truth future $\mathbf{Y}$.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate DSSER on three multivariate forecasting benchmarks, including PEMS04 (Cao et al., 2025), METR-LA (Lee & Ko, 2024), and China-AQI (Yu et al., 2025b). PEMS04 and METR-LA are traffic forecasting datasets, while China-AQI is an air-quality dataset. Following previous studies, we use the past 12 steps to forecast the next 12 steps with identical splits for all methods. The main evaluation considers random missingness at rates of 25%, 50%, 75%, and 90%. Missing masks are fixed for training and evaluation, with identical masks used across methods. Missing masks are generated by randomly masking a specified percentage of points from historical observations at the sensor-time level. The mask is applied to the target observation channel, while temporal covariates remain unchanged. Multiple sensors may have missing observations simultaneously under the same missing rate. During training, DSSER jointly optimizes student branches constructed under these missing rates, while generalization to unseen missing rates and temporally varying missingness is evaluated in Sec. 4.3. We compare DSSER with representative forecasting and missing-data methods, including STID (Shao et al., 2022), STID+GinAR+ (Yu et al., 2025a), Merlin (Yu et al., 2025b), SADI (Islam et al., 2025), Glocal-IB (Yang et al., 2025), T1 (Park et al., 2026), iTransformer (Liu et al., 2024), PatchTST (Nie et al., 2023), ImputeFormer (Nie et al., 2024), and S4M (Peng et al., 2025). We report root mean squared error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE), where lower values indicate better forecasting performance. Experimental details are provided in Appendix A.

4.2 MAIN RESULTS

Table 1 reports the forecasting results under random missingness. DSSER achieves the lowest RMSE, MAPE, and MAE across all three datasets and four missing rates, yielding the best result in all 36 reported cases. Beyond this advantage, relative improvements in RMSE and MAE increase as missing rates rise from 25% to 90%. The corresponding relative MAE reduction over the best-performing non-DSSER baseline increases from 1.47% to 5.07% on PEMS04, from 1.25% to 4.99% on METR-LA, and from 2.31% to 4.69% on China-AQI. The larger gains under severe missingness indicate that DSSER retains its forecasting advantage as the available historical observations become increasingly sparse. Appendix G provides additional analysis of missingness-induced forecast drift.

4.3 ROBUSTNESS UNDER UNSEEN-RATE AND DYNAMIC-MISSING CONDITIONS

We further evaluate the robustness of DSSER under missing-rate shifts during inference. We consider two evaluation settings, namely unseen-rate and dynamic-missing conditions. For unseen-rate evaluation, the missing rate remains fixed within each test condition but takes values not used during training. Specifically, training uses missing rates of 25%, 50%, 75%, and 90%, while testing is conducted at 15%, 40%, 60%, 85%, and 95%. This setting evaluates robustness to missing-rate values not encountered during training. As shown in Figure 3, DSSER achieves the lowest MAE and RMSE across all five unseen-rate settings on both datasets, while Figure 4(a) provides a detailed comparison on PEMS04. Additional METR-LA results are provided in Appendix D. For dynamic-missing evaluation, the missing rate changes across consecutive temporal segments according to four predefined schedules (S1–S4), rather than remaining fixed throughout the test sequence. This setting evaluates robustness to temporally varying missingness. As shown in Figure 4(b), DSSER consistently outperforms all compared baselines across the four dynamic-missing schedules on PEMS04, demonstrating robustness under temporally varying missingness.

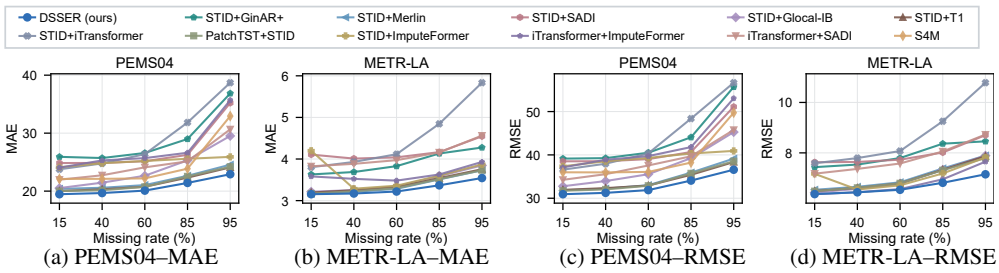
4.4 ABLATION STUDY

We conduct ablation studies to evaluate the contributions of three key components in DSSER, namely missingness-aware spatiotemporal evidence restoration, RGER, and CSUC. As shown in Table 2, the full DSSER generally achieves better performance than the ablated variants across tested missing rates on both datasets, with larger gaps under higher missingness. Specifically, removing missingness awareness from evidence restoration generally reduces forecasting accuracy, indicating the importance of conditioning evidence recovery on the current observation pattern. Removing RGER affects performance, demonstrating the contribution of reliability-guided routing and

Table 1: Forecasting performance on PEMS04, METR-LA, and China-AQI under random missingness at rates of 25%, 50%, 75%, and 90%. Best results are bolded and second-best results are underlined.

