Parity Calibration

Youngseog Chung[1] **Aaron Rumack**[1] **Chirag Gupta**[1]

[1]Machine Learning Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Abstract

In a sequential regression setting, a decision-maker may be primarily concerned with whether the future observation will increase or decrease compared to the current one, rather than the actual value of the future observation. In this context, we introduce the notion of parity calibration, which captures the goal of calibrated forecasting for the increase-decrease (or “parity") event in a time-series. Parity probabilities can be extracted from a forecasted distribution for the output, but we show that such a strategy leads to theoretical unpredictability and poor practical performance. We then observe that although the original task was regression, parity calibration can be expressed as binary calibration. Drawing on this connection, we use an online binary calibration method to achieve parity calibration. We demonstrate the effectiveness of our approach on real-world case studies in epidemiology, weather forecasting, and model-based control in nuclear fusion.

1 INTRODUCTION

Many tasks in the scope of prediction and decision making are sequential in nature. A weather forecaster who uses some procedure to make predictions for tomorrow, may find that tomorrow's events falsify these predictions. A good forecaster must then update their model before using it on the following days. In this paper we study the sequential forecasting setting where the goal is to make predictions about a sequence of real-valued outcomes $y_1, y_2, \ldots \in \mathcal{Y} \subseteq \mathbb{R}$ using informative covariates $\mathbf{x}_1, \mathbf{x}_2, \ldots \in \mathcal{X}$. In the presence of inherent stochasticity or insufficient data, forecasters who provide rich predictions in the form of complete distributions over the output allow us to reason about the inherent uncertainties in the data stream [Gneiting et al., 2007]. If a distributional prediction is available, a downstream decision-maker can account for risks that were unknown at the time of forecasting.

Often, a distributional forecast for the real-valued y_t takes the form of a predictive cdf (cumulative distribution function) for y_t, which in this paper we typically denote as $\hat{F}_t : \mathcal{Y} \to [0, 1]$. We sometimes write $\hat{F}_t$ as $\hat{F}_t(\cdot|\mathbf{x}_t)$ or $\hat{F}_t(\cdot|\mathbf{x}_t, y_{t-1}, \mathbf{x}_{t-1}, \ldots, y_1, \mathbf{x}_1)$; this overloaded notation allows us to be succinct when defining what it means for $\hat{F}_t$ to be calibrated, but explicit when it is necessary to stress that $\hat{F}_t$ depends on all available knowledge. We also refer to $\hat{F}_t$'s as regression forecasts, as it models a continuous distribution over the real-valued output.

In this paper, we are interested in the question: can we forecast whether the future outcome y_{t+1} will be greater or less than the current outcome y_t? To motivate this question, consider a hospital in the midst of a fast moving pandemic such as COVID-19. It may be difficult for the hospital to comprehend absolute numbers of patients requiring hospitalization. However, relative numbers are perhaps easier to interpret: hospitals know the situation today, and would like to know if it is going to worsen or improve tomorrow.

A domain expert (e.g. epidemiologist) may have produced a regression forecast $\hat{F}_t$ for y_t. The downstream user (e.g. hospital) can then extract from $\hat{F}_t$ a natural implied probability of the next observation decreasing:

$$\text{for } t \geq 2,\ \hat{p}_t = \hat{F}_t(y_{t-1} \mid \mathbf{x}_t). \tag{1}$$

The hope of the hospital is that the forecasted probabilities $\hat{p}_t$ are parity calibrated, as defined next.

Definition 1 (Parity calibration)**.** *The forecasts* $\{\hat{p}_t \in [0, 1]\}_{t=2,\ldots,T}$ *are said to be parity calibrated if*

$$\frac{\sum_{t=2}^{T} \mathbb{1}\{y_t \leq y_{t-1}\}\mathbb{1}\{\hat{p}_t = p\}}{\sum_{t=2}^{T} \mathbb{1}\{\hat{p}_t = p\}} \to p, \forall p \in [0, 1]. \tag{2}$$

In words, whenever a parity calibrated forecaster predicts with probability p that $y_t \leq y_{t-1}$, the event $\mathbb{1}\{y_t \leq y_{t-1}\}$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

actually occurs with empirical frequency p (in the long run). To avoid confusion with usage of the term "parity" in fairness literature, we remark that our context is purely in comparing two consecutive values.

Our first contribution is showing that even if $\hat{F}_t$ is calibrated (based on some accepted notions of calibration), the seemingly reasonable strategy mentioned above (1) can have devastating and unpredictable behavior (Section 1.1). Yet, it stands to reason that the expert's rich forecast $\hat{F}_t$ should be used in some way. Our second contribution is a methodology for doing this (Sections 2 and 3). Our main methodology described in Section 2.2 is based on the key observation that although the parity calibration problem is derived from a regression problem, it naturally reduces to a problem of forecasting binary events.

1.1 REGRESSION CALIBRATION DOES NOT GIVE PARITY CALIBRATION

A popular notion of calibration in regression is *probabilistic calibration* [Gneiting et al., 2007]. The sequence $\hat{F}_1, \hat{F}_2, \ldots$ is said to be probabilistically calibrated if

$$\frac{1}{T}\sum_{t=1}^{T} F_t(\hat{F}_t^{-1}(p)) \to p, \ \forall p \in [0,1], \tag{3}$$

where F_t denotes the ground truth distribution. Probabilistic calibration is also referred to as *quantile calibration*, since it focuses on the quantile function being valid. In other works, it has also been referred to as average calibration [Zhao et al., 2020, Chung et al., 2021b, Sahoo et al., 2021], or simply calibration [Kuleshov et al., 2018, Cui et al., 2020, Charpentier et al., 2022, Marx et al., 2022]. We will henceforth refer to this notion as *quantile calibration*.

Another notion of calibration in regression is *distributional calibration* [Song et al., 2019], which assesses the convergence of the full distribution of the observations to the predictive distribution. A distribution calibrated forecaster satisfies $\forall p \in [0,1], \ \forall F \in \mathcal{F}$,

$$\frac{\sum_{t=1}^{T} \mathbb{1}\{\hat{F}_t = F\} F_t(\hat{F}_t^{-1}(p))}{\sum_{t=1}^{T} \mathbb{1}\{\hat{F}_t = F\}} \to p, \tag{4}$$

where $\mathcal{F}$ is the space of distributions predicted by $\hat{F}_t$. However, distributional calibration is an idealistic notion that cannot be achieved in practice [Song et al., 2019].

