SAUC: Sparsity-Aware Uncertainty Calibration for Spatiotemporal Prediction with Graph Neural Networks

Dingyi Zhuang[1], **Yuheng Bu**[2], **Guang Wang**[3], **Shenhao Wang**[1,2], **Jinhua Zhao**[1]
[1]MIT, [2]University of Florida, [3]Florida State University
{dingyi,jinhua}@mit.edu, {buyuheng,shenhaowang}@ufl.edu, guang@cs.fsu.edu

Abstract

Quantifying uncertainty is essential for achieving robust and reliable predictions. However, existing spatiotemporal models predominantly predict deterministic values, often overlooking the uncertainty in their forecasts. Particularly, high-resolution spatiotemporal datasets are rich in zeros, posing further challenges in quantifying the uncertainty of such sparse and asymmetrically distributed data. This paper introduces a novel post-hoc Sparsity-aware Uncertainty Calibration (SAUC) method, calibrating the uncertainty in both zero and non-zero values. We modify the state-of-the-art deterministic spatiotemporal Graph Neural Networks (GNNs) to probabilistic ones as the synthetic models in the pre-calibration phase. Applied to two real-world spatiotemporal datasets of varied granularities, extensive experiments demonstrate SAUC's capacity to adeptly calibrate uncertainty, effectively fitting the variance of zero values and exhibiting robust generalizability. Specifically, our empirical experiments show a 20% of reduction in calibration errors in zero entries on the sparse traffic accident and urban crime prediction. The results validate our method's theoretical and empirical values, demonstrating calibrated results that provide reliable safety guidance, thereby bridging a significant gap in uncertainty quantification (UQ) for sparse spatiotemporal data.

1 Introduction

Spatiotemporal GNNs have been instrumental in leveraging spatiotemporal data for various applications, from weather forecasting to urban planning [25, 29, 16]. However, most of the models are narrowly focused on deterministic predictions, which overlooks the inherent uncertainties associated with the spatiotemporal phenomena. For various safety-related tasks, such as traffic accident prediction, it is crucial to quantify the prediction uncertainty to prevent such phenomena from happening. Despite its importance, uncertainty quantification (UQ) remains relatively understudied in the spatiotemporal context, which can lead to erroneous predictions with severe social consequences [2, 1, 17].

Several recent studies explored UQ for spatiotemporal models [22, 5]. However, the existing techniques often overlook two critical elements: the sparsity and the asymmetrically distributed nature of the high-resolution spatiotemporal data. The sparsity issue, characterized by a preponderance of zeroes, becomes particularly salient in any spatiotemporal prediction when spatial or temporal resolutions are high [30]. Prior spatiotemporal GNNs with basic UQ functionality typically assume data as NB distributions or Gaussian distributions for regression tasks, and learning the distribution parameters for spatiotemporal predictions [6, 30]. However, even though such models could learn data distributions, they typically lack the evaluation of the UQ quality with effective calibration metrics. As Kuleshov et al. [7] mentioned, most deep learning models are not fully calibrated for high UQ quality, leaving a significant research gap to be filled.

Temporal Graph Learning Workshop @ NeurIPS 2023, New Orleans.

This paper introduces a sparsity-aware uncertainty calibration (SAUC) method, calibrating model outputs to align with true data distributions, which can be adapted to any spatiotemporal GNN model. Inspired by the zero-inflated model, we partition zero and non-zero predictions and apply separate Quantile Regression (QR) models to ensure the discrepancy of prediction and true values lie within the PIs. This post-hoc approach, unlike uncertainty-aware spatiotemporal prediction models [15], offers flexibility in applying it to existing models without necessitating architectural or loss function changes or full model retraining. Consistent with prior studies on regression task calibration, we formulate a modified calibration metric tailored for PI-based calibration within asymmetric distributions. To investigate our model's generalizability, we adapt two prevalent spatiotemporal GNN models with negative binomial distributions to suit sparse settings without compromising prediction accuracy. After being tested on two real-world spatiotemporal datasets, our SAUC demonstrates promising calibration results, thus assisting in reliable decision-making and risk assessment.

The main contributions of this paper include:

- We design a novel sparsity-aware uncertainty calibration (SAUC) method for predicting sparse spatiotemporal data, which is adaptive to existing GNN models and applicable to any probabilistic outputs.
- We address the calibration of asymmetric distributions, such as NB, via quantile regression, diverging from traditional mean-variance UQ approaches, and we also design novel PI-based calibration metrics.
- We conduct experiments on different real-world sparse datasets to demonstrate promising calibration results of our SAUC, especially for zero values. We observe a roughly 20% reduction in calibration errors relative to leading baseline models, underscoring its efficacy for safety-critical decision-making.