Dataset	Method	25%			50%			75%			90%		
		RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE
PEMS04	**DSSER (ours)**	**30.99**	**13.63**	**19.50**	**31.38**	**13.67**	**19.75**	**32.64**	**14.23**	**20.55**	**34.92**	**15.16**	**21.89**
	STID (Shao et al., 2022)	36.94	16.26	23.07	37.60	16.50	23.55	38.35	16.63	24.05	39.98	17.47	25.40
	STID+GinAR+ (Yu et al., 2025a)	39.04	19.22	25.69	39.74	19.63	26.03	42.22	20.98	27.82	46.32	21.64	30.50
	STID+Merlin (Yu et al., 2025b)	31.41	14.01	19.79	32.02	14.22	20.24	33.70	14.73	21.30	36.63	15.54	23.06
	STID+SADI (Islam et al., 2025)	32.71	15.05	20.76	33.43	15.23	21.25	35.24	15.74	22.40	38.16	16.52	24.15
	STID+Glocal-IB (Yang et al., 2025)	33.07	15.26	20.77	34.59	15.85	21.93	37.05	17.01	23.74	40.48	18.35	26.27
	STID+T1 (Park et al., 2026)	32.07	14.13	20.12	32.58	14.36	20.50	33.95	15.09	21.42	36.62	16.20	23.06
	STID+iTransformer (Liu et al., 2024)	31.49	13.99	19.86	32.17	14.19	20.35	33.91	14.70	21.44	36.78	15.46	23.11
	PatchTST+STID (Nie et al., 2023)	31.75	14.23	19.98	32.36	14.37	20.39	34.16	14.87	21.57	36.92	15.71	23.24
	STID+ImputeFormer (Nie et al., 2024)	37.87	17.46	24.45	38.91	17.69	25.01	39.94	17.83	25.44	40.58	17.87	25.69
	iTransformer+ImputeFormer	37.19	16.46	23.85	38.70	17.13	24.86	40.03	17.73	25.62	41.88	18.33	26.51
	iTransformer+SADI	34.15	15.39	21.79	36.50	16.21	23.43	38.27	16.79	24.30	39.74	17.58	25.25
	S4M (Peng et al., 2025)	35.81	15.64	21.99	35.87	15.58	22.06	36.48	15.87	22.55	39.10	16.69	24.61
METR-LA	**DSSER (ours)**	**6.38**	**9.14**	**3.15**	**6.47**	**9.17**	**3.19**	**6.67**	**9.59**	**3.29**	**6.92**	**10.10**	**3.43**
	STID (Shao et al., 2022)	7.13	10.18	3.74	8.21	11.85	4.53	9.51	14.07	5.20	10.23	15.93	5.38
	STID+GinAR+ (Yu et al., 2025a)	7.46	10.98	3.64	7.65	11.75	3.75	8.20	13.32	4.02	8.42	13.77	4.18
	STID+Merlin (Yu et al., 2025b)	6.56	9.48	3.20	6.73	9.78	3.27	7.09	10.43	3.42	7.63	11.43	3.64
	STID+SADI (Islam et al., 2025)	7.50	10.97	3.94	7.66	11.16	3.99	7.90	11.61	4.09	8.38	12.57	4.32
	STID+Glocal-IB (Yang et al., 2025)	6.45	9.37	3.22	6.64	9.77	3.31	7.05	10.67	3.49	7.55	11.66	3.72
	STID+T1 (Park et al., 2026)	6.53	9.43	3.21	6.69	9.72	3.28	7.05	10.39	3.43	7.56	11.33	3.63
	STID+iTransformer (Liu et al., 2024)	8.67	12.13	4.33	8.85	12.66	4.55	10.34	14.71	5.57	15.08	22.39	10.17
	PatchTST+STID (Nie et al., 2023)	6.53	9.43	3.19	6.69	9.70	3.26	7.04	10.31	3.40	7.53	11.19	3.61
	STID+ImputeFormer (Nie et al., 2024)	6.48	9.63	3.29	6.65	9.94	3.38	6.99	10.59	3.53	7.51	11.69	3.76
	iTransformer+ImputeFormer	6.46	9.92	3.54	6.61	10.12	3.52	6.87	10.58	3.59	7.33	11.53	3.78
	iTransformer+SADI	7.22	10.66	3.82	7.46	10.98	3.93	7.83	11.44	4.06	8.21	12.04	4.23
	S4M (Peng et al., 2025)	6.89	10.56	3.62	6.96	10.68	3.63	7.12	10.89	3.69	7.74	12.10	4.00
China-AQI	**DSSER (ours)**	**25.68**	**29.41**	**14.79**	**26.29**	**29.79**	**15.21**	**27.54**	**31.85**	**16.11**	**28.32**	**33.41**	**16.68**
	STID (Shao et al., 2022)	32.76	32.57	19.12	40.47	38.46	24.95	49.07	48.49	32.57	52.08	52.60	35.38
	STID+GinAR+ (Yu et al., 2025a)	27.95	36.41	17.17	29.19	38.49	17.98	32.61	41.78	19.97	38.67	40.93	23.25
	STID+Merlin (Yu et al., 2025b)	26.04	31.06	15.14	26.82	32.12	15.68	28.32	34.73	16.79	29.31	36.33	17.50
	STID+SADI (Islam et al., 2025)	28.56	36.97	17.45	29.20	35.39	17.66	31.28	40.26	19.34	32.63	41.68	20.18
	STID+Glocal-IB (Yang et al., 2025)	27.22	35.10	16.41	27.91	35.66	16.83	29.01	37.41	17.63	29.78	38.11	18.10
	STID+T1 (Park et al., 2026)	26.63	31.60	15.60	27.35	33.25	16.21	28.82	35.70	17.31	29.75	36.77	17.94
	STID+iTransformer (Liu et al., 2024)	27.15	30.04	15.76	28.04	31.30	16.38	30.07	34.17	17.75	31.43	35.97	18.67
	PatchTST+STID (Nie et al., 2023)	26.43	33.25	15.68	26.88	33.95	16.01	28.71	33.23	16.92	30.65	33.45	18.09
	STID+ImputeFormer (Nie et al., 2024)	26.06	32.03	15.28	26.84	33.03	15.81	28.36	35.57	16.93	29.39	36.79	17.61
	iTransformer+ImputeFormer	26.60	33.11	15.73	27.31	34.06	16.24	28.88	36.10	17.33	29.95	37.39	18.05
	iTransformer+SADI	28.33	35.11	17.16	28.45	35.30	17.19	30.07	38.41	18.33	31.13	39.62	18.93
	S4M (Peng et al., 2025)	31.96	40.35	19.49	31.99	40.58	19.54	32.17	40.70	19.61	32.39	40.47	19.69

