

Time-aware Confidence Calibration Analysis in Recommender Systems Methods

Joel Machado Pires
joelpires@ufba.br
Institute of Computing (IC) - Federal
University of Bahia (UFBA)
Salvador, Bahia

Eduardo Ferreira da Silva
eduardoferreira@ufba.br
Institute of Computing (IC) - Federal
University of Bahia (UFBA)
Salvador, Bahia

Mayki dos Santos Oliveira
maykioliveira@ufba.br
Institute of Computing (IC) - Federal
University of Bahia (UFBA)
Salvador, Bahia

Frederico Araújo Durão
fdurao@ufba.br
Institute of Computing (IC) - Federal
University of Bahia (UFBA)
Salvador, Bahia

Marcelo G. Manzato
mmanzato@icmc.usp.br
Instituto de Ciências Matemáticas e
de Computação, Universidade de São
Paulo
São Carlos, São Paulo

ABSTRACT

Recommender Systems (RS) are typically evaluated using accuracy and ranking metrics, but these metrics do not show whether the distribution of predicted ratings resembles the distribution of observed ratings. This aspect is especially important in confidence-aware recommendation, where confidence calibration indicates whether predictions assigned higher confidence are indeed more reliable. This work investigates the relationship among distributional fidelity, predictive performance, and confidence calibration in RS under temporal recommendation settings. We evaluate CPMF, LBD, OrdRec, PRLIGHTGCN, and MC Dropout on the MovieLens 1M dataset using an incremental time-series cross-validation protocol, and analyze the effects of Wasserstein-based distribution regularization. The results show that models produce characteristic prediction distributions that remain relatively stable over time but do not converge toward the empirical rating distribution, even as additional interactions become available. Moreover, strong recommendation accuracy does not necessarily imply statistical fidelity to the observed data. Confidence calibration evolves over time and depends strongly on model architecture, while modifications to the global prediction distribution have limited effects on user-item confidence calibration. Overall, distributional fidelity emerges as a complementary dimension for evaluating trustworthy and confidence-aware recommender systems.

KEYWORDS

confidence, distribution, calibration, recommender systems

1 INTRODUCTION

Recommender systems (RS) are central to how users navigate large-scale digital environments, and their evaluation has long been dominated by predictive accuracy and ranking metrics. While these metrics quantify how closely predictions match individual observed

ratings, they do not capture whether the overall distribution of model outputs reflects the statistical properties of observed user behavior. This distributional dimension is rarely examined but becomes especially important in confidence-aware recommendation, where models are expected to produce not only accurate predictions but also reliable uncertainty estimates.

Recommendation models are typically evaluated using predictive accuracy metrics, such as root mean squared error (RMSE) and mean absolute error (MAE), and ranking metrics, such as normalized discounted cumulative gain (NDCG) and mean average precision (MAP). Although these metrics are important, they do not fully characterize the behavior of a recommendation model. In particular, they do not indicate whether the distribution of predicted ratings resembles the distribution of observed ratings [15], for example, whether the model preserves the relative frequency of low, medium, and high ratings. As a consequence, a model may achieve competitive accuracy or ranking performance while producing predictions whose global distribution differs substantially from the empirical distribution of user behavior [2, 4, 15, 17, 19].

This limitation is particularly relevant in the context of confidence estimation and uncertainty-aware recommendation. In this setting, models do not only predict ratings or rankings, but also provide some indication of how reliable each prediction is expected to be. Therefore, confidence calibration becomes an important evaluation aspect: predictions assigned higher confidence should be more reliable, or equivalently, should be associated with lower prediction error. However, if a model produces distorted prediction distributions, it is unclear whether its confidence estimates remain reliable, especially when user interactions evolve over time.

Prior work has proposed probabilistic extensions of collaborative filtering architectures to estimate prediction uncertainty through explicit distribution modeling. Representative examples include OrdRec (OrdRec) [10], confidence-aware probabilistic matrix factorization (CPMF), confidence-aware bayesian probabilistic matrix factorization (CBPMF) [16], learned beta distributions (LBD) [8], probabilistic recommender graph attention networks (PRGAT), and probabilistic recommender light graph convolutional networks (PRLIGHTGCN) [12]. These approaches incorporate prior distributions or probabilistic assumptions into recommendation models to

estimate confidence intervals or uncertainty scores associated with predictions. Nevertheless, recent empirical studies have reported several limitations in these approaches, including degraded predictive accuracy, convergence instability, confidence miscalibration, and strong dependency on dataset characteristics and architectural design [5, 12].

More broadly, several studies have raised concerns regarding the reliability, fairness, calibration, diversity, and statistical behavior of RS [2, 4, 15, 17, 19]. Recent work has shown that recommendation quality cannot be fully characterized by predictive accuracy alone, as issues such as popularity bias, fairness, and calibration may significantly affect the alignment between recommendations and user interests [15]. However, while these studies highlight the importance of going beyond accuracy, little attention has been given to the temporal behavior of prediction distributions in confidence-aware recommendation. In particular, it remains underexplored whether recommendation models preserve the empirical distribution of observed ratings over time, whether distributional distortions affect predictive and ranking performance, and how such distortions relate to confidence calibration.

To address these gaps, this work presents an experimental study investigating the relationship among distributional fidelity, predictive performance, and confidence calibration in RS under temporal recommendation settings. We evaluate CPMF, LBD, OrdRec, PRLIGHTGCN, and MC Dropout under an incremental time-series cross-validation protocol, analyzing how predicted rating distributions evolve over time relative to the empirical distribution of observed ratings, and whether distributional distortions are associated with changes in predictive performance, ranking quality, and confidence calibration. To further investigate this relationship, we employ a Wasserstein-based distribution regularization strategy to examine how controlled modifications to the global prediction distribution affect model behavior.

