Residual Distribution Predictive Systems

Sam Allen* Enrico Pescara† Johanna Ziegel‡

Abstract

Conformal predictive systems are sets of predictive distributions with theoretical out-of-sample calibration guarantees. The calibration guarantees are typically that the set of predictions contains a forecast distribution whose prediction intervals exhibit the correct marginal coverage at all levels. Conformal predictive systems are constructed using conformity measures that quantify how well possible outcomes conform with historical data. However, alternative methods have been proposed to construct predictive systems with more appealing theoretical properties. We study an approach to construct predictive systems that we term Residual Distribution Predictive Systems. In the split conformal setting, this approach nests conformal predictive systems with a popular class of conformity measures, providing an alternative perspective on the classical approach. In the full conformal setting, the two approaches differ, and the new approach has the advantage that it does not rely on a conformity measure satisfying fairly stringent requirements to ensure that the predictive system is well-defined; it can readily be implemented alongside any point-valued regression method to yield predictive systems with out-of-sample calibration guarantees. The empirical performance of this approach is assessed using simulated data, where it is found to perform competitively with conformal predictive systems. However, the new approach offers considerable scope for implementation with alternative regression methods.

1 Introduction

For forecasts to be useful for decision making, they must align statistically with what actually materialises. Such forecasts are said to be *calibrated*. For example, if a binary event is predicted to occur with a certain probability, then it should manifest with this probability. Similarly, prediction intervals should contain the outcome with the desired coverage level, and full predictive distributions should yield calibrated prediction intervals. We focus here on probabilistic forecasts for real-valued outcomes, where the goal is to issue predictive distributions that are as informative as possible subject to being calibrated (Gneiting et al., 2007).

Forecasting is typically performed out-of-sample, with predictions issued for future outcomes that have not yet been observed. Ideally, a forecasting procedure would issue calibrated predictions for out-of-sample outcomes by construction, possibly under some weak assumptions on the data generating process. However, designing a forecasting procedure with such out-of-sample calibration guarantees is generally infeasible (Vovk et al., 2022). Alternatively, conformal prediction has recently become popular in the machine learning literature since it goes one step in this direction by issuing a set of probabilistic predictions (i.e. a credal set) that is guaranteed to contain an out-of-sample calibrated prediction, under the assumption of exchangeable data (Gammerman et al., 1998; Shafer and Vovk, 2008; Vovk et al., 2017; Fontana et al., 2023). If the size of the credal set is small, then any forecast within the set should be approximately calibrated. If the size of the credal set is large, then it suggests there is large epistemic uncertainty in the forecast (Gammerman and Vovk, 2002; Hüllermeier and Waegeman, 2021).

When issuing probabilistic forecasts for real-valued outcomes, the credal sets are generally referred to as *predictive systems*. While they have received less attention in the literature than conformal prediction

*Institute of Statistics, Karlsruhe Institute of Technology, Karlsruhe, Germany. sam.allen@kit.edu
†Seminar for Statistics, ETH Zurich, Zurich, Switzerland. epescara@student.ethz.ch
‡Seminar for Statistics, ETH Zurich, Zurich, Switzerland. ziegel@stat.math.ethz.ch

methods for classification and regression tasks, conformal predictive systems contain considerably more information; they directly provide prediction intervals with out-of-sample coverage guarantees at any coverage level; see for example Vovk et al. (2017).

The classical conformal prediction framework is based on conformity measures (or non-conformity scores), which measure how well possible covariate-observation pairs *conform* with previously observed data. Allen et al. (2025) demonstrate that this approach essentially corresponds to a particular example of a more general framework that constructs credal sets with out-of-sample calibration guarantees using forecasting procedures that are calibrated in-sample. van der Laan and Alaa (2025) show a similar result in the context of point forecasting. An obvious question is what other forecasting procedures are calibrated in-sample. Allen et al. (2025) study in detail two approaches when interest is on probabilistic forecasts of real-valued outcomes: binning (or analogue) prediction methods, and isotonic distributional regression (IDR; Henzi et al., 2021). The latter approach provides a generalisation of the Venn-Abers predictors of Vovk et al. (2015).

In this work, we analyse a simple alternative forecasting procedure that is commonly employed in practice. The approach fits a regression model to some training data, and uses this to obtain a point forecast for the new outcome. This point forecast is then dressed with the residuals in a training data set to obtain a (discrete) forecast distribution. The resulting forecast distributions satisfy a popular notion of forecast calibration in-sample, and we therefore demonstrate how it can be used to obtain predictive systems that are guaranteed to contain a calibrated out-of-sample predictive distribution. We refer to these as *Residual Distribution Predictive Systems*.

This approach shares similarities with the classical conformal prediction framework, since conformity measures are typically constructed using regression residuals. We demonstrate that in a split conformal setting (Vovk et al., 2018b), these two approaches actually coincide, offering an alternative perspective on the original conformal prediction system framework. In the full conformal setting, the two approaches differ, and while the classical approach requires that the conformity measure satisfies a fairly strict monotonicity condition (Vovk et al., 2017), the novel approach does not. This facilitates the construction of conformal predictive systems using any regression method, without the need to check that the monotonicity condition is satisfied. However, in practice, some regression methods can yield predictive systems with excessively large thickness. In practice, we find that the two approaches yield predictive systems with similar predictive ability, though there is additional scope to apply the Residual Distribution Predictive System with more flexible regression models.

The following section defines predictive systems and notions of calibration, and gives the construction result for predictive systems with out-of-sample calibration guarantees. The classical conformal predictive systems are introduced in Section 3, while the residual distribution predictive systems that we study are introduced in Section 4. Some simulation results are provided in Section 5, before we conclude in Section 6. Code to reproduce the simulation results is available at `https://github.com/sallen12/RDPS_rep`.

2 Predictive Systems

Suppose our goal is to predict an outcome $Y_{n+1} \in \mathbb{R}$ using a training data set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$ and a new covariate $x_{n+1} \in \mathcal{X}$. A prediction for Y_{n+1} is typically obtained using state-of-the-art statistical and machine learning models, which take the training data and the new covariate as inputs, and output a prediction for the unknown outcome. In the following, we assume that the predictions are probability distributions on $\mathbb{R}$, and we refer to such a model as a *forecasting procedure*. The training data can be represented via the empirical distribution $\hat{\mathbb{P}}_n = (1/n)\sum_{i=1}^n \delta_{(x_i,y_i)}$, and we denote the forecasting procedure corresponding to the training data $\hat{\mathbb{P}}_n$ and covariate $x \in \mathcal{X}$ as the function $G_x^{\hat{\mathbb{P}}_n} : \mathbb{R} \to [0,1]$. In practice, the forecasting procedure is evaluated at $x = x_{n+1}$ to get a prediction for Y_{n+1}. Throughout, we use lower and upper case to denote deterministic and random elements, respectively, and we denote the underlying probability measure by P.

Ideally, the forecasting procedure should issue a forecast for Y_{n+1} that is calibrated, in the sense that the

forecast is statistically compatible with the corresponding outcome. There are several different ways to define forecast calibration, though it is most common in practice to assess *probabilistic calibration*, which corresponds to the prediction intervals derived from the predictive distributions achieving the correct marginal coverage (Dawid, 1984; Diebold et al., 1998; Gneiting et al., 2007). Following Allen et al. (2025), we define probabilistic calibration as follows.

Definition 1. A (random) forecast distribution G is *probabilistically calibrated* for $Y \in \mathbb{R}$ if

$$\mathsf{P}(G(Y) \leq \alpha) \leq \alpha \leq \mathsf{P}(G(Y-) < \alpha) \quad \text{for all } \alpha \in (0,1), \tag{1}$$

where $G(y-) = \lim_{z \uparrow y} G(z)$.