2 Related Work

2.1 Spatiotemporal Prediction

Spatiotemporal prediction through deep learning offers a powerful tool for various applications. Among them, GNNs have become prevalent in recent years [25, 29, 27, 26]. Most existing models focus on predicting the deterministic values without considering the associated uncertainty, which leaves a noticeable research gap concerning UQ. This gap limits the predictive reliability because of overlooked uncertainty.

Both Bayesian and Frequentist methods address spatiotemporal UQ [22]. The Bayesian approach, utilizing techniques like Laplace approximation [23] and MCMC [14], has computational and approximation challenges. Conversely, the Frequentist method emphasizes Mean-Variance Estimation for efficiency and flexibility, accommodating various models [22]. Still, sparse data applications face issues with data dependencies and distribution constraints [24, 3]. Sparse data often deviates from model assumptions, preferring NB distributions, and metrics like MSE can be skewed by non-zero sparse data values.

The use of zero-inflated distributions with spatiotemporal neural networks has been introduced to handle zero instances and non-normal distribution in sparse data [30]. However, this approach has limitations in extreme scenarios and non-time-series contexts, impacting predictive accuracy and system robustness [6]. Therefore, the urgent need to quantify and calibrate uncertainty in the sparse component of spatiotemporal data is evident, yet remains largely unexplored.

2.2 Uncertainty Calibration

Although relatively understudied in the spatiotemporal context, UQ has been examined by many machine learning studies. Among all the UQ approaches, calibration is a post-hoc method that aims to match the predictive probabilities of a model with the true probability of outcomes, developing robust mechanisms for uncertainty validation [13].

A wealth of techniques has been developed to calibrate pre-trained classifiers, mitigating the absence of inherent calibration in many uncertainty estimators [9, 4, 18]. Calibration for regression tasks, however, has received less attention [20, 7]. Post-hoc calibration methods for classifications, such as temperature scaling, Platt scaling, and isotonic regression, have successfully been adapted for

regression [8]. Chung et al. [2] further demonstrated quantile regression-based calibration methods and their extensions. These post-hoc methods offer greater flexibility and wider applicability to various probabilistic outputs compared to in-training methods, without retraining the models [15, 5].

In contexts like crime prediction with sparse spatiotemporal data [29], the need for uncertainty calibration is undervalued. While traditional methods favor symmetric distributions like Gaussian, sparse data often suits asymmetric NB distributions. Recognizing the zero-inflated nature of this data, it's crucial to differentiate confidence between zero and non-zero portions. Thus, we introduce the SAUC method to handle asymmetric distributions and align true targets with estimated PIs [20].

3 Problem Formulation

3.1 Spatiotemporal GNNs for Prediction

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ where $\mathcal{V}$ represents the set of nodes (locations), $\mathcal{E}$ the edge set, and $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ the adjacency matrix describing the relationship between nodes. We denote the spatiotemporal dataset $\mathcal{X} \in \mathbb{R}^{|\mathcal{V}|\times t}$ where t is the number of time steps. The objective is to predict the target value $\mathcal{Y}_{1:|\mathcal{V}|,t:t+k}$ of future k time steps given all the past data up to time t, which is $\mathcal{X}_{1:|\mathcal{V}|,1:t}$. The full dataset is partitioned by timesteps into training, calibration (i.e., validation), and testing sets. The three data partitions are denoted as $\mathcal{X}_T$, $\mathcal{X}_S$, and $\mathcal{X}_U$, respectively. For example, $\mathcal{X}_{1:|\mathcal{V}|,1:t}$ is denoted as $\mathcal{X}_T$. The associated target values are represented as $\mathcal{Y}_T$, $\mathcal{Y}_S$, and $\mathcal{Y}_U$, with subscripts indicating the corresponding sets.

The objective of the spatiotemporal GNN models is to design model f_θ, parametrized by θ, which yields:

$$\hat{\mathcal{Y}}_d = f_\theta(\mathcal{X}_d; \mathcal{G}), \forall\ d \in \{T, S, U\}, \tag{1}$$

where $\hat{\mathcal{Y}}_d$ is the predicted target value, and the subscript represents the corresponding data set. Note that θ is fixed once f_θ is trained on set T.