The experiments are conducted using the MovieLens 1M dataset under an incremental time-series cross-validation protocol that simulates temporal recommendation scenarios while preventing future interactions from being observed during training. The evaluation combines predictive performance metrics, including RMSE, NDCG, and MAP, with distributional and calibration analyses based on Kullback–Leibler divergence, prediction-distribution inspection, confidence-error correlation, and Expected Calibration Error.

The main contributions of this work are as follows:

- We analyze the temporal evolution of prediction distributions in confidence-aware recommendation models.
- We investigate the relationship among distributional fidelity, predictive performance, ranking quality, and confidence calibration.
- We evaluate whether Wasserstein-based distribution regularization can modify prediction distributions and how these modifications affect accuracy, ranking performance, and calibration.
- We provide empirical evidence that strong predictive performance does not necessarily imply distributional fidelity, and that changes in the global prediction distribution have limited effects on user-item confidence calibration.

The remainder of this paper is organized as follows. Section 2 presents confidence in the context of RS and the experimented confidence-aware RS methods. Section 3 reviews prior work on confidence-aware and distribution-aware recommender systems. Section 4 describes the experimental setup and evaluation protocol. Section 5 presents the experimental results. Section 6 discusses the findings and answers the research questions. Finally, Section 7 concludes the paper and outlines future research directions.

2 BACKGROUND

Confidence- and uncertainty-aware recommendation models can be analyzed at different levels of granularity, including global, user, item, and user-item levels. At each level, a model induces a distribution over rating values: the global level considers all predicted ratings, while finer-grained levels condition this distribution on specific users, items, or user-item pairs. Figure 1 illustrates these levels and their corresponding distributional views.

Because these levels capture different aspects of model behavior, it is important to distinguish distributional fidelity from confidence calibration. Distributional fidelity refers to whether the distribution of predicted ratings resembles the distribution of observed ratings at a given level. Confidence calibration, in turn, refers to whether the confidence assigned to a prediction reflects its actual reliability, so that higher-confidence predictions should generally be associated with lower prediction error. Therefore, global distributional alignment does not necessarily imply calibrated confidence at the user-item level, and modifying the global prediction distribution may not improve calibration at finer levels of granularity.

Note that, in RS, interactions between a user and an item are typically observed only once. Consequently, the true user-item interaction distribution cannot be directly estimated, as repeated observations of the same interaction are generally not available. To address this limitation, studies at this level of confidence estimation commonly impose assumptions about the underlying uncertainty, often by modeling it with a predefined probability distribution. This section presents the confidence-aware recommendation methods evaluated in this work: OrdRec [10], CPMF [16], LBD [8], PRLIGHTGCN [12], and MC Dropout [6]. All these methods minimize the negative log loss function using a gradient descent-based algorithm,

$$\mathcal{L} = -\frac{1}{|\mathcal{R}|} \sum_{(u,i) \in \mathcal{R}} \log P(r_{ui} | \Theta), \quad (1)$$

where $\mathcal{R}$ is the set of observed user-item interactions and Θ is the models' parameters.

2.1 OrdRec

Koren and Sill (2011) propose OrdRec, a collaborative filtering (CF) method that models user ratings as ordinal rather than numerical outcomes. For each user–item interaction, a latent preference score predicted by a matrix factorization (MF) model is assumed to follow a Normal distribution and is mapped to discrete rating levels through user-specific, monotonically increasing thresholds parameterized by learned scalars. The probability that a rating does not exceed a given level is modeled using a logistic function:

$$P(r_{ui} \leq r | \Theta) = \frac{1}{1 + e^{y_{ui} - t_r^u}}, \quad (2)$$

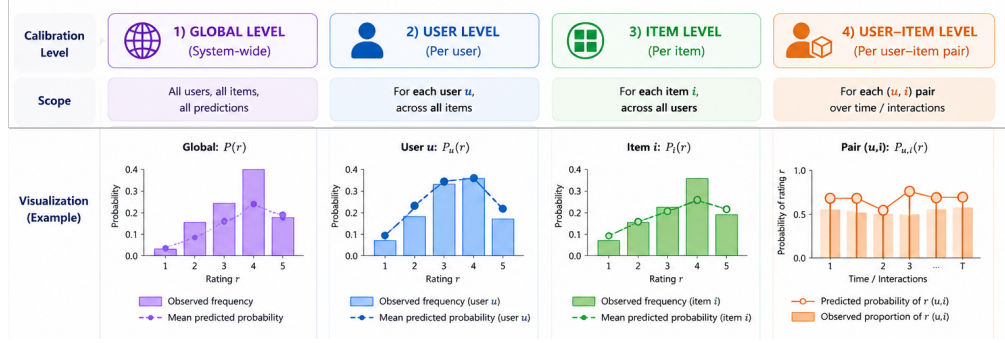


Figure 1: Different levels of confidence modeling and calibration. The global level considers all model predictions and estimates the probability of observing a rating (r), regardless of the user or item, i.e., $P(r)$. The user level estimates the probability of observing a rating (r) given a user (u), i.e., $P(r | u)$. Similarly, the item level estimates the probability of observing a rating (r) given an item (i), i.e., $P(r | i)$. Finally, the user-item level estimates the probability of observing a rating (r) given both a user (u) and an item (i), i.e., $P(r | u, i)$.

where y_{ui} is the predicted latent score and Θ denotes all model parameters. Model training minimizes the negative log-likelihood of observed ratings using stochastic gradient descent.

OrdRec also introduces a post-training calibration step to improve ranking quality. After fitting the ordinal model, a vector of ranking weights is learned for each rating level. For item pairs where the observed user preference is known, the difference between their predicted rating probability distributions is computed, and the weights are optimized to maximize a pairwise ranking objective:

$$F(\mathcal{T} | \mathbf{w}) = \sum_{(i,j) \in \mathcal{T}} f(\mathbf{w}^T \Delta P_u(i, j)), \quad (3)$$

where $\mathcal{T}$ denotes the set of correctly ordered item pairs and f is the sigmoid function.

