On the Calibration of Isotonic Distributional Regression

Tobias Biegert
Karlsruhe Institute of Technology

Johannes Resin
Goethe University Frankfurt

Alexander Jordan
Heidelberg Institute for Theoretical Studies

Sebastian Lerch
Marburg University

Abstract

Isotonic distributional regression (IDR) has recently been proposed as a non-parametric technique for probabilistic forecasting via the estimation of conditional distributions under order restrictions, based on continuous outcomes from deterministic model outputs. IDR has been widely used and multiple studies report promising performance and well-calibrated forecasts across application domains. We document a peculiar out-of-sample artifact of IDR, where probability integral transform (PIT) values are rational fractions with empirical frequencies resembling Thomae's function, an effect which thus far seems to have been overlooked. This phenomenon is demonstrated empirically using meteorological data and a simulation study. We further provide a theoretical explanation linking the spiking behavior to the discrete structure of the isotonic fits, and discuss how this artifact can be mitigated via smoothing approaches.

1 INTRODUCTION

IDR (Henzi et al., 2021) estimates the conditional distribution of a real-valued outcome Y given a predictor X based on training data of the form $(x_i, y_i), i = 1, ..., n$, under a stochastic order constraint. IDR assumes that the conditional distribution of the outcome given by the cumulative distribution function (CDF) $F_x(y) = P(Y \leq y|X = x)$ increases in stochastic order as x increases in partial order, i.e., if $x \preceq x'$ then $F_x(y) \geq F_{x'}(y)$ for all $y \in \mathbb{R}$.

In the following, we will restrict our attention to IDR for totally-ordered univariate model outputs. This special case has been introduced as EasyUQ by Walz et al. (2024a). The appeal of IDR and EasyUQ lies in their in-sample optimality, conceptual simplicity, and lack of tuning parameters. If $x_1 < \cdots < x_n$, the IDR solution

$$\widehat{F}_{x_j}(y) = \min_{k=1,\ldots,j} \max_{l=j,\ldots,n} \frac{1}{l-k+1} \sum_{i=k}^{l} \mathbb{1}\{y_i \leq y\}, \quad (1)$$

is optimal relative to a comprehensive class of proper scoring rules for probabilistic forecasts (Henzi et al., 2021).

IDR and EasyUQ have been used in various studies across application domains, including weather (Henzi et al., 2021; Schulz and Lerch, 2022; Ageet et al., 2023; Rasheeda Satheesh et al., 2023; Walz et al., 2024b; Bülte et al., 2025; Rasheeda Satheesh et al., 2025; Kalita et al., 2026) and energy (Gneiting et al., 2023; Lipiecki et al., 2024). These studies report promising performance, and, in most cases, well-calibrated forecasts obtained via IDR.

It is standard practice to assess the calibration of distributional predictions via a PIT histogram, which should show approximately uniform PIT values (e.g., Dawid, 1984; Diebold et al., 1998). In this note, we demonstrate an interesting artifact in the PIT values of IDR forecasts that appears to have been overlooked, possibly because calibration diagnostics typically rely on PIT histograms with a relatively coarse binning (usually 10–20 bins). When evaluating IDR forecasts, the PIT values show a peculiar pattern with pronounced spikes at simple rational fractions. Upon closer investigation, the empirical frequencies of the PIT values resemble *Thomae's function* (Bartle and Sherbert, 2011), which is defined by $f(x) = 0$ if x is irrational, and $f(x) = 1/q$ if $x = p/q$ is a rational number written as a reduced fraction with $q > 0$.

We provide theoretical background on calibration results for IDR (Section 2), demonstrate the PIT phenomenon in data and simulation examples (Section 3), and provide a theoretical explanation (Section 4). We conclude with a discussion of a remedy via smoothing (Section 5) and practical implications (Section 6).

