PAPER • **OPEN ACCESS**

Towards instance-wise calibration: local amortized diagnostics and reshaping of conditional densities (LADaR)

To cite this article: Biprateep Dey *et al* 2025 *Mach. Learn.: Sci. Technol.* **6** 045058

IOP Publishing Mach. Learn.: Sci. Technol. 6 (2025) 045058 https://doi.org/10.1088/2632-2153/ae1f05

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
24 December 2024

REVISED
20 October 2025

ACCEPTED FOR PUBLICATION
11 November 2025

PUBLISHED
2 December 2025

PAPER

Towards instance-wise calibration: local amortized diagnostics and reshaping of conditional densities (LADaR)

Biprateep Dey[1,2,3,4,5,6], **David Zhao**[7], **Brett H Andrews**[1,2], **Jeffrey A Newman**[1,2], **Rafael Izbicki**[8] and **Ann B Lee**[7,*]

[1] Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA, United States of America
[2] Pittsburgh Particle Physics, Astrophysics, and Cosmology Center (PITT PACC), University of Pittsburgh, Pittsburgh, PA, United States of America
[3] Department of Statistical Sciences, University of Toronto, Toronto, Canada
[4] Canadian Institute for Theoretical Astrophysics (CITA), University of Toronto, Toronto, Canada
[5] Dunlap Institute for Astronomy & Astrophysics, University of Toronto, Toronto, Canada
[6] Vector Institute, Toronto, Canada
[7] Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, United States of America
[8] Department of Statistics, Federal University of São Carlos (UFSCar), São Carlos, Brazil
[*] Author to whom any correspondence should be addressed.

E-mail: annlee@andrew.cmu.edu

Keywords: conditional density estimation, local coverage diagnostics, calibrated distributions, reliable uncertainty quantification, posterior approximations

Abstract

Key science questions, such as galaxy distance estimation and weather forecasting, often require knowing the full predictive distribution of a target variable Y given complex inputs $\mathbf{X}$. Despite recent advances in machine learning and physics-based models, it remains challenging to assess whether an initial model is calibrated for all $\mathbf{x}$, and when needed, to reshape the densities of y toward 'instance-wise' calibration. This paper introduces the local amortized diagnostics and reshaping of conditional densities (LADaR) framework and proposes a new computationally efficient algorithm (`Cal-PIT`) that produces interpretable local diagnostics and provides a mechanism for adjusting conditional density estimates (CDEs). `Cal-PIT` learns a single interpretable local probability–probability map from calibration data that identifies where and how the initial model is miscalibrated across feature space, which can be used to morph CDEs such that they are well-calibrated. We illustrate the LADaR framework on synthetic examples, including probabilistic forecasting from image sequences, akin to predicting storm wind speed from satellite imagery. Our main science application involves estimating the probability density functions of galaxy distances given photometric data, where `Cal-PIT` achieves better instance-wise calibration than all 11 other literature methods in a benchmark data challenge, demonstrating its utility for next-generation cosmological analyzes[9].

1. Introduction

In recent decades, many scientific fields have progressed from computing point predictions (or a single best guess of a quantity of interest) to developing full predictive distributions, or more specifically, *conditional density estimates (CDEs) and generative models* of a response/'target' variable $Y \in \mathbb{R}$ given covariates/features $\mathbf{X} \in \mathbb{R}^d$. This paradigm shift is evident in various disciplines, such as in astrophysics (e.g. Mandelbaum *et al* 2008, Malz and Hogg 2022), in weather forecasting (e.g. Gneiting 2008, Ravuri *et al* 2021, Li *et al* 2024), in financial risk management (e.g. Timmermann 2000), and in epidemiological projections (e.g. Alkema *et al* 2007).

[9] Code available as a Python package here: https://github.com/lee-group-cmu/Cal-PIT.

The paradigm shift has been driven by two main factors. First, advances in measurement technology across engineering, physical and biological sciences are producing data with unprecedented depth, richness, and scope. To fully exploit these data in subsequent analyzes, we need precise estimates of the uncertainty in a response variable Y given observable data $\mathbf{X}$ (see section 1.2 for two applications from the physical sciences that motivated this work). Second, we are experiencing a rapid growth of high-capacity machine learning algorithms that allow the quantification of uncertainty for complex, high-dimensional inputs of different modalities. Two examples of such data sets come from (1) large astronomical surveys that collect images and spectroscopic data for tens of millions of stars, galaxies and other astrophysical objects (York *et al* 2000, Gaia Collaboration *et al* 2016, Dey *et al* 2019, DESI Collaboration *et al* 2022) and (2) earth-observing satellites for environmental and climate science (see, e.g. NASA's Earth Observing System[10] and next-generation Earth System Observatory[11]). For the latter, the dimension d of the input space (representing, e.g. the number of image pixels or different spatial locations) is usually several orders of magnitude larger than 10^6. In addition to enabling uncertainty quantification (UQ) for complex data, modern machine learning methods allow us to 'amortize' the computation; that is, to perform the compute-intensive training process only once, which allows for very fast inference and scaling to massive data sets.

Machine learning methods for UQ include a growing range of approaches. Explicit CDE methods directly model $f(y|\mathbf{x})$, using tools like mixture density networks (Bishop 1994), kernel mixture networks (Ambrogioni *et al* 2017), normalizing flows (Papamakarios *et al* 2019, Kobyzev *et al* 2021), and other nonparametric estimators (Izbicki and Lee 2016, 2017, Dalmasso *et al* 2020). Implicit CDEs and generative models—such as VAEs (Kingma and Welling 2013), conditional GANs (Mirza and Osindero 2014), diffusion models (Sohl-Dickstein *et al* 2015, Ho *et al* 2020, Dhariwal and Nichol 2021, Nichol and Dhariwal 2021, Ho and Salimans 2022), and transformer-based generators (Vaswani *et al* 2017, Radford *et al* 2019)—represent uncertainty through learned stochastic mappings. Other strategies include quantile regression (QR) (Amerise 2018, Fasiolo *et al* 2021, Feldman *et al* 2021, Lim *et al* 2021, Chung *et al* 2021a) and ensemble-based methods, such as dropout and deep ensembles (Gal and Ghahramani 2016, Lakshminarayanan *et al* 2017, Rahaman *et al* 2021).

The goal of this paper is not to add to this list, but rather to provide the scientist with a unified interpretable framework for deciding whether an initial model of the predictive distribution is accurate with respect to (conditional on) relevant features, and if not, suggest a mechanism for reshaping CDEs. Figure 1 shows a schematic diagram of our local amortized diagnostics and Reshaping (LADaR) approach. The starting point is an initial CDE—which could, e.g. be derived from pre-trained neural networks on massive generic data (so-called foundation models) or physics-based models such as numerical weather prediction models. LADaR addresses three key questions: (1) Does the initial model need to be improved with respect to relevant features? (2) Where in the feature space might it need to be improved? (3) How can it be improved? For the third question, we propose a reshaping step that adjusts the initial CDEs while leveraging its existing strengths. LADaR is particularly relevant when there are insufficient observational data to independently fit a purely ML-based CDE model, or when the scientist needs to tie results to physical processes in the native feature space (defined by, e.g. individual spectra or specific sequences of satellite imagery) to trust predictions and stated uncertainties.

1.1. Trustworthy UQ

For a conditional density estimator to be useful, its predicted distribution $\widehat{F}(y|\mathbf{x})$ (with density function $\widehat{f}(y|\mathbf{x})$) must closely match the true $F(y|\mathbf{x})$ for each value of the input $\mathbf{x}$. This property, known as *local or instance-wise calibration*, ensures that predicted probabilities reflect true frequencies for individual cases, and not just on average.

Instance-wise UQ is essential in many practical applications; e.g. in astrophysical studies, for predicting the physical properties of individual galaxies from measured fluxes; in weather forecasts, for predicting the probability of rainfall based on current environmental conditions; and in medical research, for estimating a drug's efficacy for individuals of specific demographics. Instance-wise calibration also promotes algorithmic fairness by avoiding systematic over- or under-prediction of risks for certain groups (Kleinberg *et al* 2016, Zhao *et al* 2020), and enables well-calibrated prediction sets (remark 2).

Unfortunately, off-the-shelf CDE methods can be very far from calibrated. This is because they minimize losses that do not target calibration directly—such as KL divergence (Kullback and Leibler 1951), integral probability metrics (Papamakarios *et al* 2019, Dalmasso *et al* 2020), or the pinball loss (Koenker and Hallock 2001). As shown by Guo *et al* (2017) and Chung *et al* (2021b), many ML methods

[10] https://eospso.nasa.gov/.

[11] https://science.nasa.gov/earth-science/missions/earth-system-observatory/.

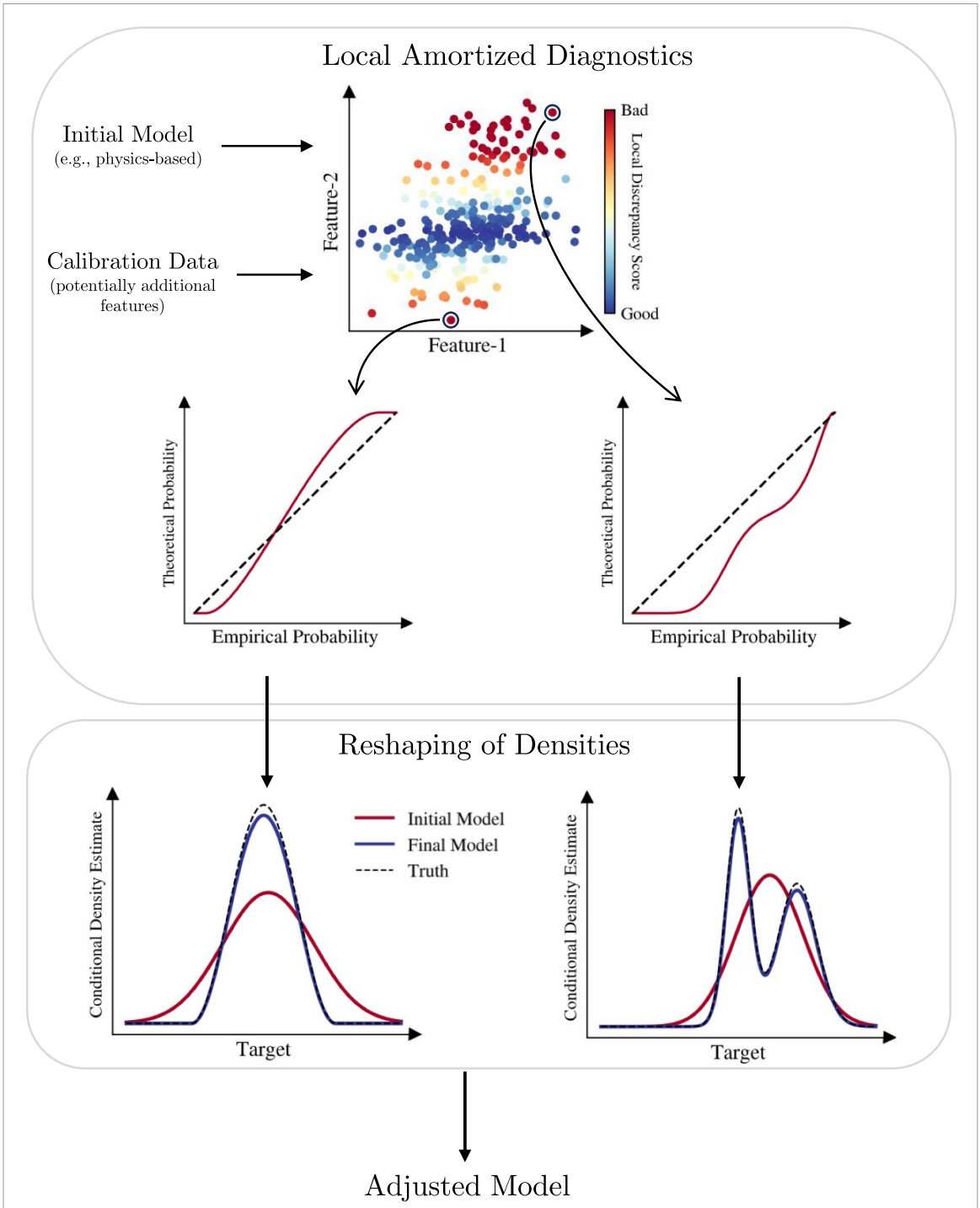


Figure 1. Schematic representation of the LADaR approach. Our approach starts with an initial (e.g. physics-based or large pre-trained) model of the predictive distribution of a target quantity. We then assess the quality of the initial conditional density estimates (CDEs) on an individual basis across the feature space using calibration data, and reshape the densities if deemed necessary. The goal is not to replace the initial model with a different end-to-end density estimator, but rather to adjust it, ensuring both calibration and insight into its potential failure modes (see figure 2 for how to interpret P–P plots). The LADaR approach is particularly relevant when there are insufficient observational data to independently fit a purely machine-learning-based CDE, or when it is important to tie predictions to the underlying physical processes (encoded by the chosen feature space) to establish trust in machine-learning methods. Our framework is fully 'amortized' over both features $\mathbf{x}$ and response variable y, which means that once we have trained LADaR to learn the map between the initial CDE model and the CDE of the calibration data, no additional training is required for new data.

prioritize accuracy and sharpness over calibration. To address this, new loss functions have been proposed to balance calibration and sharpness (Chung *et al* 2021b) or decouple coverage from sharpness (Feldman *et al* 2021).

Finally, in terms of *diagnostics*, many common metrics for assessing calibration, like the probability integral transform (PIT; Gan and Koehler 1990) and simulator-based calibration (SBC; Talts *et al* 2018), only evaluate *marginal* calibration—that is, average coverage over all $\mathbf{x}$'s (equation (2)). This weaker

notion is often referred to simply as 'calibration' (Gneiting and Katzfuss 2014, Kuleshov *et al* 2018). However, as pointed out by Schmidt *et al* (2020), PIT can be optimal even when the model ignores $\mathbf{x}$ entirely. More generally, errors across the feature space can cancel out, leading to deceptively good marginal results (Jitkrittum *et al* 2020a, Luo *et al* 2021, Zhao *et al* 2021). For instance, Theorem 1 in Zhao *et al* (2021) show that even models based on $F(y|g(\mathbf{x}))$—for any function g— can pass marginal tests, despite discarding relevant features.

1.2. Well-calibrated CDEs are essential for the physical sciences

Our trustworthy CDE work is motivated by two main applications in astronomy and weather forecasting:

(i) Photometric redshift estimation of galaxies. Estimating galaxy distances, via a measurable proxy called redshift, is a fundamental task for studies of astrophysics and cosmology. While spectroscopy can precisely measure redshift, this method is too resource-intensive for the billions of galaxies detected by modern astrophysical imaging surveys, so galaxy redshifts must be predicted from imaging data alone. In this context, the response variable y is the galaxy's redshift (by convention denoted by z), and the predictors are photometric/imaging data $\mathbf{x}$. The predictions, called photometric redshifts (photo-z's), are inherently probabilistic. Downstream science applications rely on an accurate estimate of the conditional density for each galaxy's redshift. The scientific requirements are extremely strict: to avoid biasing cosmological results, the errors in the moments of the redshift distributions for large ensembles of galaxies must be within 0.1%–0.3% of the truth (The LSST Dark Energy Science Collaboration *et al* 2018).

Our proposed photo-z use case is to adopt a physics-based photo-z model to produce initial estimates of PDFs, and then use the LADaR framework to assess the initial CDEs and reshape them if necessary. Furthermore, the interpretability of the LADaR diagnostics will be valuable for helping astrophysicists improve both physics and machine learning-based photo-z models.

(ii) Probabilistic forecasting of the intensity of tropical cyclones (TCs) from satellite imagery. TCs are highly organized rotating storms that are among the most costly natural disasters in the United States. TC intensity forecasts have improved in recent years, but these improvements have been relatively slow during the last decade compared to improvements in track forecasts, particularly at 24 h lead times (DeMaria *et al* 2014). The latest generation of geostationary satellites (GOESs), such as GOES-16, now provides unprecedented spatio-temporal resolution of TC structure and evolution (Schmit *et al* 2017). A broad range of recent work involving neural networks has explored the wealth of information from GOES imagery for TC short-term intensity prediction (e.g. Olander *et al* 2021, Griffin *et al* 2022). In this context, the response variable Y is the TC's intensity (wind speed) at time $t+\tau$ for a lead time τ of up to 24 h, and $\mathbf{x}$ could represent environmental predictors and a sequence of images at the current time t and preceding time points. In section 4.2, we present a TC-inspired synthetic example that highlights the efficacy of our method in diagnosing and recalibrating intensity forecasts with high-dimensional sequence data as inputs.

1.3. Our contribution

To ensure reliable UQ with CDEs, it is essential to have (i) interpretable diagnostics that can assess instance-wise calibration and failure modes of an initial model across the entire feature space of reference data, and (ii) computationally efficient methods that can reshape CDEs so that they are approximately calibrated for every $\mathbf{x}$. The initial model can, for example, represent the best approximation to the true conditional density according to a physics-motivated or a mathematical model, or from a data-driven model limited to a set of easily accessible input features or data sources.

The goal is to morph the initial model towards the true distribution of the quantity of interest by leveraging calibration data and machine-learning techniques, when such an adjustment is deemed to be necessary by the scientist. This work offers two primary contributions:

- From a methodological perspective, we present a unified framework for interpretable diagnostics and reshaping of entire CDEs through a single probability–probability (P–P) map learned from calibration data $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, which implicitly encodes the true distribution $F(y|\mathbf{x})$. Our approach is fully 'amortized', which means that once a regression model is trained to learn the mapping, no additional training is required for new data. We refer to the general framework of LADaR of CDEs as LADaR, and call our proposed algorithm `Cal-PIT`. The first prototype code of `Cal-PIT` occurred in Dey *et al* (2021); the full ready-to-use and modifiable implementation is now available as a Python package at https://github.com/lee-group-cmu/Cal-PIT.