During inference, the model outputs a full probability distribution over rating levels. The predicted rating is given by the expected value of this distribution, while a confidence estimate is derived from its standard deviation. For ranking tasks, items are scored using the calibrated ranking function $s_{ui} = \mathbf{w}^T \mathbf{p}_{ui}$ and sorted accordingly.

2.2 CPMF

Confidence-aware MF extends standard matrix factorization by modeling not only the expected rating, but also the uncertainty associated with each prediction. This study considers the CPMF variant. Assuming that the rating r_{ui} follows a Normal distribution, the mean and variance are given by

$$\mu_{ui} = \mathbf{e}_u \cdot \mathbf{e}_i \quad \text{and} \quad v = \frac{1}{\gamma_{U_i} \cdot \gamma_{V_j} \cdot \alpha_p}, \quad (4)$$

where γ_{U_i} , γ_{V_j} , and α_p are learned scalar parameters associated with users, items, and the confidence level. Instead of only predicting a point estimate, the model defines a prediction interval within which the rating is expected to lie with a given confidence level p :

$$[\mu_{ui} - Z(p/2) \cdot \sqrt{v}; \mu_{ui} + Z(p/2) \cdot \sqrt{v}], \quad (5)$$

where $Z(p/2)$ denotes the quantile of the standard Normal distribution corresponding to the two-tailed confidence level p .

2.3 Learned Beta Distributions

LBD represents each user-item interaction as a Beta distribution whose parameters are produced by a matrix factorization-like architecture. The mean parameter, μ_{ui} , is obtained from the cosine similarity between user and item embeddings, while the confidence term is defined as

$$v_{ui} = \|\mathbf{e}_u + \mathbf{e}_i\|_2. \quad (6)$$

The Beta shape parameters are constructed by scaling the mean and its complement by the confidence term. They are further adjusted using learned global, user-specific, and item-specific bias terms, with a small constant added to ensure numerical stability. The predicted rating corresponds to the expected value of the resulting Beta distribution, rescaled to the original rating range.

To train the model, the normalized rating interval $[0, 1]$ is discretized into S adaptive bins whose boundaries depend on both the user and the item. Bin widths are determined by learned user and item embeddings through an exponential weighting scheme that is normalized to form a valid partition of the interval. The probability mass assigned to each bin is computed as the difference between the Beta cumulative distribution function evaluated at consecutive bin boundaries.

2.4 PRLIGHTGCN

PRLIGHTGCN integrates probabilistic confidence estimation into graph-based recommendation. It builds on light graph convolutional networks (LIGHTGCN) to estimate the mean rating while incorporating uncertainty modeling techniques inspired by CPMF. The goal is to obtain a more stable confidence-aware graph recommender.

Similarly to CPMF, PRLIGHTGCN assumes that ratings follow a Normal distribution for each user-item pair (u, i) . The standard deviation of this distribution is defined as

$$\sigma_{ui} = \frac{1}{\sqrt{\gamma_u \cdot \gamma_i \cdot v'}}, \quad (7)$$

where γ_u , γ_i , and v' are learned parameters associated with users, items, and a global uncertainty component, respectively. The mean

μ_{ui} is estimated by a LIGHTGCN model; a graph attention networks (GAT)-based variant, PRGAT, also exists but is not evaluated in this work. The confidence estimate is obtained from the probability mass around the predicted value:

$$P(r_{ui} = \mu_{ui} | \Theta) = P(r_{ui} \leq \mu_{ui} + \delta_r) - P(r_{ui} \leq \mu_{ui} - \delta_r), \quad (8)$$

where δ_r is a small tunable scalar that defines the interval around the predicted rating.

The final objective combines the negative log-likelihood term with an mean squared error (MSE)-based term:

$$\mathcal{L}_{PRGAT} = \mathcal{L}_{NLL} + \mathcal{L}_{MSE}, \quad (9)$$

where

$$\mathcal{L}_{NLL} = \frac{1}{|\mathcal{R}|} \sum_{(u,i) \in \mathcal{R}} -\log(P(r_{ui} = \mu_{ui} | \Theta)), \quad (10)$$

and

$$\mathcal{L}_{MSE} = \frac{1}{|\mathcal{R}|} \sum_{(u,i) \in \mathcal{R}} \delta_{mse} \cdot \text{MSE}(r_{ui}, \mu_{ui}). \quad (11)$$

The parameter δ_{mse} controls the trade-off between the standard recommendation objective and the confidence-aware objective.

2.5 MC Dropout

MC Dropout applies dropout at inference time to approximate Bayesian uncertainty estimation without modifying the model architecture [6]. By performing multiple stochastic forward passes with dropout active, the model produces a distribution of predictions for each user-item pair. The mean of this distribution serves as the predicted rating, while the variance provides a confidence estimate: lower variance indicates higher confidence in the prediction.

3 RELATED WORK

Prior studies have investigated uncertainty and confidence estimation in recommender systems, showing that estimating reliable confidence scores remains a challenging task. Coscrato and Bridge [5] conducted an extensive empirical study comparing several classes of uncertainty estimators for recommender systems, including OrdRec and CPMF. Their results revealed important concerns regarding the reliability and consistency of uncertainty estimates for rating prediction. In particular, several methods exhibited inconsistent behavior across uncertainty-error evaluation metrics, performing well according to one correlation measure while performing poorly according to others, such as Spearman correlation or UPI. The authors also introduced two uncertainty estimation approaches, ENSEMBLE and EB-Linear; however, these methods did not outperform simpler information-based estimators in most evaluation metrics.

More recently, Pires et al. [12] explored confidence-aware recommendation models such as OrdRec, CBPMF, CPMF, and LBD. Their study identified convergence issues, dataset dependency, confidence miscalibration, and degraded predictive accuracy resulting from the integration of confidence estimation into recommendation models. To address these limitations, they proposed a confidence-integration approach based on graph neural recommendation models, introducing PRGAT and PRLIGHTGCN. Although these models improved the stability of confidence-aware recommendation, the results still

indicated remaining issues related to confidence miscalibration and dataset dependency.