Method references are provided at their first occurrence within each dataset block and apply to subsequent combinations containing the same component.



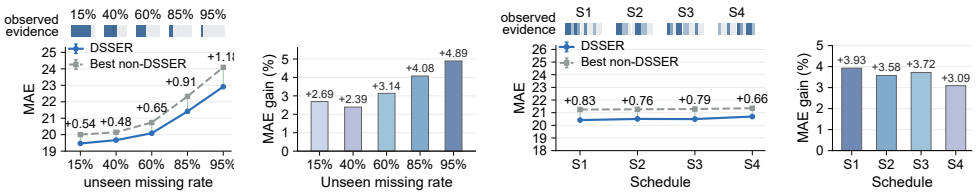
(a) PEMS04–MAE (b) METR-LA–MAE (c) PEMS04–RMSE (d) METR-LA–RMSE

Figure 3: Performance under unseen missing rates on PEMS04 and METR-LA. Results are reported for MAE and RMSE under missing rates of 15%, 40%, 60%, 85%, and 95%.

bounded residual correction to the controlled utilization of restored evidence. Replacing CSUC with fixed calibration affects performance across missing conditions, supporting the role of uncertainty-aware supervision and adaptive calibration during training and inference.

4.5 Reliability Diagnostic Analysis

To examine whether the reliability signals used by RGER reflect forecasting difficulty, we conduct a reliability diagnostic analysis. We construct a composite evidence-risk score by combining four

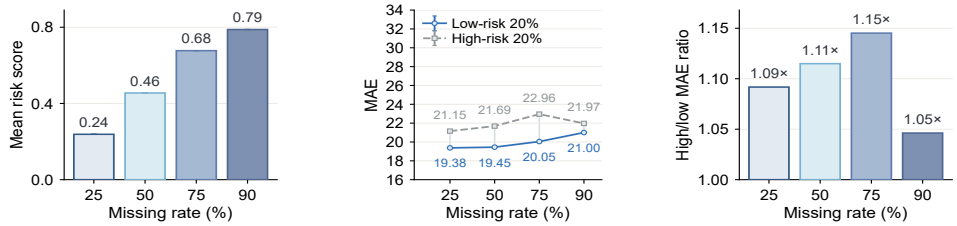


(a) Unseen Missing Rates (b) Dynamic Missingness

Figure 4: Robustness under unseen-rate and dynamic-missing shifts on PEMS04. (a) Unseen-rate evaluation at missing rates not used during training. (b) Dynamic-missing evaluation, where the missing rate varies across consecutive temporal segments.

Table 2: Ablation study of DSSER under random missingness. Best results are bolded.

Dataset	Variant	25%			50%			75%			90%		
		RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE
PEMS04	**DSSER (ours)**	**30.99**	13.63	**19.50**	**31.38**	**13.67**	**19.75**	**32.64**	**14.23**	**20.55**	**34.92**	**15.16**	**21.89**
	w/o missingness-aware restoration	31.08	**13.52**	19.54	31.58	13.73	19.89	32.97	14.28	20.81	35.99	15.37	22.60
	w/o RGER	31.05	13.75	19.52	31.53	13.99	19.86	32.78	14.51	20.67	35.14	15.45	22.13
	w/o CSUC	31.20	13.65	19.62	31.69	13.84	19.95	33.09	14.38	20.87	36.09	15.44	22.63
METR-LA	**DSSER (ours)**	**6.38**	9.14	**3.15**	**6.47**	**9.17**	**3.19**	**6.67**	**9.59**	**3.29**	**6.92**	**10.10**	**3.43**
	w/o missingness-aware restoration	6.47	9.23	3.17	6.61	9.51	3.23	6.95	10.19	3.38	7.45	11.27	3.61
	w/o RGER	**6.38**	**9.08**	3.16	6.50	9.33	3.21	6.78	9.88	3.34	7.12	10.70	3.52
	w/o CSUC	6.50	9.28	3.17	6.65	9.57	3.24	7.00	10.24	3.38	7.52	11.29	3.60



(a) Risk score (b) MAE comparison (c) MAE ratio

Figure 5: Reliability diagnostic analysis of DSSER under different missing rates.

normalized factors, including historical missingness, recent missingness, routing-state factor, and forecasting position. Specifically, for each sample, node n, and forecasting horizon h, the risk score is computed as $r_{n,h} = 0.45m_n^{hist} + 0.35m_n^{recent} + 0.15\tilde{g}_{n,h} + 0.05p_h$, where m_n^{hist} and m_n^{recent} denote historical and recent missing rates, respectively, $\tilde{g}_{n,h}$ denotes the normalized form of the routing-state factor $g_{n,h}$, and p_h denotes the normalized forecasting position. The weights are fixed, with larger contributions assigned to missingness-related factors. Forecasts are ranked according to this score, and the highest- and lowest-risk 20% subsets are compared under different missing-rate conditions. As shown in Figure 5, the high-risk subset consistently exhibits larger MAE, with the high-/low-risk MAE ratio above 1.0 across missing rates, indicating that the proposed score can effectively distinguish forecasts vulnerable to incomplete observations.