With emerging interests in quantifying the data and model uncertainty of spatiotemporal GNNs, researchers shift towards designing probabilistic GNNs to predict not only mean values but also PIs [30, 6, 22]. Previous probabilistic spatiotemporal GNN studies often assume $\mathcal{X}$ and $\mathcal{Y}$ encompassing independent and identically distributed Gaussian random variables X_i and Y_i, which are assumed as NB distributions to accommodate the sparse and discrete data in this study. The NB distribution of each predicted data point is characterized by shape parameter μ and dispersion parameter α. The set of predicted μ and α corresponding to all elements of $\mathcal{Y}$ are denoted as $\hat{M}$ and $\hat{H}$. Therefore, Equation 1 can be rewritten as:

$$(\hat{M}_d, \hat{H}_d) = f_\theta(\mathcal{X}_d; \mathcal{G}), \forall d \in \{T, S, U\}. \tag{2}$$

3.2 Prediction Interval Calibration

Uncertainty calibration aims to refine the outputs of prediction models (e.g., GNNs) such that the predicted distributions align with the true distributions. Traditional calibration models generally assume Gaussian distributions, which can be succinctly summarized into the mean and variance parameters. However, when dealing with sparse data, adopting alternative distributions such as the NB distribution is more suitable due to their asymmetric nature. Thus, our calibration methodology focuses on the prediction interval, naturally extending the previous variance-oriented studies.

The spatiotemporal prediction models generate outputs $\hat{M}_d, \hat{H}_d, \forall d \in \{T, S, U\}$, in which $\hat{\mu}_i \in \hat{M}_d$ is the mean value of the predicted distribution. We then construct the prediction intervals of the NB distributions as $\hat{\mathcal{I}}_i = [\hat{l}_i, \hat{u}_i]$ with $\hat{l}_i$ and $\hat{u}_i$ denoting the 5th- and 95th-percentiles, respectively. Notably, our discussion limits the prediction intervals to the 5% and 95% range. Based on Kuleshov et al. [7], our target is to ensure a prediction model f_θ to be calibrated, i.e.,

$$\frac{\sum_{i=1}^{|U|} \mathbb{I}\{y_i \leq \hat{F}_i^{-1}(p)\}}{|U|} \xrightarrow{|U|\to\infty} p,\ \forall p \in \{.05, .95\}, \tag{3}$$

where p stands for the targeting probability values, $\mathbb{I}$ denotes the indicator function, and $\hat{F}_i(\cdot)$ is the cumulative density function (CDF) of the NB distributions parameterized by $\hat{\mu}_i$ and $\hat{\alpha}_i$. The primary

goal is to ensure that, as the dataset size increases, the predicted CDF converges to the true one. Such an alignment is congruent with quantile definitions, suggesting, for instance, that 95% of the genuine data should fall beneath the 95% percentile. Rather than striving for a perfect match across $\forall p \in [0, 1]$, we mainly focus on calibrating the lower bound and upper bound.

Empirically, we leverage the calibration set S to calibrate the uncertainty in data. The empirical CDF for a given value $p \in \hat{F}_i(y_i)$ is computed [7, 10]:

$$\hat{\mathcal{P}}(p) = \frac{|\{y_i|\hat{F}_i(y_i) \leq p, i = 1, \ldots, |S|\}|}{|S|}. \tag{4}$$

The calibration process entails mapping a calibration function $\mathcal{C}$, such as isotonic regression or Platt scaling, to the point set $\{(p, \hat{\mathcal{P}}(p))\}_{i=1}^{|S|}$. Consequently, the composition $\mathcal{C} \circ \hat{F}$ ensures that $\hat{\mathcal{P}}(p) = p$ within the calibration dataset and thus can be applied to the test set to enhance prediction reliability.

4 Sparsity-Aware Uncertainty Calibration

4.1 QR for Uncertainty Calibration.

Unlike traditional regression techniques that target the conditional mean, QR predicts specific quantiles, which are essential for calibrating PIs. Given $\hat{M}_S$ and $\mathcal{Y}_S$ from the calibration set S, the linear QR model can be expressed as:

$$Q_{\mathcal{Y}_S|\hat{M}_S}(p) = \hat{M}_S^T \beta(p), \tag{5}$$

where $\beta(p)$ signifies the coefficient associated with quantile p. The goal here is to minimize the Pinball loss [2]. The calibrated PIs correspond to $[Q(.05), Q(.95)]$.

QR is adept at handling sparse datasets, addressing heteroscedasticity and capturing non-linear relationships. Sparsity introduces heteroscedasticity, where error variance differs across observations. Without assuming constant variance, QR manages datasets with variance fluctuations, often arising from zeros. Moreover, when combined with basis expansion techniques, QR extends its capability to capture intricate patterns across quantiles, making it suited for the zero-inflated parts of the outcomes.

4.2 SAUC Method.

Our SAUC method is a two-step post-hoc calibration tailored for sparse data. Drawing inspiration from zero-inflated models, predictions in the calibration set S are bifurcated based on **indices**: $I_Z = \{\hat{M}_S < 0.5\}$ for values nearing zero, and $I_{NZ} = \{\hat{M}_S \geq 0.5\}$ for non-zero predictions. Following this partitioning, two QR models, Q_Z and Q_{NZ}, are trained separately, calibrating quantiles $p = \{.05, .95\}$ to fine-tune the mean and PI predictions. The calibrated model is then applied to predict model outputs on the testing set. A detailed procedure is provided in Algorithm 1.