Other works have addressed calibration from different application perspectives. Wu [17] investigated trustworthy graph learning in recommender systems, focusing on reliability, explainability, and fairness. The author proposed LogiRec, a framework based on first-order logical reasoning and probabilistic embeddings to capture asymmetrical and higher-order product relationships. In the context of reliability, this work addressed rarity and miscalibration in minority classes by introducing the expected individual calibration error. In a different domain, Chen et al. [4] explored confidence calibration and quantification in medication recommendation, proposing distribution-based approaches for a multi-label clinical setting. However, this methodology is tailored to clinical safety and does not address distribution-based confidence integration in regression-oriented recommender systems.

Although these studies provide important evidence that confidence estimation and calibration are difficult to achieve in recommender systems, they are mostly based on static evaluation settings or focus on specific domains and tasks. In contrast, this work investigates how confidence calibration evolves over time as new user interactions become available. We also analyze how the distributions of predicted and observed ratings evolve under a temporal evaluation protocol, whether distributional mismatches are associated with predictive and ranking performance, and whether Wasserstein-based distribution regularization can modify prediction distributions and affect confidence calibration. These analyses address an underexplored aspect of confidence-aware recommendation: the relationship between temporal distributional fidelity, recommendation quality, and user-item confidence calibration.

4 EXPERIMENTAL EVALUATION

This section describes the experimental protocol used to investigate the relationship among predictive performance, distributional fidelity, and confidence calibration under temporal recommendation settings. Our evaluation is guided by the following research questions:

- **RQ1:** How do the prediction distributions of the evaluated recommendation models evolve over time?
- **RQ2:** Do recommendation models reproduce the empirical distribution of observed ratings?
- **RQ3:** Can models achieve competitive accuracy while producing distorted prediction distributions?
- **RQ4:** How does confidence calibration evolve over time across different recommendation models?
- **RQ5:** How does Wasserstein-based distribution regularization affect predictive performance, ranking quality, distributional fidelity, and confidence calibration?

The code of this work is publicly available online¹.

4.1 Dataset

The experiments are conducted on the MovieLens 1M [7] dataset, which contains 1,000,209 explicit ratings provided by 6,040 users

¹<https://github.com/joel021/rating-pred-dynamic-conf-analysis>

on 3,706 movies, on a scale from 1 to 5. Ratings were collected between April 2000 and February 2003. The availability of timestamp information makes this dataset suitable for temporal evaluation, allowing us to simulate the incremental arrival of user interactions over time.

4.2 Temporal Evaluation Protocol

We employ an incremental time-series cross-validation protocol. The interaction data are sorted globally by timestamp and partitioned into $K + 1 = 7$ sequential temporal folds $F_0, F_1, \dots, F_6$. For each evaluation step $k \in \{1, \dots, 6\}$, the training set is defined as

$$\mathcal{D}_{train}^{(k)} = \bigcup_{i=0}^{k-1} F_i, \quad (12)$$

while the test set is given by $\mathcal{D}_{test}^{(k)} = F_{k+1}$.

In this work, we perform $K = 7$ evaluation rounds of approximately equal size. Therefore, it begins with fold 0 for training and fold 1 for testing. Consequently, the end uses folds 0-5 for training and fold 6 for tests. Then the results will show charts with folds ranging from 0 to 5.

4.3 Methods

We evaluate the following recommendation models: CPMF, LBD, OrdRec, PRLIGHTGCN, MC Dropout, and MF. The MF model factorizes the user-item interaction matrix into two low-dimensional embedding matrices, estimating ratings as the dot product between user and item latent vectors [9]. It is included as a non-probabilistic baseline and is used in the controlled regularization analysis. For all models, the embedding dimensions were set according to the configurations reported in their original papers: 512 for OrdRec, 20 for both CPMF and CBPMF, 522 for LBD, and 64 for both MC Dropout and PRLIGHTGCN models.

The confidence-aware models are used to analyze the relationship between uncertainty estimation, confidence calibration, and prediction distributions. Each model is evaluated under its original objective function. In addition, we evaluate Wasserstein-regularized variants by adding a distribution-matching term to the original loss:

$$\mathcal{L} = \mathcal{L}_{model} + \lambda_w \cdot W(\text{SoftHist}(\hat{r}), \text{SoftHist}(r)), \quad (13)$$

where $\mathcal{L}_{model}$ is the original model loss, $\hat{r}$ denotes the predicted ratings, r denotes the observed ratings, $\text{SoftHist}(\cdot)$ denotes a differentiable histogram approximation, $W(\cdot, \cdot)$ is the Wasserstein distance between the predicted and observed rating distributions, and λ_w controls the strength of the distributional regularization.

4.4 Metrics

Predictive performance is evaluated using RMSE. Ranking performance is evaluated using NDCG and MAP [1, 18]. Distributional fidelity is evaluated by constructing histogram-based distributions with shared bins for predicted and observed ratings, and by measuring their discrepancy using Kullback-Leibler divergence. Confidence calibration is evaluated using expected calibration error (ECE) [11] and confidence-error Pearson correlation [13]. Lower ECE indicates better calibration, while a stronger negative confidence-error correlation indicates that higher-confidence predictions tend to be associated with lower prediction errors. All metrics are tracked

across temporal folds to analyze temporal stability, distributional drift, and changes in calibration behavior.

4.5 Simulation 1: Temporal Analysis with and without Distribution Regularization

The first simulation analyzes the temporal behavior of the evaluated models under their original objective functions and under Wasserstein-regularized variants. For this experiment, the regularization coefficient is fixed at $\lambda_w = 1$ for all Wasserstein-regularized models.

This simulation aims to investigate whether the prediction distributions produced by confidence-aware recommendation models approximate the empirical distribution of observed ratings over time. It also analyzes whether enforcing global distributional alignment affects predictive performance, ranking quality, and confidence calibration.