- From an application perspective, `Cal-PIT` is uniquely positioned to provide diagnostics and ensure that photo-z CDEs are *locally* calibrated (i.e. not only as a full ensemble), which will be necessary for the astrophysics community to meet the stringent photo-z requirements for next generation-astronomical surveys. Figure 7 and table 1 demonstrate the full potential of `Cal-PIT` applied to a key benchmark photo-z data set, where `Cal-PIT` outperforms the current state-of-the-art for diagnostics and estimation of photo-z CDEs. Crucially, `Cal-PIT` can (i) accurately reshape biased probability distributions and (ii) reshape unimodal distributions into multimodal distributions—both common failure modes for common photo-z estimation methods. Furthermore, `Cal-PIT` has the flexibility to be used with *high-dimensional* and *dependent sequence* data. Section 4.2 shows `Cal-PIT` applied to probabilistic forecasting with sequences of images as inputs, akin to predicting the wind speed of TCs from satellite imagery.

2. Related work

Goodness-of-fit tests and calibration. Goodness-of-fit of conditional density models to observed data can be assessed by two-sample tests (e.g. Andrews 1997, Stute and Zhu 2002, Moreira 2003, Jitkrittum *et al* 2020b). Such tests are useful for deciding whether a conditional density model needs to be improved, but do not provide any means to correct discrepancies. One way to recalibrate CDEs (proposed, e.g. by Bordoloi *et al* 2010) is to first assess how the marginal distribution of PIT values differs from a uniform distribution by diagnostic tools (Cook *et al* 2006, Freeman *et al* 2017, D'Isanto and Polsterer 2018, Talts *et al* 2018), and then apply corrections to bring them into agreement. However, by construction, such recalibration schemes only improve marginal calibration. In this work, we instead build on Zhao *et al* (2021), which proposes a version of PIT that is estimated throughout the *entire* input feature space, allowing us to directly assess and target conditional coverage.

QR. QR intervals converge to the oracle $C^*_\alpha(\mathbf{X}) = \left[F^{-1}(0.5\alpha|\mathbf{X}), F^{-1}(1-0.5\alpha|\mathbf{X})\right]$ (Koenker and Bassett 1978, Taylor and Bunn 1999). Even though the prediction interval $C^*_\alpha(\mathbf{X})$ satisfies conditional validity,

$$\mathbb{P}\left(Y \in C_\alpha\left(\mathbf{X}\right) | \mathbf{X} = \mathbf{x}\right) = 1-\alpha, \quad \forall \mathbf{x} \in \mathcal{X},$$

the standard pinball loss can yield highly miscalibrated UQ models for finite data sets (Feldman *et al* 2021, Chung *et al* 2021b). New loss functions have been proposed to address this problem (Feldman *et al* 2021, Chung *et al* 2021b). Our approach also provides calibrated prediction regions, but is more general, yielding full CDEs and not only prediction intervals.

Conformal inference. Conformal prediction methods have the appealing property of producing prediction sets with finite-sample marginal validity, $\mathbb{P}(Y \in C(\mathbf{X})) \geqslant 1-\alpha$, as long as the data are exchangeable (Vovk *et al* 2005, Lei *et al* 2018). However, there is no guarantee that conditional validity is satisfied, even approximately. More recent efforts have addressed approximate conditional validity (Romano *et al* 2019, Izbicki *et al* 2020, 2022, Chernozhukov *et al* 2021, Cabezas *et al* 2025) by designing conformal scores with an approximately homogeneous distribution throughout $\mathcal{X}$. Unfortunately, it is difficult to check whether these methods provide good conditional coverage in practice. Our method, on the other hand, provides estimates of the full CDE, and not only prediction bands. Finally, calibrated CDEs imply calibrated prediction bands, but not vice versa.

3. Methodology

Objective and notation. Our LADaR goal is to reshape an initial (often simple) cumulative distribution $\hat{F}(y|\mathbf{x})$, or equivalently, its conditional density $\hat{f}(y|\mathbf{x})$, to achieve approximate instance-wise calibration with respect to some implicit (often more complex but not explicitly known) target distribution $F(y|\mathbf{x})$. Instance-wise calibration is defined as

$$\hat{F}\left(y|\mathbf{x}\right) = F\left(y|\mathbf{x}\right), \quad \text{for all } y, \text{at every } \mathbf{x}, \tag{1}$$

and is sometimes also referred to as conditional or local calibration. Instance-wise or conditional calibration implies marginal calibration

$$\hat{F}\left(y\right) = F\left(y\right), \quad \text{for all } y, \tag{2}$$

whereas the reverse implication is not true.

To achieve instance-wise calibration, we assume the availability of i.i.d. calibration data $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ from $F_{\mathbf{X},Y}(\mathbf{x}, y)$, the joint distribution of $(\mathbf{X}, Y)$, where $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ and $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}$. We assume that the joint distribution is a product $F_{\mathbf{X},Y}(\mathbf{x}, y) = F(y|\mathbf{x})F(\mathbf{x})$ of the target distribution $F(y|\mathbf{x})$ and some sampling distribution $F(\mathbf{x})$ with support over the entire feature space $\mathcal{X}$.

In this paper, we propose a solution to the problem of diagnosing and ensuring local calibration of conditional densities based on PITs. We refer to the algorithm and the code as `Cal-PIT`. The details are as follows.

3.1. Overview of the `Cal-PIT` algorithm

The `Cal-PIT` algorithm first computes interpretable diagnostics using regression that identifies the failure modes of the initial conditional density model and pinpoints the location of poorly calibrated instances in a potentially high-dimensional feature space. The same regression function used for diagnostics is then used to continuously transform the potentially misspecified CDE into a new CDE that is approximately calibrated for all $\mathbf{x}$.

`Cal-PIT` builds on the observation that an estimate of a cumulative distribution function (CDF), $\widehat{F}$, is calibrated for every instance $\mathbf{x}$, if and only if its PIT value conditional on $\mathbf{x}$, defined by $\mathrm{PIT}(Y;\mathbf{X}) := \hat{F}(Y|\mathbf{X})$, where $(\mathbf{X}, Y)$ is drawn from $F_{\mathbf{X},Y}$, is uniformly distributed (Zhao *et al* 2021, corollary 1). As a result, if a CDE is well-calibrated, the cumulative distribution function of the PIT (hereafter PIT-CDF), defined as the cumulative distribution of the PIT random variable evaluated at $\gamma \in (0, 1)$,

$$r^{\hat{f}}(\gamma;\mathbf{x}) := \mathbb{P}\left(\mathrm{PIT}\left(Y;\mathbf{X}\right) \leqslant \gamma \mid \mathbf{x}\right), \tag{3}$$

will be approximately γ for all $\mathbf{x} \in \mathcal{X}$ and $\gamma \in (0, 1)$. In other words, the PIT-CDF will then correspond to the CDF of a uniform random variable for all $\mathbf{x}$. The PIT-CDF provides valuable information as to whether $\hat{F}$ is miscalibrated, and if so, for what instances $\mathbf{x}$, for what types of deviations and to what extent. Specifically, local P–P plots—which graph the PIT-CDF value $r^{\hat{f}}(\gamma;\mathbf{x})$, the empirical probability, against γ, the theoretically expected probability, for fixed $\mathbf{x}$—offer valuable information on how close the probability distribution $\hat{F}(Y|\mathbf{X})$ is to $F(Y|\mathbf{X})$ at different locations $X = \mathbf{x}$ in the feature space. Figure 2 shows a schematic diagram of some P–P plots and how to interpret them.

However, in practice, because the distribution of the PIT statistic depends on the true conditional distribution of $Y|\mathbf{x}$, the PIT-CDF is unknown. Section 3.2 describes how one can estimate the PIT-CDF across the feature space from calibration data using a regression method suitable for the problem at hand. Our proposed approach is *amortized*, in the sense that one can train on $\mathbf{x}$ and γ jointly, after which the function PIT-CDF can be evaluated for any new values of $\mathbf{x}$ and γ. Finally, section 3.3 describes how the learnt PIT-CDF itself defines a push-forward map (equation (5)) that reshapes the densities so as to achieve approximate local calibration. Algorithm 1 summarizes the details of the `Cal-PIT` method.

3.2. Estimating the PIT-CDF

We observe that the PIT-CDF in equation (3) is the regression (conditional mean) of a binary random variable $W^\gamma := \mathbb{I}(\mathrm{PIT}(Y;\mathbf{X}) < \gamma)$ on $\mathbf{X}$; that is,

$$r^{\hat{f}}(\gamma;\mathbf{x}) = \mathbb{E}\left[W^\gamma \mid \mathbf{x}\right], \tag{4}$$

where the expectation $\mathbb{E}[\cdot \mid \mathbf{x}]$ denotes an average with respect to the (unknown) target distribution $F(y|\mathbf{x})$. The above expression indicates that we can estimate the PIT-CDF across the entire feature space $\mathcal{X}$, as well as for different quantiles $\gamma \in (0, 1)$, via regression methods.

`Cal-PIT` is implemented as follows: first, we augment the calibration data $\mathcal{D}$ by drawing multiple quantile values $\gamma_{i,1}, \ldots, \gamma_{i,K} \sim U(0, 1)$ for each calibration data point $(i = 1, \ldots, n)$ and some chosen hyperparameter K. Next, we define the random variable

$$W_{i,j} := \mathbb{I}\left(\mathrm{PIT}\left(Y_i;\mathbf{X}_i\right) \leqslant \gamma_{i,j}\right).$$

Finally, we train a suitable regression method using the augmented calibration sample $\mathcal{D}' = \{(\mathbf{X}_i, \gamma_{i,j}, W_{i,j})\}_{i,j}$ to predict $W_{i,j}$ with $(\mathbf{X}_i, \gamma_{i,j})$ as inputs, for $i = 1, \ldots, n$ and $j = 1, \ldots, K$. The computed regression function is an estimate of $\mathbb{P}\left(\mathrm{PIT}(Y;\mathbf{X}) \leqslant \gamma \mid \mathbf{x}\right)$. Since $r^{\hat{f}}(\gamma;\mathbf{x})$ is a non-decreasing function of γ, we typically choose monotonic neural networks (Wehenkel and Louppe 2019) minimizing the binary cross-entropy loss (Good 1952) as our regression method, especially for applications with complex and

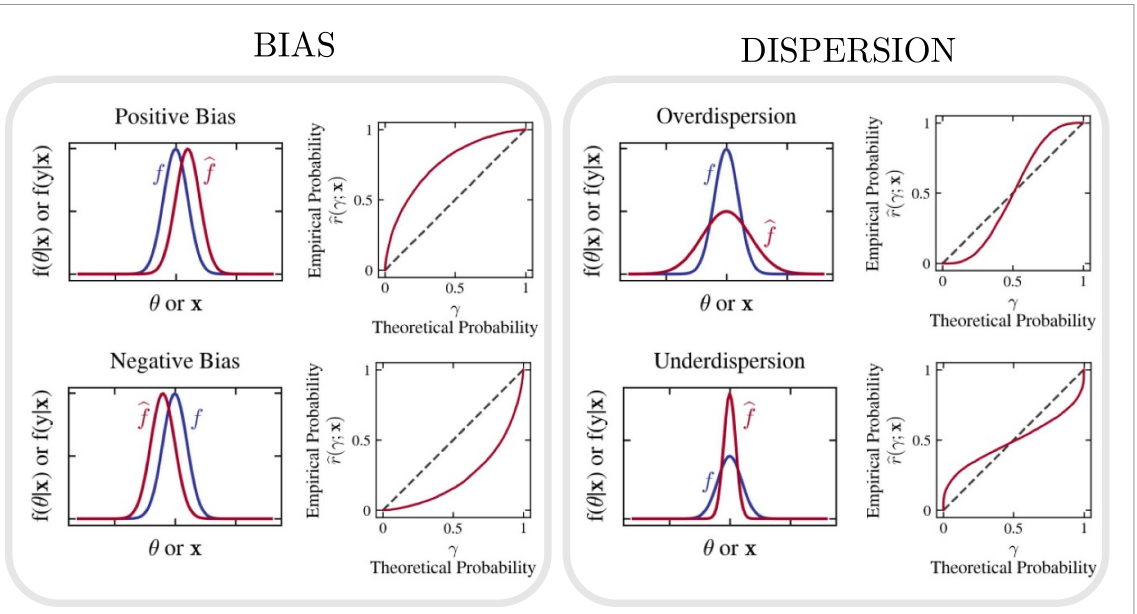


Figure 2. Interpretable diagnostics. P–P plots are commonly used to assess how well a probability density model fits actual data. Such plots display, in a clear and interpretable way, effects like *bias* (*left panel*) and *dispersion* (*right panel*) in an estimated distribution $\widehat{f}$ vis-a-vis the true data-generating distribution f. Our framework yields an amortized approach to constructing local P–P plots for comparing Bayesian posteriors $\widehat{f}(\theta|\mathbf{x})$ or predictive densities $\widehat{f}(y|\mathbf{x})$ at *any* location $\mathbf{x}$ of the feature space $\mathcal{X}$. Reproduced from Zhao *et al* 2021. CC BY 4.0 . An interactive version of this figure can be found at: https://lee-group-cmu.github.io/cal-pit-paper/fig_1_interactive/.

high-dimensional inputs of different modality. This loss function enforces $\mathbb{P}(\mathrm{PIT}(Y;\mathbf{X}) \leqslant \gamma \mid \mathbf{x})$ to be well estimated (Dawid and Musio 2014).

3.3. Reshaping conditional densities by mapping probabilities to probabilities

`Cal-PIT` uses the estimated PIT-CDF to reshape the initial CDE $\widehat{f}$ into a new CDE $\widetilde{f}$ that is approximately locally consistent across the *entire* feature space.

Our procedure for morphing one probability density into a new 'recalibrated' density works as follows: Consider a fixed evaluation point $\mathbf{x}$ and any $y_0 \in \mathcal{Y}$. Let $\gamma := \hat{F}(y_0|\mathbf{x})$. If the regression is perfectly estimated (that is, $\widehat{r}^{\widehat{f}} = r^{\widehat{f}}$), then, as long as both F and $\widehat{F}$ are continuous and $\widehat{F}$ dominates F (see assumptions 1 and 2 in appendix E for details), it holds that

$$r^{\widehat{f}}(\gamma;\mathbf{x}) := \mathbb{P}\left(\hat{F}(Y|\mathbf{x}) \leqslant \gamma \mid \mathbf{x}\right) = \mathbb{P}(Y \leqslant y_0 \mid \mathbf{x}) = F(y_0|\mathbf{x})\,.$$

In other words, the regression function $r^{\widehat{f}}$ changes the initial CDE so that the probability of observing the response variable Y below y_0 is now indeed $F(y_0|\mathbf{x})$ rather than $\hat{F}(y_0|\mathbf{x})$.

It follows directly that for fixed $\hat{F}$,

$$r^{\hat{f}}\left(\hat{F}(y|\mathbf{x})\,;\mathbf{x}\right) = \mathbb{P}\left(\hat{F}(Y|\mathbf{x}) \leqslant \hat{F}(y|\mathbf{x}) \mid \mathbf{x}\right) = \mathbb{P}(Y \leqslant y \mid \mathbf{x}) = F(y|\mathbf{x})$$

The above result suggests that we can use the estimated regression, $\hat{r}^{\hat{f}}$, which is an approximation of the PIT-CDF, $r^{\hat{f}}$, to transform the original distribution $\hat{F}$ with density $\hat{f}$ into a new 'recalibrated' conditional distribution $\widetilde{F}$ with density $\widetilde{f}$:

Definition 1 (recalibrated CDE). The recalibrated CDE of Y given $\mathbf{x}$ is defined through the P–P map,

$$\widetilde{F}(y|\mathbf{x}) := \hat{r}^{\hat{f}}\left(\hat{F}(y|\mathbf{x})\,;\mathbf{x}\right), \tag{5}$$

where $\hat{r}^{\hat{f}}$ is the regression estimator of the PIT-CDF (equation (3)).

If the PIT-CDF is well-estimated, then the new CDE will achieve instance-wise calibration. The next theorem shows that, under some assumptions, we can directly relate the quality of the recalibration (or how close the 'recalibrated' distribution is to the target distribution in a mean-squared-error sense) to the mean-squared-error of the regression estimator:

Algorithm 1. `Cal-PIT`.

Require: initial CDE $\hat{f}(y|\mathbf{x})$ evaluated at $y \in G$; calibration set $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$; oversampling factor K; evaluation points $\mathcal{V} \subset \mathcal{X}$; nominal miscoverage level α, flag `HPD` (true if computing HPD sets)
Ensure: new distribution $\widetilde{F}(y|\mathbf{x})$, `Cal-PIT` interval $C(\mathbf{x})$, new density estimate $\widetilde{f}(y|\mathbf{x})$, for all $\mathbf{x} \in \mathcal{V}$

// **Learn PIT-CDF from augmented and upsampled calibration data** $\mathcal{D}'$
Set $\mathcal{D}' \leftarrow \emptyset$
for i in $\{1, \ldots, n\}$ **do**
  for j in $\{1, \ldots, K\}$ **do**
    Draw $\gamma_{i,j} \sim U(0,1)$
    Compute $W_{i,j} \leftarrow \mathbb{I}(\text{PIT}(Y_i; \mathbf{X}_i) \leqslant \gamma_{i,j})$
    Let $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{(\mathbf{X}_i, \gamma_{i,j}, W_{i,j})\}$
  end for
end for
Use $\mathcal{D}'$ to learn $\hat{r}^{\hat{f}}(\gamma; \mathbf{x}) := \hat{\mathbb{P}}(\text{PIT}(Y; \mathbf{x}) \leqslant \gamma \mid \mathbf{x})$ via a regression of W on $\mathbf{X}$ and γ, which is monotonic w.r.t. γ.