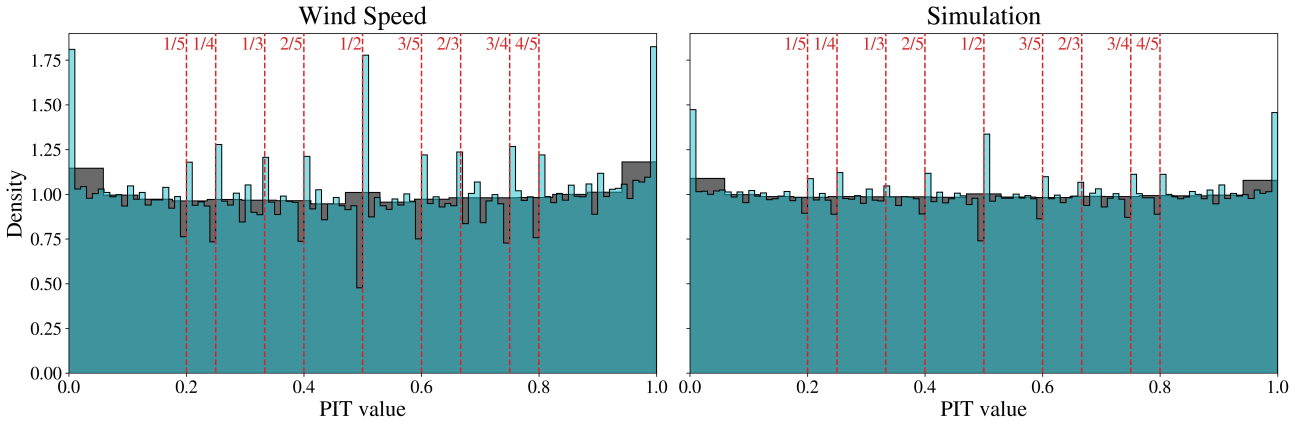

Figure 1: 10-m wind speed (left) and simulation (right) out-of-sample PIT histograms for EasyUQ forecasts with coarse (17-bin; gray) and fine (100-bin; cyan) binning. Coarse binning largely masks systematic over-representation at simple rational PIT values, while finer binning reveals pronounced spikes.

2 BACKGROUND

Following standard practice (Gneiting and Ranjan, 2013), we assess *probabilistic calibration* using the (randomized) PIT,

$$Z_F \;=\; \lim_{y \uparrow Y} F(y) \;+\; V\Big(F(Y) - \lim_{y \uparrow Y} F(y)\Big),$$

where F is a CDF-valued random quantity (i.e., a probabilistic forecast), Y denotes the realized outcome, and $V \sim \mathrm{Unif}(0,1)$ is independent of (F, Y). If F is continuous at Y, then $Z_F = F(Y)$. Probabilistic calibration of F corresponds to $Z_F \sim \mathrm{Unif}(0,1)$, and is commonly assessed using PIT histograms.

The IDR solution (1) consists of discrete distributions $\widehat{F}_{x_j}$ with possible jumps at the unique values $\widetilde{y}_1 < \cdots < \widetilde{y}_k$ of the outcomes $y_1, \ldots, y_n$ in the training set. In-sample (i.e., with respect to the empirical distribution of the training data) the IDR solution enjoys good calibration properties. Henzi et al. (2021, Theorem 2) show that the IDR solution is threshold calibrated. While this does not imply probabilistic calibration (Gneiting and Resin, 2023), IDR satisfies a somewhat weaker version of probabilistic calibration (see Arnold and Ziegel, 2025, Appendix D).

To obtain an out-of-sample predictive distribution for a new value of the predictor x, we follow Henzi et al. (2021) and interpolate the IDR solution (1) linearly, which produces another discrete CDF, $\widehat{F}_x$. Below, we demonstrate that the discreteness of the predictive distributions induces systematic spikes in PIT values at simple rational fractions out-of-sample. Thus, the IDR predictions fail to be probabilistically calibrated in a rather peculiar way. This lack of calibration has thus far gone unnoticed, likely because coarse binnings in PIT histograms can mask the observed spikes.