Recently, Sahoo et al. [2021] paired calibration with the notion of threshold decisions and proposed *threshold calibration*. Forecasts are said to be threshold calibrated if,

$$\frac{\sum_{t=1}^{T} \mathbb{1}\{\hat{F}_t(y_0) \leq \alpha\} F_t(\hat{F}_t^{-1}(p))}{\sum_{t=1}^{T} \mathbb{1}\{\hat{F}_t(y_0) \leq \alpha\}} \to p,$$
$$\forall y_0 \in \mathcal{Y}, \ \forall \alpha \in [0,1], \forall p \in [0,1].$$

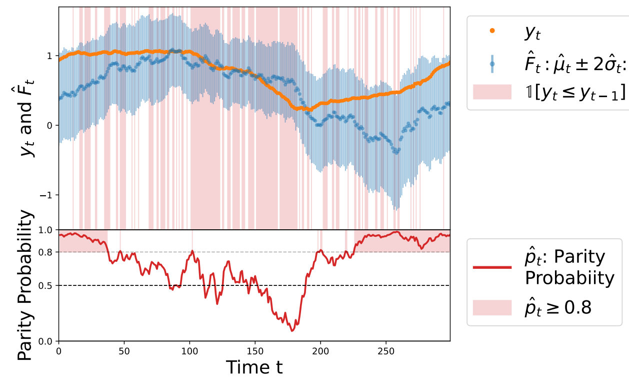


Figure 1: Snapshot of the first 300 points from one of our experiment datasets (Pressure from Section 3.2) shows a quantile calibrated forecaster that is highly parity miscalibrated. (**top**) The expert forecasts $\hat{F}_t$ are Gaussians, expressed in the plot as prediction intervals $[\hat{\mu}_t - 2\hat{\sigma}_t, \hat{\mu}_t + 2\hat{\sigma}_t]$. This prediction interval almost always contains y_t and its reliability diagram in Figure 4 (plot titled "Quantile Calibration") confirms that $\hat{F}_t$ is in fact quantile calibrated when considering the full timeseries. (**bottom**) For $t \in [0, 40]$ and $t \in [230, 300]$, the parity probabilities $\hat{p}_t = \hat{F}_t(y_{t-1})$ assign ≥ 0.8 probability (red shaded areas) to $\mathbb{1}\{y_t \leq y_{t-1}\}$. But y_t actually decreases with much lower frequency during these timesteps as can be seen from the top figure. The parity miscalibration when considering the full timeseries is confirmed by Figure 4 (plot titled "Prehoc").

Sahoo et al. [2021] show that distribution calibration implies threshold calibration, but the converse may not hold.

A common aspect of the aforementioned notions of calibration is that they all assess how well-aligned the predictive quantiles are to their empirical counterparts. The key difference among the notions is the conditioning over which this assessment is performed.

Since calibration is regarded as a desirable quality of distributional forecasts, one may wonder whether a calibrated $\hat{F}_t$ is sufficient for parity calibration of the implied probabilities as per Eq. (1). We show that this is *not* the case with the following examples.

Synthetic example. Let $\mathcal{N}_-$ and $\mathcal{N}_+$ denote the standard normal distributions truncated at 0, with density functions $f_-(x) = \mathbb{1}\{x < 0\}\sqrt{2/\pi}e^{-x^2/2}$ and $f_+(x) = \mathbb{1}\{x \geq 0\}\sqrt{2/\pi}e^{-x^2/2}$ respectively. Let F_- and F_+ be the cdfs of $\mathcal{N}_-$ and $\mathcal{N}_+$. Suppose the target sequence $(Y_t)_{t=1}^{\infty}$ is distributed as

$$Y_t \sim \begin{cases} \mathcal{N}_- & \text{if } t \text{ is odd,} \\ \mathcal{N}_+ & \text{if } t \text{ is even.} \end{cases}$$

Consider the following predictive cdf targeting Y_t,

$$\hat{F}_t = \frac{1}{2}F_- + \frac{1}{2}F_+ = \begin{cases} \frac{1}{2}F_-(y), \text{ if } y < 0, \\ 0.5 + \frac{1}{2}F_+(y), \text{ if } y \geq 0. \end{cases}$$

We note that when $y < 0, \frac{1}{2}F_-(y) \in [0, 0.5)$, and when $y \geq 0.5, 0.5 + \frac{1}{2}F_+(y) \in [0.5, 1]$. It can be verified that the corresponding quantile function is

$$\hat{F}_t^{-1}(p) = \begin{cases} F_-^{-1}(2p), \text{ if } p < 0.5 \\ F_+^{-1}(2p-1), \text{ if } p \geq 0.5. \end{cases}$$

We verify that $\hat{F}_t$ is quantile calibrated (following Eq. (3)).

When t is odd, $F_t = F_-$.

- $\forall p \in [0, 0.5)$, $F_t(\hat{F}_t^{-1}(p)) = F_-(F_-^{-1}(2p)) = 2p$.
- $\forall p \in [0.5, 1]$, $\hat{F}_t^{-1}(p) = F_+^{-1}(2p-1) \geq 0$, thus $F_t(\hat{F}_t^{-1}(p)) = F_-(F_+^{-1}(2p-1)) = 1$.

When t is even, $F_t = F_+$.

- $\forall p \in [0, 0.5)$, $\hat{F}_t^{-1}(p) = F_-^{-1}(2p) < 0$, thus $F_t(\hat{F}_t^{-1}(p)) = F_+(F_-^{-1}(2p)) = 0$.
- $\forall p \in [0.5, 1]$, $F_t(\hat{F}_t^{-1}(p)) = F_+(F_+^{-1}(2p-1)) = 2p - 1$.

Therefore, for $p \in [0, 0.5)$, $\frac{1}{T}\sum_{t=1}^{T} F_t(\hat{F}_t^{-1}(p)) = \frac{1}{T}\sum_{t \text{ is odd}} 2p = p + o(\frac{1}{T}) \to p$, and the same can be verified for $p \in [0.5, 1]$, showing that $\hat{F}_t$ *is quantile calibrated.*

We can easily show that $\hat{F}_t$ is also distribution and threshold calibrated. Since $\hat{F}_t$ is constant for all t, following Eq. (4), the space of predicted distributions is a singleton. Thus, measuring distribution calibration is equivalent to measuring quantile calibration, and $\hat{F}_t$ *is distribution calibrated*. Since distribution calibration implies threshold calibration [Sahoo et al., 2021], $\hat{F}_t$ *is threshold calibrated.*

However, as we show next, $\hat{F}_t$ is not parity calibrated.

When t is odd, $Y_t \sim F_-$ and $Y_{t-1} \sim F_+$. Thus $Y_t < Y_{t-1}$ whereas $\hat{p}_t = \hat{F}_t(Y_{t-1}) \geq 0.5$.

When t is even, $Y_t \sim F_+$ and $Y_{t-1} \sim F_-$. Thus $Y_t > Y_{t-1}$ whereas $\hat{p}_t = \hat{F}_t(Y_{t-1}) < 0.5$.

Therefore, $\forall \hat{p}_t \geq 0.5$, $\mathbb{1}\{y_t \leq y_{t-1}\} = 1$ and $\forall \hat{p}_t < 0.5$, $\mathbb{1}\{y_t \leq y_{t-1}\} = 0$, thus $\hat{F}_t$ *is parity miscalibrated* for all $\hat{p}_t \in (0, 1)$, i.e. all $\hat{p}_t \neq 0$ or 1. $\square$

Intuitively, the sequential aspect of predictions and observations is central to the notion of parity calibration, whereas traditional notions of calibration effectively treat the datapoints as an i.i.d. or exchangeable batch of points. Figure 1 provides a visualization of how this pitfall can be manifested in a practical example.

