Evaluating Probabilistic Forecasts Using PIT

Rebecca Silva, Casey Gibson, Nick Reich

UMass-Amherst Department of Biostatistics and Epidemiology

July 20, 2018

Abstract

The Probability Integral Transform (PIT) is a useful metric to assess the validity of a predicted distribution. This summer the PIT method was used to evaluate five ensemble forecasts for seasonal influenza in the United States. This paper discusses the theory and application of PIT, the evaluation of ensemble forecasts using PIT, and the interpretation of PIT in the context of public health. Although none of the models can be validated by PIT, the visual assessment of each model, as a whole and separated by target, give insight into how models forecasts perform. The shapes of each histogram can help specify mispecifications in our models, but before cementing any misspecification, they must be supported by complementary metrics. The PIT metric is a helpful step in evaluating a model but it is not an all-encompassing metric of validation, especially in the context of the influenza ensemble models where other factors contribute to a models effectiveness.

1. Introduction

Since the 2013-2014 influenza season, the CDC has run a Forecast the Influenza Season Collaborative Challenge(FluSight competition) where teams interested in forecasting seasonal influenza in the United States submit real-time models. These models are watched throughout a season and evaluated based on a log score metric at the end of a season. In 2017, a team of influenza forecasters formed the FluSight Network, a collaborative multi-institutional and multi-disciplinary group whose goal was to build a real-time ensemble model using a diverse array of 21 models. Of the five ensemble models, the four that used past performance to determine the weights of the component models performed better than any other forecast model by itself. The FluSight Network used a log score metric to select the best performing ensemble model, which they then submitted to the CDC for the 2017/2018 season [5].

Another commonly known metric that can be used to evaluate ensemble models is that of the Probability Integral Transform. The theory of PIT says that data from any continuous distribution can be transformed into having a standard uniform distribution using its cumulative distribution function (CDF). In other words, if we plug data into its own CDF, we get outputs, known as PIT values, that are uniformly distributed. We can use this theory in testing the validity of a model or distribution for which the true distribution of each observed value is unknown and unobservable. After obtaining the PIT values, we can most easily understand their distribution by plotting the values in a histogram. Thus, the PIT method can be reduced to the visual inspection of the shape of the histogram. If the PIT histogram looks uniform, the next step is a uniformity test like the Kolmogorov-Smirnov or Anderson-Darling test [4], or visual assessments of the autocorrelation (ACF) and partial correlation functions (PCF) of the PIT values [2]. The ACF and PCF functions help assess the dependence of PIT values since we want the values to be independent as well as uniform. Moreover, even if by appearance the histogram looks uniform and tests confirm uniformity, it is important to note that uniformity from the PIT method is a necessary but not sufficient condition to validate a model.

On the contrary, if the PIT histogram displays a shape other than a uniform histogram, the shape given by the histogram can sometimes hint at misspecifications in a model. Although specific shapes like a symmetric hump or U-shape may indicate less variation or more variation, respectively, from the true distribution than the model predicts it is important not to interpret the shapes to mean one thing. These shapes, along with trends like skewness, can also indicate conditional bias. As we will find in this paper, the PIT can be used to learn about the behavior of a model and can encourage further analysis as a result of visual diagnostics.

2. PIT Background

PIT Theory

Given data from any continuous distribution, we can transform the data into having a standard uniform distribution using its own cumulative distribution function.

Theorem. *Probability Integral Transform: Let F_X be the CDF of a continuous distribution from which we have random variable X. Then,*

$$F_X(X) \sim U(0,1)$$

.

Proof. Let $Y = F_X(X)$, where X is a continuous random variable with CDF, F_X. Take $t \in [0,1]$. Then,

$$\begin{aligned}
F_Y(t) &= P(Y < t) && \text{by definition of a CDF, } F_Y \\
&= P(F_X(X) < t) && \text{by definition of } Y \\
&= P(X < {F_X}^{-1}(t)) && \text{take inverse of } F_X \\
&= F_X({F_X}^{-1}(t)) && \text{by definition of a CDF, } F_X \\
&= t
\end{aligned}$$

Thus Y has a uniform distribution on the interval [0,1] by CDF of $U(0,1)$[1]. □

Application of PIT

A common application of PIT is testing whether a set of values from an unknown distribution can be accurately modeled by a predicted distribution. Given the PIT theory, if we plug the set of values into the CDF of the predicted distribution, the output, otherwise known as the PIT values, should be uniformly distributed if the predicted distribution matches the true distribution.