5 Conclusion

This paper studies multivariate time-series forecasting under missing observations by explicitly considering the availability and reliability of spatiotemporal evidence. Instead of reconstructing missing values, DSSER formulates incomplete forecasting as an evidence utilization problem and addresses missingness-induced forecast drift through dynamic sparse spatiotemporal evidence restoration, RGER, and CSUC. Experiments on three public traffic and air-quality benchmarks demonstrate that DSSER consistently improves forecasting performance under random missingness and remains robust under unseen missing rates and dynamic missingness. Future work will explore more complex missing patterns and broader spatiotemporal evidence restoration strategies.

AI USE STATEMENT

In this work, generative AI tools were used to assist with language polishing and improve the readability of the manuscript. The authors reviewed the AI-assisted revisions to ensure that they accurately reflected the intended meaning and did not introduce unsupported claims. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work uses only publicly available traffic and air-quality benchmark datasets and does not involve human subjects or personally identifiable information.

REPRODUCIBILITY STATEMENT

Section 4.1 describes the datasets, missingness protocol, baselines, and evaluation metrics, while Appendix A provides the implementation and training details, hyperparameter settings, and evaluation protocol.

REFERENCES

Abdulrahman Al-Fakih, A. Koeshidayatullah, Tapan Mukerji, Sadam Al-Azani, and SanLinn I. Kaka. Well log data generation and imputation using sequence based generative adversarial networks. *Scientific Reports*, 15:11000, 2025.

Lingxiao Cao, Bin Wang, Guiyuan Jiang, Yanwei Yu, and Junyu Dong. Spatiotemporal-aware trend-seasonality decomposition network for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11463–11471, 2025.

Hui Chen, Viet Luong, Lopamudra Mukherjee, and Vikas Singh. SimpleTM: A simple baseline for multivariate time series forecasting. In *International Conference on Learning Representations*, pp. 15431–15456, 2025.

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *International Conference on Learning Representations*, pp. 10527–10545, 2024a.

Zhichao Chen, Haoxuan Li, Fangyikang Wang, Odin Zhang, Hu Xu, Xiaoyu Jiang, Zhihuan Song, and Hao Wang. Rethinking the diffusion models for missing data imputation: A gradient flow perspective. In *Advances in Neural Information Processing Systems*, volume 37, pp. 112050–112103, 2024b.

MinGyu Choi and Changhee Lee. Conditional information bottleneck approach for time series imputation. In *International Conference on Learning Representations*, pp. 51096–51117, 2024.

Leyan Deng, Defu Lian, Chenwang Wu, and Enhong Chen. Learning from highly sparse spatio-temporal data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 94022–94046, 2024.

Wen-Shu Fan, Su Lu, Xin-Chun Li, De-Chuan Zhan, and Le Gan. Revisit the essence of distilling knowledge through calibration. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12882–12894, 2024.

Shikai Fang, Qingsong Wen, Yingtao Luo, Shandian Zhe, and Liang Sun. BayOTIDE: Bayesian online multivariate time series imputation with functional decomposition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12993–13009, 2024.

Tania Gupta, Richa Bhatia, and Sachin Sharma. An ensemble-based enhanced short and medium term load forecasting using optimized missing value imputation. *Scientific Reports*, 15:21857, 2025.

Md Imtiaz Hossain, Sharmen Akhter, Choong Seon Hong, and Eui-Nam Huh. Single teacher, multiple perspectives: Teacher knowledge augmentation for enhanced knowledge distillation. In *International Conference on Learning Representations*, pp. 91583–91594, 2025.

Xinyue Hu, Zhibin Duan, Bo Chen, and Mingyuan Zhou. Enhancing uncertainty estimation and interpretability with Bayesian non-negative decision layer. In *International Conference on Learning Representations*, pp. 304–326, 2025a.

Zhangyi Hu, Jiemin Wu, Hua Xu, Mingqian Liao, Ninghui Feng, Bo Gao, Songning Lai, and Yutao Yue. IMTS is worth time $\times$ channel patches: Visual masked autoencoders for irregular multivariate time series prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24805–24822, 2025b.

Mohammad Rafid Ul Islam, Prasad Tadepalli, and Alan Fern. Self-attention-based diffusion model for time-series imputation in partial blackout scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17564–17572, 2025.

Nishant Jain, Karthikeyan Shanmugam, and Pradeep Shenoy. Learning model uncertainty as variance-minimizing instance weights. In *International Conference on Learning Representations*, pp. 10184–10210, 2024.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17868–17876, 2025.

Marine Le Morvan and Gaël Varoquaux. Imputation for prediction: Beware of diminishing returns. In *International Conference on Learning Representations*, pp. 41174–41213, 2025.

Hyunwook Lee and Sungahn Ko. TESTAM: A time-enhanced spatio-temporal attention model with mixture of experts. In *International Conference on Learning Representations*, pp. 3011–3029, 2024.