Song et al. [19] noted quantile regressions might misrepresent true moments due to global averaging. Our SAUC method overcomes this by segmenting data into zero and non-zero values. The prevalent zero segment demands precise calibration given its real-world implications. For non-zero data, the goal is to align quantiles with actual event frequencies, focusing on relative risks. Though we detail SAUC using Equation 2 for GNNs, it's applicable to any spatiotemporal prediction. The code is available on an anonymous Github repository[1].

4.3 Calibration Metrics for Asymmetric Distributions

While calibration errors are clearly established for classification tasks using Expected Calibration Error [13, 4, 9], the definition is much less studied in the regression context, where prediction errors and intervals supersede accuracy and confidence. Based on Equation 3 and 4, we aim to calibrate the length of the prediction interval with the variance, i.e.,

$$\left[\mathbb{E}[(\hat{\mu}_i - Y_i)^2 | \hat{\mathcal{I}}_i = \mathcal{I}_i]\right]^{1/2} = c|\mathcal{I}_i|, \ \forall i \in \{1, \ldots, |U|\}. \tag{6}$$

[1] `https://github.com/AnonymousSAUC/SAUC`

Algorithm 1 SAUC Method

Require: $\hat{M}_S$, $\mathcal{Y}_S$, N, $\hat{M}_U$, $p \leftarrow \{.05, .95\}$

Define the bin thresholds $\mathcal{T}_S$ using percentiles of $\mathcal{Y}_S$.
Initialize U^*, L^* in the same size as $\hat{M}_U$.
for each bin i in $[1, N]$ **do**
  Extract indices I of instances in bin i based on $\mathcal{T}_S$.
  Split I to I_{NZ} for $\hat{M}_S \geq 0.5$ and I_Z for $\hat{M}_S < 0.5$.
  for each set i in $\{I_{NZ}, I_Z\}$ **do**
    for each value q in p **do**
      Fit $Q(q)$ on $\hat{M}_S[i]$ and $\mathcal{Y}_S[i]$ based on Equation 5 and obtain $\beta(q)$.
    end for
    $L^*[i] \leftarrow \hat{M}_U^T \beta(.05)$, $U^*[i] \leftarrow \hat{M}_U^T \beta(.95)$
  end for
end for
$\mathcal{I}^* \leftarrow [L^*, U^*]$.
return $\mathcal{I}^*$.

Note that $\mathcal{I}_i = [F_i^{-1}(.05), F_i^{-1}(.95)]$ is the realization of the predicted 5%-95% confidence interval and $|\mathcal{I}_i|$ denotes its width. Calibrating the model's standard errors with the 90% confidence interval width reveals a linear relationship with coefficient c, which varies per data point. Using the z-score of Gaussian distribution to approximately compute the coefficient c, the width of 90% confidence interval is roughly $2 \times 1.645 = 3.29$ of distribution variance, inversely proportional to c. Empirically, the ratio of predicted $\hat{\mathcal{I}}$ to distribution variance is between 3.0 and 3.4 in our experiments, aligning with theory, yielding $c \approx 0.303$. Therefore, we could diagnose the calibration performance by comparing value differences on both sides of the equation. Specifically, we partition the **indices** of test set samples into N evenly sized bins based on sorted predicted PI width, denoted as $\{B_j\}_{j=1}^N$. Each bin corresponds to a PI width axis interval: $[\min_{i \in B_j}\{|\mathcal{I}_i|\}, \max_{i \in B_j}\{|\mathcal{I}_i|\}]$. Note that the intervals are non-overlapping and their boundary values are increasing. Based on the discussion of Levi et al. [10] and Equation 6, for j-th bin, we define the Expected Normalized Calibration Error (ENCE) based on Root Mean Squared Error (RMSE) and Mean Prediction Interval Width (MPIW):

$$ENCE = \frac{1}{N} \sum_{j=1}^{N} \frac{|c \cdot MPIW(j) - RMSE(j)|}{c \cdot MPIW(j)}, \tag{7}$$

where $RMSE(j) = \sqrt{\frac{1}{|B_j|} \sum_{i \in B_j} (\hat{\mu}_i - y_i)^2}$, and $MPIW(j) = \frac{1}{|B_j|} \sum_{i \in B_j} |\hat{\mathcal{I}}_i|$. This normalizes calibration errors across bins, sorted and divided by $c \cdot MPIW$, and is particularly beneficial when comparing different temporal resolutions and output magnitudes across datasets. The $c \cdot MPIW$ and observed $RMSE$ per bin should approximate equality in a calibrated forecaster.