If the forecast distribution is continuous, probabilistic calibration implies that the forecast probability integral transform $G(Y)$ follows a standard uniform distribution. Stronger notions of calibration have also been proposed, which generally correspond to (1) conditioned on more information (Gneiting and Ranjan, 2013; Tsyplakov, 2014; Gneiting and Resin, 2023), though probabilistic calibration remains a useful notion that can be leveraged in decision-making contexts to achieve asymptotically efficient decisions (see Vovk et al., 2022, Section 7.7). We can similarly define the calibration of a forecasting procedure. In this case, it is important to distinguish between *in-sample* and *out-of-sample* calibration.

Definition 2. Let $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$ be a random sample with empirical distribution $\hat{\mathbb{P}}_n = (1/n)\sum_{i=1}^n \delta_{(X_i,Y_i)}$. A forecasting procedure G is *in-sample probabilistically calibrated* if $G_{X_i}^{\hat{\mathbb{P}}_n}$ is a probabilistically calibrated forecast for Y_i under $\hat{\mathbb{P}}_n$, for all $i = 1, \ldots, n$. That is,

$$\hat{\mathbb{P}}_n\big(G_{X_i}^{\hat{\mathbb{P}}_n}(Y_i) \leq \alpha\big) \leq \alpha \leq \hat{\mathbb{P}}_n\big(G_{X_i}^{\hat{\mathbb{P}}_n}(Y_i-) < \alpha\big) \quad \text{for all } \alpha \in (0,1) \text{ almost surely.}$$

Analogously, a forecasting procedure G could be defined as out-of-sample calibrated if, given a random covariate X_{n+1}, $G_{X_{n+1}}^{\hat{\mathbb{P}}_n}$ is a probabilistically calibrated forecast for Y_{n+1} under P. Since most forecasting tasks are out-of-sample, we generally desire forecasts that are out-of-sample calibrated. However, deriving a forecasting procedure that is guaranteed to issue out-of-sample calibrated forecasts is generally not possible (Vovk et al., 2022). Instead, conformal prediction has become popular since it provides a framework for obtaining credal sets that are guaranteed to contain an out-of-sample calibrated forecast, under the assumption of exchangeability. When dealing with probabilistic forecasts for real-valued outcomes, these credal sets are typically referred to as *conformal predictive systems*. We define a predictive system as follows.

Definition 3. A *predictive system* is a set $\Pi \subseteq \mathbb{R} \times [0,1]$ of the form

$$\Pi = \{(y, \tau) \in \mathbb{R} \times [0,1] \mid \Pi_\ell(y) \leq \tau \leq \Pi_u(y)\},$$

where the lower and upper bounds $\Pi_\ell, \Pi_u : \mathbb{R} \to [0,1]$ are increasing functions satisfying $\Pi_\ell(y) \leq \Pi_u(y)$ for all $y \in \mathbb{R}$, and

$$\lim_{y \to -\infty} \Pi_\ell(y) = 0, \qquad \lim_{y \to \infty} \Pi_u(y) = 1.$$

The *thickness* of the predictive system Π is defined as

$$\mathrm{th}(\Pi) = \inf\{\varepsilon > 0 \mid \Pi_u(y) - \Pi_\ell(y) \leq \varepsilon \text{ for all but finitely many } y \in \mathbb{R}\}.$$

A predictive system is therefore comprised of lower and upper bounds Π_ℓ, Π_u, which can be interpreted as two stochastically ordered (possibly defective) distribution functions, with the region between the two bounds defining a set of probabilistic forecasts. A predictive distribution is said to be contained in a predictive system Π if it lies between the system's lower and upper bounds, Π_ℓ and Π_u. Conformal predictive systems are constructed such that they are guaranteed to contain an out-of-sample probabilistically calibrated forecast distribution for Y_{n+1}. In practice, this guarantee is only useful if the difference between the bounds (the thickness) is not too large.

An obvious question is if and how we can construct predictive systems that are guaranteed to contain an out-of-sample calibrated forecast distribution. Allen et al. (2025) describe how they can be obtained via a general framework that leverages in-sample calibrated forecasting procedures. In particular, given a procedure G, training data $(x_1, y_1), \ldots, (x_n, y_n)$, and covariate x_{n+1}, a predictive system can be defined by the bounds

$$\Pi^{\hat{\mathbb{P}}_n}_{\ell, x_{n+1}}(y) = \inf_{y' \in \mathbb{R}} G^{\hat{\mathbb{P}}_{n+1}(y')}_{x_{n+1}}(y), \qquad \Pi^{\hat{\mathbb{P}}_n}_{u, x_{n+1}}(y) = \sup_{y' \in \mathbb{R}} G^{\hat{\mathbb{P}}_{n+1}(y')}_{x_{n+1}}(y), \quad y \in \mathbb{R}, \tag{2}$$

where $\hat{\mathbb{P}}_{n+1}(y')$ is used to denote the empirical distribution of $(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y')$, the training data along with the new covariate x_{n+1} and an assumed value $y' \in \mathbb{R}$ of the new outcome. These bounds essentially correspond to the pointwise lowest and highest values that the predictive distribution can attain when the forecasting procedure is trained using the training data augmented by the new pair (x_{n+1}, y'), for all possible y'. If a large amount of training data is available, then this additional training data point will typically have less influence on the learned predictive distribution, resulting in a predictive system whose lower and upper bounds are close together (and vice versa).

Allen et al. (2025, Theorem 1) demonstrate that if the forecasting procedure G is in-sample calibrated, then the predictive system Π generated by G is guaranteed to contain an out-of-sample calibrated forecast for Y_{n+1}.

Theorem 4 (Allen et al. (2025))**.** *Let $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{X} \times \mathbb{R}$ be exchangeable with empirical distribution $\hat{\mathbb{P}}_n = (1/n) \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$. If G is an in-sample probabilistically calibrated forecasting procedure, then the predictive system defined by the bounds at* (2) *satisfies*

$$\mathsf{P}\left(\Pi^{\hat{\mathbb{P}}_n}_{\ell, X_{n+1}}(Y_{n+1}) \leq \alpha\right) \leq \alpha \leq \mathsf{P}\left(\Pi^{\hat{\mathbb{P}}_n}_{u, X_{n+1}}(Y_{n+1}-) < \alpha\right) \quad \textit{for all } \alpha \in (0, 1). \tag{3}$$

Remark 5. Equation (3) in Theorem 4 continues to hold almost surely when we replace the probability measure P with the (random) empirical distribution $\hat{\mathbb{P}}_{n+1}$. This statement is stronger, and can be used to test the assumption of exchangeability (Vovk et al., 2022, Section 8).

The bounds at (2) therefore provide a general framework for constructing predictive systems based on existing, well-understood forecasting procedures. In a split conformal setting (see below), the classical conformal prediction framework corresponds to a particular choice of forecasting procedure G, provided in the following section. More generally, predictive systems with out-of-sample calibration guarantees can be obtained from any in-sample calibrated forecasting procedure. Allen et al. (2025) use this to introduce predictive systems based on binning prediction methods and isotonic distributional regression (IDR; Henzi et al., 2021), for example. A further in-sample calibrated forecasting procedure is given in Allen et al. (2025, Example 1) but not studied in detail. This procedure is the main focus of this paper (see Section 4).

Remark 6 (Split conformal prediction)**.** In practice, calculating the lower and upper bounds of the predictive system at (2) can be prohibitively computationally expensive, since this may require refitting the forecasting procedure for every possible new outcome $y' \in \mathbb{R}$. Instead, if sufficient data is available, a more efficient approach is to divide the training data $(x_1, y_1), \ldots, (x_n, y_n)$ into an *estimation set* $(x_1, y_1), \ldots, (x_N, y_N)$ and a *calibration set* $(x_{N+1}, y_{N+1}), \ldots, (x_n, y_n)$, for $N < n$, and to estimate some parameters of the forecasting procedure on the estimation set. It then often becomes clear for which y' the bounds of the predictive system at (2) are attained, and they can be obtained easily using the calibration data. In this case, the model parameters would not need to be relearned for every possible $y' \in \mathbb{R}$, substantially simplifying the calculation of the predictive system. This approach is often referred to as a *split conformal* framework, in contrast to the *full conformal* framework previously described (Vovk et al., 2018b). The results of Theorem 4 hold in both the split and conformal settings (Allen et al., 2025), and, unless specified otherwise, we adopt the notation of the full conformal setting throughout.