Xvyuan Liu, Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Chenjuan Guo, Bin Yang, and Jilin Hu. ASTGI: Adaptive spatio-temporal graph interactions for irregular multivariate time series forecasting. In *International Conference on Learning Representations*, pp. 23940–23958, 2026a.

Xvyuan Liu, Xiangfei Qiu, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Jilin Hu, and Bin Yang. Rethinking irregular time series forecasting: A simple yet effective baseline. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23873–23881, 2026b.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, pp. 11116–11140, 2024.

Zhiding Liu, Mingyue Cheng, Guanhao Zhao, Jiqian Yang, Qi Liu, and Enhong Chen. Improving time series forecasting via instance-aware post-hoc revision. In *Advances in Neural Information Processing Systems*, volume 38, pp. 35604–35629, 2025.

Ivan Marisca, Cesare Alippi, and Filippo Maria Bianchi. Graph-based forecasting with missing data through spatiotemporal downsampling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34846–34865, 2024.

Vahid Naghashi, Mounir Boukadoum, and Abdoulaye Banire Diallo. A multiscale model for multivariate time series forecasting. *Scientific Reports*, 15:1565, 2025.

Tong Nie, Guoyang Qin, Wei Ma, Yuewen Mei, and Jian Sun. ImputeFormer: Low rankness-induced transformers for generalizable spatiotemporal imputation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2260–2271, 2024.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Dongik Park, Hyunwoo Ryu, Suahn Bae, Keondo Park, and Hyung-Sin Kim. T1: One-to-one channel-head binding for multivariate time-series imputation. In *International Conference on Learning Representations*, pp. 39341–39367, 2026.

Jing Peng, Meiqi Yang, Qiong Zhang, and Xiaoxiao Li. S4M: S4 for multivariate time series forecasting with missing values. In *International Conference on Learning Representations*, pp. 49369–49395, 2025.

Xiangfei Qiu, Xingjian Wu, Hanyin Cheng, Xvyuan Liu, Chenjuan Guo, Jilin Hu, and Bin Yang. DBLoss: Decomposition-based loss function for time series forecasting. In *Advances in Neural Information Processing Systems*, volume 38, pp. 27741–27768, 2025.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 4454–4458, 2022.

Minsuk Song, Ryun-Han Koo, Jangsaeng Kim, Chang-Hyeon Han, Jiyong Yim, Jonghyun Ko, Si-jung Yoo, Duk-hyun Choe, Sangwook Kim, Wonjun Shin, and Daewoong Kwon. Ferroelectric NAND for efficient hardware bayesian neural networks. *Nature Communications*, 16:6879, 2025.

Chenxi Sun, Moxian Song, Derun Cai, Baofeng Zhang, Hongyan Li, and Shenda Hong. A review of deep learning methods for irregularly sampled medical time series data. *Health Data Science*, 6:0456, 2026.

Ege Onur Taga, Muhammed Emrullah Ildiz, and Samet Oymak. TimePFN: Effective multivariate time series forecasting with synthetic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20761–20769, 2025.

Zihao Tang, Zheqi Lv, Shengyu Zhang, Yifan Zhou, Xinyu Duan, Fei Wu, and Kun Kuang. AuG-KD: Anchor-based mixup generation for out-of-domain knowledge distillation. In *International Conference on Learning Representations*, pp. 13048–13071, 2024.

Hao Wang, Zhengnan Li, Haoxuan Li, Xu Chen, Mingming Gong, Bin Chen, and Zhichao Chen. Optimal transport for time series imputation. In *International Conference on Learning Representations*, pp. 828–852, 2025.

Vijaya Krishna Yalavarthi, Kiran Madhusudhanan, Randolf Scholz, Nourhan Ahmed, Johannes Burchert, Shayan Jawed, Stefan Born, and Lars Schmidt-Thieme. GraFITi: Graphs for forecasting irregularly sampled time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16255–16263, 2024.

Vijaya Krishna Yalavarthi, Randolf Scholz, Stefan Born, and Lars Schmidt-Thieme. Probabilistic forecasting of irregularly sampled time series with missing values via conditional normalizing flows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21877–21885, 2025.

Jie Yang, Kexin Zhang, Guibin Zhang, Philip S. Yu, and Kaize Ding. Glocal information bottleneck for time series imputation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 104452–104484, 2025.

Chengqing Yu, Fei Wang, Zezhi Shao, Tangwen Qian, Zhao Zhang, Wei Wei, Zhulin An, Qi Wang, and Yongjun Xu. GinAR+: A robust end-to-end framework for multivariate time series forecasting with missing values. *IEEE Transactions on Knowledge and Data Engineering*, 37(8): 4635–4648, 2025a.

Chengqing Yu, Fei Wang, Chuanguang Yang, Zezhi Shao, Tao Sun, Tangwen Qian, Wei Wei, Zhulin An, and Yongjun Xu. Merlin: Multi-view representation learning for robust multivariate time series forecasting with unfixed missing rates. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 3633–3644, 2025b.