5 Experiments

5.1 Data Description

Two spatiotemporal datasets are used: (1) Chicago Traffic Crash Data (CTC) sourced from 277 police beats between January 1, 2016 and January 1, 2023; (2) Chicago Crime Records (CCR) derived from 77 census tracts spanning January 1, 2003 to January 1, 2023. Despite both CTC and CCR data originating from the Chicago area, their disparate reporting sources lead to different spatial units: census tracts for CCR and police beats for CTC. Both datasets use the first 60% timesteps for training, 20% for calibration and validation, and 20% for testing.

The temporal resolutions of the datasets are varied to demonstrate the ubiquitous sparsity issue and its practical significance in spatiotemporal analysis. We design the cases with 1-hour, 8-houir, 1-day, and 1-week for each dataset, and designate the 1-hour and 8-hour cases of Crash and Crime datasets as sparse instances, due to a higher prevalence of zeros.

For adjacency matrices, we calculate geographical distances d_{ij} between centroids of regions i and j, be they census tracts or police beats. This distance is then transformed into a similarity measure, $\mathbf{A}_{ij} = e^{-d_{ij}/0.1}$, where 0.1 is a scaling parameter, thus forming an adjacency matrix.

5.2 Modified Spatiotemporal Models

In order to demonstrate our model's ability to generalize on spatiotemporal prediction models, we modify two existing popular spatiotemporal forecasting models: Spatio-temporal Graph Convolution Network (STGCN) [27] and Graph WaveNet (GWN) [26]. They are designed for numerical value outputs. We modify their last layer from single numerical output to the location parameter μ and the dispersion parameter α of negative binomial distributions. We renamed the modified model as STGCN-NB and GWN-NB respectively. Besides, we modify the loss function from mean square error loss to the loss defined in Equation 8:

$$\begin{aligned}\mathcal{L}(\mu, \alpha, y) = &- \left[y \cdot \log\left(\frac{\mu + \epsilon}{\mu + \alpha + 2\epsilon}\right)\right] - \Gamma(y + \alpha + \epsilon) + \Gamma(y + 1) \\ &+ \Gamma(\alpha + \epsilon) - \alpha \cdot \log\left(\frac{\alpha + \epsilon}{\mu + \alpha + 2\epsilon}\right) + \lambda \cdot ||\alpha||^2\end{aligned}, \tag{8}$$

where y is the target variable, μ and α are model outputs, Γ is the gamma function, λ is the regularization parameter, and ϵ is a small constant added to improve numerical stability. This loss function is derived from the likelihood of NB distribution controlled by μ and α.

The loss function aligns the predicted Negative Binomial distribution with true targets, penalizing deviations in mean (μ) and incorporating variability via the dispersion parameter (α). A regularization term curbs overfitting by limiting large α values. Modified models maintain accuracy and often perform better with sparse datasets due to the NB distribution's fit for discrete data. Detailed comparisons can be found in the supplemental materials.

5.3 Baseline Calibration Methods

The SAUC will be compared to the baseline post-hoc calibration methods proven to be effective in the regression tasks using cross-sectional data. The baseline methods include:

- *Isotonic regression* [12], a non-parametric technique, which ensures non-decreasing predictive probabilities, assuming higher model accuracy with increased predicted probabilities;
- *Temperature scaling* [8], which scales the model's output using a learned parameter, whose efficacy for regression models depends on the output's monotonicity in relation to confidence;
- *Histogram binning* [28], a method partitioning predicted values into bins, calibrating each independently according to observed frequencies;
- *Platt Scaling* [7], which applies a scaling transformation to the predicted values with the function derived from minimizing observed and predicted value discrepancies, providing linear adaptability but potentially inadequate for complex uncertainty patterns;
- *Quantile regression* [2], aforementioned.

5.4 Calibration Results Comparison

The calibration performance of the initial models and our proposed calibration methods is presented in Table 1, assessed via ENCE on two adapted spatiotemporal prediction models. The "zero-only targets" refers to computing the ENCE exclusively on the true target values that are zeroes.