// **Map initial CDE into a new CDE by applying learnt PIT-CDF**
for $\mathbf{x} \in \mathcal{V}$ **do**
  // **Construct recalibrated CDE**
  Compute $\widehat{F}(y|\mathbf{x}) \leftarrow \text{cumsum}(\widehat{f}(y|\mathbf{x}))$ for $y \in G$
  Let $\widetilde{F}(y|\mathbf{x}) \leftarrow \hat{r}^{\hat{f}}\left(\widehat{F}(y|\mathbf{x}); \mathbf{x}\right)$ for $y \in G$
  Apply interpolating (or smoothing) splines to obtain $\widetilde{F}(\cdot|\mathbf{x})$ and $\widetilde{F}^{-1}(\cdot|\mathbf{x})$
  Differentiate $\widetilde{F}(y|\mathbf{x})$ to obtain new distribution $\widetilde{f}(y|\mathbf{x})$ for $y \in G$
  Renormalize $\widetilde{f}(y|\mathbf{x})$ according to Izbicki and Lee (2016), section 2.2

  // **Construct Cal-PIT interval with conditional coverage** $1-\alpha$
  Compute $C(\mathbf{x}) \leftarrow [\widetilde{F}^{-1}(0.5\alpha|\mathbf{x}); \widetilde{F}^{-1}(1-0.5\alpha|\mathbf{x})]$.
  if `HPD` **then**
    Obtain HPD sets $C(\mathbf{x}) = \{y : \widetilde{f}(y|\mathbf{x}) \geqslant \widetilde{t}_{\mathbf{x},\alpha}\}$, where $\widetilde{t}_{\mathbf{x},\alpha}$ is such that $\int_{y \in C_\alpha(\mathbf{x})} \widetilde{f}(y|\mathbf{x}) dy = 1-\alpha$
  end if
end for
return $\widetilde{F}(y|\mathbf{x})$, $C(\mathbf{x})$, $\widetilde{f}(y|\mathbf{x})$, for all $\mathbf{x} \in \mathcal{V}$

Theorem 1 (performance of the recalibrated CDE). *Under assumptions 1–3 (appendix E),*

$$\mathbb{E}\left[\int\int \left(\widetilde{F}(y|\mathbf{x}) - F(y|\mathbf{x})\right)^2 \mathrm{d}P(y,\mathbf{x})\right] = K\mathbb{E}\left[\int\int \left(\widehat{r}^{\widehat{f}}(\gamma;\mathbf{x}) - r^{\widehat{f}}(\gamma;\mathbf{x})\right)^2 \mathrm{d}\gamma \mathrm{d}P(\mathbf{x})\right].$$

The rate of convergence of $\widetilde{F}(y|\mathbf{x})$ to the target distribution $F(y|\mathbf{x})$ is given by Corollary 1.

Algorithm 1 details the `Cal-PIT` procedure for computing the PIT-CDF from calibration data, and for constructing recalibrated CDEs and prediction intervals. In practice, for each $\mathbf{x}$ of interest, we first evaluate $\widetilde{F}(y|\mathbf{x})$ across a grid G of y-values, and then use linear or spline-based interpolation scheme to calculate the derivatives to finally obtain $\widetilde{f}(y|\mathbf{x})$, our estimate of the recalibrated CDE at $\mathbf{x}$.

Remark 1. If the initial model is good, then r is easy to estimate; for instance, $\widehat{f} = f$ implies a constant function $r^{\widehat{f}}(\gamma;\mathbf{x}) = \gamma$. However, $\widehat{f}$ needs to have support on the entire range of the target variable y across the feature space $\mathcal{X}$. Depending on the application, a viable initial model could, for example, be an estimate of the marginal distribution $f(y)$, a uniform distribution with finite support (as in Experiment 2 of appendix B, Example 3), an initial fit of the density with a Gaussian distribution (as in the TC application in section 4.2), or a nonparametric density estimate (as in Experiment 1 of B, Example 3). In the photo-z application in section 5, we use a weighted sum of the marginal distribution $f(y)$ and a Gaussian model for $f(y|\mathbf{x})$. The Gaussian model was obtained from a widely popular photo-z method (`GPz`; Almosallam *et al* 2016); the marginal distribution was then added to expand the support of the fitted Gaussian distribution.

Remark 2 (CDEs and prediction sets). As a by-product of conditional distributions, one can derive various quantities of interest, such as moments, kurtosis, prediction intervals, or even more general prediction bands; such as highest predictive density (HPD) regions $\{y : f(y|\mathbf{x}) > c\}$, where f is the conditional density associated to F; see appendix D for details on how to compute HPD regions. By construction, locally calibrated CDEs yield prediction bands with approximately correct *conditional* coverage. That is, suppose that

$C_\alpha(\mathbf{X})$ is a $(1-\alpha)/100\%$ prediction band derived from the CDF $\widehat{F}$. Local calibration of $\widehat{F}$ then implies that the prediction bands $C_\alpha(\mathbf{X})$ have approximate nominal coverage

$$\mathbb{P}\left(Y \in C_\alpha\left(\mathbf{X}\right) | \mathbf{X} = \mathbf{x}\right) = 1 - \alpha, \tag{6}$$

for *every* instance $\mathbf{x} \in \mathcal{X}$. On the other hand, it is difficult to convert prediction bands and quantile estimates to entire CDEs without additional assumptions. That is, calibrated CDEs imply calibrated prediction bands but not vice versa. For example, theorem 2 in appendix E shows that a `Cal-PIT` prediction interval at $\mathbf{x}$, defined as

$$C_\alpha\left(\mathbf{x}\right) := \left[\widetilde{F}^{-1}\left(0.5\alpha|\mathbf{x}\right), \widetilde{F}^{-1}\left(1-0.5\alpha|\mathbf{x}\right)\right], \tag{7}$$

achieves asymptotic conditional coverage, even if the initial CDE $\hat{f}(y|\mathbf{x})$ is not consistent.

4. Synthetic examples

4.1. Example 1: diagnostics and reshaping of CDEs via P–P maps

This example illustrates the LADaR framework with `Cal-PIT`: we start with an initial model $f_0(y|\mathbf{x})$ (in this case, a Gaussian density with correct mean and fixed variance). Then, via a PIT-CDF regression (equation (4)), we learn the local diagnostics which can be visualized via P–P plots similar to figure 2. Finally, we reshape the initial densities to better fit the calibration data by applying the same learned P–P map (that is, the PIT-CDF transformation) to the initial densities (equation (5)).

As an illustration, we create a 'skewed' data setting. The data are drawn from the family of sinh-arcsinh normal distributions (Jones and Pewsey 2009, 2019), where the *skewed* data follow

$$Y_A|X \sim \text{sinh-arcsinh}\left(\mu = X, \sigma = 2 - |X|, \gamma = X, \tau = 1\right).$$

We start with an initial Gaussian model given by

$$Y|X \sim \mathcal{N}\left(\mu = X, \sigma = 2\right),$$

and we learn the PIT-CDF function $\widehat{r}^{f_0}(\gamma;\mathbf{x})$ from a calibration set of $n = 10000$ pairs of (X, Y).

The top panel of figure 3 shows 'Local Amortized Diagnostics' for the skewed setting: the first row graphs a *local discrepancy score* (LDS) across the feature space (see Kodra *et al* 2023 for an example use of the global analog), where the LDS is defined as

$$D\left(\mathbf{x}\right) := \frac{1}{|G|}\sum_{\gamma \in G}\left(\widehat{r}^{f_0}\left(\gamma;\mathbf{x}\right) - \gamma\right)^2, \tag{8}$$

for a set $G \subset [0,1]$ of γ values. The LDS is a one-number summary that estimates the amount of discrepancy between the initial model and the true density in terms of coverage: a large value of $D(\mathbf{x})$ indicates that f_0 is miscalibrated at the evaluation point $\mathbf{x}$. The PIT-CDF function $\widehat{r}^{f_0}$ then provides more detailed information on *how* the initial model $f_0(y|\mathbf{x})$ might deviate from the true density $f(y|\mathbf{x})$ at $\mathbf{x}$, as illustrated by the shape of the P–P plots in the second row. Top panel II ('Reshaping of Densities') shows examples of morphing the initial density f_0 (blue) into an approximation $\widetilde{f}$ (red) of the final density defined by equation (5). For illustrative purposes, we show intermediate curves $s\widetilde{f} + (1-s)f_0$ for a few different values $s \in [0,1]$.

Finally, because we know the true data-generating distribution F, we can directly assess the quality of the reshaped densities $\widetilde{f}$ by first generating MC samples from the true distribution at each evaluation point $\mathbf{x}$, and then computing a local version of the continuous rank probability score (CRPS). More specifically: CRPS is a proper scoring rule commonly used to evaluate probabilistic predictions (Matheson and Winkler 1976). The local CRPS loss at a point $(\mathbf{x}, y)$ is typically defined as

$$L_{\text{CRPS}}\left(\widetilde{f};\mathbf{x},y\right) = \int_{-\infty}^{\infty}\left(\widetilde{F}(t|\mathbf{x}) - \mathbb{I}\left(y \leqslant t\right)\right)^2 \mathrm{d}t, \tag{9}$$

which checks whether a single draw $y \sim F(Y|\mathbf{x})$ (from the unknown true distribution F) is consistent with the estimated distribution $\widetilde{F}(y|\mathbf{x})$. However, for our synthetic examples, we can generate an entire

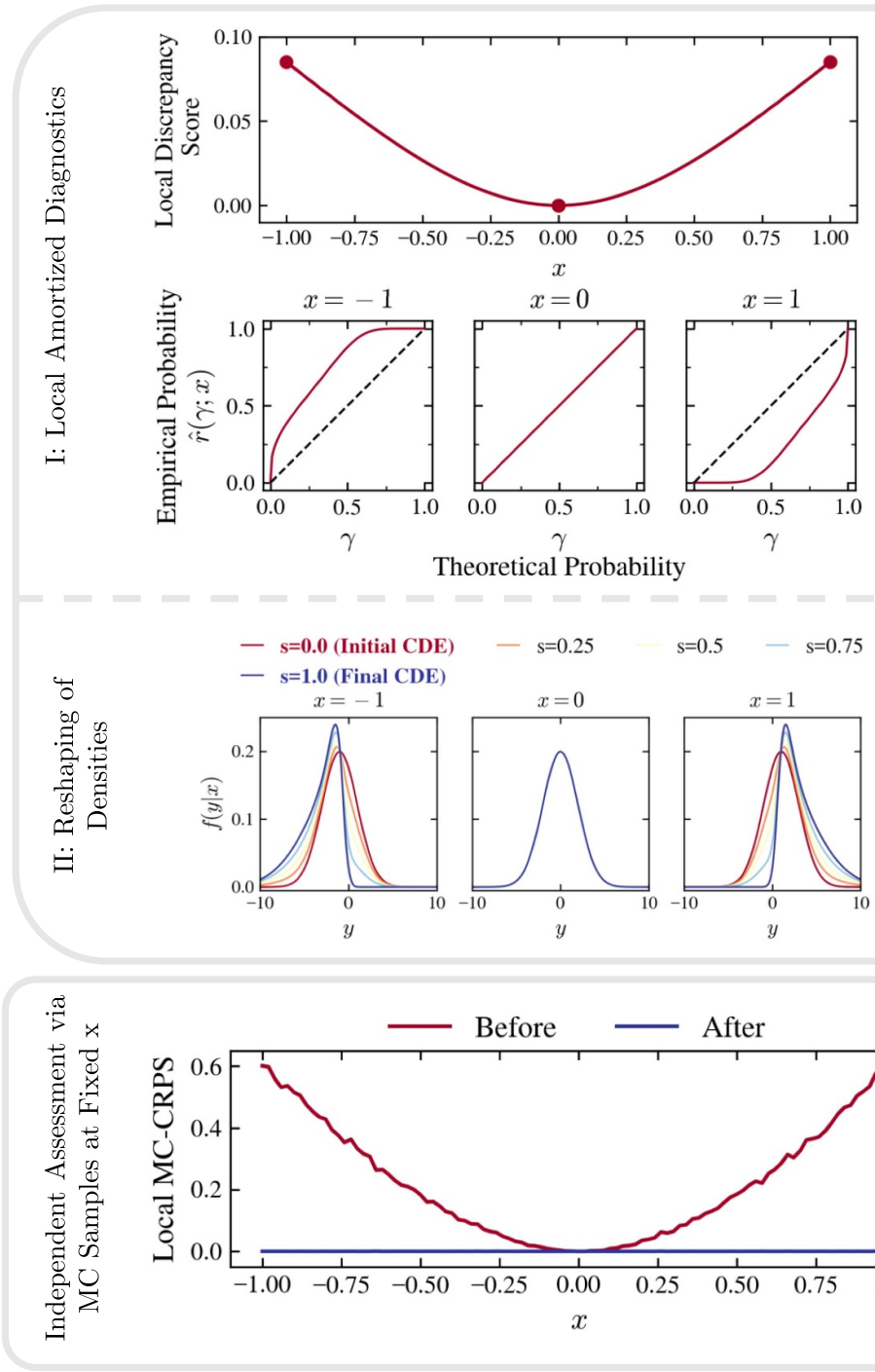


Figure 3. Illustration of LADaR framework: example 1 skewed data. Initial CDE is Gaussian, but the true distribution is skewed. *Top panel (I):* local discrepancy score across the input space (first row) and examples of diagnostic P–P plots (second row). `Cal-PIT` identifies that the model is *positively/negative biased* relative to calibration data at $X = -1$ / $X = 1$ but well-estimated at $X = 0$. The diagnostics define a family of P–P maps for reshaping the initial densities to fit the calibration data across the feature space. *Top panel (II):* continuous morphing of densities via `Cal-PIT`, illustrated at the three evaluation points, from the initial Gaussian distributions (**red**; $s = 0$) to the final distributions (**blue**; $s = 1$). For illustrative purposes, we have included intermediate values of s to show the morphing of distributions. *Bottom panel:* independent assessment of final results by computing a local Monte Carlo version of the continuous ranked probability score (MC-CRPS) at fixed x before and after `Cal-PIT`.

MC sample $Y_1, \ldots, Y_B \sim F(y|\mathbf{x})$ (from the known true distribution F) at *any* fixed evaluation point $\mathbf{x}$ for some chosen large value B. We then define the local Monte Carlo CRPS (MC-CRPS) loss at fixed $\mathbf{x}$ as

$$L_{\text{MC}-\text{CRPS}}\left(\widetilde{f}; \mathbf{x}, f\right) = \int \left(\widetilde{F}(t|\mathbf{x}) - \frac{1}{B}\sum_{b=1}^{B} I(Y_b < t)\right)^2 \mathrm{d}t. \tag{10}$$

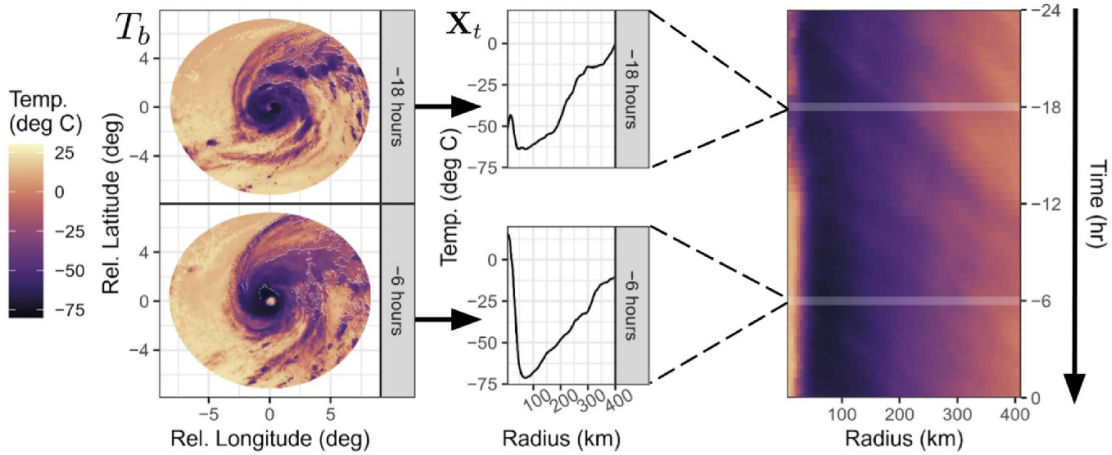


Figure 4. TC satellite images. Left: a sequence of TC-centered cloud-top temperature images from GOES. Center: we represent each GOES image with a radial profile of azimuthally-averaged cloud-top temperatures. Right: the 24 h sequence of consecutive radial profiles, sampled every 30 minutes, defines a structural trajectory $\mathbf{S}_{<t}$ or Hovmöller diagram. Reproduced with permission from McNeely *et al* (2023b). © 2025 Project Euclid, Duke University Press initiative.

For large B, equation (10) is close to zero when $\widetilde{F}(\cdot|\mathbf{x})$ is a good estimate of $F(\cdot|\mathbf{x})$. Furthermore, equation (10) is, up to a constant that does not depend on $\widetilde{F}$, approximately the same as $\mathbb{E}\left[L_{\text{CRPS}}(\widetilde{f};\mathbf{x},Y)\mid\mathbf{x}\right]$, the conditional mean of the CRPS loss given $\mathbf{X}=\mathbf{x}$ (see appendix C for more details). The bottom panel of figure 3 shows the local MC-CRPS results before and after applying `Cal-PIT` for the 'skewed' setting. The corresponding results for a 'kurtotic' setting can be found in appendix A.

4.2. Example 2: probabilistic nowcasting with high-dimensional sequence data as inputs

Our next synthetic example is motivated by short-term forecasting of the intensities of TC with high-resolution satellite images. This application is challenging both because of the high-dimensional nature of spatio-temporal satellite data and because the intensities are auto-correlated in time. Figure 4, right, shows an example of a 24 h sequence $\mathbf{S}_{<t}$ of consecutive radial profiles (one-dimensional functions) extracted from GOES infrared imagery (Janowiak *et al* 2017).

Infrared imagery, as observed by GOES, measures the cloud top temperature, which is a proxy for the strength of convection (the key component of the mechanism through which TCs extract energy from the ocean). Hence, each computed sequence $\mathbf{S}_{<t}$ can be seen as a summary of the spatio-temporal evolution of the convective structure of the TC leading up to time t, where patterns in $\mathbf{S}_{<t}$ signaling strengthening/weakening convection are predictors of intensifying/weakening storms; that is, they predict changes in the intensities of the TC, I_τ, for $\tau \geqslant t$.

As a proof-of-concept of our LADaR framework, we create a synthetic example with the same format as actual TC data. The details are described in supplementary material S3[12]. Figure 5 shows an example of a simulated storm. On the left, we have a toy Hovmöller diagram of the evolution of the 'convective structure' $\{(\mathbf{X}_t)\}_{t\geqslant 0}$, with each row representing the radial profile $\mathbf{X}_t \in \mathbb{R}^{120}$ of temperature as a function of radial distance from the storm center; time evolution is top-down in hours. On the right, we have $\{Y_t\}_{t\geqslant 0}$, the simulated 'TC intensities' at corresponding times t. The trajectory $\mathbf{S}_{<t} := (\mathbf{X}_{t-47}, \mathbf{X}_{t-46}, \ldots, \mathbf{X}_t)$ represents the 24 h history of the convective structure (48 radial profiles). We simulate 8000 'storms' according to a fitted TC length distribution. Sequence data $\{(\mathbf{S}_{<t}, Y_t)\}$ from the same storm are shifted by 30 minutes; therefore, they are strongly correlated. Sequence data from different storms, on the other hand, are independent.