The implication is that methods designed to achieve traditional notions of calibration in regression cannot be expected to provide parity calibration. The following section introduces the posthoc binary calibration framework that can instead be used to achieve parity calibrated forecasts.

2 PARITY CALIBRATION VIA BINARY CALIBRATION

Define the *parity outcomes* as

$$\text{for } t \geq 2,\ \widetilde{y}_t := \mathbb{1}\{y_t \leq y_{t-1}\}, \tag{5}$$

and observe that the parity calibration condition (Eq. (2)) is equivalently written as,

$$\frac{\sum_{t=2}^{T} \widetilde{y}_t \mathbb{1}\{\hat{p}_t = p\}}{\sum_{t=2}^{T} \mathbb{1}\{\hat{p}_t = p\}} \to p, \forall p \in [0, 1]. \tag{6}$$

Thus parity calibration is in fact targeting the binary sequence $\widetilde{y}_t$, instead of y_t. In this section, we show how this connection allows us to leverage powerful techniques from the rich literature of binary calibration that goes back four decades [DeGroot and Fienberg, 1981, Dawid, 1982, Foster and Vohra, 1998]. Of specific interest to us will be a class of methods that have been proposed for *posthoc calibration* of machine learning (ML) classifiers, which we review next.

2.1 POSTHOC BINARY CALIBRATION

Let $f : \mathcal{X} \to [0, 1]$ be a binary classifier that takes as input a feature vector in feature space $\mathcal{X}$ and outputs a score in $[0, 1]$. Suppose a feature-label pair (X, Y) is drawn from some distribution P over $\mathcal{X} \times \{0, 1\}$. Then, f is said to be calibrated (in the binary sense) if

$$P(Y = 1 \mid f(X)) = f(X). \tag{7}$$

The terms on either side of the equal sign are random variables and the equality is understood almost-surely. The connection between (6) and (7) is evident: $\hat{p}_t$ is like $f(X)$, conditioning on the random variable $f(X)$ is akin to using indicators in the numerator/denominator, and $\widetilde{y}_t$ is like Y.

We do not expect ML models to be calibrated "out-of-the-box". So, if f is a logistic regression or neural network trained on some training data, it is unlikely to satisfy an approximate version of (7) on unseen data. Posthoc calibration techniques transform f to a function that is better calibrated by using a so-called *calibration dataset* $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_c, y_c)\}$. $\mathcal{D}_{\text{cal}}$ is a set of points on which f was not trained—in practice $\mathcal{D}_{\text{cal}}$ is often just the validation dataset. $\mathcal{D}_{\text{cal}}$ is used to a learn a mapping $m : [0, 1] \to [0, 1]$ so that $m \circ f$ is better calibrated than f. By way of an example, we now introduce the popular Platt scaling technique [Platt, 1999] that will be central to this paper (henceforth, Platt scaling is referred to as PS). Given a pair of real numbers $(a, b) \in \mathbb{R}^2$, the PS mapping $m^{a,b} : [0, 1] \to [0, 1]$ is defined as,

$$m^{a,b}(z) = \text{sigmoid}(a \cdot \text{logit}(z) + b).$$

Here $\text{logit}(z) = \log(\frac{z}{1-z})$ and $\text{sigmoid}(z) = 1/(1 + e^{-z})$ are inverses of each other. Thus PS is a logistic model on top of the f-induced one-dimensional feature $\text{logit}(f(x)) \in [0, 1]$, instead of on the raw feature $x \in \mathcal{X}$. In the posthoc setting, (a, b) are set to the values that minimize log-loss (equivalently cross entropy loss) on $\mathcal{D}_{\text{cal}}$:

$$(\widehat{a}, \widehat{b}) = \underset{(a,b) \in \mathbb{R}^2}{\arg\min} \sum_{(\mathbf{x}_s, y_s) \in \mathcal{D}_{\text{cal}}} l(m^{a,b}(f(\mathbf{x}_s)), y_s), \tag{8}$$

where $l(p, y) = -y \log p - (1 - y) \log(1 - p)$.

We briefly note some other popular posthoc calibration methods. These broadly fall under two categories: parametric scaling methods such as beta scaling [Kull et al., 2017], temperature scaling [Guo et al., 2017], and PS [Platt, 1999]; and nonparametric methods such as binning [Zadrozny and Elkan, 2001, Gupta et al., 2020, Gupta and Ramdas, 2021], isotonic regression [Zadrozny and Elkan, 2002], and Bayesian binning [Naeini et al., 2015].

2.2 PARITY CALIBRATION USING ONLINE VERSIONS OF PLATT SCALING (PS)

To achieve parity calibration using posthoc techniques, we start with a base cdf predictor $G : \mathcal{X} \to \Delta(\mathcal{Y})$ derived from an expert—such as an epidemiologist, a weather forecaster, or a stock trader. Here, $\Delta(\mathcal{Y})$ refers to the space of distributions over $\mathcal{Y}$. If the expert is an ML engineer, such a G can be obtained using Gaussian processes [Rasmussen, 2004] or probabilistic neural networks [Nix and Weigend, 1994, Lakshminarayanan et al., 2017], among other methods. The test-stream occurs after G has been trained and fixed. This G gives us a $\hat{F}_t$ as described in the introduction: $\hat{F}_t = G(\mathbf{x}_t)$. Recall that the strategy Eq. (1) is to forecast $\hat{p}_t = \hat{F}_t(y_{t-1})$. If $\hat{F}_t$ were the true cdf of y_t given the past, the above $\hat{p}_t$ would be the true probability of $\widetilde{y}_t = 1$, and thus the most useful parity forecast possible.

However, in Section 1.1 we showed that we must modify $\hat{p}_t$ in order to achieve parity calibration. We propose using PS to perform this modification (any posthoc calibration method can be used; we focus on PS in this paper). A natural possibility would be to use an initial part of the test-stream to learn fixed PS parameters once, as described in the previous subsection. However, real-world regression sequences (weather, stocks, etc) have non-stationary shifting behavior across time. Therefore, a fixed model is unlikely to remain calibrated over time.

In Algorithm 1 we outline three ways to mitigate this. Increasing Window (IW) updates the PS parameters using all datapoints until some recent time step, such as every 100 timesteps ($t = 100, 200$, etc). A related alternative, Moving Window (MW) is to use only the most recent datapoints when updating the PS parameters (instead of all the points). The third alternative is Online Platt Scaling (OPS) based on our own recent work [Gupta and Ramdas, 2023].