4.6 Simulation 2: Controlled Regularization Intensity in Matrix Factorization

To better understand the effect of distributional regularization under controlled perturbations, we conduct a second simulation using the MF model. The MF model is selected for this analysis because, unlike the probabilistic models, it does not impose prior distributional constraints on its predictions, making it more susceptible to distribution-matching regularization and thus a more informative setting for studying the effect of varying λ_w .

In this simulation, we vary the regularization coefficient as follows:

$$\lambda_w \in \{0, 0.0001, 0.001, 0.01, 0.5, 1.0\}. \quad (14)$$

This allows us to examine how different levels of distribution matching affect predictive performance, ranking quality, distributional fidelity, and confidence calibration under the same temporal evaluation protocol.

5 RESULTS

Figure 2 shows that confidence calibration varies substantially over time, as indicated by the differences across folds 0 to 4. The figure also sheds light on the sensitivity of confidence calibration to distribution distortions. Specifically, the application of Wasserstein regularization appears to have little or no impact on calibration performance. Because this regularization acts on the global distribution, the results suggest that either (i) the models have already reached a near-optimal approximation of the global distribution, making further improvements negligible, or (ii) the global distribution has limited influence on confidence calibration at the user-item level. These findings indicate that improving global distribution alignment alone may not be sufficient to enhance calibration at finer levels of granularity. Note that OrdRec and PRLIGHTGCN exhibit the best calibration performance among the evaluated models. OrdRec reduced the ECE and increased the magnitude of the negative correlation, reaching approximately -0.2 in the final fold. Similarly, PRLIGHTGCN achieved a lower RMSE while maintaining a stable ECE and strengthening the negative correlation to approximately -0.3. The application of Wasserstein regularization kept the same trend.

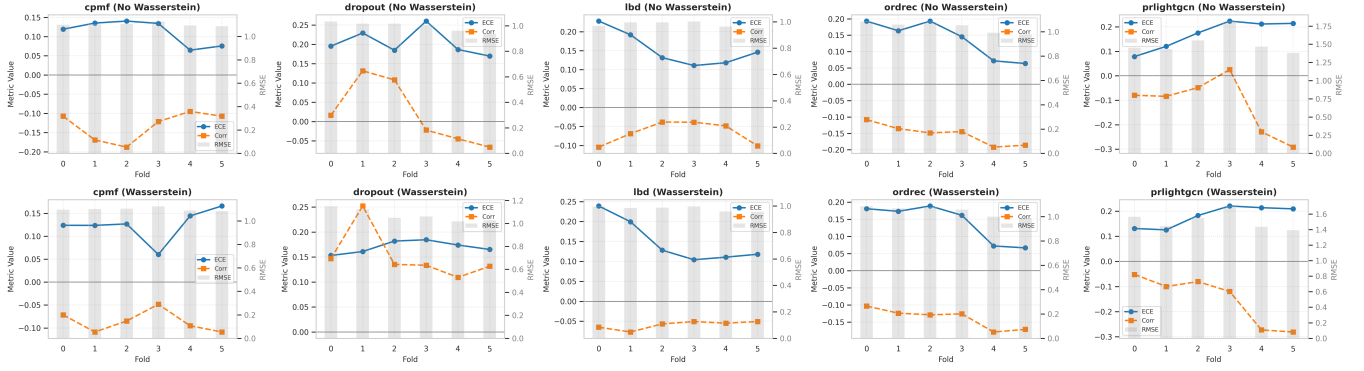


Figure 2: ECE error and confidence-error correlation across folds for confidence-aware models with and without Wasserstein regularization, evaluated on the Movie Lens 1M dataset. Higher negative correlation is desired, where lower ECE is better. These two measures evaluate the user-item level of confidence calibration.

Figure 3 presents the RMSE performance over a narrower range, allowing for a more detailed analysis than that provided by Figure 2. Similar to the confidence calibration results, we observe performance variations across folds. However, distribution distortion appears to have little effect on predictive performance, as the models exhibit similar behavior regardless of whether Wasserstein regularization is applied. In the case of the MF model, we observe a considerable drop in RMSE error when the Wasserstein regularization is applied. Compared to the other models, MF does not impose prior distributional constraints on its predictions. The remaining models may already partially constrain the prediction distribution through their probabilistic assumptions, making additional distribution alignment less impactful.

Figure 5 shows that the Dropout and MF models are the most affected by the distributional regularization. This suggests either that the remaining models already impose distributional constraints that influence the global rating distribution or that their architectures are inherently more resistant to distribution matching. Notably, none of the models fully reproduces the observed distributions, as shown in Figure 4, although the Dropout and MF models exhibit distributions that more closely resemble the observed ones after the application of the regularization.

Figure 6 shows that the introduction of the distribution constraint leads to a sharp reduction in RMSE, which quickly saturates as the constraint strength increases. However, this improvement in RMSE does not translate into better NDCG performance, a phenomenon also reported in previous RS studies [14]. Contrary to expectations, applying Wasserstein regularization yields higher Kullback–Leibler (KL) divergence values, which also appear largely insensitive to the distortion intensity. This outcome is consistent with the fact that Wasserstein distance and KL divergence characterize different aspects of distribution similarity. Wasserstein distance measures the optimal transport cost required to transform one distribution into another, while KL divergence measures discrepancies in probability mass allocation [3]. As a consequence, optimizing a model with respect to Wasserstein distance does not necessarily improve, and may even worsen, the distributional alignment measured by KL divergence.

6 DISCUSSION

This section discusses the experimental findings by addressing each research question defined in Section 4.

RQ1: How do the prediction distributions of the evaluated recommendation models evolve over time? The results, particularly those presented in Figure 5, show that each model imposes a characteristic distributional shape on its predictions and maintains this shape as additional interactions are incorporated over time. Rather than adapting to evolving data, the models tend to produce stable output distributions specific to their architecture. This suggests that the inductive biases embedded in each model — such as the ordinal thresholds in OrdRec or the Beta parameterization in LBD — strongly determine the shape of the prediction distribution, independently of the volume of training data.