Our goal is to 'nowcast' the conditional distribution $Y_t|\mathbf{S}_{<t}$, where Y_t is the intensity at time t. Here we illustrate how `Cal-PIT` can diagnose and improve an initial convolutional mixture density network (ConvMDN) model. In our example, we perform training, calibration, and testing on different simulated 'storms': first, we fit an initial CDE (ConvMDN; D'Isanto and Polsterer 2018), which estimates $f(y|\mathbf{s})$ as a unimodal Gaussian, based on a train set with 8000 points, $\{(\mathbf{S}_{<t}, Y_t)\}$ (see supplementary material S3 for details). Next, we apply `Cal-PIT` to learn $\hat{\pi}^{\hat{f}}(\gamma;\mathbf{s})$ using 8000 calibration points. (Note that the data

[12] Supplementary materials: https://lee-group-cmu.github.io/cal-pit-paper/supplementary_material.pdf.

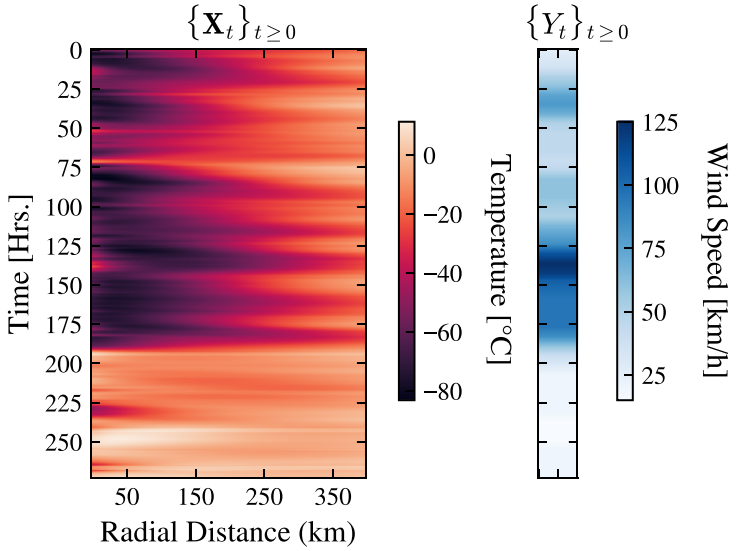


Figure 5. Synthetic data in Example 2. Simulated radial profiles $\{\mathbf{X}_t\}_{t\geqslant 0}$ and intensities $\{Y_t\}_{t\geqslant 0}$ for an example TC. Left: each row represents the radial profile $\mathbf{X}_t$ of temperature as a function of radial distance from the storm center at time t Our predictors are 48 h overlapping sequences $\{\mathbf{S}_t\}_{t\geqslant 0}$ with data from the same 'storm' being highly dependent. Right: the target response, here shown as a time series $\{Y_t\}_{t\geqslant 0}$ of simulated TC intensities.

within the same storm are highly dependent; hence, the effective train or calibration sample sizes are much smaller than the nominal values.) Because we have access to the data-generating distribution, we can assess the performance of CDEs before and after reshaping densities by MC samples at 4000 test points.

Figure 6 summarizes the results. With the LADaR framework (top panel), we are able to identify regions in a high-dimensional space of sequence data where our initial CDE of $Y_t|\mathbf{S}_{<t}$ is a poor fit. In the upper left panel, each point corresponds to a 24 h structural trajectory $\mathbf{S}_{<t}$ or a sequence of radial profiles visualized in a reduced dimensionality space using principal component analysis (PCA); the points are color-coded by the LDS between the initial model and the true distribution of the calibration data according to `Cal-PIT`. Three specific examples of input sequences are also shown. After applying the estimated P–P map via `Cal-PIT` to all CDEs, we obtain near instance-wise calibration according to an independent MC assessment (bottom panel).

4.3. Example 3: prediction sets

The novelty of our method lies in the fact that we can construct full CDEs with approximate instance-wise coverage. Nevertheless, as a by-product, we can also construct prediction sets with approximate conditional coverage. We have included an additional synthetic example in appendix B to demonstrate that prediction sets derived from `Cal-PIT` are also competitive with sets from conformal inference and QR.

5. Main application: reshaping CDEs of galaxy photometric redshifts

Many astrophysical studies depend on knowing the distances to external galaxies. Geometric distances to galaxies are incredibly difficult to measure, so astrophysicists typically use the redshift of light emitted from a galaxy as a proxy for its distance, where the spectral energy distribution (the intensity of light as a function of wavelength) is shifted to longer (redder) wavelengths due to the cosmological expansion of space. Redshifts can be precisely measured using spectroscopy to identify spectral features that occur at known wavelengths, but obtaining spectroscopic redshifts is resource-intensive. A far more efficient approach is to estimate redshifts from imaging data (i.e. photo-z's), but even with measurements at several wavelengths, imaging data produce a less precise localization of these features (and hence more uncertain photo-z's) due to a much coarser wavelength binning of photons. In particular, upcoming multi-billion dollar imaging projects like the Rubin Observatory's Legacy Survey of Space and Time (LSST; Ivezić *et al* 2019), the Nancy Grace Roman Space Telescope (Akeson *et al* 2019), and the Euclid Mission (Laureijs *et al* 2011) will make key cosmological measurements using weak gravitational lensing

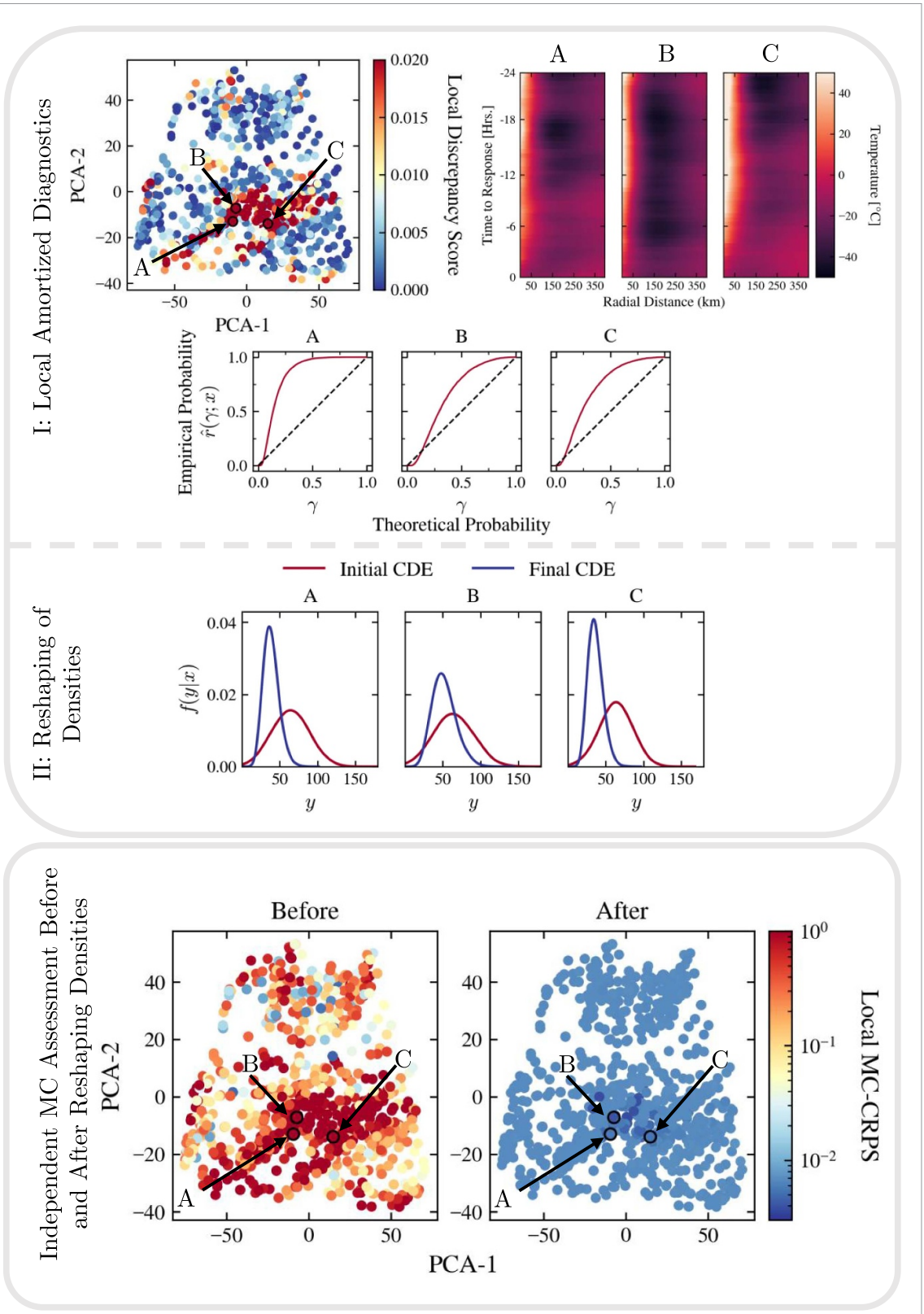


Figure 6. Example 2: probabilistic now casting with high-dimensional sequence data as inputs. *Top panel I: local amortized diagnostics.* First row: two-dimensional PCA map of sequence data. One point in the map represents a 24 h structural trajectory $\mathbf{S}_{<t}$ or sequence of radial profiles; the points are color-coded by the LDS between the initial model and calibration data according to `Cal-PIT`. Points A–C represent three examples of inputs $\mathbf{S}_{<t}$ where the initial model appears to perform the worst (i.e. high LDS). Second row: P–P plots help reveal the nature of the discrepancy; the initial model appears positively biased and overdispersed at the three locations. *Top panel II: reshaping of densities.* The density of $Y_t|\mathbf{S}_{<t}$ before (**red**) and after (**blue**) applying the P–P map. *Bottom panel: independent MC assessment.* For synthetic data, we can compute the CRPS locally for simulated MC samples at fixed $\mathbf{S}_{<t}$. The local MC CRPS scores are shown before (left) and after (right) reshaping the densities. After applying the P–P map, the CDEs are well-calibrated for all inputs $\mathbf{S}_{<t}$. Reproduced with permission from McNeely *et al* (2023b). © 2025 Project Euclid, Duke University Press initiative.

(see, e.g. Mandelbaum 2018 for an overview), a method that relies on well-calibrated photo-z's of millions of galaxies. The demands on the accuracy of photo-z CDEs for these projects are extremely stringent: discrepancies in the moments of redshift distributions for samples that are instrumental in measuring cosmological parameters must be less than approximately 0.1% to prevent degradation of subsequent physical analyzes (The LSST Dark Energy Science Collaboration *et al* 2018).

However, calibrating photo-z CDEs remains tricky because galaxies span a wide range of intrinsic properties and spectral energy distributions (Conroy 2013), which leads to different combinations of redshift and intrinsic spectral energy distribution producing nearly identical observed imaging data. This problem is further complicated by measurement errors and the coarseness of the spectral information available from imaging data. Thus, the estimation of photo-z's is inherently probabilistic with often non-trivial (e.g. non-Gaussian or bimodal) distributions. These distributions cannot be accurately captured by point estimates or prediction sets and must be quantified using full predictive distributions (Benítez 2000, Mandelbaum *et al* 2008, Malz and Hogg 2022), which `Cal-PIT` is uniquely suited to estimate.

Most photo-z estimation methods fall into two main classes: physics-inspired methods that find the combination of redshift and spectral energy distribution that best matches the data (e.g. Arnouts *et al* 1999, Brammer *et al* 2008), and (ii) data-driven methods that learn a non-linear mapping between the input imaging data and redshift (e.g. Beck *et al* 2016, Dalmasso *et al* 2020, Zhou *et al* 2021, Dey *et al* 2022). No class of method is clearly the best for all imaging data sets, with the physics-based methods typically performing better when training data are sparse and the data-driven methods typically doing better when training data densely sample parameter space. Previous studies have used global metrics to reshape probability distributions (e.g. Euclid Collaboration *et al* 2021, Kodra *et al* 2023), including PIT-based recalibration schemes (see, for instance, Bordoloi *et al* 2010, section 3). Regardless, no method guarantees correct *local* calibration of uncertainty estimates, a more stringent requirement that is the focus of `Cal-PIT`.

To showcase the effectiveness of our LADaR approach, we utilize the data set from Schmidt *et al* (2020), which has been used as a reference for assessing photo-z CDE prediction techniques. This data set was developed by assigning realistic spectral energy distributions to galaxies in a dark matter-only simulation (DeRose *et al* 2019) to mimic their appearance in LSST imaging data. The input features consist of logarithmic measurements of intensity of observed galaxy light (spatially-integrated across the image) in a given wavelength range (corresponding to a photometric filter) called apparent *magnitudes* and the differences between them called *colors*. Additionally, uncertainty estimates for these measurements were also provided. For the Schmidt *et al* (2020) data challenge, the participants were given an unbiased 'training set' of ∼44 000 instances (galaxies) to which they applied 11 different physics-inspired and data-driven photo-z approaches. The photo-z methods were then evaluated on an unseen 'test set' of ∼400 000 instances (galaxies). For this exercise, the training set was perfectly representative of the test set. Schmidt *et al* (2020) also evaluated the performance of a method that simply predicted the marginal distribution of redshifts in the training set (i.e. the same prediction for every galaxy in the data set), which they called `trainZ`. Although this naive estimate does not contain any meaningful information about the redshift of any individual galaxy, Schmidt *et al* (2020) demonstrated that it can perform well on many commonly used metrics that check for marginal calibration.

Reassuringly, Schmidt *et al* (2020) found that `trainZ` performed very poorly on the conditional density estimate (CDE) loss (Izbicki and Lee 2017), a metric of conditional coverage. The CDE loss is a proper scoring technique and the *conditional* analog of the root-mean-square-error for probabilistic regression. Given an estimate $\widetilde{f}$ of f, it is defined as the L^2 distance between $\widetilde{f}$ and f,

$$L(f,\widetilde{f}) = \int\int \left[f(y|\mathbf{x}) - \widetilde{f}(y|\mathbf{x}))\right]^2 \mathrm{d}y\mathrm{d}P(\mathbf{x}), \tag{11}$$

where $\mathrm{d}P(\mathbf{x})$ is the marginal distribution of features $\mathbf{x}$. The CDE loss cannot be evaluated directly as it depends on the unknown true density $f(y|\mathbf{x})$, but it can be estimated up to a constant (K_f, dependent on $f(y|\mathbf{x})$) by

$$\widehat{L}\left(f,\widetilde{f}\right) = \widetilde{\mathbb{E}}_{\mathbf{x}}\left[\int \widetilde{f}(y|\mathbf{x})^2\,\mathrm{d}y\right] - 2\widetilde{\mathbb{E}}_{\mathbf{x},y}\left[\widetilde{f}(y|\mathbf{x})\right] + K_f,$$

which is sufficient for relative comparisons between methods. The CDE loss was the only conditional metric identified and tested in the Schmidt *et al* (2020) challenge, so we use it as our main metric for the assessment of `Cal-PIT` in the context of photo-z's and note that `Cal-PIT` is independent of the CDE loss.

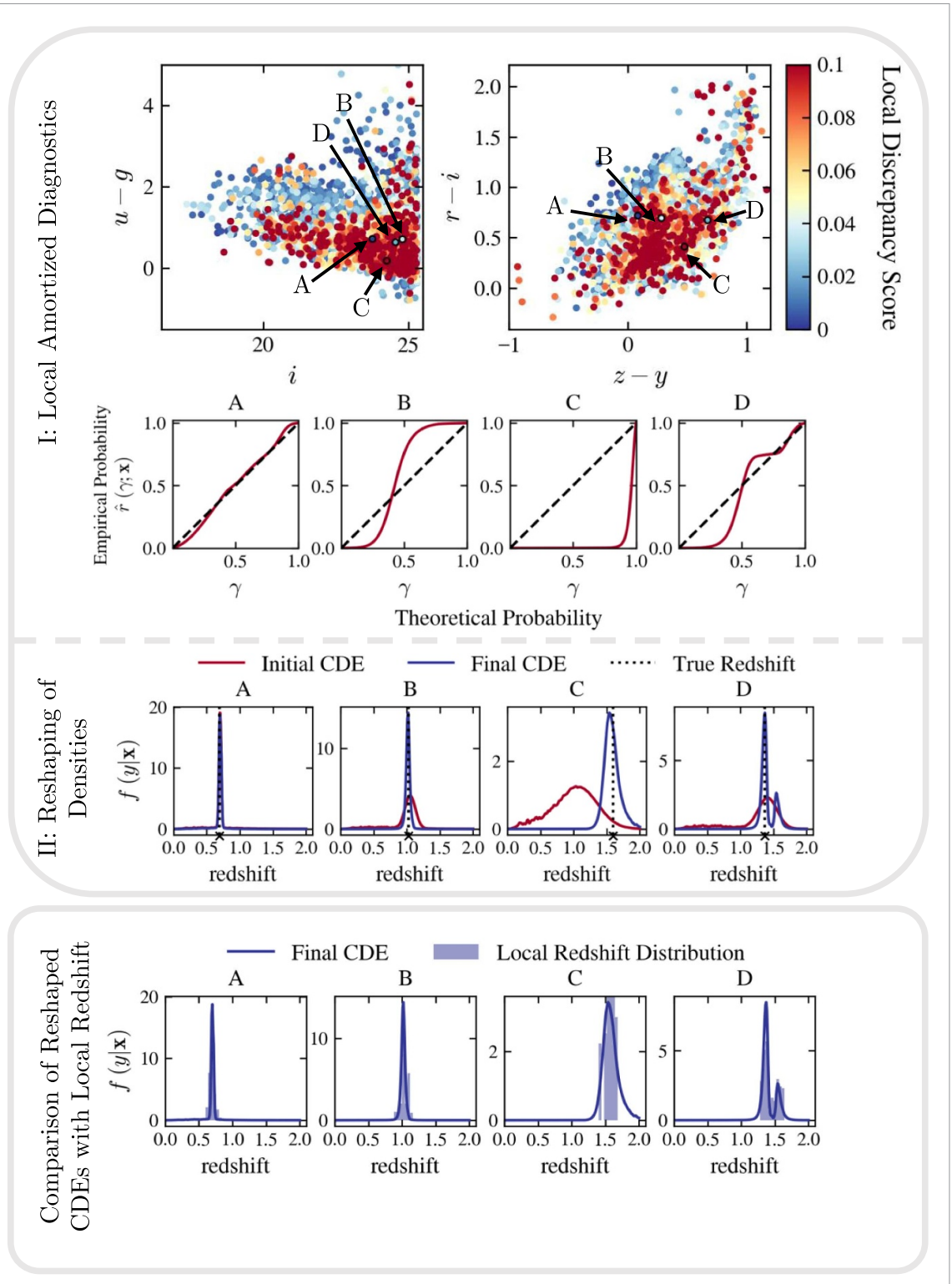


Figure 7. Photo-z application. *Top panel I: local amortized diagnostics.* First row: the local discrepancy score for the initial model shown in two projections of the feature space. The first figure shows galaxy i-band magnitude and $u-g$ color space, and the second figure shows $z-y$ color and $r-i$ color space. The points labeled A–D correspond to the four galaxies for which we show the diagnostics and reshaping. Second row: diagnostic P–P plots of the initial model (modified `GPz` CDEs; Almosallam *et al* 2016) for four galaxies from the LSST-DESC Photo-z Data Challenge (Schmidt *et al* 2020) test set. *Top panel II: reshaping of densities.* Photo-z CDEs for the corresponding galaxies before (**red**) and after (**blue**) reshaping the densities via `Cal-PIT`; the true (spectroscopic) redshift is shown as a vertical dotted black line and a cross. `Cal-PIT` can correct for bias and over-/under-dispersion. Most impressively, it can recover accurate bimodal CDEs even if the initial estimate was unimodal. *Bottom row:* comparison of the final reshaped CDEs (**blue** line) with the local 'nearest-neighbor' distribution (**blue** shaded histogram) of true redshifts of other galaxies with similar imaging properties. `Cal-PIT` accurately approximates the local redshift distribution for unimodal and multimodal redshift distributions. Further, the inferred CDEs are bimodal only when the histograms are bimodal.