Table 1 denotes superior performance within each dataset with bold type and underlines the second-best results. Our findings reveal that the SAUC method typically outperforms other models, though it might fall short in aggregated cases, such as those with a 1-week resolution. Interestingly, the SAUC method achieved approximately a 23% reduction in ENCE compared to the second-best model when using full observations and also about a 20% reduction in sparsity entries. The separation of zero and non-zero data during calibration proved particularly beneficial, effectively addressing the issues of heteroscedasticity and zero inflation, commonly found in real-world datasets. However, we observed that calibration methods sometimes do not improve, or even worsen the calibration errors. It is mainly

STGCN-NB Outputs

Method / Interval	Full observations: Before calibration	Histogram binning	Isotonic regression	Temp. scaling	Platt scaling	QR	SAUC	Zero-only targets: Before calibration	Histogram binning	Isotonic regression	Temp. scaling	Platt scaling	QR	SAUC
CCR_1h	1.160	0.410	0.517	0.887	0.443	0.235	**0.198**	1.167	0.361	0.386	0.473	0.477	0.273	**0.267**
CCR_8h	1.389	0.344	0.417	0.385	0.373	0.392	**0.336**	1.065	0.568	0.578	0.427	0.162	0.175	**0.127**
CCR_1d	0.413	0.593	0.546	0.364	0.294	0.297	**0.192**	1.266	1.047	0.979	1.167	0.840	0.843	**0.738**
CCR_1w	0.855	0.784	0.773	0.765	0.579	0.272	**0.242**	0.354	0.362	0.332	0.287	0.345	0.155	**0.113**
CTC_1h	0.578	0.227	0.200	1.390	0.452	0.388	**0.165**	3.211	0.415	0.363	2.042	0.646	0.416	**0.078**
CTC_8h	0.323	0.320	0.255	0.570	0.297	0.241	**0.233**	0.330	0.247	0.276	0.551	0.230	0.196	**0.184**
CTC_1d	0.722	0.050	**0.011**	0.862	0.060	0.048	0.047	0.434	0.128	**0.027**	5.451	0.152	0.379	0.150
CTC_1w	2.279	1.227	1.527	0.914	2.265	0.378	**0.365**	1.823	2.384	2.009	1.160	1.343	0.767	**0.755**

GWN-NB Outputs

Method / Interval	Full observations: Before calibration	Histogram binning	Isotonic regression	Temp. scaling	Platt scaling	QR	SAUC	Zero-only targets: Before calibration	Histogram binning	Isotonic regression	Temp. scaling	Platt scaling	QR	SAUC
CCR_1h	1.200	1.022	0.889	1.149	0.834	0.604	**0.493**	2.042	1.796	1.274	1.710	1.356	0.476	**0.375**
CCR_8h	0.670	0.636	0.612	0.696	0.655	0.597	**0.566**	1.398	0.389	0.492	1.179	0.478	0.308	**0.186**
CCR_1d	0.997	0.662	0.633	0.581	0.562	0.587	**0.504**	1.158	0.583	0.580	0.725	0.747	0.644	**0.563**
CCR_1w	0.858	0.938	0.935	**0.611**	0.931	0.857	0.829	0.335	0.218	0.219	**0.161**	0.219	0.250	0.172
CTC_1h	0.868	0.816	0.891	0.859	0.634	0.832	**0.290**	1.300	0.712	0.561	1.001	0.143	0.998	**0.139**
CTC_8h	0.931	0.416	0.388	1.381	0.169	0.226	**0.148**	2.213	0.720	0.659	1.995	0.421	0.776	**0.221**
CTC_1d	0.523	0.420	0.448	0.571	0.523	0.432	**0.139**	0.250	0.464	0.405	0.248	0.484	0.439	**0.210**
CTC_1w	0.408	0.115	**0.114**	0.259	0.236	0.241	0.239	0.466	0.479	**0.436**	0.457	0.451	0.476	0.450

Table 1: ENCE of the calibration models. Bold fonts mark the best and underlines denote the second-best calibration results.

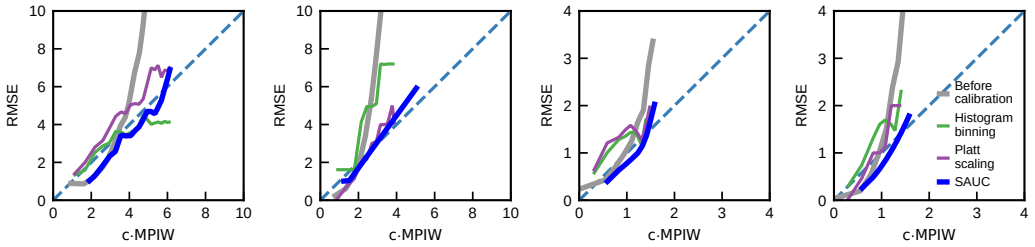
due to overfitting on the calibration set, a problem particularly evident in temperature scaling and isotonic regression methods.

On a general note, histogram binning and isotonic regression showed similar calibration results among baseline models as Equation 4 is non-decreasing, which makes the two methods easier to fit. Temperature scaling provides better calibration results in coarse temporal resolution scenarios and Platt scaling excels in sparse cases. QR consistently ranks as the second-best model in many cases, and our SAUC method has a close performance as QR in aggregated non-sparse cases but excels notably in sparse cases.