Algorithm 1 Platt scaling (PS) variants for parity calibration

Input: Any base forecaster $G : \mathcal{X} \to \Delta(\mathcal{Y})$, covariate-outcome pairs $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots \in \mathcal{X} \times \mathcal{Y}$, update-frequency `uf`, moving-window-size `ws`.
Output: PS forecasts $(\hat{p}_t^{\text{IW}}, \hat{p}_t^{\text{MW}}, \hat{p}_t^{\text{OPS}})_{t=2}^{\infty}$
Initialize IW, MW, OPS parameters:
  $(a^{\text{IW}}, b^{\text{IW}}) = (a^{\text{MW}}, b^{\text{MW}}) = (a^{\text{OPS}}, b^{\text{OPS}}) \leftarrow (1, 0)$
for $t = 2$ **to** T **do**
  $\widetilde{y}_t = \mathbb{1}\{y_t \leq y_{t-1}\}$
  $\hat{p}_t = G(\mathbf{x}_t)[y_{t-1}]$
  $\hat{p}_t^{\text{IW}} \leftarrow \text{sigmoid}(a^{\text{IW}} \cdot \text{logit}(\hat{p}_t) + b^{\text{IW}})$
  $\hat{p}_t^{\text{MW}} \leftarrow \text{sigmoid}(a^{\text{MW}} \cdot \text{logit}(\hat{p}_t) + b^{\text{MW}})$
  $\hat{p}_t^{\text{OPS}} \leftarrow \text{sigmoid}(a^{\text{OPS}} \cdot \text{logit}(\hat{p}_t) + b^{\text{OPS}})$
  if t is a multiple of `uf` **then**
    $(a^{\text{IW}}, b^{\text{IW}}) \leftarrow$ optimal PS parameters based on (8) setting $\mathcal{D}_{\text{cal}} = (\mathbf{x}_s, \widetilde{y}_s)_{s=1}^{t}$
    $(a^{\text{MW}}, b^{\text{MW}}) \leftarrow$ optimal PS parameters based on (8) setting $\mathcal{D}_{\text{cal}} = (\mathbf{x}_s, \widetilde{y}_s)_{s=t-\texttt{ws}+1}^{t}$
  end if
  $(a^{\text{OPS}}, b^{\text{OPS}}) \leftarrow \text{OPS}((\mathbf{x}_1, \widetilde{y}_1), \ldots, (\mathbf{x}_t, \widetilde{y}_t))$
  (OPS is Algorithm 2 in Appendix D)
end for

In the following section, we compare these online versions of Platt scaling on three real-world sequential prediction tasks. We find that OPS performs better than the base model, MW, and IW, across multiple settings. Further, while MW and IW involve re-fitting the PS parameters from scratch, OPS makes a constant time update at each step, hence the overall computational complexity of OPS is $O(T)$.

Brief note on theory and limitations of OPS. OPS satisfies a regret bound with respect to the Platt scaling class for log-loss [Gupta and Ramdas, 2023, Theorem 2.1]. This means that the OPS forecasts do as well as forecasts of the single best Platt scaling model in hindsight. However, we note that OPS could fail if the best Platt scaling model is itself not good. This limitation can be overcome by combining OPS with a method called calibeating, as discussed in Gupta and Ramdas [2023]. We do not pursue calibeating in this paper since OPS already performs well on the data we considered.

3 REAL-WORLD CASE STUDIES

We study parity calibration in three real-world scenarios: 1) forecasting COVID-19 cases in the United States, 2) forecasting weather, and 3) predicting plasma state evolution in nuclear fusion experiments. This diverse set of domains, datasets, and expert forecasters provides an attractive testbed to demonstrate the parity calibration concept and the performance of the calibration methods from Section 2.2.

In each setting, the prediction target is real-valued, and we assume an expert forecaster provides regression forecasts $\hat{F}_t$ for the target. We also refer to $\hat{F}_t : \mathcal{Y} \to [0, 1]$ as the *base regression model*. The expert forecaster implicitly provides parity probabilities $\hat{p}_t$ (following Eq. (1)). We refer to $\hat{p}_t$ as the *prehoc* probabilities, in contrast to the *posthoc* probabilities that the calibration methods produce. We calibrate $\hat{p}_t$ with the calibration methods from Section 2.2 to produce the posthoc probabilities $\hat{p}'_t$. Each calibration method requires a set of hyperparameters, which we tune with a validation set. Details regarding hyperparameter tuning are provided in Appendix C.

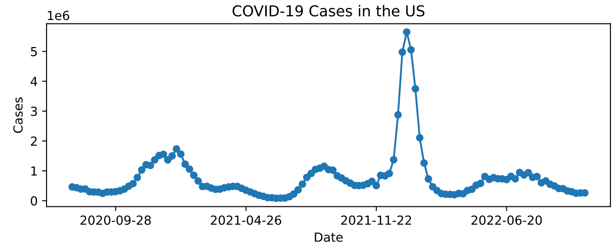
Metrics. Given a test dataset $\mathcal{D}_{\text{test}} = \{\mathbf{x}_t, y_t\}_{t=1}^T$, we initially assess the quantile calibration of $\hat{F}_t$ and the parity calibration of $\hat{p}_t$ and $\hat{p}'_t$ by visualizing the reliability diagrams and measuring calibration errors.

To assess quantile calibration of $\hat{F}_t$, we produce the reliability diagram using the Uncertainty Toolbox [Chung et al., 2021a], which takes a finite set of quantile levels $\mathcal{P} = \{p_i \in [0, 1]\}$, computes the empirical coverage of the predictive quantile $\hat{F}_t^{-1}(p_i)$ as $p_{i,\text{obs}} = \frac{1}{T}\sum_{t=1}^T \mathbb{1}\{y_t \leq \hat{F}_t^{-1}(p_i)\}$, and plots each p_i against $p_{i,\text{obs}}$. Calibration error is then summarized into a single scalar with Quantile Calibration Error (QCE), which is computed as $\frac{1}{|\mathcal{P}|}\sum_i \mid p_{i,\text{obs}} - p_i \mid$. In our experiments, we set $\mathcal{P}$ to be 100 equi-spaced quantile levels in $[0, 1]$.

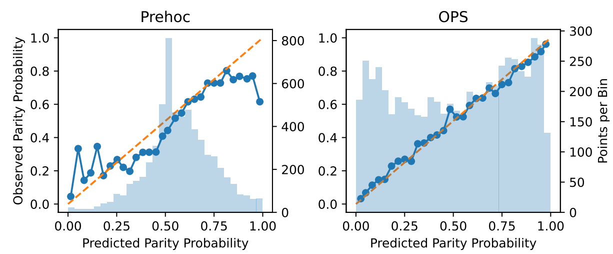
To assess parity calibration of a parity probability $\hat{p}_t$, we follow the standard method of producing reliability diagrams in binary calibration [DeGroot and Fienberg, 1981, Niculescu-Mizil and Caruana, 2005]. Noting that $\hat{p}_t$ is a predicted probability of the binary parity outcome $\widetilde{y}_t := \mathbb{1}\{y_t \leq y_{t-1}\}$, we first bin $\hat{p}_t$ into a finite set of fixed width bins $\mathcal{B} = \{B_m\}$, then for each bin B_m, we compute the average outcome as $\text{obs}(B_m) = \frac{1}{|B_m|}\sum_{t:\hat{p}_t \in B_m} \mathbb{1}\{\widetilde{y}_t = 1\}$ and the average prediction as $\text{pred}(B_m) = \frac{1}{|B_m|}\sum_{t:\hat{p}_t \in B_m} \hat{p}_t$, and finally, we plot $\text{pred}(B_m)$ against $\text{obs}(B_m)$ to produce the reliability diagram. Parity Calibration Error (PCE) summarizes the diagram following the standard definition of (ℓ_1-)expected calibration error (ECE): $\sum_m \frac{|B_m|}{T} \mid \text{obs}(B_m) - \text{pred}(B_m) \mid$. In our experiments, we set $\mathcal{B}$ to be 30 fixed-width bins: $[0, \frac{1}{30}), [\frac{1}{30}, \frac{2}{30}), \ldots [\frac{29}{30}, 1]$.