RQ2: Do recommendation models reproduce the empirical distribution of observed ratings? No. As shown in Figure 4, the observed rating distribution remains relatively stable throughout the evaluation period. Although one would expect predicted distributions to progressively approximate the observed distribution as additional data become available, this convergence is not observed in practice. None of the evaluated models fully reproduces the empirical rating distribution, and the gap between predicted and observed distributions persists across all temporal folds. This indicates that distributional fidelity is not an emergent property of training on more data, but rather a characteristic that depends on model design.

RQ3: Can models achieve competitive accuracy while producing distorted prediction distributions? Yes. Although the models achieve competitive RMSE performance, as shown in Figure 3, their predicted rating distributions differ substantially from each other and from the observed distribution, as illustrated in Figure 5. Furthermore, the distributional regularization experiments show that improving the alignment between predicted and observed distributions does not necessarily degrade — and may even improve — predictive accuracy. The MF model, for instance, exhibited a consistent and substantial reduction in RMSE after regularization, while remaining the model most affected in terms of distribution matching. These results confirm that strong predictive performance does

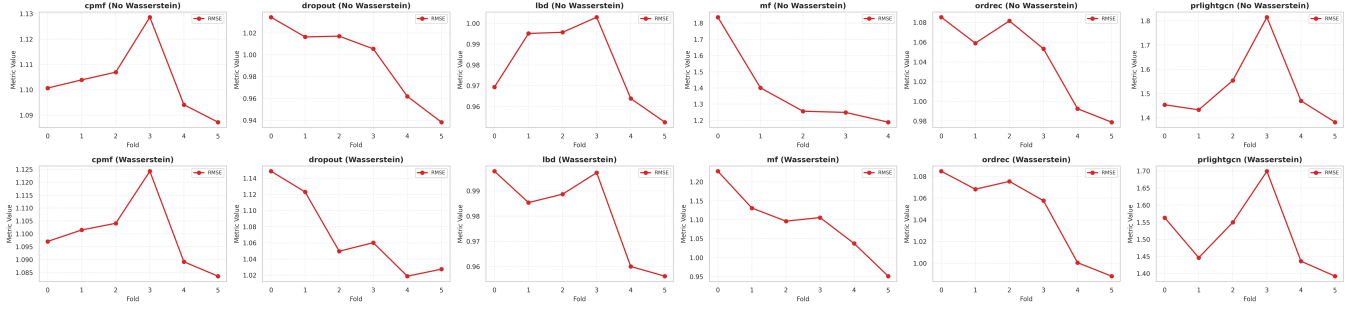


Figure 3: Error performance of the models over time evaluated in the Movie Lens 1M.

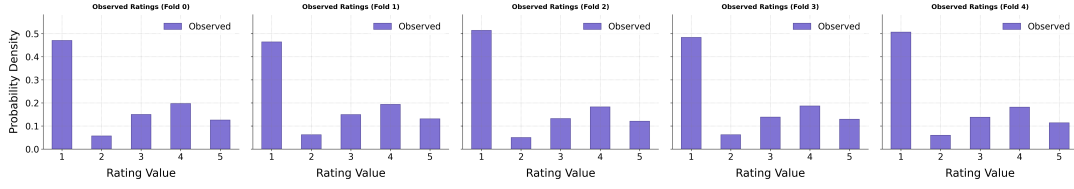


Figure 4: Observed rating distributions.

not imply distributional fidelity: models with equivalent accuracy can differ substantially in their statistical alignment with observed user behavior.

RQ4: How does confidence calibration evolve over time across different recommendation models? User-item confidence calibration varies considerably over time and is strongly dependent on model architecture, as shown in Figure 2. Among the evaluated models, only OrdRec and PRLIGHTGCN exhibit a clear and consistent temporal trend in calibration: OrdRec progressively reduced its ECE and strengthened the negative confidence-error correlation, reaching approximately -0.2 in the final fold, while PRLIGHTGCN maintained a stable ECE and strengthened the correlation to approximately -0.3 . The remaining models exhibit substantial fluctuations across folds without a clear trend. Importantly, these temporal changes in user-item calibration are largely independent of the evolution of the global rating distribution, which remains stable throughout the evaluation for all models. This dissociation suggests that calibration improvements in OrdRec and PRLIGHTGCN arise from architecture-specific mechanisms rather than from changes in global distributional behavior.

RQ5: How does Wasserstein-based distribution regularization affect predictive performance, ranking quality, distributional fidelity, and confidence calibration? The effects of Wasserstein-based regularization are inconsistent across evaluation dimensions, as shown in Figure 6. In terms of predictive performance, regularization leads to a sharp reduction in RMSE for the MF model, which quickly saturates as λ_w increases. However, this improvement does not transfer to ranking quality: NDCG and MAP remain largely unaffected or slightly degraded, consistent with prior findings in the RS literature [14]. Regarding distributional fidelity, the regularization shifts the predicted distributions of the MF and MC Dropout models toward the observed distribution, but does not eliminate the mismatch. Counterintuitively, KL

divergence increases under Wasserstein regularization, which is explained by the fact that Wasserstein distance and KL divergence capture different aspects of distributional similarity: the former measures optimal transport cost, while the latter measures discrepancies in probability mass allocation [3]. Finally, confidence calibration at the user-item level is largely insensitive to the application of Wasserstein regularization, reinforcing the finding that global distributional alignment does not imply calibration at finer levels of granularity.

7 CONCLUSION

This work presented an experimental study investigating the relationship among distributional fidelity, predictive performance, and confidence calibration in recommender systems under a temporal recommendation setting, a dimension that conventional accuracy and ranking metrics do not capture. The results showed that confidence-aware recommendation models produce characteristic prediction distributions that remain stable over time and do not converge toward the empirical rating distribution, even as additional interactions become available (RQ1, RQ2). Furthermore, models can achieve competitive predictive performance while producing substantially distorted output distributions, confirming that strong RMSE does not imply distributional fidelity (RQ3). These findings suggest that conventional evaluation metrics capture only part of model behavior and that distributional fidelity should be considered a complementary dimension for assessing recommendation quality.