APPENDIX

A ADDITIONAL EXPERIMENTAL DETAILS

All reported comparisons are conducted using the same data splits, missingness masks, 12-step history-to-12-step forecast windows, validation-MAE checkpoint selection, and evaluation protocol for RMSE, MAE, and MAPE. The evaluated datasets, PEMS04, METR-LA, and China-AQI, contain 307, 207, and 1,300 time series variables, respectively. Unless otherwise stated, DSSER experiments use seed 3407 and are trained with Adam using a learning rate of 2×10^{-4}, weight decay of 10^{-4}, batch size of 16, and 104 epochs. We use $R = 5$ stochastic forward passes to estimate teacher uncertainty for CSUC. The four missing-rate branches are jointly optimized, where the pretrained teacher provides guidance to the student model through knowledge distillation. During student training, the pretrained teacher is evaluated without gradient updates, with dropout enabled for stochastic forward passes. The student checkpoint is selected using the validation MAE averaged equally across four missing rates of 25%, 50%, 75%, and 90%. The time-of-day embedding uses 288 slots for PEMS04 and METR-LA and 24 slots for China-AQI.

Test predictions and targets are inverse-transformed to the original data scale before evaluation. Metrics are computed over the concatenated test predictions across samples, forecast horizons, and nodes. MAE and RMSE exclude targets within 5×10^{-5} of zero. MAPE additionally treats targets with absolute values below 10^{-4} as zero and excludes these targets. MAPE is reported as a percentage. The main coefficient settings used in the reported experiments are summarized in Table A1.

Table A1: Hyperparameter settings used in the reported experiments.

Category	Parameter	Value
Objective weighting	ω_{rep}	1.0
	ω_{pred}	1.0
	λ_{kd}	2.0
	λ_{corr}	0.05
	λ_{MIT}	0.05
	λ_{ORT}	0.01
	λ_{cl}	1.0
Contrastive calibration	τ_c	1.0
	ψ_0	0.8
	τ_w	0.1
	ζ	1.0
Calibration function	$(\kappa_0, \kappa_1, \kappa_2)$	$(1.0, 1.0, 1.0)$
Forecast-position weighting	γ	0.1
RGER routing	$s_{\max}$	0.35
	ρ_0	0.45
	η	8.0

B ADDITIONAL FORECASTING RESULTS

To further evaluate the robustness of DSSER under missing observations, we conduct additional experiments on the PEMS08 and PEMS-BAY traffic forecasting benchmarks. PEMS08 and PEMS-BAY contain 170 and 325 sensors, respectively. Following the same experimental settings, we evaluate all methods under random missingness at rates of 25%, 50%, 75%, and 90%. All methods use consistent preprocessing, chronological splits, 12-step input/output windows, fixed masks, and training budgets for fair comparison. Table B1 reports the corresponding forecasting performance. DSSER consistently achieves the best RMSE and MAE results across different missing rates on both datasets, while maintaining competitive MAPE performance. These results further demonstrate the robustness of DSSER under varying missing conditions.

Table B1: Forecasting performance on PEMS08 and PEMS-BAY under random missingness at rates of 25%, 50%, 75%, and 90%. Best results are bolded and second-best results are underlined.

Dataset	Method	25%			50%			75%			90%		
		RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE
PEMS08	**DSSER (ours)**	**24.76**	10.49	**15.55**	**25.28**	10.65	**15.90**	**27.03**	11.41	**16.94**	**31.68**	12.76	**19.29**
	STID (Shao et al., 2022)	26.26	12.07	16.69	27.12	12.30	17.25	29.97	13.04	18.90	35.49	14.13	21.58
	STID+GinAR+ (Yu et al., 2025a)	37.53	17.32	24.59	38.56	17.73	25.18	40.63	18.85	26.61	46.58	18.99	29.55
	STID+Merlin (Yu et al., 2025b)	25.35	10.70	15.66	25.96	10.88	16.01	28.30	11.42	17.14	33.83	12.42	19.54
	STID+SADI (Islam et al., 2025)	35.50	13.22	21.91	37.38	13.68	22.77	39.65	14.32	24.05	41.76	15.15	25.52
	STID+Glocal-IB (Yang et al., 2025)	26.26	11.65	16.75	27.16	12.02	17.40	29.67	12.97	18.99	34.95	14.27	21.72
	STID+T1 (Park et al., 2026)	25.02	10.37	15.66	25.75	10.60	16.14	28.00	11.44	17.47	33.08	12.90	19.90
	STID+iTransformer (Liu et al., 2024)	25.16	10.92	15.85	25.90	11.19	16.32	28.32	11.88	17.64	34.02	12.97	20.23
	PatchTST+STID (Nie et al., 2023)	25.01	**10.26**	15.56	25.59	**10.45**	15.94	27.54	**10.99**	17.07	32.67	**12.22**	19.51
	STID+ImputeFormer (Nie et al., 2024)	28.09	11.84	17.83	28.93	12.24	18.34	31.11	12.87	19.52	35.29	13.90	21.63
	iTransformer+ImputeFormer	31.73	12.41	20.05	34.34	13.15	21.47	37.58	13.99	23.16	39.34	14.29	23.88
	iTransformer+SADI	33.70	12.68	21.21	36.42	13.42	22.71	39.31	14.37	24.35	41.98	15.47	26.17
	S4M (Peng et al., 2025)	34.41	13.15	21.05	34.51	13.12	21.06	35.67	13.34	21.52	38.76	14.09	23.12
PEMS-BAY	**DSSER (ours)**	**3.7743**	**3.9147**	**1.6910**	**3.8526**	**4.0429**	**1.7437**	**4.0893**	**4.3885**	**1.8752**	**4.4893**	**4.9793**	**2.1096**
	STID (Shao et al., 2022)	4.0836	4.4390	1.8231	4.2139	4.6231	1.8978	4.5280	5.0713	2.0920	4.9776	5.7094	2.3856
	STID+GinAR+ (Yu et al., 2025a)	4.6981	5.3135	2.2018	4.8137	5.6286	2.3102	5.1655	6.2993	2.5539	5.3678	6.5265	2.7085
	STID+Merlin (Yu et al., 2025b)	3.8644	4.0513	1.7064	3.9892	4.2081	1.7695	4.3030	4.6082	1.9301	4.8002	5.2359	2.2074
	STID+SADI (Islam et al., 2025)	5.0249	5.7429	2.4250	5.0893	5.8093	2.4610	5.1735	5.9367	2.5151	5.3527	6.2398	2.6142
	STID+Glocal-IB (Yang et al., 2025)	3.8533	4.1453	1.7426	4.0011	4.3587	1.8290	4.3527	4.8644	2.0187	4.8215	5.4893	2.3010
	STID+T1 (Park et al., 2026)	3.8722	4.0695	1.7290	3.9943	4.2363	1.7935	4.2766	4.6107	1.9387	4.7535	5.2232	2.1929
	STID+iTransformer (Liu et al., 2024)	3.8484	4.0547	1.7034	3.9763	4.2190	1.7660	4.2844	4.6127	1.9250	4.7786	5.2384	2.2039
	PatchTST+STID (Nie et al., 2023)	3.9045	4.1478	1.7340	4.0219	4.2926	1.7971	4.3165	4.6774	1.9607	4.7985	5.2875	2.2323
	STID+ImputeFormer (Nie et al., 2024)	3.8988	4.1916	1.7631	4.0358	4.3765	1.8346	4.3723	4.8076	2.0072	4.8606	5.4349	2.2764
	iTransformer+ImputeFormer (Nie et al., 2024)	4.7729	5.7247	2.4630	4.8454	5.8294	2.4838	5.0363	6.2569	2.6838	5.3659	6.9192	3.0269
	iTransformer+SADI (Islam et al., 2025)	4.9905	5.6768	2.4144	5.0620	5.7624	2.4541	5.1630	5.9108	2.5164	5.3849	6.2640	2.6262
	S4M (Peng et al., 2025)	4.7671	5.5355	2.3540	4.8097	5.5707	2.3626	5.0033	5.8390	2.4889	5.2824	6.2740	2.6815