3 EMPIRICAL EVIDENCE

We analyze global gridded 10-m wind speed forecasts obtained by applying IDR to point predictions from an AI-based weather model, which we evaluate out-of-sample against reanalysis data within the WeatherBench 2 framework (Rasp et al., 2024). All forecasts use a 3-day lead time for 00/12 UTC initialization times. We use data from the years 2018-2019 and 2021-2022 to fit IDR on each grid point individually and evaluate out-of-sample on 2020. The underlying grid has a resolution of 64×32 with 732 test time steps in 2020.

The left panel of Figure 1 shows PIT histograms for the same set of out-of-sample PIT values using a fine and a coarse binning. With the fine binning, PIT values are clearly clustered near simple rational fractions such as 1/2, 1/3, 2/3, 1/4, and 3/4, with additional clusters near 0 and 1, and thus deviate from the expected uniform distribution under probabilistic calibration. In contrast, when the same PIT values are displayed with a coarser binning, as is common in the literature, the spikes are largely hidden, and deviations from uniformity are much less apparent.

To illustrate that the observed spiking behavior is not specific to the data example, we replicate the simulation study of Walz et al. (2024a). In each iteration, we generate $n = 5\,000$ predictor–outcome pairs with $X \sim \mathrm{Unif}(0, 10)$, and

$$Y \mid X \sim \mathrm{Gamma}\big(\text{shape} = \sqrt{X},\ \text{scale} = \sigma(X)\big),$$

where $\sigma(X) = \min\{\max\{X, 2\}, 8\}$. We fit IDR on an 80% training split and compute PIT values on the remaining 20% test split. We repeat this procedure $N = 2\,000$ times to obtain 2 million out-of-sample PIT

values in total. As shown in the right panel of Figure 1, the resulting PIT values exhibit the same qualitative concentration near simple rational fractions as in the weather forecasting example.

Figure 2 shows empirical frequencies of individual PIT values. Here, we observe a distinct pattern resembling Thomae's function, with additional clusters at 0 and 1. The same pattern emerges in the weather application (not shown). The effect becomes clearly visible only when multiple IDR fits are used for prediction (in the simulation, one for each of the N repeats; in the application, one for each grid point). With a single IDR fit, the clustering is noisier due to randomness in the training set and individual fits.

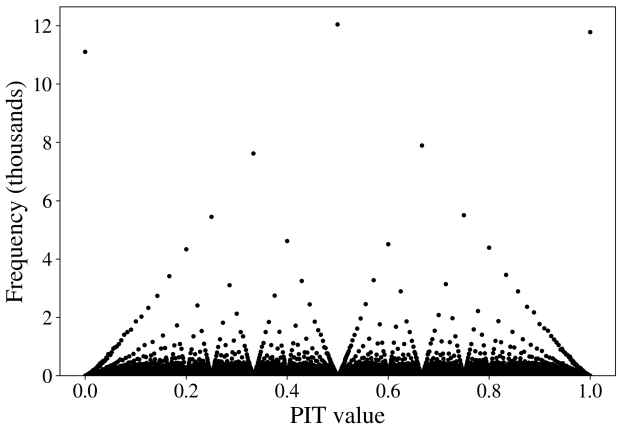

Figure 2: Empirical frequencies of unique PIT values under standard EasyUQ in the simulation study (test set). The distribution exhibits pronounced spikes at simple rational fractions (e.g., 1/2, 1/3, 2/3, 1/4, 3/4), yielding a Thomae's function-like pattern.

4 THEORETICAL CONSIDERATIONS

The spiking behavior can be traced to the pool block structure of isotonic regression. In the univariate predictor case, EasyUQ estimates the conditional CDF $F_x(y) = P(Y \le y \mid X = x)$ by fitting, for each threshold y, an isotonic regression for the binary exceedance response $\mathbb{1}\{y_i \le y\}$ against the predictor values x_i (Henzi et al., 2021; Walz et al., 2024a). For fixed y, isotonic regression pools the training points based on the predictor values x_i according to the PAV (pool-adjacent-violators) algorithm and fits a single value within each pool. If a new predictor value x is within the interval $[x_k, x_l]$ spanned by the predictor values in a pool $B = \{k, k+1, \dots, l\}$ (assuming that the training set is ordered according to the x_i) of size $|B| = l-k+1$, the fitted value equals the pool average,

$$\widehat{F}_x(y) = \frac{1}{|B|} \sum_{i \in B} \mathbb{1}\{y_i \le y\},$$

and therefore any such value is a rational number of the form $r/|B|$. EasyUQ assembles these threshold-wise fits into a discrete conditional CDF $y \mapsto \widehat{F}_x(y)$ that is a step function in y (Walz et al., 2024a).