We begin with an example from the normal distribution. Although the true distribution is usually not known, for the purposes of this example, we know the true distribution to be the standard normal distribution, $N(0,1)$. However, the true distribution is unknown to the modeler. The modeler looks at the observed set of values and predicts the true distribution to be N(0,3).

Let X be a sample of 10,000 random variables from N(0,1). We plug the sample into the CDF of N(0,3) and get a set of PIT values. To see if the PIT values have a uniform distribution, we plot them on a histogram. Figure 1 shows this progression.

First, we sample from N(0,1), so we treat the distribution N(0,1), as the true distribution unknown to the modeler. The middle plot shows a comparison between the two CDFs. As we plug in the sample into the CDF

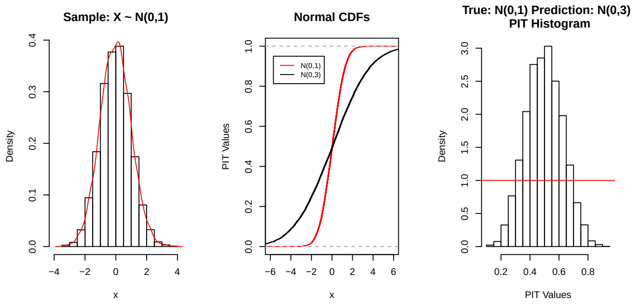


Figure 1: Normal Example

of N(0,3), we get PIT outputs on the y-axis, which are then modeled on the x-axis in the right plot. The right plot shows a hump-shape in the histogram which implies the predicted distribution does not display a uniform distribution. The CDF of N(0,3) is less steep around 0 due to a larger variance, so when we plug in samples from N(0,1) into this CDF, we get more outputs around .5 and less outputs around 0 and 1.0 than expected. This results in a larger density around .5 and smaller density on the tails as displayed by a hump-shaped histogram. When we plug the true data into the true distribution we see a uniform distribution in Figure 2 (as we should).

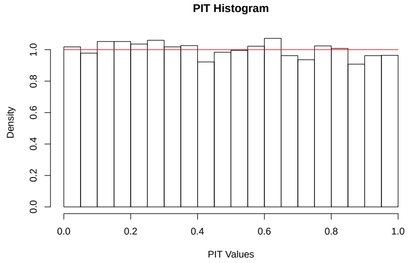


Figure 2: Normal Example

3. Ensemble Models

From One Distribution to Many

Up to now, we have only seen the PIT method used with data coming from the same distribution. However, the metric also holds with models that have many different distributions. Using the PIT metric for density forecasts was proposed and proved in Diebold et al, (1997)[2].

Although harder to visualize when working with many distributions and only using one sample from each, the PIT theory still holds. Given one sample value from a predictive distribution, the PIT value corresponding to that value represents the probability that a random variable from the distribution is less than the sample value. If each predicted distribution is equal to the true distribution, then the PIT value of each sampled value will accurately give a probability and those probabilities will be uniformly distribution no matter how diverse the distributions are. If the reader wants to witness for themselves an example of samples all from different distributions, they can use the R code from *many_dist_ex.R* [7]. We see that PIT values from 10,000 different variations of normal and exponential distributions are uniformly distributed.

FluSight Network Models

As mentioned in the introduction, the ensemble models built by the FluSight Network consist of 21 models submitted for the FluSight competition run by the CDC. To give the reader a better understanding of what these models are trying to accurately forecast, we present an overview of the structure of the models. Since 2010, for each week during a season of influenza (early November to mid April), the CDC has collected data from all 11 HHS regions of the U.S. for seven targets. The seven targets of interest are: four weighted percentages of doctor reported cases of influenza-like-illness (wILI) corresponding to the subsequent 4 weeks of the season, the peak weighted influenza-like-illness (wILI), and two week numbers corresponding to the season onset and the week in which the maximum wILI occurs. For each model and season, the forecast is contained in a text file, which gives a predictive distribution of each target in each region. The five ensemble models incorporate the 21 models in different ways by using different types of weights. The five different types of weights are described in Table 1.

For the purpose of evaluating the ensemble models using PIT, we focus on the five targets that predict a continuous wILI percentage which are the four subsequent week-ahead targets and the peak wILI percentage. Although the predictive distributions are discretized by .1 to assign probabilities to corresponding wILI percentages, we will treat these target values as continuous.