Method references are provided at their first occurrence within each dataset block and apply to subsequent combinations containing the same component.

C Additional Dynamic Missingness Evaluation

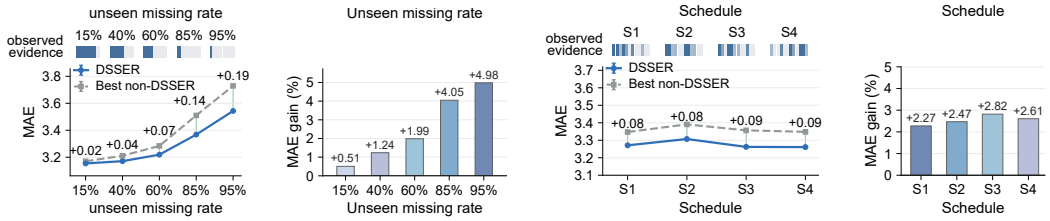
To evaluate robustness under temporally varying missingness, we construct dynamic missing scenarios by varying the missing rate across consecutive test segments. Specifically, the test period is divided into four consecutive segments, and each segment is assigned a predefined missing rate. We consider four dynamic missing schedules, denoted as S1–S4, whose missing-rate configurations are summarized in Table C1. For each segment, missing entries are randomly selected according to the assigned missing rate. Given a historical input with N nodes and T time steps, missing masks are generated at the sensor-time level and fixed during evaluation for fair comparison across methods.

Table C1: Dynamic missingness schedules for robustness evaluation.

Schedule	Segment 1	Segment 2	Segment 3	Segment 4
S1	25%	50%	75%	90%
S2	25%	90%	25%	90%
S3	50%	25%	90%	75%
S4	90%	75%	50%	25%

D Additional Robustness Results

To complement the PEMS04 robustness analysis in Figure 4, we report the corresponding results on METR-LA under the same unseen-rate and dynamic-missing settings. Figure D1 presents the robustness evaluation results under two settings, including unseen missing rates and dynamic missingness schedules. Under unseen missing rates, DSSER outperforms the best-performing non-DSSER baseline in each test condition, yielding an average relative MAE reduction of 2.55%. Under dynamic missingness, DSSER also achieves better performance than the best-performing non-DSSER baseline across all schedules (S1–S4), with an average relative MAE reduction of 2.54%. These results show that the robustness trends observed on PEMS04 are consistent with those on METR-LA.



(a) Unseen Missing Rates (b) Dynamic Missingness

Figure D1: Additional robustness results on METR-LA under (a) unseen missing rates and (b) dynamic missingness schedules (S1–S4). Relative MAE gains are measured against the best-performing non-DSSER baseline in each condition.

E Additional Reliability Analysis

Sensitivity to the Number of MC Passes. We repeat the teacher-uncertainty diagnostic with $R = 20$ stochastic forward passes using the same test set and teacher checkpoint. As shown in Table E1, the results remain close to those obtained with the default $R = 5$ setting. Spearman correlation changes from 0.692 to 0.697, while the high-/low-uncertainty MAE ratio changes from $2.65\times$ to $2.67\times$. These results indicate that $R = 5$ provides a stable estimate of teacher uncertainty while avoiding the additional cost of a larger number of MC passes.