Table 1. Comparison of the CDE loss values for `Cal-PIT` and the methods benchmarked in the LSST-DESC Photo-z Data Challenge (Schmidt *et al* 2020). In terms of the CDE loss, `Cal-PIT` performs better than all of the other methods tested, including `FlexZBoost`, which is specifically optimized to minimize the CDE loss. Reproduced from Schmidt *et al* (2020). CC BY 4.0 .

Photo-z algorithm	CDE loss
`ANNz2` (Sadeh *et al* 2016)	−6.88
`BPZ` (Benítez 2000)	−7.82
`Delight` (Leistedt and Hogg 2017)	−8.33
`EAZY` (Brammer *et al* 2008)	−7.07
`FlexZBoost` (Izbicki and Lee 2017)	**−10.60**
`GPz` (Almosallam *et al* 2016)	−9.93
`LePhare` (Arnouts *et al* 1999)	−1.66
`METAPhoR` (Cavuoti *et al* 2017)	−6.28
`CMNN` (Graham *et al* 2018)	−10.43
`SkyNet` (Graff *et al* 2014)	−7.89
`TPZ` (Carrasco Kind and Brunner 2013)	−9.55
`trainZ` (Schmidt *et al* 2020)	−0.83
`Cal-PIT`	**−10.80**

For a fair comparison, we adopt the same training and test sets from Schmidt *et al* (2020) and use the former as our calibration set to learn the local PIT-CDF . Among the methods compared by Schmidt *et al* (2020), we use the density estimates from `GPz` (Almosallam *et al* 2016) as our initial model. `GPz` uses sparse Gaussian processes to estimate the CDEs. Although, `GPz` produces Gaussian density estimates, it is commonly recognized that photo-z conditional densities can have non-Gaussian characteristics such as long tails or bimodalities. To expand the support of the initial distributions, we took a weighted sum of the marginal distribution of redshifts in the calibration set and the `GPz` outputs with weights 0.1 and 0.9, respectively, as our initial CDEs. We used monotonic neural networks to learn the PIT-CDF from an input feature set of one galaxy magnitude and five colors along with their measurement uncertainties. We then use the same features to diagnose and reshape the initial densities. Finally, we assess the quality of our reshaped CDEs with the CDE loss.

Figure 7 showcases how `Cal-PIT` is a powerful tool for diagnosing and reshaping photo-z CDEs. The top row of panel I displays a subset of the test data points in two projections (left: $u-g$ color vs. i-band magnitude; right: $r-i$ color vs. $z-y$ color) of feature space with the points color-coded by the LDS. Four individual galaxies are highlighted, and their diagnostic P–P plots are shown in the second row of panel I. The first P–P plot shows an instance where the initial model was good and no substantial reshaping is necessary. The second P–P plot shows an instance where the initial guess is overdispersed, whereas the third shows an instance where the initial guess was heavily biased. The last P–P plot demonstrates a case where the P–P plot has multiple steep sections, indicating that initial model failed to express a bimodal density.

Panel II shows the initial CDE (red), the reshaped CDE (blue), and the true redshift (dotted black line and cross). `Cal-PIT` leverages the information contained in the diagnostics (i.e. the P–P plots from panel I) to reshape the initial CDEs and even recover bimodal CDEs from unimodal input CDEs (with the true redshift being in one of the modes). Figure 7 (bottom row) provides a clear (though not statistically rigorous) demonstration that the CDEs from `Cal-PIT` are indeed meaningful. Since we do not know the 'ground truth' distributions for this data set, we have to rely on indirect methods to assess the quality. Specifically, we use the distribution of true redshifts of other galaxies with similar imaging data. We identify those counterparts by searching for other galaxies in the test set whose magnitudes and colors (rescaled by subtracting the mean and dividing by the standard deviation (SD) for each feature) lie within a Euclidean distance of 0.5 units of our selected galaxies. Figure 7 (bottom row) shows their redshift distribution as weighted histograms, where the weights are inversely proportional to the euclidean distance to each neighbor, together with their predicted CDEs. When CDEs are unimodal, the nearest-neighbor histograms are also unimodal with similar widths. Even more impressively, when our inferred CDEs are bimodal, the nearest-neighbor histograms show matching bimodal distributions, indicating that not only did `Cal-PIT` correctly find the mode with the true redshift, but also correctly identified the other redshift solution with similar imaging properties.

Finally, table 1 shows that `Cal-PIT` achieves a lower CDE loss than any of the methods in the LSST-DESC Photo-z data challenge (Schmidt *et al* 2020). The values of the CDE loss for all methods except

Cal-PIT come from Schmidt *et al* (2020), whereas the value for Cal-PIT was obtained by running our algorithm on the same train and test sets. As expected, there is a major improvement in the value of the CDE loss (from -9.93 to -10.80) from our input distribution (i.e. GPz) to our Cal-PIT-reshaped distributions. Moreover, Cal-PIT outperforms all other photo-z methods tested by Schmidt *et al* (2020), including FlexZBoost (Izbicki and Lee 2017), which was designed to minimize the CDE loss. Although the improvement over FlexZBoost is not dramatic, Cal-PIT guarantees proper calibration, which FlexZBoost does not. Because Cal-PIT outperforms state-of-the-art photo-z prediction methods on independent metrics while ensuring proper calibration, it is perhaps the most promising method for meeting the exacting photo-z requirements of next generation imaging surveys.

6. Discussion

There has been a growing interest in conditional density and generative models (see Chen *et al* 2022 and references therein)—however, there are few tools for assessing whether these methods yield trustworthy instance-wise UQ.

Our proposed solution, LADaR with Cal-PIT, draws on the success of high-capacity predictive algorithms, such as deep neural networks, to recalibrate CDEs in complex data settings with interpretable results and a minimum of assumptions.

Cal-PIT first assesses whether an initial conditional density model $\widehat{F}(\cdot|\mathbf{x})$ is well calibrated for all inputs $\mathbf{x}$ with respect to calibration data, and then provides a mechanism for morphing the initial densities toward the distribution $F(\cdot|\mathbf{x})$ of the reference data. Any transformation is valid as long as both $\widehat{F}(\cdot|\mathbf{x})$ and $F(\cdot|\mathbf{x})$ are continuous functions and $\widehat{F}(\cdot|\mathbf{x})$ dominates $F(\cdot|\mathbf{x})$-that is, $\widehat{F}$ assigns positive probability to any region where F does. Under these conditions (see E for details), the recalibrated distribution is well defined, and the conditional PIT fully characterizes the conditional CDF of the target variable. This flexibility explains why a unimodal distribution can be transformed into a bimodal one, as seen in the photo-z example. Individually calibrated CDEs also automatically return conditionally calibrated prediction sets. Our method does not impose shape constraints on the recalibrated density. Cal-PIT also does not require exchangeability. Instead, it only requires stationarity (to ensure that the regression function remains stable over time) and a form of weak dependence (to allow the regression method to effectively learn from new data). Therefore the method can be applied to (stationary) probabilistic time series forecasting.

Although we focus on prediction problems, our approach also applies to Bayesian inference, where the goal is to estimate intractable posterior distributions $F(\theta|\mathbf{x})$. This includes Simulation-Based Inference (SBI; Cranmer *et al* 2020), which approximates posteriors using simulations instead of explicit likelihoods (Beaumont *et al* 2002, Papamakarios and Murray 2016, Lueckmann *et al* 2017, Greenberg *et al* 2019, Izbicki *et al* 2019). Cal-PIT can assess and recalibrate such estimates $\widehat{F}(\theta|\mathbf{x})$—whether obtained via MCMC or neural methods-toward the true posterior. For implicit models like MCMC, for a fixed $\mathbf{x} \in \mathcal{X}$ and $\theta \in \Theta$, we draw $\theta_1, \ldots, \theta_L \sim \widehat{F}(\cdot|\mathbf{x})$ and approximate $\mathrm{PIT}(\theta;\mathbf{x})$ using $L^{-1}\sum_{i=1}^{L}\mathbb{I}(\theta_i \leqslant \theta)$. Unlike (SBC; Talts *et al* 2018), which focuses on marginal validity, Cal-PIT enables instance-wise recalibration and reveals local failure modes. Recent methods (Linhart *et al* 2024, Torres *et al* 2024, Wehenkel *et al* 2024) also offer local diagnostics or data-driven calibration, but Cal-PIT uniquely combines feature-space interpretability with an amortized P–P map to correct individual CDEs.

Finally, Cal-PIT can potentially be extended to multivariate output vectors $\mathbf{Y}$ by the decomposition $f(\mathbf{y}|\mathbf{x}) = \prod_i f(y_i|\mathbf{x}, \mathbf{y}_{<i})$; thus performing Cal-PIT corrections on autoregressive components of the conditional distribution. This is a particularly promising direction for deep autoregressive generative models (Van den Oord *et al* 2016, van den Oord *et al* 2016, Vaswani *et al* 2017, Hoogeboom *et al* 2021). We are currently investigating whether Cal-PIT can improve structural forecasts for short-term TC intensity guidance (McNeely *et al* 2023a). Refer to recent work by Linhart *et al* (2022) for a multivariate extension of Cal-PIT specific to normalizing flows. Other open problems include fast sampling from recalibrated conditional distributions to generate ensemble forecasts in real time, and extending Cal-PIT to classification tasks (Wald and Globerson 2017, Kull *et al* 2019).

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

The authors would like to thank Tria McNeely for helpful discussions and for preparing the TC data that were used to fit the TC-inspired model. A B L thanks Jing Lei and Larry Wasserman for valuable comments on the P–P map. B D is a postdoctoral fellow at the University of Toronto in the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship Program, a program of Schmidt Sciences. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC Award HEP-ERCAP0022859. This work is supported in part by NSF DMS-2053804, NSF PHY-2020295, and the C3.ai Digital Transformation Institute. The efforts of BD and JAN were supported by Grant DE-SC0007914 from the U.S. Department of Energy Office of Science, Office of High Energy Physics. BD, BHA, and JAN acknowledge the support of the National Science Foundation under Grant No. AST-2009251. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. RI is grateful for the financial support of CNPq (422705/2021-7 and 305065/2023-8) and FAPESP (2023/07068-1) initiatve.

Appendix A. Example 1: synthetic example (kurtotic setting)

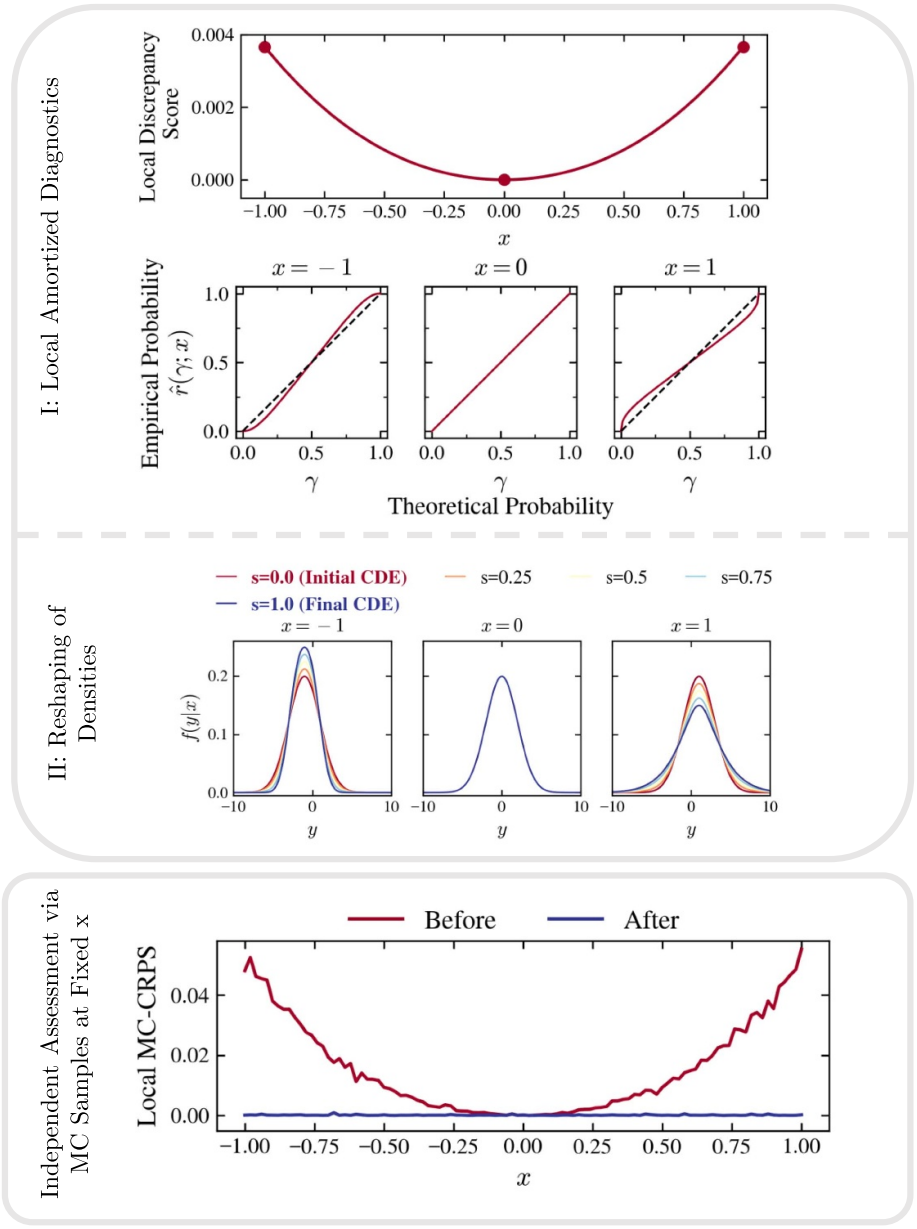


Figure A1. Illustration of LADaR framework: example 1 kurtotic data. Initial CDE is Gaussian, but the true distribution is kurtotic. *Top panel I:* Local discrepancy score across the input space (first row) and examples of diagnostic P–P plots (second row). `Cal-PIT` identifies that the model is *over-/under-dispersed* relative to calibration data at $X = -1$ / $X = 1$ but well-estimated at $X = 0$. The diagnostics define a family of P–P maps for reshaping the initial densities so as to fit the calibration data across the feature space. *Top panel II:* continuous morphing of densities via `Cal-PIT`, illustrated at the three evaluation points, from the initial Gaussian distributions (**red**; $s = 0$) to the final distributions (**blue**; $s = 1$). For illustrative purposes, we have included intermediate values of s to show the morphing of distributions. *Bottom panel:* Independent assessment of final results by computing a local MC version of the MC-CRPS at fixed x before and after `Cal-PIT`.

Figure A1 presents the LADaR approach and the results for the 'kurtotic' setting in Example 1. The data are drawn from the sinh-arcsinh normal distribution and follow $Y_B|X \sim \text{sinh-arcsinh}(\mu = X, \sigma = 2, \gamma = 0, \tau = 1 - X/4)$. The initial model is Gaussian given by $Y|X \sim \mathcal{N}(\mu = X, \sigma = 2)$, and we learn the PIT-CDF function $\widehat{r}^{\mathcal{J}_0}(\gamma; \mathbf{x})$ from a calibration set of $n = 10000$ pairs of (X, Y).

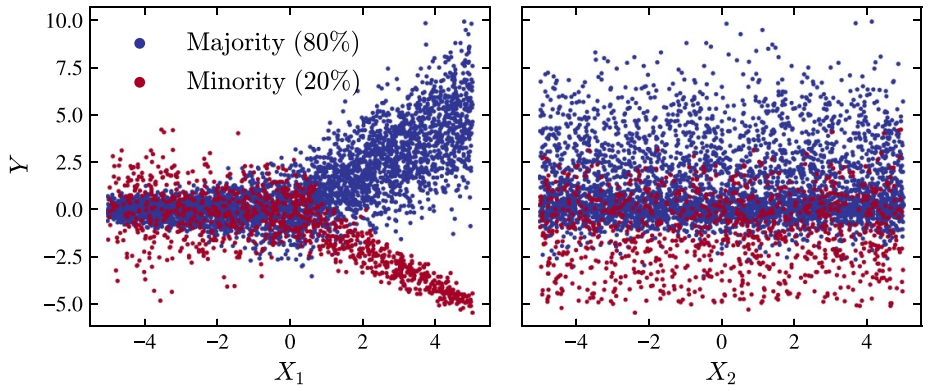


Figure B1. Visualization of one random instance of the data used for Example 1. There are two covariates (X_1, X_2), and a target variable Y. The analytic form of the true data distribution is defined in supplementary material S2. The data set consists of two groups with different spreads. Y splits into two branches for $X_1 > 0$; that is, the true CDE is bimodal in this region.