Model	No. of weights	description
Equal weights (EW)	1	Every model gets same weight.
Constant weights (CW)	21	Every model gets a single weight, not necessarily the same.
Target-type-based weights (TTW)	42	Two sets of weights, one for seasonal targets and one for weely wILI targets.
Target-based weights (TW)	147	Seven sets of weights, one for each target separately.
Target-and-region-based weights (TRW)	1,617	Target-based weights estimated separately for each region.

Table 1: Ensemble Models.[6]

To understand the predictive distributions in each ensemble model we first discuss one distribution. Since we are dealing with ensemble models, the predictive distribution is different based on the season, evaluation week, region, and target. There is one observed value for each distribution and therefore one PIT value per distribution. Since we are using the PIT method on seven seasons, 10 regions, and 5 targets for 30 evaluation weeks per season, we obtain 10,500 PIT values per ensemble model.

The predictive distribution is given on a scale of 0 to 13.1 with bins separated by .1 wILI, and one bin from 13.1 to 100 to cover the small probability that the wILI would exceed 13.1. Figure 3 shows the probability distribution for the 1 week ahead target from the Constant weight model at evaluation week 1 during season 2010/2011 at HHS Region 1. Note that evaluation week 1 corresponds to calendar weeks and not season weeks, so week 1 in season 2010/2011 corresponds to the first week of January 2011 whereas the first evaluation week of the season is in the first week of November 2010. On the left of Figure 3, we see the probability distribution function and use an empirical cumulative summation function to obtain the cumulative distribution function on the right. The observed value for that week was 1.2 wILI which corresponds to a .774 PIT value.

4. PIT Histograms

Now that we understand how each PIT value is obtained, we put all the PIT values together for each ensemble model to study the shape of the PIT histogram. Figure 4 displays all five ensemble model PIT histograms.

There is a notable difference in shape between the first row and second row. The Equal weight and Constant weight models show a more hump-shaped histogram which can indicate that the observed data varied less than

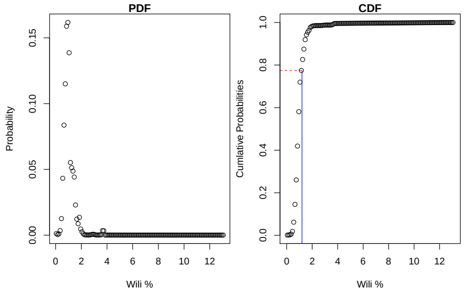


Figure 3: Constant Weight, Evaluation week 1, 1-week-ahead, HHS Region 1, Distribution