3 Conformal Predictive Systems

Vovk et al. (2017) initially defined conformal predictive systems using conformity measures. A conformity measure is a real-valued function A that quantifies how much a prospective covariate-outcome pair conforms

with the training data. The conformity measure is often of the form

$$A(\hat{\mathbb{P}}_n, (x, y)) := y - \hat{y}_x, \tag{4}$$

for $(x, y) \in \mathcal{X} \times \mathbb{R}$, where $\hat{y}_x \in \mathbb{R}$ is a prediction for y computed from x and the training data $(x_1, y_1), \ldots, (x_n, y_n)$. The prediction could be obtained using ordinary least squares regression, for example, or more flexible alternatives. A predictive system corresponding to A is defined by the bounds

$$\Pi^{\hat{\mathbb{P}}_n}_{\ell, x_{n+1}}(y) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}\{A(\hat{\mathbb{P}}_{n+1}(y), (x_i, y_i)) < A(\hat{\mathbb{P}}_{n+1}(y), (x_{n+1}, y))\}, \tag{5}$$

$$\Pi^{\hat{\mathbb{P}}_n}_{u, x_{n+1}}(y) = \frac{1}{n+1} \left(1 + \sum_{i=1}^{n} \mathbb{1}\{A(\hat{\mathbb{P}}_{n+1}(y), (x_i, y_i)) \leq A(\hat{\mathbb{P}}_{n+1}(y), (x_{n+1}, y))\} \right). \tag{6}$$

We refer to these in the following as *Conformal Predictive Systems*.

In the full conformal setting, it is sufficient to define the conformity measure $A(\hat{\mathbb{P}}_n, (x, y))$ only at pairs (x, y) that are in the support of $\hat{\mathbb{P}}_n$ (see (5) and (6)). However, calculating the lower and upper bounds of the predictive system is generally computationally expensive and often practically infeasible; for example, for a conformity measure of the form (4), it requires fitting a regression model to $(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y)$ for all possible values of $y \in \mathbb{R}$. In the split conformal setting, $\hat{\mathbb{P}}_{n+1}(y)$ in (5) and (6) is replaced with $\hat{\mathbb{P}}_N$, the empirical distribution of the data in the estimation set, $(x_1, y_1), \ldots, (x_N, y_N)$. Since this does not depend on y, the lower and upper bounds of the prediction system are generally much easier to calculate. In this case, the conformal predictive system can also be expressed as the predictive system generated using the bounds at (2) with the forecasting procedure

$$G^{\hat{\mathbb{P}}_n}_x(y) = \frac{1}{n-N} \sum_{i=N+1}^{n} \mathbb{1}\{A(\hat{\mathbb{P}}_N, (x_i, y_i)) \leq A(\hat{\mathbb{P}}_N, (x, y))\}, \tag{7}$$

where $\hat{\mathbb{P}}_N$ is the empirical distribution of the estimation data; this empirical distribution $\hat{\mathbb{P}}_N$ does not change when $\hat{\mathbb{P}}_n$ is replaced with $\hat{\mathbb{P}}_{n+1}(y)$ in (2). This representation is useful to study the connection with the Residual Distribution Predictive Systems introduced in Section 4.

The forecasting procedure at (7) outputs valid predictive distributions when $A(\hat{\mathbb{P}}_N, (x_{n+1}, y))$ is increasing in y. In this case, the forecasting procedure is in-sample probabilistically calibrated. Hence, conformal predictive systems are guaranteed to contain an out-of-sample probabilistically calibrated forecast distribution in the split conformal framework. This is also true in the full conformal framework, but only when a stronger constraint on the conformity measure A is satisfied. Importantly, for certain choices of A, the lower and upper bounds at (5) and (6) may not be increasing functions of y, which violates the definition of a predictive system as a set of predictive distribution functions. The following lemma provides a sufficient condition on the conformity measure A to ensure that the bounds are increasing functions of y (Vovk et al., 2022).

Lemma 7. *Suppose that, for any $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$ and any $x_{n+1} \in \mathcal{X}$, the function*

$$y \mapsto A(\hat{\mathbb{P}}_{n+1}(y), (x_{n+1}, y)) - A(\hat{\mathbb{P}}_{n+1}(y), (x_i, y_i)) \tag{8}$$

is increasing, for all $i = 1, \ldots, n$. Then, when $\Pi^{\hat{\mathbb{P}}_n}_{\ell, x_{n+1}}$ and $\Pi^{\hat{\mathbb{P}}_n}_{u, x_{n+1}}$ are defined using (5) and (6), the functions $y \mapsto \Pi^{\hat{\mathbb{P}}_n}_{\ell, x_{n+1}}(y)$ and $y \mapsto \Pi^{\hat{\mathbb{P}}_n}_{u, x_{n+1}}(y)$ are increasing.

The requirement at (8) is referred to as the *monotonicity condition*. If this is satisfied, then the indicator $\mathbb{1}\{A(\hat{\mathbb{P}}_{n+1}(y), (x_i, y_i)) \leq A(\hat{\mathbb{P}}_{n+1}(y), (x_{n+1}, y))\}$ is an increasing function of y, and hence the lower and upper bounds of the predictive system at (5) and (6) must also be increasing in y. The thickness of the resulting predictive system is then at most $1/(n+1)$.

Remark 8. The monotonicity condition is most relevant in the full conformal setting. In the split conformal framework, the first argument of the conformity measure is fixed at $\hat{\mathbb{P}}_N$. Hence, for conformity measures of the form (4), the parameters of the regression model would be obtained from the estimation data set, in which case (8) simplifies to $y \mapsto y - \hat{y}_{n+1} - y_i + \hat{y}_i$ (with $\hat{y}_i = \hat{y}_{x_i}$ for $i = 1, \ldots, n+1$). The latter three terms are independent of y, so this is trivially an increasing function of y. Any conformity measure in this form therefore immediately yields valid predictive distributions and predictive systems in the split conformal setting. The same is not true in the full conformal setting, as elucidated by the following example.

Example 9 (Least Squares Prediction Machine)**.** Consider the conformity measure at (4) where $\hat{y}_x$ is the least squares prediction for y, trained using $(x_1, y_1), \ldots, (x_n, y_n)$ and evaluated at x. Vovk et al. (2017) term the corresponding predictive system the *Least Squares Prediction Machine* (LSPM). However, according to Vovk et al. (2017, Proposition 6), this conformity measure does not always satisfy the monotonicity condition at (8). Instead, it is common to replace (4) with the studentised conformity measure

$$A(\hat{\mathbb{P}}_n, (x, y)) := \frac{y - \hat{y}_x}{\sqrt{1 - h_{(x,y)}}}, \tag{9}$$

where $h_{(x,y)}$ is the leverage of (x, y) in the regression fit. It is not restrictive to assume that this leverage term exists, since the conformal predictive systems in the full conformal setting only require evaluating the conformity measure at a covariate-outcome pair that is in the support of the first argument of A. In the split conformal setting, this studentised conformity measure cannot be used, since the conformity measure must be evaluated at pairs (x, y) that are not in the support. However, from Remark 8, the standard non-normalised conformity measure at (4) can be employed in this case without issue. The studentised conformity measure at (9) satisfies the monotonicity condition, unlike its non-normalised counterpart at (4), and using this within (5) and (6) yields the *studentised LSPM* of Vovk et al. (2017) (see also Vovk et al., 2022, Section 7.3). Similar results holds for kernel regression (Vovk et al., 2018a).