5.5 Reliability Diagram Evaluation

Utilizing Equations 6 and 7, we refine the conventional reliability diagram to evaluate the calibration efficacy of a model [8]. This involves partitioning predictions into bins based on $\mathcal{I}$, after which we compute the RMSE and MPIW employing the calibrated μ^* and $|\mathcal{I}^*|$ for each respective bin.

Figure 1 presents a comparative reliability diagram of STGCN-NB outputs prior to calibration, alongside Isotonic regression, Platt scaling, and our SAUC technique, each applied to different segments of the CCR_8h dataset. We selectively feature representative non-parametric and parametric baseline calibration approaches based on their performance. The diagonal dashed line symbolizes the calibration ideal: a closer alignment to this line indicates superior calibration.



(a) Full observations (b) Zero-only targets (c) Full observations (d) Zero-only targets

Figure 1: Reliability diagrams of different calibration methods applied to STGCN-NB outputs on different components of CCR_8h and CCR_1h data. (a) & (b) show the calibration performance on CCR_8h dataset and (c) & (d) reflect the results on CCR_1h dataset. Results closer to the diagonal dashed line are considered better.

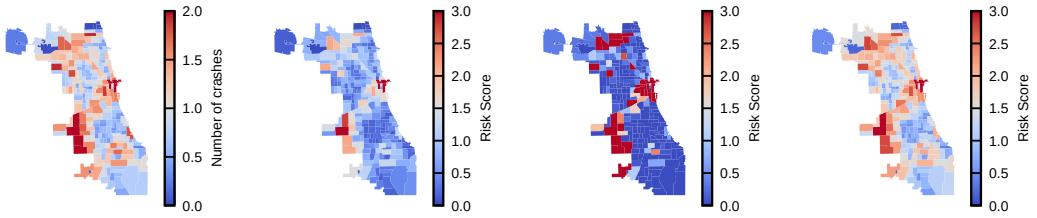
Examining Figure 1(a), both baseline models and SAUC calibrate well across all observations. However, focusing on the reliability diagram for ground-truth zeros in Figure 1(b), there's a deviation from the ideal line. In contrast, SAUC aligns well with the ideal for actual zero data, consistent with the CCR_1h dataset in Figures 1(c) and 1(d). Comparing Figures 1(a) and 1(c), calibration is less optimal in finer resolutions. Still, SAUC performs well, especially for zero predictions, due to our unique calibration and QR's ability to handle heteroscedasticity. While current probabilistic spatiotemporal models mainly focus on refining MPIWs for accuracy, they often overlook safety [22, 30, 6]. Our method ensures both accuracy and reliability.

5.6 Spatial Evaluation

To evaluate the use of SAUC outputs, we introduce Risk Score (RS), an enhancement of the traditional metric defined as the product of *event probability* and *potential loss magnitude* [11, 21]:

$$RS = \hat{\mu} \times |\hat{\mathcal{I}}|. \tag{9}$$

This metric integrates anticipated risk ($\hat{\mu}$) and prediction uncertainty ($|\hat{\mathcal{I}}|$), attributing higher RS to regions with both high incident frequency and uncertainty. This RS metric assigns higher risk to predictions that are characterized by both large predicted mean values and high uncertainty.



(a) Avg. number of crashes

(b) RS distribution using prediction before calibration

(c) RS distribution using calibrated prediction via Platt scaling

(d) RS distribution using using calibrated prediction via SAUC

Figure 2: Traffic crash accident distributions and calibration using CTC_8h data. All the data are averaged over the temporal dimension.

Using the CTC_8h dataset, we showcase SAUC's significance in assessing traffic accident risks. Figure 2 contrasts average crash numbers and RS values over time by region, both pre and post-calibration. Before calibration, Figure 2(b) shows RS in areas like northern and southern Chicago lacked clarity due to frequent incidents in Figure 2(a), influenced by tight PIs reducing RS. Platt scaling in Figure 2(c) identifies some risky zones but overlooks key crash areas in southern Chicago. However, SAUC predictions in Figure 2(d) reveal spatial patterns aligning closely with Figure 2(a), proving its value in urban safety monitoring.

Models should integrate uncertainty to convey consistent severity levels, enhancing reliability in crime and accident forecasts. The variance between zero and one incident is crucial for risk management. Current probabilistic studies often emphasize exact values and tight confidence bounds, occasionally overlooking variability around zero incidents.