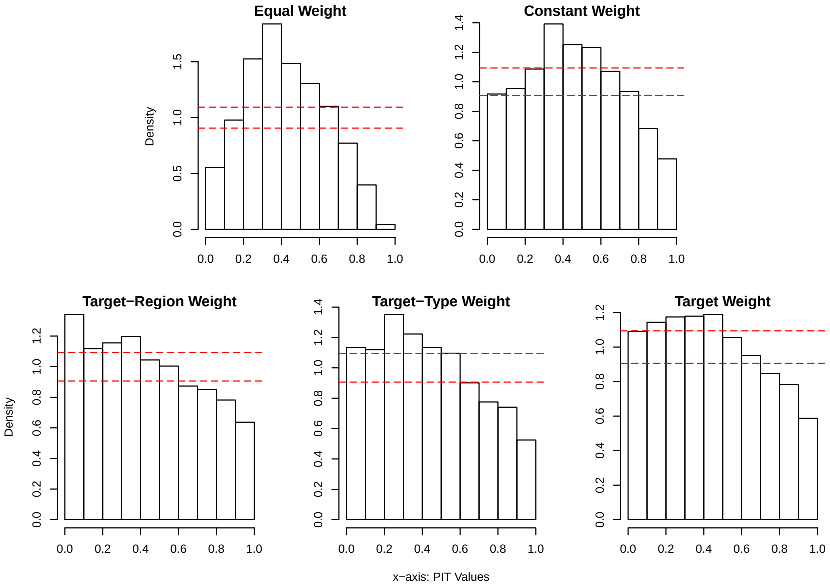


Figure 4: Ensemble PIT Histograms

the models predicted. A common trend with all the histograms is that there are more PIT values under .5 than PIT values above .5. Since a PIT value of .5 corresponds to an observed value at the estimated median of the distribution, a higher density of PIT values below .5 implies that there are more observed values below the median than above. We see a common right skew in all the ensemble model histograms and that the last three percentiles, corresponding to PIT values of .7 to 1.0, have a low density in all the models. A low density signifies fewer observations with PIT values of .7 to 1.0, which means there are fewer observations in the right tail of their respective distributions than expected by the distributions. Fewer values in the tail could suggest a misspecification of bias in the ensemble models where the models tend to overestimate the target values. This could be checked in a bias measurement which we discuss further in section 5. It is possible that if large outbreaks have occurred in the past that are taken into account in the building of the models, the models may have distributions that allow for higher wILI values even if most seasons never reach those values. Furthermore, it could be that the models purposely have wider right tails then they need in order to hedge against large outbreaks. We consider this idea further in the discussion.

Targets

It is interesting to look further at the PIT values for each ensemble model. Considering we are dealing with so many PIT values, we can not assume the general shape in each model is present when divided up by target and region. We first look at one model, separated by targets. Then we look at each target separated by ensemble model. Figure 6 shows the Target-Type based weights model separated by targets. The red dashed lines represent a 95 % confidence interval under the null hypothesis of U(0,1), considering the number of breaks and PIT values.

Although we see a consistent pattern of low right tails and slight right skews, we would expect more of the histograms to follow the shape of the main Target-Type model with all of its PIT values. The new shapes seen in the 1-week-ahead and Season peak percentage histograms may indicate that the ensemble models' performance differs between targets. With the 1-week-ahead target, the histogram shape is indicating that the observed values are not reaching the end tails of the models' distributions as much as predicted. In other words, this could be some form of overdispersion where the model expects a larger variance than what is observed. It makes sense

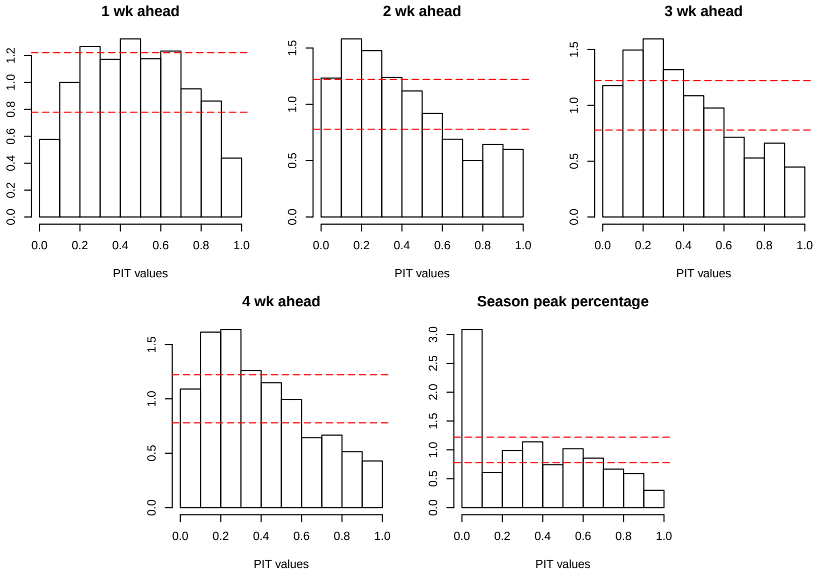
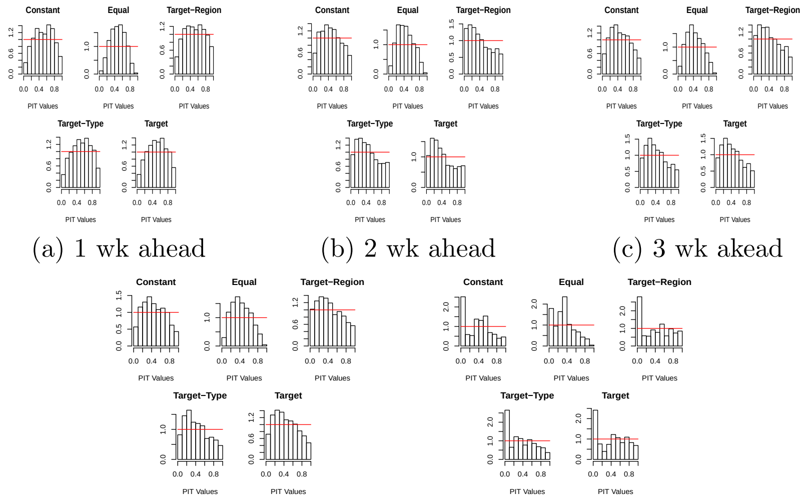