Regarding confidence calibration, the results revealed that calibration evolves over time and is strongly dependent on model architecture, with only OrdRec and PRLIGHTGCN exhibiting consistent improvement across temporal folds (RQ4). Wasserstein-based regularization modified the global prediction distribution of MF and MC Dropout, reduced RMSE, but had limited effects on ranking quality and did not improve user-item confidence calibration (RQ5). This

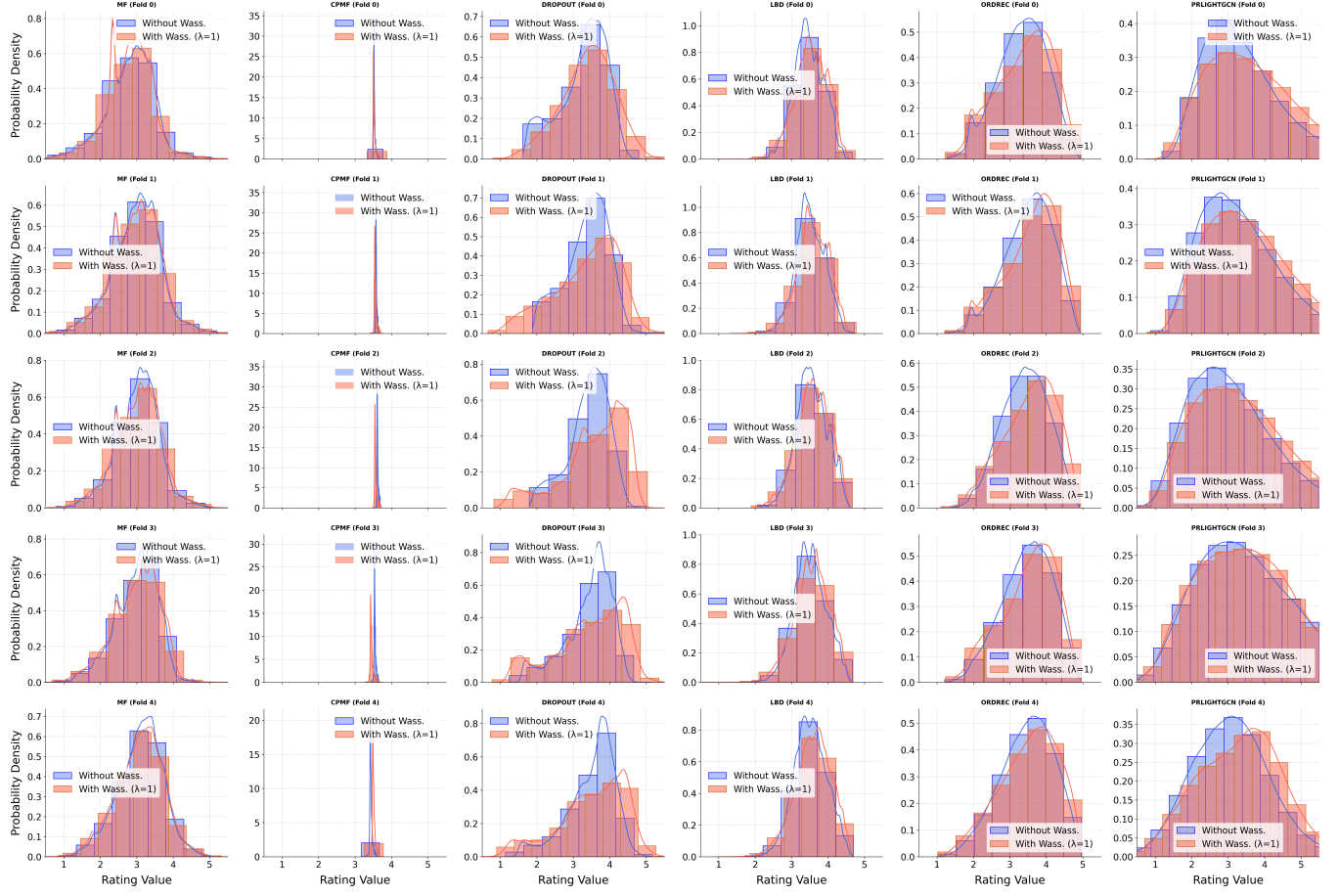


Figure 5: Distribution of the predicted ratings with and without Wasserstein regularization ($\lambda_w = 1$) across the models in the last evaluated fold, evaluated in the Movie Lens 1M dataset.

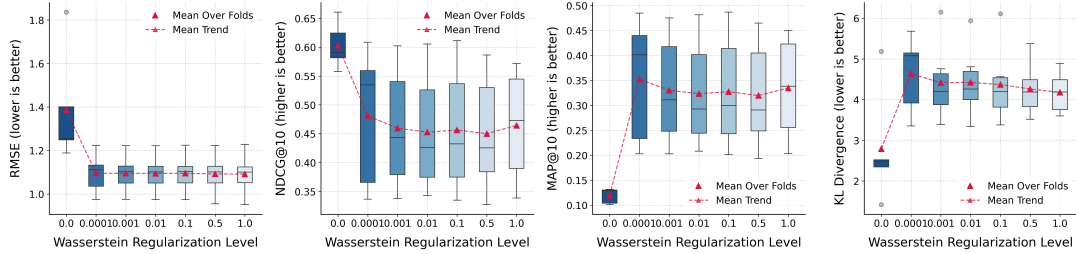


Figure 6: Average metric values across folds for different Wasserstein regularization levels (λ_w) in the loss function of the MF model, evaluated on the Movie Lens 1M dataset. The first value of Wasserstein regularization level is 0.0001, and the last is 1.0, for each metric, RMSE, NDCG@10, MAP@10, and KL Divergence.

indicates that global distributional alignment alone is not sufficient to improve calibration at finer levels of granularity.