6 Conclusion

Current spatiotemporal GNN models mainly yield deterministic predictions, overlooking data uncertainties. Given the sparsity and asymmetry in high-resolution spatiotemporal data, quantifying uncertainty is challenging. We introduce the SAUC method, which calibrates uncertainties for both zero and non-zero values, converting GNN deterministic models to probabilistic ones. SAUC addresses the zero-inflated nature of these datasets and introduces new calibration metrics for sparse, asymmetric distributions. Tests on real-world datasets show SAUC reduces calibration errors by 20% for zero-only targets. SAUC enhances GNN models and aids risk assessment in sparse datasets where accurate predictions are critical due to safety concerns.

References

[1] H Abdo, Jean-Marie Flaus, and F Masse. 2017. Uncertainty quantification in risk assessment-representation, propagation and treatment approaches: application to atmospheric dispersion modeling. *Journal of Loss Prevention in the Process Industries* 49 (2017), 551–571.

[2] Youngseog Chung, Willie Neiswanger, Ian Char, and Jeff Schneider. 2021. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. *Advances in Neural Information Processing Systems* 34 (2021), 10971–10984.

[3] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2021. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342* (2021).

[4] Sebastian Gruber and Florian Buettner. 2022. Better uncertainty calibration via proper scores for classification and beyond. *Advances in Neural Information Processing Systems* 35 (2022), 8618–8632.

[5] Kexin Huang, Ying Jin, Emmanuel Candes, and Jure Leskovec. 2023. Uncertainty Quantification over Graph with Conformalized Graph Neural Networks. `https://doi.org/10.48550/arXiv.2305.14535` arXiv:2305.14535 [cs, stat].

[6] Xinke Jiang, Dingyi Zhuang, Xianghui Zhang, Hao Chen, Jiayuan Luo, and Xiaowei Gao. 2023. Uncertainty Quantification via Spatial-Temporal Tweedie Model for Zero-inflated and Long-tail Travel Demand Prediction. *arXiv preprint arXiv:2306.09882* (2023).

[7] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*. PMLR, 2796–2804.

[8] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems* 32 (2019).

[9] Ananya Kumar, Percy S Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. *Advances in Neural Information Processing Systems* 32 (2019).

[10] Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. 2022. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors* 22, 15 (2022), 5540.

[11] Kwok-suen Ng, Wing-tat Hung, and Wing-gun Wong. 2002. An algorithm for assessing the risk of traffic accident. *Journal of safety research* 33, 3 (2002), 387–410.

[12] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*. 625–632.

[13] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring Calibration in Deep Learning.. In *CVPR workshops*, Vol. 2.

[14] Tim Pearce, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. 2018. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International conference on machine learning*. PMLR, 4075–4084.

[15] Moshiur Rahman. 2023. Uncertainty-Aware Traffic Prediction using Attention-based Deep Hybrid Network with Bayesian Inference. *International Journal of Advanced Computer Science and Applications* 14, 6 (2023).

[16] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. 2020. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637* (2020).

[17] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[18] Maohao Shen, Yuheng Bu, Prasanna Sattigeri, Soumya Ghosh, Subhro Das, and Gregory Wornell. 2023. Post-hoc Uncertainty Learning Using a Dirichlet Meta-Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 9772–9781.

[19] Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. 2019. Distribution calibration for regression. In *International Conference on Machine Learning*. PMLR, 5897–5906.

[20] Jayaraman J Thiagarajan, Bindya Venkatesh, Prasanna Sattigeri, and Peer-Timo Bremer. 2020. Building calibrated deep models via uncertainty matching with auxiliary interval predictors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6005–6012.

[21] David Vose. 2008. *Risk analysis: a quantitative guide*. John Wiley & Sons.

[22] Qingyi Wang, Shenhao Wang, Dingyi Zhuang, Haris Koutsopoulos, and Jinhua Zhao. 2023. Uncertainty Quantification of Spatiotemporal Travel Demand with Probabilistic Graph Neural Networks. *arXiv preprint arXiv:2303.04040* (2023).

[23] Ying Wu and JQ James. 2021. A bayesian learning network for traffic speed forecasting with uncertainty quantification. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–7.

[24] Ying Wu, Yongchao Ye, Adnan Zeb, James JQ Yu, and Zheng Wang. 2023. Adaptive Modeling of Uncertainties for Traffic Forecasting. *arXiv preprint arXiv:2303.09273* (2023).

[25] Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun. 2021. Inductive graph neural networks for spatiotemporal kriging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4478–4485.

[26] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. 2019. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121* (2019).

[27] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2017. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875* (2017).

[28] Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, Vol. 1. 609–616.

[29] Xiangyu Zhao, Wenqi Fan, Hui Liu, and Jiliang Tang. 2022. Multi-type urban crime prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 4388–4396.

[30] Dingyi Zhuang, Shenhao Wang, Haris Koutsopoulos, and Jinhua Zhao. 2022. Uncertainty quantification of sparse travel demand prediction with spatial-temporal graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4639–4647.