For a new test pair (x, y), the randomized PIT coincides with the ordinary PIT $\widehat{F}_x(y)$ with probability one under continuity of Y, since the probability of y matching a jump point of the CDF is zero. Hence, when x falls within a pool B of the isotonic regression at threshold y, the PIT value $\widehat{F}_x(y)$ is a fraction of the form $r/|B|$, as above. We argue that the frequency of a reduced fraction p/q as PIT value is proportional to the ratio of counting numbers divisible by the denominator q, under the assumption that the remainder of $|B|$ modulo q is roughly uniformly distributed and r given $|B|$ is approximately uniform across $0, 1, \dots, |B|$. Thus, when considering the test pair (x, y) and training sample $(x_1, y_1), \dots, (x_n, y_n)$ to be iid random pairs with the same distribution to capture the randomness in both IDR fits and test samples, we obtain (for small, coprime positive integers p, q, assuming for simplicity that n is divisible by q)

$$\begin{aligned} P\big(\tfrac{r}{|B|} = \tfrac{p}{q}\big) &= \sum_{i=1}^{n/q} P(|B| = iq)\, P(r = ip \mid |B| = iq) \\ &\approx \sum_{i=1}^{n/q} P(|B| = iq)\, \tfrac{1}{iq+1} =: S_0(q). \end{aligned}$$

The last sum is a partial sum of

$$\begin{aligned} S &= \sum_{i=1}^{n} P(|B| = i)\, \frac{1}{i+1} \\ &= \sum_{j=0}^{q-1} \underbrace{\sum_{i=1}^{n/q} P(|B| = iq - j)\, \frac{1}{iq - j + 1}}_{=: S_j(q)}. \end{aligned}$$

As long as the partial sums are all approximately equal, i.e., $S_j(q) \approx S_0(q)$ for all j, which should be the case as long as the distribution of $|B|$ is sufficiently broad, we have $S \approx qS_0(q)$, and thus empirical frequencies will be roughly proportional to $\frac{1}{q}$, explaining the Thomae's function-like pattern observed in Figure 2.

Note that there is a small (about 1.1% in the simulation) chance of falling in between pools, in which case the PIT value is not a simple fraction (due to interpolation between fitted values). Furthermore, PIT values of zero or one may arise with any pool size or when a new observation is outside the range of training observations ($y < \widetilde{y}_1$ or $y > \widetilde{y}_k$).