The studentisation of the LSPM is reminiscent of a strategy to *localise* conformal predictive systems (Papadopoulos et al., 2008; Lei et al., 2018). The conformity measure at (4) is often adapted to use a standardised residual, where the residual is divided by an estimate of the uncertainty in the prediction that is similarly obtained from $\hat{\mathbb{P}}_n$. In doing so, the conformity measure can generally better adapt to regions of the outcome space where the performance of the prediction is less reliable. More generally, we could apply any strictly increasing transformation $\hat{f}_x$ to the residuals, $A(\hat{\mathbb{P}}_n, (x, y)) = \hat{f}_x(y - \hat{y}_x)$, where $\hat{f}_x$ is obtained from $\hat{\mathbb{P}}_n$ and x. However, in general, there is still no guarantee that these local and transformed conformity measures satisfy the monotonicity condition at (8), and details would therefore have to be worked out for specific cases. In the next section, we introduce predictive systems that coincide with conformal predictive systems in the split conformal setting, and that are more intuitive in the full conformal setting since they naturally avoid the monotonicity condition.

4 Residual Distribution Predictive Systems

4.1 Definition

Theorem 4 holds for any in-sample calibrated forecasting procedure. An open question is what forecasting procedures are in-sample calibrated. We study the following approach.

Definition 10. Let $\hat{y}_i \in \mathbb{R}$ denote a point-prediction for y_i, based on x_i and the training data $\hat{\mathbb{P}}_n$, for $i = 1, \ldots, n$, and denote by $\hat{\varepsilon}_i = y_i - \hat{y}_i$ the corresponding residuals. Let $\hat{y}_x \in \mathbb{R}$ similarly denote a point-prediction for a new outcome, obtained from $\hat{\mathbb{P}}_n$ and a covariate $x \in \mathcal{X}$. The *Residual Distribution* forecasting procedure is defined as

$$G_x^{\hat{\mathbb{P}}_n}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{y}_x + \hat{\varepsilon}_i \leq y\}, \quad y \in \mathbb{R}. \tag{10}$$

The *Residual Distribution Predictive System* (RDPS) is the predictive system Π generated by this forecasting procedure using the bounds defined at (2).

The forecasting procedure at (10) fits a regression model to the training data, and calculates the in-sample residuals. It then takes an out-of-sample prediction for the new outcome, and dresses this prediction using the training residuals to obtain a discrete predictive distribution. If the regression model yields inaccurate forecasts, then the residuals will generally be large, leading to a highly dispersed distribution, whereas more accurate forecasts will lead to more concentrated predictive distributions. An illustration of this using simulated data is provided in Figure 1.

Just as residuals can be replaced with standardised residuals to obtain a locally adaptive conformity measure, the RDPS can similarly be defined using standardised residuals. More generally, we can again apply any strictly increasing transformation to the residuals.

Definition 11. Let $\hat{y}_i \in \mathbb{R}$ denote a prediction for y_i, and let $\hat{f}_i : \mathbb{R} \to \mathbb{R}$ be a strictly increasing function, based on x_i and the training data $\hat{\mathbb{P}}_n$, for $i = 1, \ldots, n$. Denote by $\hat{\varepsilon}_i = y_i - \hat{y}_i$ the corresponding residuals. Let $\hat{y}_x \in \mathbb{R}$ similarly denote a point-prediction for a new outcome, and let $\hat{f}_x : \mathbb{R} \to \mathbb{R}$ be a strictly increasing function, both obtained from $\hat{\mathbb{P}}_n$ and a covariate $x \in \mathcal{X}$. The *generalised Residual Distribution* forecasting procedure is defined as

$$G_x^{\hat{\mathbb{P}}_n}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{ \hat{y}_x + \hat{f}_x^{-1}(\hat{f}_i(\hat{\varepsilon}_i)) \leq y \right\}, \quad y \in \mathbb{R}. \tag{11}$$

The predictive system generated by (11) is referred to as the *generalised RDPS*.

This generalised RDPS similarly calculates the in-sample residuals of a regression model, but transforms them according to the functions $\hat{f}_i$ before dressing the new prediction $\hat{y}_x$. The motivation behind this approach is that different residuals in the training data set can be rescaled depending on their relevance to the new prediction. The functions $\hat{f}_i$ therefore help to incorporate context-dependent information into the predictive distribution and, in turn, the predictive system.

Example 12. Let $\hat{f}_i(t) = t/\hat{\sigma}_i$, where $\hat{\sigma}_i > 0$ is a measure of uncertainty in the prediction $\hat{y}_i$, for $t \in \mathbb{R}$ and $i = 1, \ldots, n$. Let $\hat{y}_x$ be a prediction for a new outcome, based on the covariate x, and let $\hat{f}_x(t) = t/\hat{\sigma}_x$, where $\hat{\sigma}_x > 0$ is similarly a measure of uncertainty in $\hat{y}_x$. The generalised RDPS at (11) becomes

$$G_x^{\hat{\mathbb{P}}_n}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{ \hat{y}_x + \frac{\hat{\sigma}_x}{\hat{\sigma}_i} \hat{\varepsilon}_i \leq y \right\}, \quad y \in \mathbb{R}.$$

This corresponds to an approach whereby the residuals are standardised by a measure of uncertainty in the prediction, before being added to the new prediction $\hat{y}_x$.

Allen et al. (2025, Example 3) demonstrate that the forecasting procedure at (10) is in-sample probabilistically calibrated, and this extends easily to the generalised forecasting procedure at (11). The resulting RDPS is therefore guaranteed to contain an out-of-sample calibrated forecast distribution for Y_{n+1} (under the assumption of exchangeability). Hence, this approach shares the same calibration properties as the classic conformal predictive systems. The RDPS has the advantage that it generates valid predictive distributions and predictive systems for any choice of the regression method; it does not require that the fairly stringent monotonicity condition at (8) is satisfied, for example. This facilitates the introduction of predictive systems with out-of-sample calibration guarantees using any regression procedure. However, for some choices of the regression method in the full conformal setting, the thickness of the RDPS will be excessively high, in which case the predictive systems are not particularly useful in practice. In contrast, conformal predictive systems have a thickness of at most $1/(n+1)$.

4.2 Comparison with Conformal Predictive Systems

Consider first the split conformal setting. In this case, the following theorem demonstrates that Residual Distribution Predictive Systems are actually equivalent to conformal predictive systems with conformity measure of the form (4) when both are constructed using the same regression model. This equivalence holds more generally for the generalised RDPS and the conformal predictive systems defined using the transformed residuals.

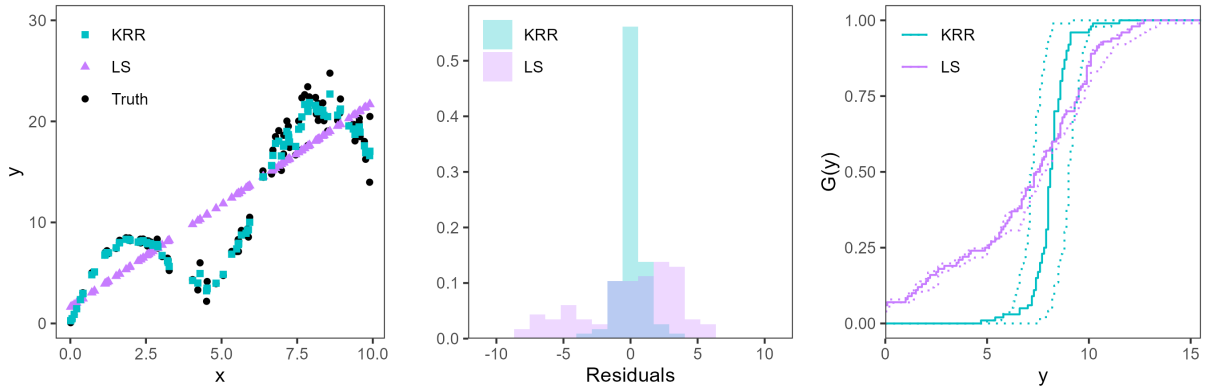

Figure 1: Left: 100 samples from the nonlinear simulation data in Section 5, along with the point forecasts obtained from ordinary least squares regression (LS) and kernel ridge regression (KRR). Centre: A histogram of the corresponding residuals. Right: The resulting predictive distributions $G^{\hat{\mathbb{P}}_n}_{x_{n+1}}$ (solid) and RDPS bounds $\Pi^{\hat{\mathbb{P}}_n}_{\ell,x_{n+1}}, \Pi^{\hat{\mathbb{P}}_n}_{u,x_{n+1}}$ (dotted) at $x_{n+1} = 2.5$. Least squares prediction yields less accurate forecasts for the nonlinear data than kernel ridge regression, resulting in higher residuals and a more dispersed predictive distribution. However, the RDPS thickness is larger for kernel ridge regression since the model is more flexible and therefore more difficult to estimate.