Appendix B. Example 3: prediction sets

`Cal-PIT`'s uniqueness stems from its ability to generate complete CDEs with approximate instance-wise coverage. Additionally, it enables the creation of prediction sets with approximate conditional coverage. Considering the extensive literature on prediction sets, we have included an additional example to demonstrate that prediction sets obtained from `Cal-PIT` can effectively compete with those derived using methods such as conformal inference or QR. We also include a comparison with normalizing flows, as they have gained popularity for density estimation in the physical sciences.

In photo-z estimation, multiple widely different distances (redshifts) can be consistent with the observed features (colors) of a galaxy. As mentioned previously, this results in conditional distributions that are multi-modal in parts of the feature space. Motivated by the photo-z application, we have modified the two-group example of Feldman *et al* (2021) to have bimodal structure due to limited predictor information. Here the target variable Y depends on three variables: X_0, X_1, X_2. Variable X_0 indicates group membership but it is not measured; that is, X_1 and X_2 are our only predictors. The missing membership information results in the CDE $f(y|x_1, x_2)$ being bimodal in the regime $X_1 > 0$ with one branch corresponding to each class. Supplementary material S2[13] details the data-generating process (DGP), and figure B1 visualizes one random instance of data drawn from $f(y|x_1, x_2)$ with the 'majority' and 'minority' groups displayed as blue versus red points.

We design two experiments for benchmarking `Cal-PIT` prediction sets against results from conformal inference, QR, and normalizing flows:

- Experiment 1 (comparison with conformal inference): for this experiment, we split a sample of total size n in two halves: the first half is used to train an initial model, and the second half is used for calibration. The empirical coverage of the final prediction sets are computed via 1000 MC simulations from the true DGP at each test point on a grid. Test points with coverage within two SDs of the nominal coverage of $1 - \alpha = 0.9$ based on 100 random realizations are labeled as having 'correct' coverage. We report the proportion of test points in the feature space with 'under-,' 'correct,' and 'over-' coverage.
- Experiment 2 (comparison with QR and normalizing flows): here we use the entire sample of size n to compute quantiles or to estimate the conditional density. As above, we use MC simulations on a grid to assess conditional coverage.

The top row of figure B2 shows results for Experiment 1. We compare 90% prediction sets for Y using `Cal-PIT (INT)` and `Cal-PIT (HPD)` (defined by equations (7) and (D.1), respectively) with prediction sets from `Reg-split` (Lei *et al* 2018), conformalized QR (`CQR`; Romano *et al* 2019), and distributional conformal prediction (`DCP`; Chernozhukov *et al* 2021). `Reg-split` and `CQR` are trained with XGBoost (Chen and Guestrin 2016). Our `Cal-PIT` methods use an initial CDE trained using FlexCode with an XGBoost regressor (Izbicki and Lee 2017, Dalmasso *et al* 2020) and monotonic neural networks (Wehenkel and Louppe 2019) for learning $\widehat{r}(\gamma; \mathbf{x})$ with binary cross entropy loss. `DCP` computes a

[13] Supplementary materials: https://lee-group-cmu.github.io/cal-pit-paper/supplementary_material.pdf.

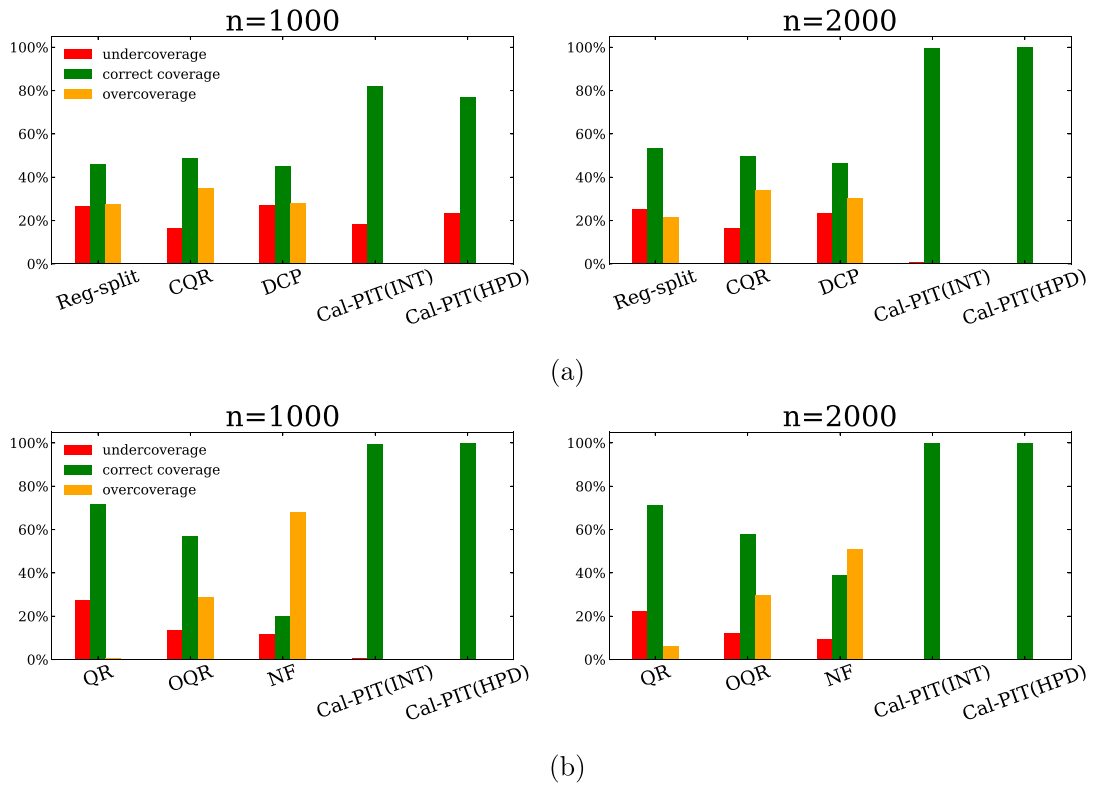
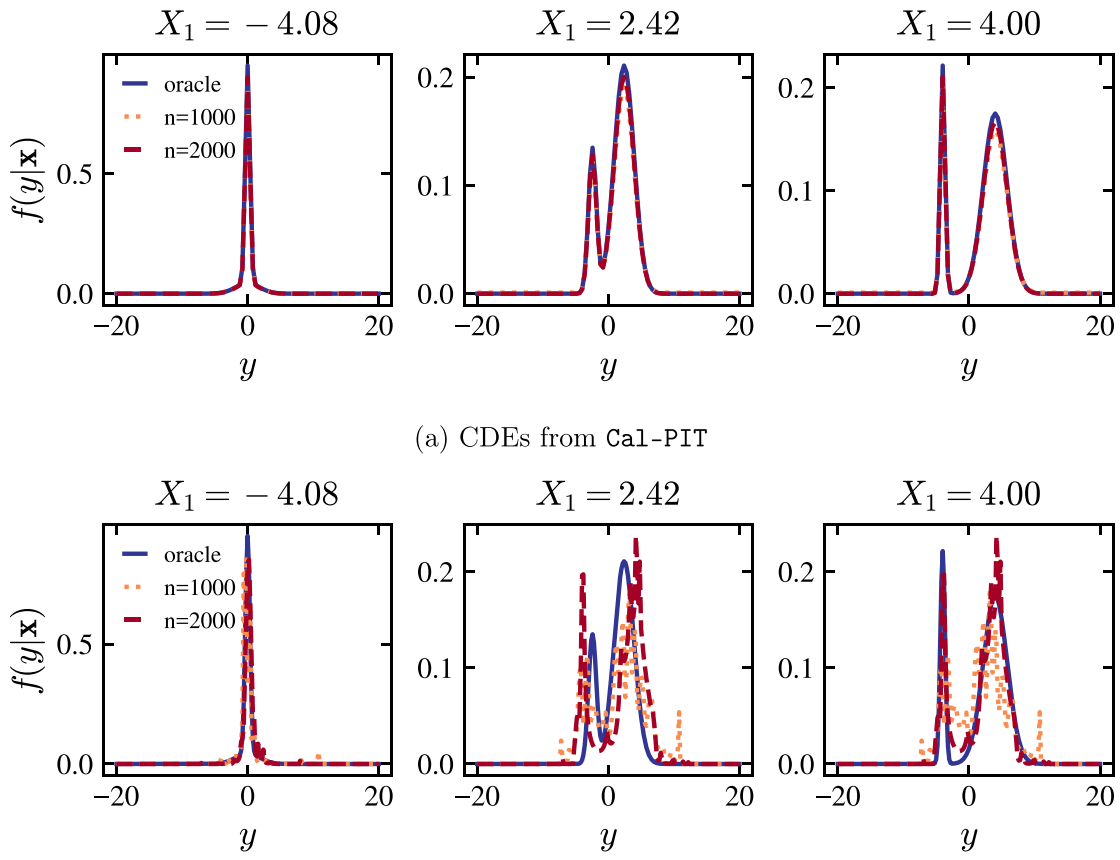


Figure B2. The proportion of test points with correct conditional coverage for (a) 'Experiment 1' with state-of-the-art conformal inference methods, using data of total size n split into a train and a calibration set, and (b) 'Experiment 2' with quantile regression and normalizing flow approaches, which use all data for training. See text for details. Only `Cal-PIT` consistently attains the nominal 90% coverage across the feature space with increasing sample size n.

conformal score based on PIT values derived from the same initial CDE as `Cal-PIT`. In terms of conditional coverage, all methods improve with increasing sample size, but only `Cal-PIT` consistently attains the nominal 90% coverage across the feature space for $n \geqslant 2000$. As the data distribution can sometimes be bimodal, the most efficient prediction sets in this feature subspace would not be single intervals (INT), but rather pairs of intervals. We can create such disjoint prediction sets using HPD regions (HPD; see appendix D for definition).

The bottom row of figure B2 shows results for Experiment 2. `Cal-PIT (INT)` and `Cal-PIT (HPD)` reshape a uniform distribution on $\mathbf{x} \in [-5, 5]$; hence, there is no need for a separate training set. The `Cal-PIT` prediction sets are then compared to output from `QR`; Koenker and Bassett 1978 trained with XGBoost and a pinball loss, orthogonal QR (`OQR`; Feldman *et al* 2021) which introduces a penalty on the pinball loss to improve conditional coverage, and normalizing flows (`NF`). We use the `PZFlow` (Crenshaw *et al* 2023) implementation of normalizing flows which has been optimized to work well out-of-the-box with tabular data and uses neural spline flows (Dinh *et al* 2014, 2016, Durkan *et al* 2019) as the backbone.

Figure B3, top row, shows some examples of calibrated CDEs from `Cal-PIT`. The estimates reveal that the true conditional density is bimodal for $X_1 > 0$; thus, the most efficient prediction sets in this feature subspace would be HPD regions. Indeed, `Cal-PIT (HPD)` yields smaller prediction sets than `Cal-PIT (INT)`; see figure S1 in supplementary material. Because HPD sets can capture the bimodality in the data while intervals cannot, this is a case where `Cal-PIT (HPD)` has better efficiency. This qualitative insight is only possible because `Cal-PIT` estimates the entire PDs. Normalizing flows also provide entire CDEs (see figure B3, bottom row) but can be difficult to train. Indeed, the normalizing flow CDEs generally deviate significantly from the oracle.



(a) CDEs from `Cal-PIT`

(b) CDEs from Normalizing Flows

Figure B3. CDEs at three different values of X_1 ($X_2 = 0$) for (a) `Cal-PIT` and (b) Normalizing Flows for Example 3. The results for $n = 1000$ and $n = 2000$ are compared to the 'oracle' probability density functions. For both sample sizes, the `Cal-PIT` CDEs are close to the oracle. Normalizing flow CDEs, on the other hand, are harder to train and a standard implementation can deviate significantly from the oracle.

Appendix C. Local CRPS scores

The conditional expectation of the CRPS loss given $\mathbf{X} = \mathbf{x}$ is

$$\mathbb{E}\left[L_{\text{CRPS}}\left(\widetilde{f}; \mathbf{X}, Y\right)|\mathbf{x}\right] = \mathbb{E}\left[\int_{-\infty}^{\infty}\left(\widetilde{F}(t|\mathbf{x}) - F(t|\mathbf{x}) + F(t|\mathbf{x}) - \mathbb{I}(Y \leqslant t)\right)^2 \mathrm{d}t \middle| \mathbf{x}\right].$$

By expanding the square and by changing the order of expectation and integration, we have:

$$\begin{aligned}\mathbb{E}\left[L_{\text{CRPS}}\left(\widetilde{f}; \mathbf{X}, Y\right)|\mathbf{x}\right] = \mathbb{E}&\left[\int_{-\infty}^{\infty}\left(\widetilde{F}(t|\mathbf{x}) - F(t|\mathbf{x})\right)^2 \mathrm{d}t \mid \mathbf{x}\right] \\ &+ 2\int_{-\infty}^{\infty}\left(\widetilde{F}(t|\mathbf{x}) - F(t|\mathbf{x})\right)\mathbb{E}\left[\left(F(t|\mathbf{x}) - \mathbb{I}(Y \leqslant t)\right)\mathrm{d}t|\mathbf{x}\right] \\ &+ \int_{-\infty}^{\infty}\mathbb{E}\left[\left(F(t|\mathbf{x}) - \mathbb{I}(Y \leqslant t)\right)^2|\mathbf{x}\right]\mathrm{d}t.\end{aligned}$$

Note that:

- The first term represents the squared distance between $\widetilde{F}$ and F and is minimized when $\widetilde{F}(\cdot|\mathbf{x}) = F(\cdot|\mathbf{x})$.
- The second term equals zero,

$$\mathbb{E}\left[F(t|\mathbf{x}) - \mathbb{I}(Y \leqslant t)|\mathbf{x}\right] = F(t|\mathbf{x}) - \mathbb{E}\left[\mathbb{I}(Y \leqslant t)|\mathbf{x}\right] = F(t|\mathbf{x}) - F(t|\mathbf{x}) = 0.$$

- The third term is a constant that does not depend on $\widetilde{F}$.

Thus,

$$\begin{aligned}\mathbb{E}\left[L_{\text{CRPS}}\left(\widetilde{f};\mathbf{X},Y\right)|\mathbf{x}\right] &= \int_{-\infty}^{\infty}\left(\widetilde{F}(t|\mathbf{x})-F(t|\mathbf{x})\right)^2\mathrm{d}t+K\\ &\approx \int\left(\widetilde{F}(t|\mathbf{x})-\frac{1}{B}\sum_{b=1}^{B}I(Y_b<t)\right)^2\mathrm{d}t+K\\ &= L_{\text{MC}-\text{CRPS}}\left(\widetilde{f};\mathbf{x},f\right)+K,\end{aligned}$$

where K does not depend on $\widetilde{F}$.

Appendix D. Cal-PIT (HPD) and Cal-HPD

Here we describe two approaches to deriving prediction sets (instead of prediction intervals) from an estimate of the conditional distribution function $f(y|\mathbf{x})$.

D.1. Cal-PIT (HPD)

Cal-PIT can also be used to compute HPDs regions instead of prediction intervals. The oracle (1-α)-level HPD set is defined as

$$\text{HPD}_\alpha(\mathbf{x}) = \{y : f(y|\mathbf{x}) \geqslant t_{\mathbf{x},\alpha}\},$$

where $t_{\mathbf{x},\alpha}$ is such that $\int_{y\in\text{HPD}_\alpha(\mathbf{x})} f(y|\mathbf{x})\mathrm{d}y = 1-\alpha$. HPDs are the smallest prediction sets that have coverage $1-\alpha$, and thus they may be more precise (smaller set size) than quantile-based intervals, while maintaining the conditional coverage at the nominal level (see appendix B for an example with a bimodal predictive distribution).

The Cal-PIT estimate of $\text{HPD}_\alpha(\mathbf{x})$ is given by

$$C_\alpha(\mathbf{x}) = \left\{y : \widetilde{f}(y|\mathbf{x}) \geqslant \widetilde{t}_{\mathbf{x},\alpha}\right\}, \tag{D.1}$$

where $\widetilde{t}_{\mathbf{x},\alpha}$ is such that $\int_{y\in C_\alpha(\mathbf{x})}\widetilde{f}(y|\mathbf{x})\mathrm{d}y = 1-\alpha$ and $\widetilde{f}$ is the Cal-PIT calibrated CDE (algorithm 1).

D.2. Cal-HPD

Alternatively, one can directly use HPD values, defined as

$$\hat{H}(y;\mathbf{x}) := \int_{\{y:\hat{f}(y'|\mathbf{x})\leqslant\hat{f}(y|\mathbf{x})\}}\hat{f}(y'|\mathbf{x})\,\mathrm{d}y',$$

to recalibrate HPD prediction sets (rather than using PIT values). The idea is to estimate the local HPD coverage at each $\mathbf{x}$, $h^{\hat{f}}(\gamma;\mathbf{x}) := \mathbb{P}(\hat{H}(Y;\mathbf{x})\leqslant\gamma|\mathbf{x})$, by regression, analogous to estimating the PIT-CDF in Cal-PIT. Let $\hat{h}^{\hat{f}}(\gamma;\mathbf{x})$ be such an estimate. The recalibrated $(1-\alpha)$-level HPD set at a location $\mathbf{x}$ is given by the $(1-\alpha^*(\mathbf{x}))$-level HPD set of the original density $\hat{f}(y|\mathbf{x})$, where $\alpha^*(\mathbf{x})$ is such that $\hat{h}^{\hat{f}}(\alpha^*(\mathbf{x});\mathbf{x}) = \alpha$. This framework however does not yield full CDEs. Moreover, although the approach corrects HPD sets, aiming for conditional coverage, the constructed sets will not be optimal if the initial model $\widehat{f}$ is far from the true data generating process f.