For the parity probabilities $\hat{p}_t$ and $\hat{p}'_t$, we additionally report sharpness and two metrics for accuracy: binary accuracy and area under the ROC curve. Sharpness (Sharp) is computed as $\sum_m \frac{|B_m|}{T} \cdot \text{obs}(B_m)^2$ and measures the degree to which the forecaster can discriminate events with different outcomes [Bröcker, 2009]. Binary accuracy (Acc) and area under the ROC curve (AUROC) are computed following their standard definitions in binary classification. Appendix A provides the full set of details on how each metric is computed. Lastly, in reporting the metrics in numeric tables, we denote each metric with their orientation, e.g. ↑ indicates that a higher value is more desirable and vice versa.

Code is available at https://github.com/YoungseogChung/parity-calibration



(a) Total COVID-19 cases in the US displays high non-stationarity.



(b) Reliability diagrams for the prehoc parity probabilities from the expert forecasts (**left**) and OPS calibrated probabilities (**right**). **Blue bars** denote the frequency of predictions in each bin.

Figure 2: The prehoc parity probabilities for the COVID-19 single-timeseries setting are miscalibrated and un-sharp. Posthoc calibration via OPS improves both aspects.

3.1 CASE STUDY 1: COVID-19 CASES IN THE US

In response to the COVID-19 pandemic, research groups across the world have created models to predict the short-term future of the pandemic. The COVID-19 Forecast Hub [Cramer et al., 2021] solicits and collects quantile forecasts of weekly incident COVID-19 cases in each US state (plus Washington D.C.), among other targets. Each week, the Hub generates an ensemble forecast from the dozens of submitted forecasts. This ensemble has proven to be more reliable and accurate than any constituent individual forecast in predicting other targets of interest (e.g. mortality [Cramer et al., 2022]). Thus, we take the ensemble forecast as the expert forecast and use its historical forecasts made between 2020-07-20 and 2022-10-24, which span a total of 119 weeks. Denoting the target y as the number of cases, there are effectively 51 timeseries, $\{y_{s,t}\}$: one for each US state $s \in$ {Alabama, Alaska, Arizona, ..., Wisconsin, Wyoming}, and $t \in \{1, \ldots, 119\}$. For any given s, t, the expert forecast is provided by the Hub as seven forecasted quantiles for the distribution of $y_{s,t}$. Therefore, we must interpolate the

	Prehoc	$\text{OPS}_{\text{alpha-order}}$	$\text{OPS}_{\text{rand100}}$
PCE $\downarrow$	0.0599	0.0216	0.0246 ± 0.0002
Sharp $\uparrow$	0.2953	0.3087	0.3090 ± 0.00002
Acc $\uparrow$	0.6309	0.6727	0.6737 ± 0.0001
AUROC $\uparrow$	0.6922	0.7355	0.7357 ± 0.00002

Table 1: In the COVID-19 single-timeseries setting, OPS improves the prehoc parity probabilities w.r.t all metrics. $\pm$ indicates mean $\pm$ 1 standard error across 100 state orders.

	Prehoc	MW	IW	OPS
PCE $\downarrow$	0.0599	0.0748	0.0406	**0.0328**
Sharp $\uparrow$	0.2953	0.2882	0.2839	**0.2993**
Acc $\uparrow$	0.6309	0.6237	0.6055	**0.6522**
AUROC $\uparrow$	0.6922	0.6622	0.6403	**0.7035**

Table 2: In the COVID-19 sequential-batch setting, OPS outperforms prehoc and alternative PS methods. Best value for each metric is in bold.

quantiles to produce $\hat{F}_t$ (see Appendix B.1 for details).

The observed targets $y_{s,t}$ are the incident number of cases actually reported from each state, for each week. Figure 2a visualizes a summary of the target timeseries: the total incident number of cases in the US ($= \sum_s y_{s,t}$). We can observe high non-stationarity, with periods of rapid increases and falls, and other periods of long monotonic trends.

3.1.1 Parity calibration of expert forecasts and OPS

Note that the underlying timeseries $\{y_{s,t}\}$ is indexed by both state and time. We transform this to a fully sequential timeseries by concatenating $\{y_{s,t}\}$ chronologically across t and in alphabetical order across s. In other words, within a given week, we observe the number of cases for the states in alphabetical order. We refer to this experiment setting as the *single-timeseries* setting.

The reliability diagram in Figure 2b (left) shows that the prehoc probabilities implied by the expert forecast ($\hat{p}_t$) are parity calibrated in the $[0.25, 0.75]$ region (i.e. higher predicted probabilities result in higher empirical frequencies), but are miscalibrated otherwise. The distribution of $\hat{p}_t$ displayed by the blue bars further indicate that $\hat{p}_t$ is centered around 0.5, an uninformative or less sharp prediction.

Figure 2b (right) displays the reliability diagram of $\hat{p}_t^{\text{OPS}}$. We observe significant improvements in both parity calibration and sharpness, i.e. $\hat{p}_t^{\text{OPS}}$ is much more dispersed compared to $\hat{p}_t$. The second column of Table 1 (labeled $\text{OPS}_{\text{alpha-order}}$) show these improvements via the PCE and Sharp metrics, and we can also observe improvement in accuracy.

One may question whether this improvement by OPS is

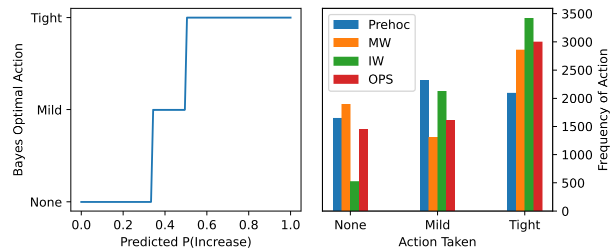


Figure 3: (Decision making on the COVID-19 dataset) (**left**) The Bayes optimal action for each predicted probability of increase in number of cases. (**right**) Frequency of each action taken by each method.

specific to the alphabetical order of states. In the third column of Table 1 (labeled $\text{OPS}_{\text{rand100}}$), we show the mean and standard error of each of the metrics across 100 different random orders of the states, and observe that the improvements provided by OPS over prehoc are fairly robust.

3.1.2 Comparing calibration methods

We perform an additional experiment to compare the performance of MW, IW and OPS. In this experiment, we assume a more realistic test setting for the data-stream. At each timestep t, we assume we observe cases from all 51 states, $\{y_{s,t}\}_{s=1}^{51}$, and update the PS parameters with this batch of data. We then fix the PS parameters and calibrate the next full batch of predictions for timestep $t+1$. This settings assumes that PS parameters are updated once per week based on all the data observed during the week. We refer to this experiment setting as the *sequential-batch* setting.