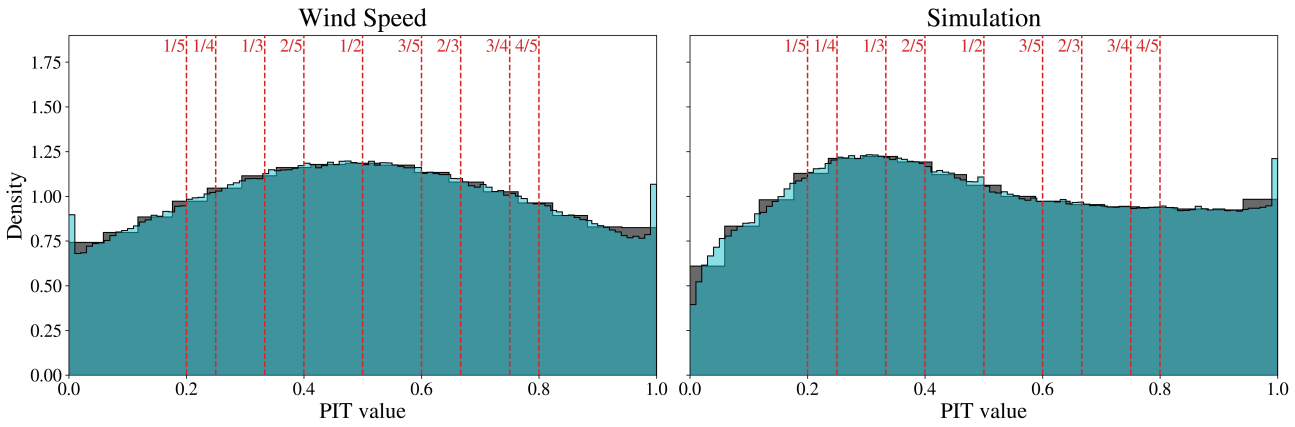

Figure 3: 10-m wind speed (left) and simulation (right) out-of-sample PIT histograms for Smooth EasyUQ forecasts with coarse (17-bin; gray) and fine (100-bin; cyan) binning.

5 SMOOTHING

Walz et al. (2024a) propose the Smooth EasyUQ approach, which supplements IDR with kernel smoothing and yields continuous predictive distributions. We use a Gaussian kernel leading to the smooth predictive CDF

$$\widetilde{F}_x(y) = \sum_{j=1}^{k} w_j(x) \Phi\left(\frac{y - \widetilde{y}_j}{h}\right),$$

where Φ denotes the standard Gaussian CDF, $\widetilde{y}_j$ denotes the jth unique training observation, and the mixture weights correspond to the jumps in the CDF of the IDR solution, $w_j(x) = \widehat{F}_x(\widetilde{y}_j) - \widehat{F}_x(\widetilde{y}_{j-1})$ where $\widehat{F}_x(\widetilde{y}_0) := 0$. The bandwidth h is selected using the one-fit log-score approach of Walz et al. (2024a): after fitting IDR once on the training data, we choose h to minimize the mean log score of the training observations under the smoothed model, using a leave-one-out style correction where the mixture weight $w_j(\cdot)$ corresponding to the left-out observation $y_i = \widetilde{y}_j$ is set to zero and the remaining weights are renormalized. For the weather application, we tune h independently on 250 randomly chosen grid points and use the median of the resulting values as a single global bandwidth applied across all grid points; we proceed analogously in the simulation study.

Empirically, the mean CRPS, a common measure of overall predictive performance (e.g., Gneiting and Raftery, 2007), is almost unchanged under smoothing: 0.4397 (IDR) vs. 0.4395 (smoothed) for the real data, and 4.1591 vs. 4.1654 in the simulation. Figure 3 shows that the smoothing approach effectively mitigates spikes in the PIT histogram. However, smoothing yields some clear deviations from uniformity in the PIT histogram that were not present with the original IDR predictions (cf. Figure 1).

6 CONCLUSIONS

We have documented a peculiar pattern in the PIT values of IDR predictions, which can be observed in both simulated and real data. While this observation does not contradict existing theory on the calibration of IDR, the observed pattern is in contrast to various previous applications of IDR, where well-calibrated forecasts were reported. While deviations of the PIT values from uniformity are to be expected because of the discreteness of IDR predictions, we were surprised by the striking Thomae-like pattern in empirical frequencies. We believe this pattern to be of a theoretical interest in its own right, hinting at interesting regularities in the distribution of the PIT values obtained from IDR.

Clearly, some smoothing is required to obtain appropriate continuous CDF predictions. Walz et al. (2024a) suggest kernel smoothing as in their Smooth EasyUQ approach as a remedy that should be preferred for continuous outcomes. Yet, many practitioners appear to still use the standard (unsmoothed) IDR. Smooth EasyUQ requires some tuning via the selection of a bandwidth (and kernel) and appears to introduce some deviations from probabilistic calibration of its own that cannot be traced to the underlying IDR solution. We also experimented with evaluating PIT values after replacing the piecewise-constant predictive CDF with a simple piecewise-linear interpolation, which effectively removed the Thomae-like pattern. Except for accumulation points at zero and one due to missing tail extrapolation, this approach produced continuous PIT values resulting in roughly uniform looking PIT histograms in our simulation. It remains an open problem how to smooth IDR in a way that produces well-calibrated predictions.