In example 3 (appendix B), we only report results for Cal-PIT(INT) and Cal-PIT(HPD); we do not report results for Cal-HPD.

Appendix E. Theoretical properties of Cal-PIT

We here describe the assumptions needed for theorem 1, and provide convergence rates. We also prove that Cal-PIT intervals achieve asymptotic conditional validity even if the initial CDE $\hat{f}$ is not consistent. The following results are conditional on $\hat{f}$; all uncertainty refers to the calibration sample. We assume in theorem 1 that the true distribution of $Y|\mathbf{x}$ and its initial estimate are continuous, and that $\widehat{F}$ places its mass on a region that is at least as large as that of F:

Assumption 1 (continuity of the cumulative distribution functions). For every $\mathbf{x}\in\mathcal{X}$, $\widehat{F}(\cdot|\mathbf{x})$ and $F(\cdot|\mathbf{x})$ are continuous functions.

Assumption 2 ($\widehat{F}$ dominates F). For every $\mathbf{x}\in\mathcal{X}$, $\widehat{F}(\cdot|\mathbf{x})$ dominates $F(\cdot|\mathbf{x})$.

We also assume that $F(\cdot|\mathbf{x})$ cannot place too much mass in regions where the initial estimate $\widehat{F}(\cdot|\mathbf{x})$ places little mass:

Assumption 3 (bounded density). There exists $K > 0$ such that, for every $\mathbf{x} \in \mathcal{X}$, the Radon-Nikodym derivative of $F(\cdot|\mathbf{x})$ with respect to $\widehat{F}(\cdot|\mathbf{x})$ is bounded above by K.

To provide rates of convergence for the recalibrated CDE, we will in addition assume that the regression method converges at a rate $O(n^{-\kappa})$:

Assumption 4 (Convergence rate of the regression method). The regression method used to estimate $\widehat{r}^{\widehat{f}}$ is such that its convergence rate is given by

$$\mathbb{E}\left[\int\int\left(\widehat{r}^{\widehat{f}}(\gamma;\mathbf{x}) - \widehat{r}^{\widehat{f}}(\gamma;\mathbf{x})\right)^2 \mathrm{d}\gamma \mathrm{d}P(\mathbf{x})\right] = O\left(\frac{1}{n^{\kappa}}\right)$$

for some $\kappa > 0$.

Many methods satisfy assumption 4 for some value κ, which is typically rated to the dimension of $\mathcal{X}$ and the smoothness of the true regression r (see for instance Györfi *et al* 2002).

Under these assumptions, we can derive the rate of convergence for $\widetilde{F}$:

Corollary 1 (convergence rate of recalibrated CDE). *Under assumptions 1–4,*

$$\mathbb{E}\left[\int\int\left(\widetilde{F}(y|\mathbf{x}) - F(y|\mathbf{x})\right)^2 \mathrm{d}P(y,\mathbf{x})\right] = O\left(\frac{1}{n^{\kappa}}\right). \tag{E.1}$$

Next, we show that with an uniformly consistent regression estimator $\hat{r}^{\hat{f}}(\gamma;\mathbf{x})$ (see Bierens 1983, Hardle *et al* 1984, Liero 1989, Girard *et al* 2014for some examples), `Cal-PIT` intervals achieve asymptotic conditional validity, even if the initial CDE $\hat{f}(y|\mathbf{x})$ is not consistent.

Assumption 5 (uniform consistency of the regression estimator). The regression estimator is such that

$$\sup_{\mathbf{x}\in\mathcal{X},\gamma\in[0,1]} |\widehat{r}^{\widehat{f}}(\gamma;\mathbf{x}) - \widehat{r}^{\widehat{f}}(\gamma;\mathbf{x})| \xrightarrow[n\longrightarrow\infty]{\text{a.s.}} 0,$$

where the convergence is with respect to the calibration set $\mathcal{D}$ only; $\widehat{f}$ is fixed.

Theorem 2 (consistency and conditional coverage of `Cal-PIT` intervals). *Let $C^*_\alpha(\mathbf{x}) = \left[F^{-1}(0.5\alpha|\mathbf{x}); F^{-1}(1-0.5\alpha|\mathbf{x})\right]$ be the oracle prediction band, and let $C^n_\alpha(\mathbf{x})$ denote the `Cal-PIT` interval. Under assumptions 1, 2 and 5,*

$$\lambda\left(C^n_\alpha(\mathbf{X})\,\Delta\,C^*_\alpha(\mathbf{X})\right) \xrightarrow[n\longrightarrow\infty]{\text{a.s.}} 0, \tag{E.2}$$

where λ is the Lebesgue measure in $\mathbb{R}$ and Δ is the symmetric difference between two sets. It follows that $C^n_\alpha(\mathbf{X})$ has asymptotic conditional coverage of $1-\alpha$ (Lei et al 2018).

See appendix F.1 for theoretical results for `Cal-PIT (HPD)`.

Appendix F. Proofs

Lemma 1. *Let G and H be two cumulative distribution functions such that G dominates H, and let μ_G and μ_H be their associated measures over $\mathbb{R}$. Then, for every fixed $y \in \mathbb{R}$,*

$$\mu_H(\{y' \in \mathbb{R} : y' \leqslant y\}) = \mu_H(\{y' \in \mathbb{R} : G(y') \leqslant G(y)\}).$$

Proof. Fix $y \in \mathbb{R}$ and let $A = \{y' \in \mathbb{R} : y' \leqslant y\}$ and $B = \{y' \in \mathbb{R} : G(y') \leqslant G(y)\}$. Because $A \subseteq B$,

$$\mu_H(A) \leqslant \mu_H(B). \tag{F.1}$$

We note that $\mu_G(B \cap A^c) = 0$. From this and the assumption that G dominates H, we conclude that $\mu_H(B \cap A^c) = 0$. It follows that

$$\begin{aligned}\mu_H(B) &= \mu_H(B \cap A) + \mu_H(B \cap A^c) \leqslant \mu_H(A) + 0 \\ &= \mu_H(A).\end{aligned} \tag{F.2}$$

From equations (F.1) and (F.2), we conclude that $\mu_H(A) = \mu_H(B)$. □

Lemma 2. *Fix $y \in \mathbb{R}$ and let $\gamma := \widehat{F}(y|\mathbf{x})$. Then, under assumptions 1 and 2, $\widetilde{F}(y|\mathbf{x}) = \widehat{\widetilde{r}}^{\widehat{f}}(\gamma;\mathbf{x})$ and $F(y|\mathbf{x}) = \widehat{r}^{\widehat{f}}(\gamma;\mathbf{x})$.*

Proof. We note that $\gamma = \widehat{F}(y|\mathbf{x})$ implies that $y = \widehat{F}^{-1}(\gamma|\mathbf{x})$. It follows then by construction,

$$\tilde{F}(y|\mathbf{x}) = \tilde{F}\left(\widehat{F}^{-1}(\gamma|\mathbf{x})\,|\mathbf{x}\right) = \widehat{\tilde{r}}^{\widehat{f}}(\gamma;\mathbf{x})\,.$$

Moreover,

$$\begin{aligned} F(y|\mathbf{x}) &= \mathbb{P}(Y \leqslant y|\mathbf{x}) \\ &= \mathbb{P}\left(\widehat{F}(Y|\mathbf{x}) \leqslant \widehat{F}(y|\mathbf{x})\,|\mathbf{x}\right) \\ &\quad \text{(Assumption 2 and Lemma 1)} \\ &= \mathbb{P}\left(\mathrm{PIT}(Y;\mathbf{x}) \leqslant \widehat{F}(y|\mathbf{x})\,|\mathbf{x}\right) \\ &= \mathbb{P}(\mathrm{PIT}(Y;\mathbf{x}) \leqslant \gamma|\mathbf{x}) \\ &= \widehat{r}^{\widehat{f}}(\gamma;\mathbf{x})\,, \end{aligned}$$

which concludes the proof. □

Proof of theorem 1. Consider the change of variables $\gamma = \widehat{F}(y|\mathbf{x})$, so that $d\gamma = \widehat{f}(y|\mathbf{x})dy$. Lemma 2 implies that $\widetilde{F}(y|\mathbf{x}) = \widehat{\widetilde{r}}^{\widehat{f}}(\gamma;\mathbf{x})$ and $F(y|\mathbf{x}) = \widehat{r}^{\widehat{f}}(\gamma;\mathbf{x})$. It follows from that and assumption 3 that

$$\begin{aligned} \int\int \left(\widetilde{F}(y|\mathbf{x}) - F(y|\mathbf{x})\right)^2 \mathrm{d}P(y,\mathbf{x}) &\leqslant K \int\int \left(\widetilde{F}(y|\mathbf{x}) - F(y|\mathbf{x})\right)^2 \widehat{f}(y|\mathbf{x})\,\mathrm{d}y P(\mathbf{x}) \\ &= K \int\int \left(\widehat{\widetilde{r}}^{\widehat{f}}(\gamma;\mathbf{x}) - \widehat{r}^{\widehat{f}}(\gamma;\mathbf{x})\right)^2 \mathrm{d}\gamma \mathrm{d}P(\mathbf{x})\,, \end{aligned}$$

which concludes the proof. □

Proof of corollary 1. Follows directly from assumption 4 and theorem 1. □

Proof of theorem 2. From lemma 2,

$$\sup_{\mathbf{x}\in\mathcal{X}, y\in\mathbb{R}} |\tilde{F}(y|\mathbf{x}) - F(y|\mathbf{x})| = \sup_{\mathbf{x}\in\mathcal{X}, \gamma\in[0,1]} |\widehat{\tilde{r}}^{\widehat{f}}(\gamma;\mathbf{x}) - \widehat{r}^{\widehat{f}}(\gamma;\mathbf{x})| \xrightarrow[n\longrightarrow\infty]{\text{a.s.}} 0,$$

where the last step follows from assumption 5. It then follows from assumption 1 that

$$\sup_{\mathbf{x}\in\mathcal{X}, \gamma\in[0,1]} |\tilde{F}^{-1}(\gamma|\mathbf{x}) - F^{-1}(\gamma|\mathbf{x})| \xrightarrow[n\longrightarrow\infty]{\text{a.s.}} 0,$$

and, in particular,

$$\sup_{\mathbf{x}\in\mathcal{X}, \alpha\in\{.5\alpha, 1-.5\alpha\}} |\tilde{F}^{-1}(\alpha|\mathbf{x}) - F^{-1}(\alpha|\mathbf{x})| \xrightarrow[n\longrightarrow\infty]{\text{a.s.}} 0,$$

from which the conclusion of the theorem follows. □

F.1. Theory for `Cal-PIT` HPD sets

For every $\mathbf{x} \in \mathcal{X}$, let $C_\alpha(\mathbf{x}) = \{y : \widetilde{f}(y|\mathbf{x}) \geqslant \widetilde{t}_{\mathbf{x},\alpha}\}$, where $\widetilde{t}_{\mathbf{x},\alpha}$ is such that $\int_{y\in C_\alpha(\mathbf{x})} \widetilde{f}(y|\mathbf{x})dy = 1 - \alpha$ be the `Cal-PIT` HPD-set. Similarly, let $\mathrm{HPD}_\alpha(\mathbf{x}) = \{y : f(y|\mathbf{x}) \geqslant t_{\mathbf{x},\alpha}\}$, where $t_{\mathbf{x},\alpha}$ is such that $\int_{y\in\mathrm{HPD}_\alpha(\mathbf{x})} f(y|\mathbf{x})dy = 1 - \alpha$ be the true HPD-set. The next theorem shows that if the probabilistic classifier is well estimated, then `Cal-PIT` HPD sets are exactly equivalent to oracle HPD sets.

Theorem 3 (fisher consistency `Cal-PIT` HPD-sets). *Fix $\mathbf{x} \in \mathcal{X}$. If $\widehat{r}(\gamma;\mathbf{x}) = r(\gamma;\mathbf{x})$ for every $\gamma \in [0,1]$, $C_\alpha(\mathbf{x}) = \mathrm{HPD}_\alpha(\mathbf{x})$ and $\mathbb{P}(Y \in C_\alpha(\mathbf{X})|\mathbf{x}) = 1 - \alpha$.*

Proof of theorem 3. Fix $y \in \mathbb{R}$ and let $\gamma = \widehat{F}(y|\mathbf{x})$, so that $y = \widehat{F}^{-1}(\gamma|\mathbf{x})$. It follows that

$$\begin{aligned} \tilde{F}(y|\mathbf{x}) &= \tilde{F}\left(\widehat{F}^{-1}(\gamma|\mathbf{x})\,|\mathbf{x}\right) = \widehat{r}(\gamma;\mathbf{x}) = r(\gamma;\mathbf{x}) \\ &= \mathbb{P}\left(\widehat{F}(Y|\mathbf{x}) \leqslant \widehat{F}(y|\mathbf{x})\,|\mathbf{x},\gamma\right) = \mathbb{P}(Y \leqslant y|\mathbf{x},\gamma) \\ &= F(y|\mathbf{x})\,, \end{aligned}$$

and therefore $\tilde{f}(y|\mathbf{x}) = f(y|\mathbf{x})$ for almost every $y \in \mathbb{R}$. It follows that $C_\alpha(\mathbf{x}) = \mathrm{HPD}_\alpha(\mathbf{x})$. The claim about conditional coverage follows from the definition of the HPD. □

F.2. Further details on experiments

We refer the reader to the online supplementary materials for details on the training of the regression model to learn the PIT-CDF function in our experiments, further remarks on Example 3 (prediction sets) results, and a description of the synthetic data generation and the training of the initial ConvMDN model in Example 2.

ORCID iDs

Biprateep Dey 0000-0002-5665-7912
Brett H Andrews 0000-0001-8085-5890
Jeffrey A Newman 0000-0001-8684-2222
Rafael Izbicki 0000-0003-0379-9690
Ann B Lee 0000-0003-4430-7621

References

Akeson R *et al* 2019 The wide field infrared survey telescope: 100 hubbles for the 2020s (arXiv:1902.05569)
Alkema L, Raftery A E and Clark S J 2007 Probabilistic projections of HIV prevalence using Bayesian melding *Ann. Appl. Stat.* **1** 229
Almosallam I A, Jarvis M J and Roberts S J 2016 GPZ: non-stationary sparse Gaussian processes for heteroscedastic uncertainty estimation in photometric redshifts *Mon. Not. R. Astron. Soc.* **462** 726
Ambrogioni L, Güçlü U, van Gerven M A J and Maris E 2017 The kernel mixture network: a nonparametric method for conditional density estimation of continuous random variables (arXiv:1705.07111)
Amerise I L 2018 Quantile Regression Estimation Using Non-Crossing Constraints *J. Math. Stat.* **14** 107
Andrews D W K 1997 A conditional Kolmogorov test *Econometrica* **65** 1097
Arnouts S *et al* 1999 Measuring and modelling the redshift evolution of clustering: the Hubble Deep Field North *Mon. Not. R. Astron. Soc.* **310** 540
Beaumont M A, Zhang W and Balding D J 2002 Approximate Bayesian computation in population genetics *Genetics* **162** 2025
Beck R, Dobos L, Budavári T, Szalay A S and Csabai I 2016 Photometric redshifts for the SDSS Data Release 12 *Mon. Not. R. Astron. Soc.* **460** 1371
Benitez N 2000 Bayesian photometric redshift estimation *Astrophys. J.* **536** 571
Bierens H J 1983 Uniform consistency of kernel estimators of a regression function under generalized conditions *J. Am. Stat. Assoc.* **78** 699
Bishop C M 1994 (available at: https://publications.aston.ac.uk/id/eprint/373/)
Bordoloi R, Lilly S J and Amara A 2010 Photo-z performance for precision cosmology *Mon. Not. R. Astron. Soc.* **406** 881
Brammer G B, van Dokkum P G and Coppi P 2008 EAZY: a fast, public photometric redshift code *Astrophys. J.* **686** 1503
Cabezas L *et al* 2025 Regression trees for fast and adaptive prediction intervals *Inf. Sci.* **686** 31
Carrasco Kind M and Brunner R J 2013 TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests *Mon. Not. R. Astron. Soc.* **432** 1483
Cavuoti S *et al* 2017 METAPHOR: a machine-learning-based method for the probability density estimation of photometric redshifts *Mon. Not. R. Astron. Soc.* **465** 1959
Chen T Y *et al* 2022 Interpretable uncertainty quantification in AI for HEP *Proc. US Community Study on the Future of Particle Physics (Snowmass 2021)*
Chen T and Guestrin C 2016 XGBoost: a scalable tree boosting system *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD'16* (ACM) pp 785–94
Chernozhukov V, Wüthrich K and Zhu Y 2021 Distributional conformal prediction *Proc. Natl Acad. Sci.* **118** e2107794118
Chung Y, Neiswanger W, Char I and Schneider J 2021b Beyond pinball loss: quantile methods for calibrated uncertainty quantification *Advances in Neural Information Processing Systems* vol 35 (Curran Associates, Inc.) (available at: https://proceedings.neurips.cc/paper/2021/file/5b168fdba5ee5ea262cc2d4c0b457697-Paper.pdf)
Chung Y, Neiswanger W, Char I and Schneider J 2021a *Advances in Neural Information Processing Systems: Annual Conf. on Neural Information Processing Systems 2021, NeurIPS 2021, (6 December–14 December 2021)* vol 31, ed M Ranzato, A Beygelzimer, Y N Dauphin, P Liang and J W Vaughan pp 10971–84 (available at: https://proceedings.neurips.cc/paper/2021/hash/5b168fdba5ee5ea262cc2d4c0b457697-Abstract.html)
Conroy C 2013 Modeling the panchromatic spectral energy distributions of galaxies *Ann. Rev. Astron. Astrophys.* **51** 393
Cook S R, Gelman A and Rubin D B 2006 Validation of software for Bayesian models using posterior quantiles *J. Comput. Graph. Stat.* **15** 675
Cranmer K, Brehmer J and Louppe G 2020 The frontier of simulation-based inference *Proc. Natl Acad. Sci.* **117** 30055
Crenshaw J F, Yan Z and vladislav doster 2023 jfcrenshaw/pzflow: v3.1.1, *Zenodo* (https://doi.org/10.5281/zenodo.7843901)
D'Isanto A and Polsterer K L 2018 Photometric redshift estimation via deep learning. Generalized and pre-classification-less, image based, fully probabilistic redshifts *Astron. Astrophys.* **609** A111
Dalmasso N *et al* 2020 Conditional density estimation tools in python and R with applications to photometric redshifts and likelihood-free cosmological inference *Astron. Comput.* **30** 100362
Dawid A P and Musio M 2014 Theory and applications of proper scoring rules *Metron* **72** 169
DeMaria M, Sampson C R, Knaff J A and Musgrave K D 2014 Is tropical cyclone intensity guidance improving? *Bull. Am. Meteorol. Soc.* **95** 387
DeRose J *et al* 2019 The buzzard flock: dark energy survey synthetic sky catalogs (arXiv:1901.02401)
DESI Collaboration *et al* 2022 Overview of the instrumentation for the dark energy spectroscopic instrument *Astron. J.* **164** 207
Dey A *et al* 2019 Overview of the DESI legacy imaging surveys *Astron. J.* **157** 168
Dey B *et al* 2021 Re-calibrating photometric redshift probability distributions using feature-space regression (arXiv:2110.15209)
Dey B *et al* 2022 Photometric redshifts from SDSS images with an interpretable deep capsule network *Mon. Not. R. Astron. Soc.* **515** 5285