The first 20 weeks of data (i.e. 20 weeks $\times$ 51 states = 1020 datapoints) were used to tune the hyperparameters of each method. The subsequent 99 weeks of data was used for testing. Table 2 displays the results of the sequential batch setting (note that the prehoc values are the same for this setting as in Table 1). OPS is the best performing method on all metrics when compared with MW, IW, and prehoc.

3.1.3 Decision-making with parity probabilities

In this section, we demonstrate the utility of OPS in a decision-making setting where parity outcomes (Eq. (5)) dictate the loss incurred. Using the same COVID-19 dataset, we assume a setting where a policymaker (i.e. the decision-maker) at each timestep must decide among three levels of restrictions for disease spread prevention: Tight, Mild, or None. For any chosen level of restriction, the loss is dictated by the parity outcome in the number of cases, and the policymaker's goal is to minimize cumulative loss. A *Bayes optimal* policymaker will always choose an action which minimizes the expected loss, calculated with a predic-

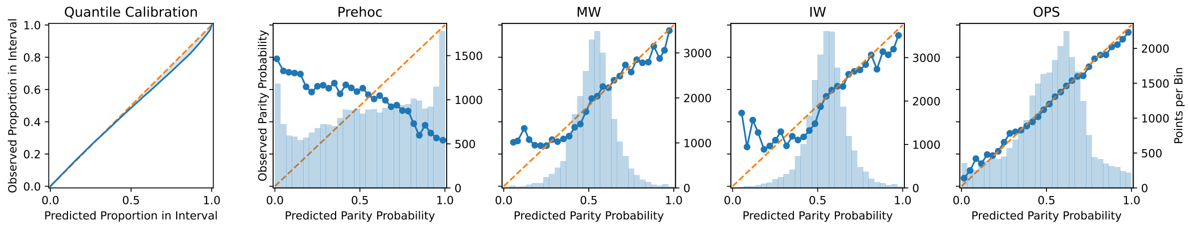


Figure 4: OPS significantly improves both parity calibration and sharpness of the base regression model predicting Pressure. The left two plots display the quantile calibration and parity calibration of the base model (Prehoc): it is nearly perfectly quantile calibrated, but terribly parity calibrated. **Blue bars** denote the frequency of predictions in each bin.

tive distribution over the loss [Lehmann and Casella, 2006]. Hence the policymaker will assess the optimality of each action based on predicted parity probabilities.

We design an exemplar loss function $l_{\text{truth, decision}}$ as follows:

# Cases	Tight = 1	Mild = 2	None = 3
Increase = 1	$l_{1,1} = 0.3$	$l_{1,2} = 0.6$	$l_{1,3} = 1$ (max)
Decrease = 2	$l_{2,1} = 0.5$	$l_{2,2} = 0.2$	$l_{2,3} = 0$ (min)

Given this loss function, the Bayes optimal action is visualized in Figure 3 (left). On computing the the cumulative loss incurred with the predicted parity probabilities, we find that OPS incurs the lowest cumulative loss.

	Prehoc	MW	IW	OPS
Loss ↓	2119	2177	2196	**2050**

Figure 3 (right) shows the frequency of each action chosen by each method. We observe that OPS chooses Mild with relatively low frequency, which is a result of sharper and more accurate parity probabilities. We further note that IW results in a worse loss than prehoc despite being better parity calibrated (Table 2). To understand this, notice that IW is also less sharp and less accurate than Prehoc. Thus calibration, while a desirable quality, is not the only aspect to assess for good uncertainty quantification—sharpness and accuracy could also affect decision making.

3.2 CASE STUDY 2: WEATHER FORECASTING

Our second case study examines weather forecasting using the benchmark Jena climate modeling dataset [2016], which records the weather conditions in Jena, Germany, with 14 different measurements, in 10 minute intervals, for the years 2009—2016. We did not have access to historical predictions from an expert weather forecaster, so instead we trained our own base regression model.

We follow the Keras tutorial on *Timeseries Forecasting for Weather Prediction*[1] to define our specific problem setup and

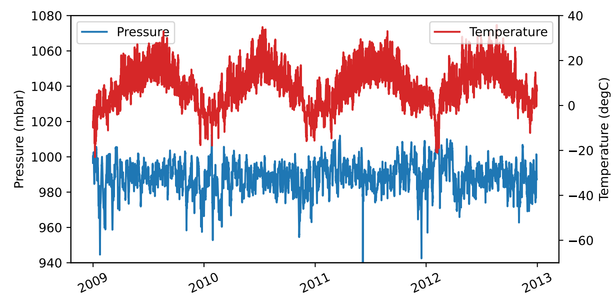


Figure 5: Snapshots of 4 years from the Temperature and Pressure timeseries display noise around a cyclical trend.

train our base regression model. In summary, the regression model is implemented with an LSTM network [Hochreiter and Schmidhuber, 1997] which predicts the mean and variance of a Gaussian distribution. We trained 7 different models that each predict one of 7 weather features: Pressure, Temperature, Saturation vapor pressure, Vapor pressure deficit, Specific humidity, Airtight, and Wind speed. Appendix B.2.1 provides more details on the problem setup.

Lastly, we note that unlike the COVID-19 data, the weather data (Figure 5) displays high levels of noise around a cyclical, repeating trend.

Results on Pressure timeseries. We first examine results from one of the 7 models predicting Pressure. Figure 4 displays quantile calibration (i.e. probabilistic calibration) of the base model, and parity calibration before and after MW, IW and OPS are applied to the prehoc parity probabilities. We first note that the base model is almost perfectly quantile calibrated, but terribly parity calibrated, which corroborates our argument from Section 1.1, that calibration in regression does not imply parity calibration. In the same plot, we can see that MW, IW and OPS are all able to improve parity calibration, but the numerical results in Table 3 show that OPS produces superior parity probabilities w.r.t. all of the metrics considered.

Binary classifiers as expert forecasters. While we have so far assumed that the expert forecaster provides regression models $\hat{F}_t$, one may argue that an expert forecaster

[1] https://keras.io/examples/timeseries/timeseries_weather_forecasting/

	QCE ↓	PCE ↓	Sharp ↑	Acc ↑	AUROC ↑
Prehoc	**0.0181±0.0026**	0.3493±0.0015	0.3019±0.0004	0.4044±0.0006	0.3525±0.0012
MW	N/A	0.0278±0.0005	0.3005±0.0004	0.6124±0.0008	0.6410±0.0012
IW	N/A	0.0322±0.0005	0.3013±0.0004	0.6147±0.0009	0.6450±0.0013
OPS	N/A	**0.0148±0.0002**	**0.3172±0.0004**	**0.6525±0.0007**	**0.7056±0.0010**

Table 3: OPS improves the overall quality of parity probabilities from the base regression model predicting Pressure. $\pm$ indicates mean $\pm$ 1 standard error, across 50 test trials. Best value for each metric is in bold.