This study has limitations that should be considered. The evaluation is conducted on a single dataset (MovieLens 1M) with a fixed number of temporal folds, and the findings may not generalize to other domains or interaction sparsity regimes. Future work should evaluate additional datasets and extend the analysis to user-level

and item-level distributional alignment. Investigating regularization strategies that directly target calibration at the user-item level, rather than the global distribution, is a promising direction for developing more trustworthy confidence-aware recommender systems.

ACKNOWLEDGMENTS

The authors acknowledge the financial support from FAPESP, CAPES, CNPq, and INCT-TILD-IAR. The authors used AI-assisted tools (Claude/Anthropic and ChatGPT/OpenAI) during the preparation of this work to support code development, data analysis, and manuscript editing. After using these tools, the authors reviewed and edited all content as needed and take full responsibility for the content of the publication. No AI-generated text was used without revision by the authors.

ETHICAL CONSIDERATIONS

This study uses only a publicly available dataset (MovieLens 1M) and did not involve direct interaction with human participants. Only the information provided in the public datasets was used, with no attempt to identify or re-identify users. Therefore, the study is consistent with Brazilian Resolution CNS No. 510/2016 and did not require evaluation by a Research Ethics Committee.

REFERENCES

- [1] Charu C. Aggarwal. 2016. *Recommender Systems: The Textbook* (1st ed.). Springer Publishing Company, Incorporated, Online.
- [2] Gabrielle Alves and Marcelo Manzato. 2024. Furando a Bolha: Nudges Digitais em Sistemas de Recomendação. In *Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web* (Juiz de Fora/MG). SBC, Porto Alegre, RS, Brasil, 11–12. https://doi.org/10.5753/webmedia_estendido.2024.244374
- [3] Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi. 2018. Information geometry connecting Wasserstein distance and Kullback–Leibler divergence via the entropy-relaxed transportation problem. *Information Geometry* 1, 1 (2018), 13–37.
- [4] Qianyu Chen, Xin Li, Yujie Fang, and Mingzhong Wang. 2025. Advancing Confidence Calibration and Quantification in Medication Recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1* (Toronto ON, Canada) (KDD '25). Association for Computing Machinery, New York, NY, USA, 106–117. <https://doi.org/10.1145/3690624.3709232>
- [5] Victor Coscrato and Derek Bridge. 2023. Estimating and Evaluating the Uncertainty of Rating Predictions and Top-n Recommendations in Recommender Systems. *ACM Trans. Recomm. Syst.* 1, 2, Article 7 (apr 2023), 34 pages. <https://doi.org/10.1145/3584021>
- [6] Yarín Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, New York, New York, USA, 1050–1059. <https://proceedings.mlr.press/v48/gal16.html>
- [7] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. <https://doi.org/10.1145/2827872>
- [8] Norman Knyazev and Harrie Oosterhuis. 2023. A lightweight method for modeling confidence in recommendations with learned beta distributions. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 306–317. <https://doi.org/10.1145/3604915.3608788>
- [9] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37. <https://doi.org/10.1109/MC.2009.263>
- [10] Yehuda Koren and Joe Sill. 2011. OrdRec: an ordinal model for predicting personalized item rating distributions. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) (RecSys '11). Association for Computing Machinery, New York, NY, USA, 117–124. <https://doi.org/10.1145/2043932.2043956>
- [11] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 29. <https://doi.org/10.1609/aaai.v29i1.9602>
- [12] Joel Machado Pires, Eduardo Ferreira da Silva, and Frederico Araújo Durão. 2026. Exploiting distribution-based confidence integration in graph neural network recommenders. *Applied Intelligence* 56, 5 (2026), 142. <https://doi.org/10.1007/s10489-026-07187-0>
- [13] Joseph Lee Rodgers and W. Alan Nicewander. 1988. Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician* 42, 1 (1988), 59–66. <https://doi.org/10.1080/00031305.1988.10475524> arXiv:<https://doi.org/10.1080/00031305.1988.10475524>
- [14] Eduardo Silva, Joel Pires, Denis Dantas, Mayki Oliveira, and Frederico Durão. 2025. A Gated Review Attention Framework for Topics in Graph-Based Recommenders. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 19–27. <https://doi.org/10.5753/webmedia.2025.15515>
- [15] Rodrigo Souza and Marcelo Manzato. 2024. Explorando Formas de Calibração e Redução do Viés de Popularidade em Sistemas de Recomendação. In *Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web* (Juiz de Fora/MG). SBC, Porto Alegre, RS, Brasil, 9–10. https://doi.org/10.5753/webmedia_estendido.2024.244380
- [16] Chao Wang, Qi Liu, Runze Wu, Enhong Chen, Chuanren Liu, Xunpeng Huang, and Zhenya Huang. 2018. Confidence-aware matrix factorization for recommender systems. In *Proceedings of the AAAI Conference on artificial intelligence*, Vol. 32. <https://doi.org/10.1609/aaai.v32i1.11251>
- [17] Longfeng Wu. 2024. Towards Trustworthy Graph Neural Networks and Their Applications in Recommender Systems. In *2024 IEEE International Conference on Big Data (BigData)*. 8250–8252. <https://doi.org/10.1109/BigData62323.2024.10825262>
- [18] Eva Zangerle and Christine Bauer. 2022. Evaluating Recommender Systems: Survey and Framework. *ACM Comput. Surv.* 55, 8, Article 170 (Dec. 2022), 38 pages. <https://doi.org/10.1145/3556536>
- [19] André Zanon, Leonardo Rocha, and Marcelo Manzato. 2025. An exploration of recommender systems explanation paradigms: generating and evaluating syntactic, semantic, and generative models with knowledge graphs: An Extended Abstract. In *Anais Estendidos do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 19–20. https://doi.org/10.5753/webmedia_estendido.2025.16351