Table E1: Analysis of teacher-uncertainty sensitivity to the number of MC passes on PEMS04.

MC passes	Spearman ρ	High/low MAE ratio
$R = 5$	0.692	$2.65\times$
$R = 20$	0.697	$2.67\times$

F Efficiency Analysis

We evaluate the accuracy–efficiency trade-off of DSSER. As shown in Figure F1, DSSER achieves the lowest average MAE while maintaining low inference latency. The additional inference overhead mainly arises from sparse evidence restoration and RGER, whereas teacher uncertainty estimation is confined to training and incurs no inference-time cost. These results demonstrate that DSSER improves forecasting robustness under missing observations with modest computational overhead.

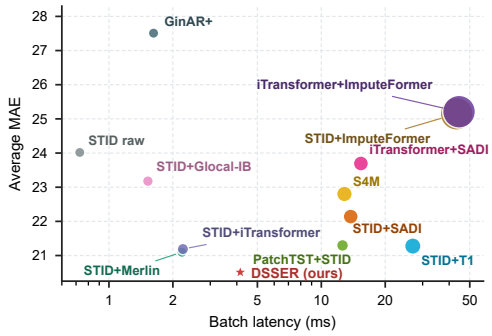


Figure F1: Accuracy–latency trade-off on PEMS04. Bubble size indicates peak GPU memory usage.

G Forecast Drift Analysis

To quantify missingness-induced forecast drift, we define the forecast drift metric as:

$$D_{\text{drift}} = \frac{1}{N_{\text{test}} N H} \sum_{i=1}^{N_{\text{test}}} \sum_{n=1}^{N} \sum_{h=1}^{H} \left| \hat{Y}^{\text{miss}}_{i,n,h} - \hat{Y}^{\text{full}}_{i,n,h} \right|, \tag{12}$$

where N_{test} denotes the number of test windows and i indexes test windows. $\hat{Y}^{\text{full}}$ and $\hat{Y}^{\text{miss}}$ denote forecasting outputs from complete and incomplete historical observations, respectively. Lower drift indicates less deviation from complete-view forecasts. We evaluate forecast drift on PEMS04 under the same missing rates as the main experiments. Table G1 reports the joint drift and forecasting accuracy. The results show that lower drift does not always imply lower forecasting error. These results demonstrate that DSSER reduces missingness-induced forecast drift while maintaining competitive forecasting accuracy under incomplete observations.

Table G1: Forecast drift and forecasting accuracy on PEMS04 under different missing rates. Lower values indicate better performance. Best values are bolded and second-best values are underlined.

	25%		50%		75%		90%	
Method	Drift	MAE	Drift	MAE	Drift	MAE	Drift	MAE
DSSER (ours)	<u>3.183</u>	**19.533**	5.642	**19.854**	<u>7.873</u>	**20.658**	**10.278**	**21.971**
STID+GinAR+ (Yu et al., 2025a)	6.043	25.684	11.929	26.025	17.648	27.822	20.307	30.503
STID+Merlin (Yu et al., 2025b)	3.349	<u>20.422</u>	<u>5.365</u>	<u>20.764</u>	8.130	<u>21.735</u>	12.256	<u>23.408</u>
STID+ImputeFormer (Nie et al., 2024)	9.362	24.451	12.141	25.015	13.711	25.440	14.660	25.691
S4M (Peng et al., 2025)	**2.865**	22.085	**3.964**	22.132	**5.783**	22.626	<u>10.649</u>	24.627

Method references are provided at their first occurrence within each dataset block and apply to subsequent combinations containing the same component.

H Random Seed Stability Analysis

To assess the sensitivity of DSSER to training randomness, we repeat the experiments on PEMS04 using three random seeds, 3407, 3408, and 3409. All other experimental settings follow the main experiments, including the data splits and missingness generation protocol, while only the seed governing model initialization and stochastic training operations is varied. Table H1 reports the mean and standard deviation of MAE, RMSE, and MAPE across the three runs at missing rates of 25%, 50%, 75%, and 90%. The results show that DSSER maintains stable performance across different random seeds, with small standard deviations for all metrics under varying missing rates. Although the prediction errors increase as the missing rate becomes higher, the performance variations across different seeds remain limited, indicating that DSSER is not sensitive to random initialization or stochastic training effects. These results demonstrate the robustness of DSSER under different missing conditions.

Table H1: Mean and standard deviation of DSSER performance under different random seeds and missing rates on PEMS04.

Missing rate	Seed	RMSE	MAPE	MAE
25%	3407	30.992	13.635	19.505
	3408	31.019	13.693	19.505
	3409	31.014	13.720	19.492
	Mean ± Std	31.009 ± 0.014	13.683 ± 0.043	19.501 ± 0.007
50%	3407	31.383	13.669	19.747
	3408	31.399	13.700	19.741
	3409	31.374	13.722	19.707
	Mean ± Std	31.385 ± 0.013	13.697 ± 0.027	19.732 ± 0.022
75%	3407	32.643	14.232	20.545
	3408	32.634	14.282	20.535
	3409	32.618	14.300	20.498
	Mean ± Std	32.631 ± 0.013	14.272 ± 0.035	20.526 ± 0.025
90%	3407	34.921	15.158	21.892
	3408	34.901	15.252	21.891
	3409	34.872	15.284	21.837
	Mean ± Std	34.898 ± 0.024	15.231 ± 0.065	21.873 ± 0.031