Theorem 13. *Suppose that, for any $x \in \mathcal{X}$, we have a prediction $\hat{y}_x \in \mathbb{R}$ and a strictly increasing function $\hat{f}_x : \mathbb{R} \to \mathbb{R}$ based on $\hat{\mathbb{P}}_N$. Define a conformity measure by $A(\hat{\mathbb{P}}_N, (x, y)) = \hat{f}_x(y - \hat{y}_x)$. Then, in the split conformal setting, the conformal predictive system defined using this conformity measure is equivalent to the generalised RDPS defined by* (11).

Proof. Denote $\hat{f}_i = \hat{f}_{x_i}$, $\hat{y}_i = \hat{y}_{x_i}$, and $\hat{\varepsilon}_i = y_i - \hat{y}_i$, for $i = N+1, \ldots, n+1$. In the split conformal setting, the forecasting procedure at (11) becomes

$$G^{\hat{\mathbb{P}}_n}_{x_{n+1}}(y) = \frac{1}{n-N} \sum_{i=N+1}^{n} \mathbb{1}\{\hat{y}_{n+1} + \hat{f}^{-1}_{n+1}(\hat{f}_i(\hat{\varepsilon}_i)) \leq y\}, \quad y \in \mathbb{R}$$
$$= \frac{1}{n-N} \sum_{i=N+1}^{n} \mathbb{1}\{\hat{f}_i(y_i - \hat{y}_i)) \leq \hat{f}_{n+1}(y - \hat{y}_{n+1})\}, \quad y \in \mathbb{R}.$$

This is exactly the forecasting procedure at (7) with $A(\hat{\mathbb{P}}_N, (x, y)) = \hat{f}_x(y - \hat{y}_x)$. Since the residual distribution and conformity measure predictive systems both correspond to the bounds at (2) defined using the same forecasting procedure, the predictive systems themselves must also be equivalent. □

Theorem 13 holds for any regression method used to obtain the predictions, and any choice of the strictly increasing functions. The forecasting procedure underlying the RDPS at (7) is obtained when $\hat{f}_x$ is the identity function for all $x \in \mathcal{X}$, while the standardised RDPS in Example 12 is recovered using the functions $\hat{f}_x(t) = t/\hat{\sigma}_x$ for $t \in \mathbb{R}$.

One immediate corollary of Theorem 13 is that the thickness of the RDPS in the split conformal framework is equal to $1/(n - N + 1)$, where the denominator is simply the size of the calibration dataset plus one. This thickness is constant and does not depend on y, in contrast to the full conformal framework, where the thickness of the RDPS varies with y (see Section 5). Hence, while the conformal and Residual Distribution predictive systems are equivalent in the split conformal setting, they differ when a full conformal framework is adopted. One benefit of the RDPS is that they do not require that any monotonicity condition is satisfied, as in (8). In principle, this allows the approach to be implemented with any regression method.

However, for some regression methods, the bounds at (2) will be too wide to be informative. We can make the following approximation to find a sufficient criterion for informative bounds. Let $\hat{y}'_i$ denote the prediction for y_i when the regression method is trained using $\hat{\mathbb{P}}_{n+1}(y')$, for $y' \in \mathbb{R}$. The superscript prime in $\hat{y}'_i$ is used to clarify that this is a function of y'. Then,

$$\begin{aligned}
\Pi^{\hat{\mathbb{P}}_n}_{u,x_{n+1}}(y) &= \sup_{y' \in \mathbb{R}} G^{\hat{\mathbb{P}}_{n+1}(y')}_{x_{n+1}}(y) \\
&= \sup_{y' \in \mathbb{R}} \frac{1}{n+1} \left(\sum_{i=1}^{n} \mathbb{1}\{\hat{y}'_{n+1} + y_i - \hat{y}'_i \leq y\} + \mathbb{1}\{\hat{y}'_{n+1} + y' - \hat{y}'_{n+1} \leq y\} \right) \\
&= \sup_{y' \in \mathbb{R}} \frac{1}{n+1} \left(\sum_{i=1}^{n} \mathbb{1}\{\hat{y}'_{n+1} - \hat{y}'_i \leq y - y_i\} + \mathbb{1}\{y' \leq y\} \right) \\
&\leq \frac{1}{n+1} \left(\sum_{i=1}^{n} \mathbb{1}\{\inf_{y' \in \mathbb{R}} (\hat{y}'_{n+1} - \hat{y}'_i) \leq y - y_i\} + 1 \right).
\end{aligned}$$

Similarly,

$$\Pi^{\hat{\mathbb{P}}_n}_{\ell,x_{n+1}}(y) = \inf_{y' \in \mathbb{R}} G^{\hat{\mathbb{P}}_{n+1}(y')}_{x_{n+1}}(y) \geq \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}\{\sup_{y' \in \mathbb{R}} (\hat{y}'_{n+1} - \hat{y}'_i) \leq y - y_i\}.$$

Hence, for the RDPS to have reasonable thickness, we want that the regression method is such that $\inf_{y' \in \mathbb{R}}(\hat{y}'_{n+1} - \hat{y}'_i)$ and $\sup_{y' \in \mathbb{R}}(\hat{y}'_{n+1} - \hat{y}'_i)$ are bounded. For example, when using quantile regression to determine $\hat{y}'_i$, we have the property that the regression line, or the regression hyperplane more generally for p-dimensional covariates, passes through at least p data points if it is unique. When y' is extreme, it is highly unlikely that the regression hyperplane would pass through (x_{n+1}, y'), so the difference $\hat{y}'_{n+1} - \hat{y}'_i$ is robust to outlying values of y'. This ensures informative predictive bounds. In contrast, for ordinary least squares regression or kernel regression, the difference $\hat{y}'_{n+1} - \hat{y}'_i$ is typically linear in y', so the bounds will typically be uninformative unless the range of y_i is bounded.

Remark 14 (Deleted RDPS)**.** One approach to obtain robust regression methods for this purpose is to simply identify and remove outliers from the augmented training data. This shares similarities with the *deleted* LSPM proposed by Vovk et al. (2017), though here we cannot simply remove the pair (x_{n+1}, y) since Theorem 4 requires that the parameter estimation procedure is permutation invariant. Instead, one could fit the regression model to the training data, and then remove the covariate-outcome pairs that yield large (absolute) residuals. The regression model can then be refit to the remaining training data. This increases the computational cost, but circumvents the sensitivity of the RDPS to outlying values. This approach is implemented in Section 5 when calculating the residuals using least squares regression and kernel ridge regression. Several other approaches exist to construct robust regression models (e.g. Salibian-Barrera, 2023), and they could similarly be employed.

4.3 Computation

Consider now the computation of the RDPS. The bounds at (2) correspond to infimum and supremum of the predictive distributions obtained for all possible $y' \in \mathbb{R}$. In practice, these could be calculated by discretising the real line, refitting the forecasting procedure for all y' on this discretised line, and then taking the pointwise minimum and maximum of the resulting predictive distributions. However, this approach is generally prohibitively computationally expensive, and the results may also depend on how the real line is discretised, particularly in the limits. Moreover, in practice, it is often not necessary to consider every possible value of y'.