	PCE ↓	Sharp ↑	Acc ↑	AUROC ↑
Prehoc	0.0258±0.0005	0.3008±0.0007	0.6069±0.0011	0.6474±0.0016
MW	0.0201±0.0005	0.3002±0.0007	0.6050±0.0012	0.6439±0.0017
IW	0.0166±0.0003	0.3003±0.0008	0.6068±0.0010	0.6456±0.0016
OPS	**0.0150±0.0001**	**0.3232±0.0006**	**0.6665±0.0007**	**0.7275±0.0007**

Table 4: While MW, IW, OPS all improve parity calibration of the base classification model for Pressure (Prehoc), OPS is the only method that improves all metrics simultaneously. $\pm$ indicates mean $\pm$ 1 standard error, across 50 test trials. Best value for each metric is in bold.

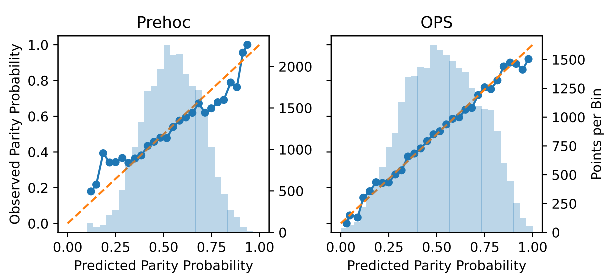


Figure 6: The base classification model for Pressure (Prehoc) is better parity calibrated than the base regression model (Figure 4 Prehoc), but OPS still improves its parity calibration and sharpness.

may be well-aware that the downstream user is primarily concerned with parity probabilities. Accordingly, the expert may choose to directly model parity probabilities in the context of a binary classification problem.

In Figure 6 and Table 4, we show results from training a base binary classifier with parity outcome labels and applying posthoc calibration. As expected, the prehoc parity probabilities of the binary classification model is significantly better calibrated than the regression model. Posthoc calibration still improves parity calibration further, especially in the case of OPS. In fact, OPS is the only method which improves all of the metrics simultaneously, while MW and IW notably worsen sharpness and AUROC. The full set of reliability diagrams is provided in Figure 10 in Appendix B.2.2.

Results across all 7 timeseries. Table 6 in Appendix B.2.2 shows each metric averaged across all 7 prediction targets: Table 6a displaying results with the base regression model, and 6b that of the base classification model. The pattern observed for the Pressure timeseries tend to hold on average across all 7 timeseries.

3.3 CASE STUDY 3: MODEL-BASED CONTROL FOR NUCLEAR FUSION

Nuclear fusion is the physical process during which atomic nuclei combine together to form heavier atomic nuclei, while releasing atomic particles and energy. Although fusion is possibly a safe, clean, and fuel-abundant technology for the future [Morse, 2018], there are various challenges to realizing fusion power, one of which is controlling nuclear fusion reactions [Humphreys et al., 2015].

Recently, model-based control methods, where a dynamics model of the system is learned and used to optimize control policies, has emerged as an effective control method for fusion devices [Abbate et al., 2023]. To the experimenter utilizing the dynamics model, it is of significant interest to know when certain signals will increase, and whether the dynamics model assigns correct probabilities to the events [Char et al., 2021]. In this section, we consider the problem of predicting the parity of β_N, which is a signal indicating reaction efficiency in a fusion device called a tokamak.

To this end, we design our empirical case study as follows. We take a pretrained dynamics model which was trained with a logged database of 10294 fusion experiments (referred to as “shots”) conducted on the DIII-D tokamak [Luxon, 2002], a device in San Diego, CA, USA. This pretrained model has been used for model-based policy optimization for deployment in actual fusion experiments on this device [Char et al., 2021, Seo et al., 2021, Abbate et al., 2021]. The model architecture is a recurrent probabilistic neural network (RPNN), which is a recurrent neural

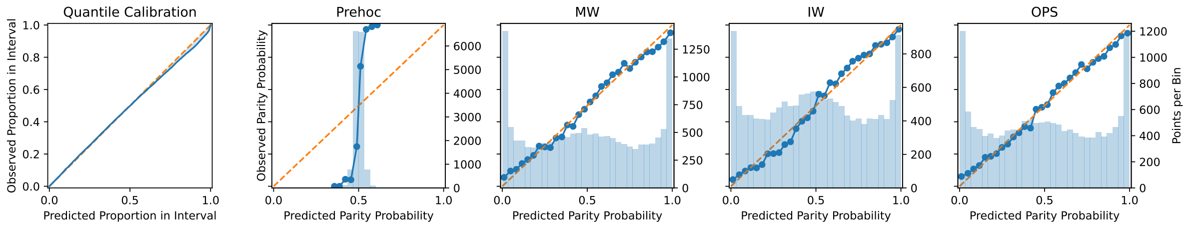


Figure 7: All methods (MW, IW, OPS) perform equally well in calibrating the Prehoc parity probabilities of the nuclear fusion dynamics model. The left two plots display the quantile calibration and parity calibration of the base dynamics model.

	QCE ↓	PCE ↓	Sharp ↑	Acc ↑	AUROC ↑
Prehoc	**0.0108±0.0003**	0.2571±0.0003	0.3243±0.0002	**0.7727±0.0003**	**0.8536±0.0002**
MW	N/A	0.0266±0.0002	0.3345±0.0002	0.7665±0.0003	0.8463±0.0002
IW	N/A	0.0291±0.0002	**0.3385±0.0002**	**0.7726±0.0003**	**0.8533±0.0002**
OPS	N/A	**0.0261±0.0002**	0.3334±0.0002	0.7629±0.0002	0.8440±0.0002

Table 5: MW, IW, and OPS all improve parity calibration and sharpness of the Prehoc fusion dynamics model predicting β_N, while maintaining roughly the same level of accuracy. $\pm$ indicates mean $\pm$ 1 standard error, across 50 test trials. Best value for each metric is in bold.

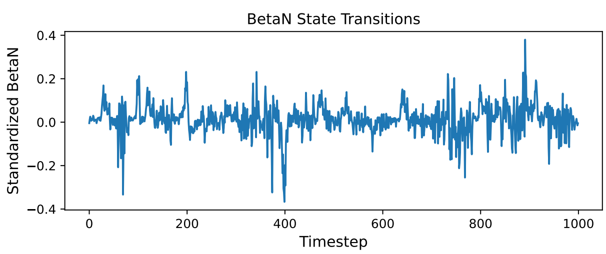


Figure 8: State transitions of the β_N signal during nuclear fusion experiments ("shots") concatenated across 50 training shots resemble trend-less noise.

network with a Gaussian output head. We refer the reader to Appendix B.3.1 for more details of the dynamics model and dataset. For testing, we allocate a set of 900 held-out test shots. On this test set, we produce the model's distributional predictions for β_N as the expert forecast. We concatenate the forecasts and the actual observed β_N values across the 900 test shots in chronological order into a single timeseries to assess parity calibration.