Acknowledgements

We thank Tilmann Gneiting, Kristof Kraus, and Johanna Ziegel for insightful comments and fruitful discussion. TB gratefully acknowledges support by the German Weather Service (Deutscher Wetterdienst) through the SPARC-ML project within the extramural research program, funding reference number 4823EMF01. JR gratefully acknowledges support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via project number 502572912. AJ is grateful for the generous support of the Klaus Tschira Foundation. SL gratefully acknowledges support by the Vector Stiftung through the Young Investigator Group "Artificial Intelligence for Probabilistic Weather Forecasting".

Code and Data Availability

Code for replicating the simulation study and the empirical analysis is available at `https://github.com/tobiasbiegert/idr-calibration`. WeatherBench 2 data are freely accessible; see Rasp et al. (2024).

References

Ageet, S., Fink, A. H., Maranan, M., and Schulz, B. (2023). Predictability of Rainfall over Equatorial East Africa in the ECMWF Ensemble Reforecasts on Short- to Medium-Range Time Scales. *Weather and Forecasting*, 38(12):2613–2630.

Arnold, S. and Ziegel, J. F. (2025). Isotonic conditional laws. *Bernoulli*, 31(2):1140–1159.

Bartle, R. G. and Sherbert, D. R. (2011). *Introduction to Real Analysis*. John Wiley & Sons, 4th edition.

Bülte, C., Horat, N., Quinting, J., and Lerch, S. (2025). Uncertainty quantification for data-driven weather models. *Artificial Intelligence for the Earth Systems*. In press.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society Series A*, 147(2):278–292.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review*, 39(4):863–883.

Gneiting, T., Lerch, S., and Schulz, B. (2023). Probabilistic solar forecasting: Benchmarks, postprocessing, verification. *Solar Energy*, 252:72–80.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782.

Gneiting, T. and Resin, J. (2023). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, 17(2):3226–3286.

Henzi, A., Ziegel, J. F., and Gneiting, T. (2021). Isotonic Distributional Regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):963–993.

Kalita, I., Vilallonga, L., and Atchade, Y. (2026). Data-Driven Rainfall Prediction at a Regional Scale: A Case Study with Ghana. *Artificial Intelligence for the Earth Systems*, 5(1):240116.

Lipiecki, A., Uniejewski, B., and Weron, R. (2024). Postprocessing of point predictions for probabilistic forecasting of day-ahead electricity prices: The benefits of using isotonic distributional regression. *Energy Economics*, 139:107934.

Rasheeda Satheesh, A., Knippertz, P., and Fink, A. H. (2025). Machine Learning Models for Daily Rainfall Forecasting in Northern Tropical Africa Using Tropical Wave Predictors. *Weather and Forecasting*, 40(10):1895–1916.

Rasheeda Satheesh, A., Knippertz, P., Fink, A. H., Walz, E.-M., and Gneiting, T. (2023). Sources of predictability of synoptic-scale rainfall during the West African summer monsoon. *Quarterly Journal of the Royal Meteorological Society*, 149(757):3721–3737.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallegue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F. (2024). WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019.

Schulz, B. and Lerch, S. (2022). Machine Learning Methods for Postprocessing Ensemble Forecasts of Wind Gusts: A Systematic Comparison. *Monthly Weather Review*, 150(1):235–257.

Walz, E.-M., Henzi, A., Ziegel, J. F., and Gneiting, T. (2024a). Easy Uncertainty Quantification (EasyUQ): Generating Predictive Distributions from Single-Valued Model Output. *SIAM Review*, 66(1):91–122.

Walz, E.-M., Knippertz, P., Fink, A. H., Köhler, G., and Gneiting, T. (2024b). Physics-Based vs Data-Driven 24-Hour Probabilistic Forecasts of Precipitation for Northern Tropical Africa. *Monthly Weather Review*, 152(9):2011–2031.