For the standard RDPS, we have that

$$G^{\hat{\mathbb{P}}_{n+1}(y')}_{x_{n+1}}(y) = \frac{1}{n+1} \left[\sum_{i=1}^{n} \mathbb{1}\{\hat{y}'_{n+1} - \hat{y}'_i \leq y - y_i\} + \mathbb{1}\{y' \leq y\} \right]. \tag{12}$$

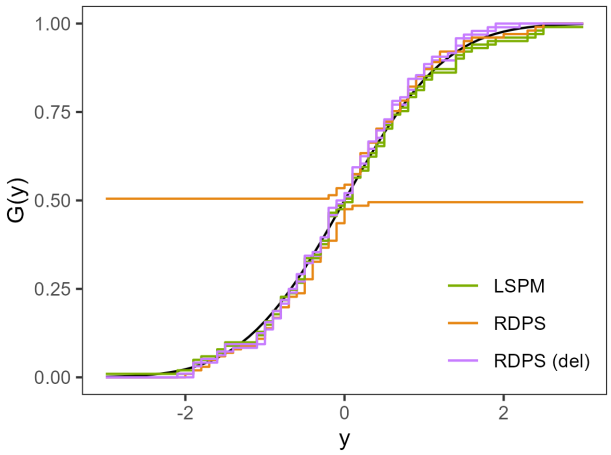

Figure 2: Lower and upper bounds of the LSPM, RDPS, and the deleted RDPS (see Remark 14) at a given x_{n+1} for the linear simulation data in Section 5 Both RDPS models are obtained using ordinary least squares regression. The black line is the true conditional distribution of the outcome at this covariate value.

In the split conformal setting, the predictions $\hat{y}'_{n+1}$ and $\hat{y}'_i$ do not depend on y', so the lower and upper bounds can be calculated by setting $\mathbb{1}\{y' \leq y\}$ to zero and one, respectively. Similarly, in the full conformal setting, if $\hat{y}'_{n+1} - \hat{y}'_i$ is an increasing function of y', for all $i = 1, \ldots, n$, then (12) is a decreasing function of y'. In this case, the lower and upper bounds of the predictive system can be computed efficiently using just two fits of the forecasting procedure: the infimum is obtained as $y' \to +\infty$, and the supremum as $y' \to -\infty$. Some examples where this is (and is not) satisfied are given in Appendix A.

Even if $\hat{y}'_{n+1} - \hat{y}'_i$ is not increasing in y', it may still be possible to efficiently implement the RDPS. This is possible, for example, when the predictions are linear functions of the observations in the training data. Some examples where this is the case include least squares regression, kernel regression, smoothing splines, k-nearest neighbours, and binning procedures. Vovk et al. (2017) and Vovk et al. (2018a) use this to derive efficient algorithms to calculate the LSPM and its kernel regression counterpart, the Kernel Ridge Regression Prediction Machine (KRRPM). Details to calculate the RDPS in this case are described in Appendix A. Alternatively, other estimation methods could be leveraged to obtain more efficient implementations of the RDPS. We provide an example below for quantile regression; details for other methods would have to be similarly worked out.

Example 15 (Quantile regression)**.** A robust alternative to ordinary least squares regression is median regression (more generally, quantile regression) (Koenker and Bassett Jr, 1978). Shen et al. (2024) recently proposed an iterative algorithm for parameter estimation in online quantile regression. We modify this slightly to obtain an approach that allows for efficient estimation of the RDPS bounds when the predictions are obtained using quantile regression. Firstly, we randomise the order of the training pairs $(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y')$, and let $(x_{(i)}, y_{(i)})$ denote the i-th pair in the randomised sequence, for $i = 1, \ldots, n+1$. Following Shen et al. (2024), an initial parameter estimate $\beta_{(0)}$ is then iteratively updated using sub-gradient descent, with

$$\beta_{(t+1)} = \beta_{(t)} - \eta_{t+1} \left(\tau - \mathbb{1}\{ y_{(t+1)} > \beta_{(t)}^\top x_{(t+1)} \} \right) x_{(t+1)}, \quad t = 0, \ldots, n,$$

where $\eta_1, \ldots, \eta_{n+1} > 0$ are stepsize parameters, and $\tau \in (0, 1)$ is the quantile level of interest. We assume that the parameters $\beta_{(0)}$ and $\eta_1, \ldots, \eta_{n+1}$ do not depend on y'. We then take the final predictions as $\hat{y}_i = \beta_{(n+1)}^\top x_i$, for $i = 1, \ldots, n+1$. All predictions are therefore derived from the final parameter in the algorithm, $\beta_{(n+1)}$. Assume that $(x_{n+1}, y') = (x_{(j)}, y_{(j)})$, i.e. (x_{n+1}, y') is the j-th training pair in the randomised sequence for $j \in \{1, \ldots, n+1\}$. Then, the parameter $\beta_{(n+1)}$, and hence the predictions $\hat{y}'_i$, only depend on y' via the indicator $\mathbb{1}\{ y' > \beta_{(j-1)}^\top x_{n+1} \}$. Hence, over all possible $y' \in \mathbb{R}$, $\hat{y}'_{n+1} - \hat{y}'_i$ takes on only two possible values,

meaning the infimum and supremum of the forecasting procedure at (12) can be obtained from just two runs of the iterative parameter estimation, which is anyway quick to implement. This can be extended further to run the iterative algorithm multiple times, and to average the resulting parameter estimates; while the RDPS is guaranteed to contain a calibrated probabilistic prediction for Y_{n+1} for any regression model, better parameter estimates will result in more informative predictive systems.

5 Simulation Examples

We compare the performance of the RDPS and conformal predictive systems when applied to two simple simulated datasets. The first case assumes that there is a linear relationship between the covariates and the outcome variable: $X_i \sim \mathcal{N}(0,1)$, and $Y_i \sim \mathcal{N}(X_i,1)$ for $i = 1,\dots,n+1$. The second case assumes that there is a nonlinear relationship between X_i and Y_i, and that the conditional variance of the outcome is a function of the covariate: $X_i \sim \text{Uniform}(0,10)$ and $Y_i = 2X_i + 5\sin(X_i) + \eta_i$, where $\eta_i \sim \mathcal{N}(0,(x_i/5)^2)$. In both cases, $(X_1,Y_1),\dots,(X_{n+1},Y_{n+1})$ are iid. A sample from the nonlinear simulated data is shown in Figure 1.

In each case, we set $n = 100$, and draw realisations of $(X_1,Y_1),\dots,(X_{n+1},Y_{n+1})$. We fit a conformal predictive system and an RDPS to the training data and X_{n+1}, and assess the predictive system in its ability to predict Y_{n+1}. This is repeated 1000 times. The predictive systems are compared via the (conservative) prediction intervals derived from them. We evaluate the $1-\alpha$-level prediction intervals for $1-\alpha$ ranging from 0.5 to 0.95. The central $1-\alpha$-level prediction interval is obtained from the predictive system using the $\alpha/2$-quantile of Π_u and the $1-\alpha/2$-quantile of Π_ℓ. While these bounds may be defective distribution functions, these quantiles exist in all cases here. Since the predictive systems contain a probabilistically calibrated predictive distribution out-of-sample, the resulting prediction intervals contain Y_{n+1} with probability $\geq 1-\alpha$. The prediction intervals are then assessed with respect to their unconditional coverage, average width, and average interval score; the interval score is a popular proper scoring rule to compare prediction intervals, and a lower score is desired (Gneiting and Raftery, 2007).

Since conformal predictive systems and the RDPS are equivalent in the split conformal setting, we restrict attention to the full conformal setting. Conformal predictive systems with a conformity measure of the form (9) have been proposed using ordinary least squares regression (the LSPM; Vovk et al., 2017) and kernel ridge regression (the KRRPM; Vovk et al., 2018a). For comparison, we therefore implement these two regression models within the RDPS. Following Vovk et al. (2018a), we employ kernel ridge regression using the Laplacian (exponential) kernel, with the regularisation parameter estimated using cross-validation on an independent sample of the simulated data. We use the deleted versions of the RDPS, as discussed in Remark 14.

While the conformal predictive systems have a fixed thickness of $1/(n+1)$, the thickness of the RDPS changes as a function of the covariate. Large covariates typically lead to larger thicknesses, aligning with the idea that the thickness provides a measure of epistemic forecast uncertainty. The distribution of the thickness of the RDPS over the 1000 repetitions is shown in Figure 3. Since kernel ridge regression is more flexible than ordinary least squares, it typically results in a larger thickness. This is especially true in the nonlinear case.