Dhariwal P and Nichol A Q 2021 *Advances in Neural Information Processing Systems: Annual Conf. on Neural Information Processing Systems 2021, NeurIPS 2021, (6 December–14 December 2021)* vol 34, ed M Ranzato, A Beygelzimer, Y N Dauphin, P Liang and J W Vaughan pp 8780–94 (available at: https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html)
Dinh L, Krueger D and Bengio Y 2014 NICE: non-linear independent components estimation (arXiv:1410.8516)
Dinh L, Sohl-Dickstein J and Bengio S 2016 Density estimation using real NVP (arXiv:1605.08803)
Durkan C, Bekasov A, Murray I and Papamakarios G 2019 Neural spline flows (arXiv:1906.04032)
Euclid Collaboration *et al* 2021 Euclid preparation. XI. Mean redshift determination from galaxy redshift probabilities for cosmic shear tomography *Astron. Astrophys.* **647** A117
Fasiolo M, Wood S N, Zaffran M, Nedellec R and Goude Y 2021 Fast calibrated additive quantile regression *J. Am. Stat. Assoc.* **116** 1402
Feldman S, Bates S and Romano Y 2021 Improving conditional coverage via orthogonal quantile regression (arXiv:2106.00394)
Freeman P, Izbicki R and Lee A B 2017 A unified framework for constructing, tuning and assessing photometric redshift density estimates in a selection bias setting *Mon. Not. R. Astron. Soc.* **468** 4556
Gaia Collaboration *et al* 2016 The Gaia mission *Astron. Astrophys.* **595** A1
Gal Y and Ghahramani Z 2016 Dropout as a Bayesian approximation: representing model uncertainty in deep learning *Int. Conf. on Machine Learning* (*Proc. Machine Learning Research)* vol 48 (PMLR) pp 1050–9 (available at: https://proceedings.mlr.press/v48/gal16.pdf)
Gan F F and Koehler K J 1990 Goodness-of-Fit Tests Based on P-P Probability Plots *Technometrics* **32** 289
Girard S, Guillou A and Stupfler G 2014 Uniform strong consistency of a frontier estimator using kernel regression on high order moments *ESAIM: Probab. Stat.* **18** 642
Gneiting T 2008 Probabilistic forecasting *J. R. Stat. Soc. A* **171** 319
Gneiting T and Katzfuss M 2014 Probabilistic forecasting *Ann. Rev. Stat. Appl.* **1** 125
Good I J 1952 Rational Decisions *J. R. Stat. Soc. B* **14** 107
Graff P, Feroz F, Hobson M P and Lasenby A 2014 SKYNET: an efficient and robust neural network training tool for machine learning in astronomy *Mon. Not. R. Astron. Soc.* **441** 1741
Graham M L *et al* 2018 Photometric redshifts with the LSST: evaluating survey observing strategies *Astron. J.* **155** 1
Greenberg D, Nonnenmacher M and Macke J 2019 Automatic posterior transformation for likelihood-free inference *Int. Conf. on Machine Learning* (PMLR) pp 2404–14 (available at: https://proceedings.mlr.press/v97/greenberg19a.html)
Griffin S M, Wimmers A and Velden C S 2022 Predicting rapid intensification in North Atlantic and Eastern North Pacific tropical cyclones using a convolutional neural network *Weather Forecast.* **37** 1333–55
Guo C, Pleiss G, Sun Y and Weinberger K Q 2017 *Proc. 34th Int. Conf. on Machine Learning (Proc. Machine Learning Research* vol 70), ed D Precup and Y W Teh (PMLR) pp 1321–30 (available at: https://proceedings.mlr.press/v70/guo17a.html)
Györfi L *et al* 2002 *A Distribution-Free Theory of Nonparametric Regression* vol 1 (Springer) (https://doi.org/10.1007/b97848)
Hardle W *et al* 1984 Uniform consistency of a class of regression function estimators *Ann. Stat.* **12** 612
Ho J, Jain A and Abbeel P 2020 Denoising diffusion probabilistic models *CoRR* (arXiv:2006.11239)
Ho J and Salimans T 2022 Classifier-free diffusion guidance, *CoRR* (arXiv:2207.12598)
Hoogeboom E *et al* 2021 Autoregressive diffusion models (arXiv:2110.02037)
Ivezić Ž *et al* 2019 LSST: from science drivers to reference design and anticipated data products *Astrophys. J.* **873** 111
Izbicki R and Lee A B 2016 Nonparametric conditional density estimation in a high-dimensional regression setting *J. Comput. Graph. Stat.* **25** 1297
Izbicki R and Lee A B 2017 Converting high-dimensional regression to high-dimensional conditional density estimation *Electron. J. Stat.* **11** 2800
Izbicki R, Lee A B and Pospisil T 2019 ABC–CDE: toward approximate Bayesian computation with complex high-dimensional data and limited simulations *J. Comput. Graph. Stat.* **28** 1
Izbicki R, Shimizu G and Stern R B 2022 CD-split and HPD-split: efficient conformal regions in high dimensions *J. Mach. Learn. Res.* **23** 1 (available at: https://www.jmlr.org/papers/volume23/20-797/20-797.pdf)
Izbicki R, Shimizu G and Stern R 2020 Flexible distribution-free conditional predictive bands using density estimators *Int. Conf. on Artificial Intelligence and Statistics* (PMLR) pp 3068–77 (available at: https://proceedings.mlr.press/v108/izbicki20a/izbicki20a.pdf)
Janowiak J, Joyce B and Xie P 2017 NCEP/CPC L3 half hourly 4km global (60S - 60N) merged IR v1 *NASA Goddard Earth Sciences Data and Information Services Center (DAAC) data set P4HZB9N27EKU* (https://doi.org/10.5067/P4HZB9N27EKU)
Jitkrittum W, Kanagawa H and Schölkopf B 2020b *Proc. 36th Conf. on Uncertainty in Artificial Intelligence (UAI) (Proc. Machine Learning Research* vol 124), ed J Peters and D Sontag (PMLR) pp 221–30 (available at: http://proceedings.mlr.press/v124/jitkrittum20a.html)
Jitkrittum W, Kanagawa H and Schölkopf B 2020a *Proc. 36th Conf. on Uncertainty in Artificial Intelligence, UAI 2020, Virtual Online, 3 August–6 August 2020) (Proc. Machine Learning Research* vol 124), ed R P Adams and V Gogate (AUAI Press) pp 221–30 (available at: http://proceedings.mlr.press/v124/jitkrittum20a.html)
Jones C and Pewsey A 2019 *The Sinh-Arcsinh Normal Distribution* (Wiley) (https://doi.org/10.1111/j.1740-9713.2019.01245.x)
Jones M C and Pewsey A 2009 Sinh-arcsinh distributions *Biometrika* **96** 761
Kingma D P and Welling M 2013 Auto-encoding variational bayes (arXiv:1312.6114)
Kleinberg J, Mullainathan S and Raghavan M 2016 Inherent trade-offs in the fair determination of risk scores (arXiv:1609.05807)
Kobyzev I, Prince S J D and Brubaker M A 2021 Normalizing flows: an introduction and review of current methods *IEEE Trans. Pattern Anal. Mach. Intell.* **43** 3964
Kodra D *et al* 2023 Optimized photometric redshifts for the cosmic assembly near-infrared deep extragalactic legacy survey (CANDELS) *Astrophys. J.* **942** 36
Koenker R and Bassett G 1978 Regression quantiles *Econometrica* **46** 33
Koenker R and Hallock K F 2001 Quantile regression *J. Econ. Persp.* **15** 143
Kuleshov V, Fenner N and Ermon S 2018 Accurate uncertainties for deep learning using calibrated regression *Int. Conf. on Machine Learning* (PMLR) pp 2796–804 (available at: https://proceedings.mlr.press/v80/kuleshov18a/kuleshov18a.pdf)
Kull M *et al* 2019 Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration *CoRR* (arXiv:1910.12656)
Kullback S and Leibler R A 1951 On Information and Sufficiency *Ann. Math. Stat.* **22** 79

Lakshminarayanan B, Pritzel A and Blundell C 2017 Simple and scalable predictive uncertainty estimation using deep ensembles *Advances in Neural Information Processing Systems* p 30 (available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf)
Laureijs R *et al* 2011 Euclid definition study report (arXiv:1110.3193)
Lei J, G'Sell M, Rinaldo A, Tibshirani R J and Wasserman L 2018 Distribution-free predictive inference for regression *J. Am. Stat. Assoc.* **113** 1094
Leistedt B and Hogg D W 2017 Data-driven, interpretable photometric redshifts trained on heterogeneous and unrepresentative data *Astrophys. J.* **838** 5
Li L, Carver R, Lopez-Gomez I, Sha F and Anderson J 2024 Generative emulation of weather forecast ensembles with diffusion models *Sci. Adv.* **10** eadk4489
Liero H 1989 Strong uniform consistency of nonparametric regression function estimates *Probab. Theory Relat. Fields* **82** 587
Lim B, Arık S O, Loeff N and Pfister T 2021 Temporal fusion transformers for interpretable multi-horizon time series forecasting *Int. J. Forecast.* **37** 1748
Linhart J, Gramfort A and Rodrigues P L 2022 Validation Diagnostics for SBI algorithms based on normalizing flows (arXiv:2211.09602)
Linhart J, Gramfort A and Rodrigues P 2024 L-c2st: local diagnostics for posterior approximations in simulation-based inference *Advances in Neural Information Processing Systems* vol 36 (available at: https://papers.neurips.cc/paper_files/paper/2023/file/b0313c2f4501a81d0e0d4a1e8fbf4995-Paper-Conference.pdf)
Lueckmann J-M *et al* 2017 Flexible statistical inference for mechanistic models of neural dynamics *Advances in Neural Information Processing Systems* p 30 (available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/addfa9b7e234254d26e9c7f2af1005cb-Paper.pdf)
Luo R *et al* 2021 Localized calibration: metrics and recalibration *CoRR* (arXiv:2102.10809)
Malz A I and Hogg D W 2022 How to obtain the redshift distribution from probabilistic redshift estimates *Astrophys. J.* **928** 127
Mandelbaum R *et al* 2008 Precision photometric redshift calibration for galaxy-galaxy weak lensing *Mon. Not. R. Astron. Soc.* **386** 781
Mandelbaum R 2018 Weak lensing for precision cosmology *Annu. Rev. Astron. Astrophys.* **56** 393
Matheson J E and Winkler R L 1976 Scoring rules for continuous probability distributions *Manage. Sci.* **22** 1087
McNeely T, Khokhlov P, Dalmasso N, Wood K M and Lee A B 2023a Structural forecasting for short-term tropical cyclone intensity guidance *Weather Forecast.* **38** 985–998
McNeely T, Vincent G, Wood K M, Izbicki R and Lee A B 2023b Detecting distributional differences in labeled sequence data with application to tropical cyclone satellite imagery *Ann. Appl. Stat.* **17** 1260
Mirza M and Osindero S 2014 Conditional generative adversarial nets *CoRR* (arXiv:1411.1784)
Moreira M J 2003 A conditional likelihood ratio test for structural models *Econometrica* **71** 1027
Nichol A Q and Dhariwal P 2021 *Proc. Machine Learning Research, Proc. 38th Int. Conf. on Machine Learning, ICML 2021, (18 July–24 July 2021)* vol 139, ed M Meila and T Zhang (PMLR) pp 8162–71 (available at: http://proceedings.mlr.press/v139/nichol21a.html)
Olander T, Wimmers A, Velden C and Kossin J P 2021 Investigation of machine learning using satellite-based advanced dvorak technique analysis parameters to estimate tropical cyclone intensity *Weather Forecast.* **36** 2161
Papamakarios G and Murray I 2016 Fast ε-free inference of simulation models with bayesian conditional density estimation *Advances in Neural Information Processing Systems* p 29 (available at: https://proceedings.neurips.cc/paper_files/paper/2016/file/6aca97005c68f1206823815f66102863-Paper.pdf)
Papamakarios G, Nalisnick E, Rezende D J, Mohamed S and Lakshminarayanan B 2019 Normalizing flows for probabilistic modeling and inference (arXiv:1912.02762)
Radford A *et al* 2019 Language models are unsupervised multitask learners (available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf)
Rahaman R *et al* 2021 Uncertainty quantification and deep ensembles *Advances in Neural Information Processing Systems* vol 34 p 20063 (available at: https://proceedings.neurips.cc/paper_files/paper/2021/file/a70dc40477bc2adceef4d2c90f47eb82-Paper.pdf)
Ravuri S *et al* 2021 Skilful precipitation nowcasting using deep generative models of radar *Nature* **597** 672
Romano Y, Patterson E and Candès E 2019 Conformalized quantile regression *Advances in Neural Information Processing Systems* vol 32 (Curran Associates, Inc.) pp 3543–53 (available at: https://papers.nips.cc/paper_files/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf)
Sadeh I, Abdalla F B and Lahav O 2016 ANNz2: photometric redshift and probability distribution function estimation using machine learning *Publ. Astron. Soc. Pac.* **128** 104502
Schmidt S J *et al* 2020 Evaluation of probabilistic photometric redshift estimation approaches for the rubin observatory legacy survey of space and time (LSST) *Mon. Not. R. Astron. Soc.* **499** 1587
Schmit T J *et al* 2017 A closer look at the ABI on the GOES-R series *Bull. Am. Meteorol. Soc.* **98** 681
Sohl-Dickstein J, Weiss E A, Maheswaranathan N and Ganguli S 2015 Deep unsupervised learning using nonequilibrium thermodynamics *CoRR* (arXiv:1503.03585)
Stute W and Zhu L-X 2002 Model checks for generalized linear models *Scan. J. Stat.* **29** 535
Talts S, Betancourt M, Simpson D, Vehtari A and Gelman A 2018 Validating Bayesian inference algorithms with simulation-based calibration (arXiv:1804.06788)
Taylor J W and Bunn D W 1999 A quantile regression approach to generating prediction intervals *Manage. Sci.* **45** 225
The LSST Dark Energy Science Collaboration *et al* 2018 The LSST dark energy science Collaboration (DESC) science requirements document (arXiv:1809.01669)
Timmermann A 2000 Density forecasting in economics and finance *J. Forecast.* **19** 231
Torres R *et al* 2024 Model-free local recalibration of neural networks (arXiv:2403.05756)
Van den Oord A *et al* 2016 Conditional image generation with pixelcnn decoders *Advances in Neural Information Processing Systems* p 29 (available at: https://proceedings.neurips.cc/paper_files/paper/2016/file/b1301141feffabac455e1f90a7de2054-Paper.pdf)
van den Oord A *et al* 2016 Pixel recurrent neural networks (available at: https://proceedings.mlr.press/v48/oord16.pdf)
Vaswani A *et al* 2017 *Advances in Neural Information Processing Systems: Annual Conf. on Neural Information Processing Systems 2017, (Long Beach, CA, USA, 4 December–9 December 2017)*, ed I Guyon, U von Luxburg, S Bengio, H M Wallach, R Fergus, S V N Vishwanathan and R Garnett pp 5998–6008 (available at: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html)
Vovk V, Gammerman A and Shafer G 2005 *Algorithmic Learning in a Random World* (Springer) (https://doi.org/10.1007/978-3-031-06649-8)

Wald Y and Globerson A 2017 *Advances in Neural Information Processing Systems: Annual Conf. on Neural Information Processing Systems 2017, (Long Beach, CA, USA, 4 December -9 December 2017)* vol 30, ed I Guyon, U von Luxburg, S Bengio, H M Wallach, R Fergus, S V N Vishwanathan and R Garnett pp 6359–68 (available at: https://proceedings.neurips.cc/paper/2017/hash/3eb414bf1c2a66a09c185d60553417b8-Abstract.html)

Wehenkel A *et al* 2024 Addressing misspecification in simulation-based inference through data-driven calibration (arXiv:2405.08719)

Wehenkel A and Louppe G 2019 *Advances in Neural Information Processing Systems: Annual Conf. on Neural Information Processing Systems 2019, NeurIPS 2019, (Vancouver, BC, Canada, 8 December–14 December, 2019)* vol 34, ed H M Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E B Fox and R Garnett pp 1543–53 (available at: https://proceedings.neurips.cc/paper/2019/hash/2a084e55c87b1ebcdaad1f62fdbbac8e-Abstract.html)

York D G *et al* 2000 The sloan digital sky survey: technical summary *Astron. J.* **120** 1579

Zhao D, Dalmasso N, Izbicki R and Lee A B 2021 *Proc. Machine Learning Research, Proc. 37th Conf. on Uncertainty in Artificial Intelligence* vol 161, ed C de Campos and M H Maathuis (PMLR) pp 1830–40 (available at: https://proceedings.mlr.press/v161/zhao21b.html)

Zhao S, Ma T and Ermon S 2020 Individual calibration with randomized forecasting *Int. Conf. on Machine Learning* (PMLR) pp 11387–97 (available at: https://proceedings.mlr.press/v119/zhao20e/zhao20e.pdf)

Zhou R *et al* 2021 The clustering of DESI-like luminous red galaxies using photometric redshifts *Mon. Not. R. Astron. Soc.* **501** 3309