Figure 7 and Table 5 indicate that the expert forecast (Prehoc) is quantile calibrated but parity miscalibrated. The accuracy metrics in Table 5 indicate that despite prehoc's poor parity calibration, the model is still highly predictive, with an AUROC > 0.85. MW, IW and OPS significantly improve parity calibration and sharpness, while maintaining roughly the same level of accuracy.

We note that the β_N timeseries, as displayed in Figure 8, tends to fluctuate rapidly, between timesteps and between shots, almost resembling white noise. The pretrained model still manages to model the signal well, and assigns correct tendencies of increases/decreases in β_N: the relibility diagram of prehoc in Figure 7 shows that although the parity probabilities are not aligned with the empirical frequencies, they predict higher probabilities for actually higher frequency events. We believe this provides for a relatively easy posthoc calibration problem, thus all methods (MW, IW, OPS) perform equally well. Hence, this case study highlights the significance of the base model's initial parity probabilities, especially in alleviating the difficulty of posthoc calibration.

4 CONCLUSION

We considered the problem of forecasting whether a continuous-valued sequence is going to increase or decrease at the next time step. Such scenarios, where relative changes are more interpretable than actual values, are ubiquitous: COVID-19 cases per day, weather, or stock prices. In this context, we proposed the notion of parity calibration. To be parity calibrated, a forecaster must predict probabilities for the outcome increasing at the next time step, and these probabilities should be calibrated in the binary sense.

A decision-maker may attempt to achieve parity calibration by using regression forecasts produced by an expert forecaster. However, this is unlikely to give parity calibration. Instead, we proposed the usage of posthoc binary calibration techniques to achieve parity calibration. Specifically, we advocated for a recently proposed online Platt scaling algorithm (OPS) in this setting. In three real-world empirical case studies, OPS consistently improves the overall quality of parity probabilities compared to the expert forecaster.

Author Contributions

YC led the project as first author. AR played a key role in acquiring and interpreting the COVID-19 data for our experiments and contributed to discussions during project development. CG played the advisory role. He contributed the initial idea, planned the project direction, and steered the execution of the initial write-up as well as consequent revisions.

Acknowledgements

We would like to thank our Ph.D. advisors—Jeff Schneider (YC), Roni Rosenfeld (AR), and Aaditya Ramdas (CG)—for enabling us to pursue this student-only work. YC is supported in part by US Department of Energy grants under contract numbers DE-SC0021414 and DE-AC02-09CH1146, and the Kwanjeong Educational Foundation. AR is supported by McCune Foundation grant FP00004784. CG is supported by the Bloomberg Data Science Ph.D. Fellowship. The authors would also like to thank the anonymous UAI reviewers for their valuable feedback.

References

Joseph Abbate, R Conlin, and E Kolemen. Data-driven profile prediction for DIII-D. *Nuclear Fusion*, 61(4): 046027, 2021.

Joseph Abbate, Rory Conlin, Ricardo Shousha, Keith Erickson, and Egemen Kolemen. A general infrastructure for data-driven control design and implementation in tokamaks. *Journal of Plasma Physics*, 89(1):895890102, 2023.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135(643):1512–1519, 2009.

Ian Char, Youngseog Chung, Mark Boyer, Egemen Kolemen, and Jeff Schneider. A model-based reinforcement learning approach for beta control. In *APS Division of Plasma Physics Meeting Abstracts*, volume 2021, pages 11–150, 2021.

Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. Natural posterior network: Deep bayesian predictive uncertainty for exponential family distributions. In *International Conference on Learning Representations*, 2022.

Youngseog Chung, Ian Char, Han Guo, Jeff Schneider, and Willie Neiswanger. Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification. *arXiv preprint arXiv:2109.10254*, 2021a.

Youngseog Chung, Willie Neiswanger, Ian Char, and Jeff Schneider. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. In *Advances in Neural Information Processing Systems*, 2021b.

Estee Y Cramer, Yuxin Huang, Yijin Wang, Evan L Ray, Matthew Cornell, Johannes Bracher, Andrea Brennen, Alvaro J Castro Rivadeneira, Aaron Gerding, Katie House, Dasuni Jayawardena, Abdul H Kanji, Ayush Khandelwal, Khoa Le, Jarad Niemi, Ariane Stark, Apurv Shah, Nutcha Wattanachit, Martha W Zorn, Nicholas G Reich, and US COVID-19 Forecast Hub Consortium. The United States COVID-19 Forecast Hub dataset. *medRxiv*, 2021. doi: 10.1101/2021.11.04.21265886.

Estee Y. Cramer, Evan L. Ray, Velma K. Lopez, et al., and Nicholas G. Reich. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the united states. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119, 2022. doi: 10.1073/pnas.2113561119.

Peng Cui, Wenbo Hu, and Jun Zhu. Calibrated reliable regression using maximum mean discrepancy. In *Advances in Neural Information Processing Systems*, 2020.

A Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Morris H DeGroot and Stephen E Fienberg. Assessing probability assessors: Calibration and refinement. Technical report, Carnegie Mellon University, 1981.

Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69(2):243–268, 2007.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Chirag Gupta and Aaditya Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International Conference on Machine Learning*, 2021.

Chirag Gupta and Aaditya Ramdas. Online Platt scaling with calibeating. In *International Conference on Machine Learning*, 2023.

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. In *Advances in Neural Information Processing Systems*, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

David Humphreys, G Ambrosino, Peter de Vries, Federico Felici, Sun H Kim, Gary Jackson, A Kallenbach, Egemen Kolemen, J Lister, D Moreau, et al. Novel aspects of plasma control in iter. *Physics of Plasmas*, 22(2):021806, 2015.

Jena Weather Station at Max Planck Institute for Biogeochemistry. Jena climate data. 2016. URL https://www.bgc-jena.mpg.de/wetter/.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, 2018.

Meelis Kull, Telmo M Silva Filho, and Peter Flach. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, 11(2):5052–5080, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

James L Luxon. A design retrospective of the DIII-D tokamak. *Nuclear Fusion*, 42(5):614, 2002.

Charles Marx, Shengjia Zhao, Willie Neiswanger, and Stefano Ermon. Modular conformal calibration. In *International Conference on Machine Learning*, 2022.

Edward Morse. *Nuclear Fusion*. Springer, 2018.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*, 2005.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *International Conference on Neural Networks*, 1994.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *ML Summer School 2003 (Canberra, Australia)*. Springer, 2004.

Roshni Sahoo, Shengjia Zhao, Alyssa Chen, and Stefano Ermon. Reliable decisions with threshold calibration. In *Advances in Neural Information Processing Systems*, 2021.

Jaemin Seo, Y-S Na, B Kim, CY Lee, MS Park, SJ Park, and YH Lee. Feedforward beta control in the kstar tokamak by deep reinforcement learning. *Nuclear Fusion*, 61(10): 106010, 2021.

Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *International Conference on Machine Learning*, 2019.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *International Conference on Machine Learning*, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *International Conference on Knowledge Discovery and Data mining*, 2002.

Shengjia Zhao, Tengyu Ma, and Stefano Ermon. Individual calibration with randomized forecasting. In *International Conference on Machine Learning*, 2020.