The coverage, width, and interval score of the four predictive systems are shown in Figure 4 for the linear case. The unconditional coverage of the prediction intervals is approximately equal to the desired coverage level in all cases. Due to the larger thickness of the RDPS with kernel ridge regression, this approach yields slightly wider, more conservative prediction intervals when the level $1-\alpha$ is close to 0.5. Nonetheless, the width and coverage of all intervals is very similar for the four methods, and they therefore all result in similar interval scores. The more parsimonious least squares approaches slightly outperform their kernel ridge regression counterparts, but there is very little to distinguish the RDPS from the classic conformal predictive systems.

The analogous results are shown for the nonlinear case in Figure 5. The conservativeness of the kernel ridge regression RDPS intervals is even more prevalent in the nonlinear case, owing again to its increased thickness. The kernel ridge regression intervals are nonetheless generally shorter than those obtained using

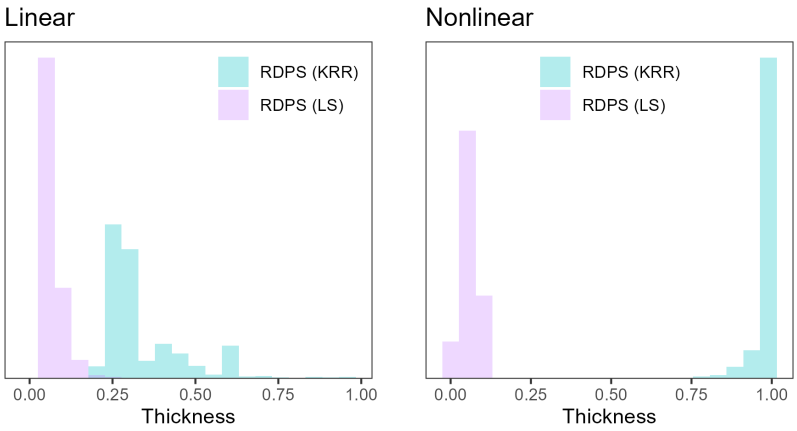

Figure 3: Histograms of the thickness of the four predictive systems.

least squares (with either type of predictive system). This is because the least squares predictions do not capture the nonlinearity in the data (Figure 1), leading to highly-dispersed predictive distributions, and thus wide prediction intervals. As a result, there is a clear difference between the interval scores assigned to the prediction intervals obtained from least squares regression and kernel ridge regression. Nonetheless, for both choice of regression model, the results are again similar for conformal predictive systems and the RDPS, with conformal predictive systems offering a slight advantage at most prediction levels.

While we restrict attention to least squares and kernel ridge regression, since these facilitate a comparison with the LSPM and KRRPM in the classic conformal prediction framework, the RDPS could also be implemented with more flexible regression models. This offers a distinct advantage over conformal predictive systems that only allow for regression models satisfying the monotonicity condition at (8).

6 Conclusions

Conformal prediction provides a means to obtain prediction sets with out-of-sample calibration guarantees. It has recently been shown that conformal predictive systems can be interpreted as one example of a more general framework to construct out-of-sample calibrated predictive systems (i.e. sets of probabilistic forecasts for real-valued outcomes). Alternative approaches have therefore been proposed that similarly

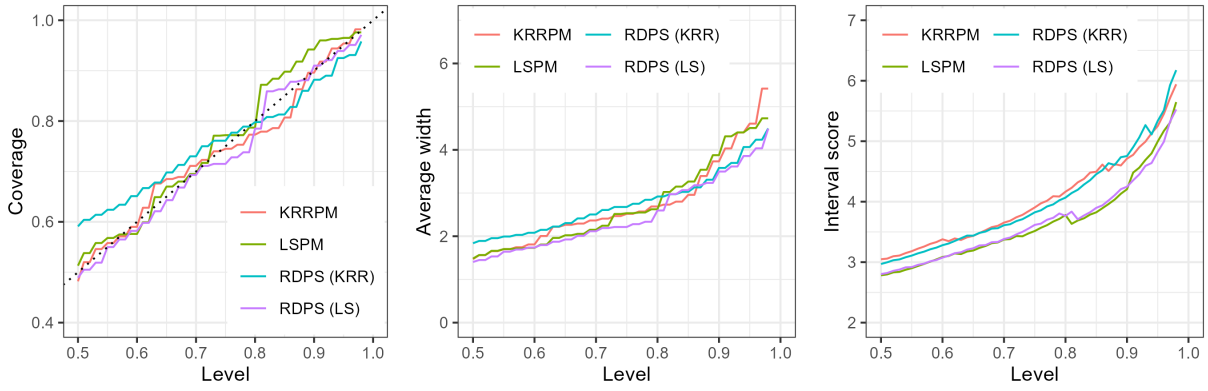

Figure 4: Coverage, average width, and average interval score of prediction intervals obtained from the three predictive systems, shown as a function of the level of the prediction interval. Results are shown for the linear simulated data.

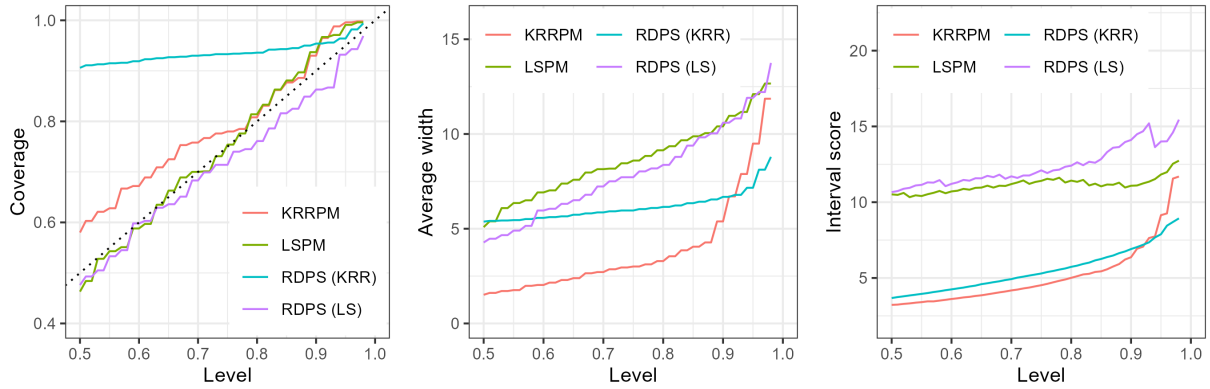

Figure 5: As in Figure 4 but for the nonlinear simulated data.

construct predictive systems that are guaranteed to contain a calibrated forecast distribution out-of-sample, possibly satisfying a stronger notion of calibration. In this work, we rigorously analyse Residual Distribution Predictive Systems, which are predictive systems obtained via a simple forecasting procedure based on a point-valued regression method and the empirical distribution of its residuals in the training data.

We demonstrate that this approach coincides with conformal predictive systems in the split conformal setting, facilitating a better understanding of classical conformal prediction. This equivalence holds for the typical conformity measures used in practice, based on (transformed) residuals of regression methods. The two approaches differ in the full conformal setting, where the validity of conformal predictive systems depends on the conformity measure satisfying a relatively stringent monotonicity condition. In contrast, the new approach can directly be applied with any method, facilitating the design of predictive systems based on more powerful forecasting methods, such as random forests and neural networks. However, while this approach can be applied alongside any regression method, we find that the thickness of the resulting predictive system can be prohibitively large in practice. This can be circumvented by using robust regression methods.

Moreover, the thickness of the residual distribution predictive systems will generally change depending on the model choice and covariate information. In contrast, the thickness of conformal predictive systems is fixed at $1/(n+1)$. It is often suggested that the size of the predictive system provides a measure of epistemic forecast uncertainty, and since the thickness of conformal predictive systems does not depend on the covariate information, it suggests that this measure of epistemic uncertainty only accounts for the epistemic uncertainty induced by the sample size. Further work is still needed to assess estimates of this second order forecast uncertainty.

Finally, conformal as well as Residual Distribution Predictive Systems satisfy a fairly weak notion of forecast calibration. While this notion is still useful in practice, other more complex methods have been proposed that generate predictive systems with stronger out-of-sample calibration guarantees. However, these methods typically require large amounts of data, whereas the two approaches discussed herein are more parsimonious, rendering them easy to implement even for small sample sizes. It is in these low data regimes that the RDPS is most beneficial; when data is rich, a split conformal approach can be applied, in which case the novel approach is generally equivalent to conformal predictive systems. Overall, conformal predictive systems remain a valid choice in standard settings, especially when monotonic conformity measures are applicable. Residual Distribution Predictive Systems, on the other hand, offer a more flexible framework that accommodates a broader range of regression techniques, making it particularly suited to scenarios involving complex or nonstandard data structures. Future research could therefore further explore the integration of other regression techniques into this framework.

References

Allen, S., Gavrilopoulos, G., Henzi, A., Kleger, G.-R., and Ziegel, J. (2025). In-sample calibration yields conformal calibration guarantees. *arXiv preprint arXiv:2503.03841*.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A*, 147:278–290.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39:863–883.

Fontana, M., Zeni, G., and Vantini, S. (2023). Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29:1–23.

Gammerman, A. and Vovk, V. (2002). Prediction algorithms and confidence measures based on algorithmic randomness theory. *Theoretical Computer Science*, 287:209–217.

Gammerman, A., Vovk, V., and Vapnik, V. (1998). Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI'98)*, pages 148–155, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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

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

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

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

Henzi, A., Ziegel, J. F., and Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B*, 85:963–993.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094–1111.

Papadopoulos, H., Gammerman, A., and Vovk, V. (2008). Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, pages 64–69.

Salibian-Barrera, M. (2023). Robust nonparametric regression: review and practical considerations. *Econometrics and Statistics*.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421.

Shen, Y., Xia, D., and Zhou, W.-X. (2024). Online quantile regression. *arXiv preprint arXiv:2402.04602*.

Tsyplakov, A. (2014). Theoretical guidelines for a partially informed forecast examiner. MPRA paper 67333, `https://mpra.ub.uni-muenchen.de/67333/`.

van der Laan, L. and Alaa, A. (2025). Generalized Venn and Venn-Abers calibration with applications in conformal prediction. *arXiv preprint arXiv:2502.05676*.

Vovk, V., Gammerman, A., and Shafer, G. (2022). Algorithmic learning in a random world.

Vovk, V., Nouretdinov, I., Manokhin, V., and Gammerman, A. (2018a). Conformal predictive distributions with kernels. *In Braverman Readings in Machine Learning. Key Ideas from Inception to Current State: International Conference Commemorating the 40th Anniversary of Emmanuil Braverman's Decease, Boston, MA, USA, April 28-30, 2017, Invited Talks*, pages 103–121.

Vovk, V., Nouretdinov, I., Manokhin, V., and Gammerman, A. (2018b). Cross-conformal predictive distributions. *Proceedings of Machine Learning Research*, 91:37–51.

Vovk, V., Petej, I., and Fedorova, V. (2015). Large-scale probabilistic predictors with and without guarantees of validity. In *Advances in Neural Information Processing Systems*, volume 28.

Vovk, V., Shen, J., Manokhin, V., and Xie, M. (2017). Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, 108:445–474.

A Computation

In this appendix, we provide examples of regression models where the prediction difference $\hat{y}'_{n+1} - \hat{y}'_i$ is an increasing function of y', for all $i = 1, \ldots, n$, possibly under some conditions.

Example 16 (OLS regression)**.** Consider an ordinary least squares regression model that issues predictions $\hat{y}'_i$ given training data $(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y')$. Let H denote the hat matrix of the regression fit, with elements $h_{i,j}$, for $i, j = 1, \ldots, n+1$. Up to an additive constant that does not depend on y', we have

$$\hat{y}'_{n+1} - \hat{y}'_i = (h_{n+1,n+1} - h_{i,n+1})y'.$$

This difference is therefore an increasing function of y' whenever $h_{n+1,n+1} \geq h_{i,n+1}$ for all $i = 1, \ldots, n$. Since $h_{i,n+1} \in [-0.5, 0.5]$, this is satisfied if $h_{n+1,n+1} \geq 0.5$. In contrast, Vovk et al. (2022, Proposition 7.4) states that the LSPM satisfies the monotonicity condition if $h_{n+1,n+1} < 0.5$. An analogous result holds for kernel regression.

Example 17 (Kernel-density estimation)**.** Consider a kernel-density prediction method that, given training data $(x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y')$, issues predictions

$$\hat{y}'_i = \frac{\sum_{j=1}^{n} \exp(-\|x_i - x_j\|)y_j + \exp(-\|x_i - x_{n+1}\|)y'}{\sum_{j=1}^{n+1} \exp(-\|x_i - x_j\|)}.$$

When predicting y_i, this essentially issues a weighted average of the observations in the training data, where the weights decrease exponentially as the covariate moves away from x_i. A fixed bandwidth parameter could also be used in the exponentials without changing the following analysis. Up to an additive constant that does not depend on y', we have

$$\begin{aligned}\hat{y}'_{n+1} - \hat{y}'_i &= y' \left\{ \frac{1}{\sum_{j=1}^{n+1} \exp(-\|x_j - x_{n+1}\|)} - \frac{\exp(-\|x_i - x_{n+1}\|)}{\sum_{j=1}^{n+1} \exp(-\|x_j - x_i\|)} \right\} \\ &\propto y' \left\{ \sum_{j=1}^{n+1} \Big[\exp(-\|x_j - x_i\|) - \exp(-(\|x_j - x_{n+1}\| + \|x_i - x_{n+1}\|)) \Big] \right\}.\end{aligned}$$

By the triangle inequality, $\exp(-\|x_j - x_i\|) \geq \exp(-(\|x_j - x_{n+1}\| + \|x_i - x_{n+1}\|))$, for all $i, j = 1, \ldots, n+1$. The factor multiplied by y' above is therefore non-negative, and hence $\hat{y}'_{n+1} - \hat{y}'_i$ is an increasing function of y', for all i. The RDPS corresponding to this regression method can therefore be calculated efficiently using two runs of the algorithm. While this approach is not robust, we can simply trim the predictions by replacing y_j and y' above with $\max\{\min\{y_j, t_1\}, t_2\}$ and $\max\{\min\{y', t_1\}, t_2\}$, for some cut-off thresholds $t_1, t_2 \in \mathbb{R}$.

Even if $\hat{y}'_{n+1} - \hat{y}'_i$ is not increasing in y', it may still be possible to efficiently implement the RDPS. This is possible, for example, when the predictions are linear functions of the observations in the training data. In particular, if $\hat{y}_i = a_i y' + b_i$, where a_i and b_i are constants in the sense that they do not depend on y' and only on $x_1, \ldots, x_{n+1}$ and $y_1, \ldots, y_n$, then we have

$$G_{x_{n+1}}^{\hat{\mathbb{P}}_{n+1}(y')}(y) = \frac{1}{n+1}\left(\sum_{i=1}^{n} \mathbb{1}\{(a_{n+1} - a_i)y' \le y - b_i)\} + \mathbb{1}\{y' \le y\}\right).$$

In this case, the indicator functions only change at the values $c_i(y) = (y - b_i)/(a_{n+1} - a_i)$, $i = 1, \ldots, n$, and $c_{n+1}(y) = y$. Hence, the bounds $\Pi_{\ell, x_{n+1}}^{\hat{\mathbb{P}}_n}(y)$ and $\Pi_{u, x_{n+1}}^{\hat{\mathbb{P}}_n}(y)$ can be calculated by evaluating $G_{x_{n+1}}^{\hat{\mathbb{P}}_{n+1}(y')}(y)$ at $y' \le c_{(1)}(y)$, $y' \in (c_{(i-1)}(y), c_{(i)}(y)]$, for $i = 2, \ldots, n+1$, and $y' > c_{(n+1)}(y)$, and taking the minimum and maximum over these $n+1$ distinct values.