Figure 5: Target-Type Weights Model by Target

that observed values for a 1-week-ahead target vary less than observed values for a 4-week-ahead target because the forecast reaches further into the future, but that is something that can be taken into account by the model, so we can not assume the variance is misspecified in this way. On the other hand, the Season peak percentage target histogram shows a high density in the first percentile, indicating that many observed values fall on the left tail of the distribution. Ultimately, by breaking up each model by target, we see we must be more careful when interpreting each model's main PIT histogram with all PIT values; the overall shape is clearly not present in each target histogram. The question then follows: is there a similar distinction between targets in all ensemble models?

Evidently, there seems to be more difference in shapes between the targets than between the models themselves. For the 2-week-ahead, 3-week-ahead and 4-week-ahead targets, we see a slight right skew where as the 1-week-ahead target gives a hump-shaped histogram and the Season peak percentage shows a notably high density in the first percentile.

Figure 6 is most interesting because we see the value in the PIT metric. Through visually understanding the patterns of observed values in each target and learning how the performance varies by target, we get insight into the



(a) 1 wk ahead (b) 2 wk ahead (c) 3 wk akead

(d) 4 wk ahead (e) Season peak %

Figure 6: Targets by Model

component models, indicating that they all tend to forecast each target similarly. From here, it may be easier to improve models by investigating certain predictors that factor into forecasts differently depending on the target.

We are careful to interpret the models separated by region because there is more variability that is hard to account for between regions. For example, some regions have more data than other regions.

5. PIT complemented by another metric

Although a true bias was not explored here, we do look into an error measurement that uses the median of each distribution. We define the error measure as: error = point estimate − observed value. The point estimate in this case is the median of the predictive distribution.

The data in Table 2 helps complement findings in the PIT histogram and shows us why many different evaluation metrics are helpful in understanding the models as a whole. In the table we note the count for positive and negative error values, and the total and mean of all the error measures.

At first, a negative average error (indicating the mean of observed values is greater than the mean of point estimates) might be surprising because we have been noting more observations below the median (PIT values $< .5$) than above, but here we are working with the median as opposed to the mean.

Models	**Count** $+error, -error$	**Sum Error**	**Avg Error**
Constant	**5678**, 4771	−2815.8	−0.267
Equal	**7185**, 3364	−1096.797	−0.104
Target-Region	**5544**, 4705	−2846.2	−0.270
Target	**5509**, 4793	−3267.9	−0.310
Target-Type	**5600**, 4715	−2781.104	−0.264

Table 2: Error.

Therefore, the observed values above the median (which makes the error measure negative), are relatively farther from the median than the observed values below the median due to the wide right tail in all distributions. If instead, we used the mean point estimate, we could understand something about possible bias in the models and suggest either a general overestimation or underestimation of the observed values based on positive or negative error, respectively.

Although we can not interpret bias from our error measure, the error measure complements our histogram shapes regarding the count of positive errors versus negative errors. The number of positive errors is greater than negative errors which translates to more observed values below the median than the number above. This is an example of how the visual quality of the PIT metric can be more effective in showing a pattern than numbers can in an analysis. Namely, the trend we notice in our histograms, of there being a higher density of PIT values below .5 as opposed to above, is represented through the number of positive and negative error measures in Table 2 as *count*, but this statistic is not an obvious one to investigate and it less informative than the visual representation we get from a PIT histogram.

6. Discussion

As discussed throughout this paper, we have to be careful in interpreting the histogram shapes. Unlike in our normal distribution example from section 2, we are working with PIT values from many different distributions so it is harder to identify the exact misspecification. For example, when we divided each model up by target, we found varied shapes between the targets, demonstrating that we cannot generalize interpretations from the main ensemble PIT histograms to all targets and regions. We can, though, think about why the trend of a right skew and low left density on each plot is seen throughout all histograms. We will discuss two hypotheses for why we are not seeing a uniform-shape histogram and why this right skew may not indicate our models are invalid.

Firstly, because there are $10,500$ different distributions, it may be unrealistic to expect a uniform-looking histogram. As we know, the PIT considers the mean, variance, and shape of a predictive distribution. Even if the mean and variance of a model are close to the true mean and variance, if the shape of the distribution is not close to that of the true distribution, the PIT histogram can look skewed. Figure 6 demonstrates an extreme example using two distributions with distinct shapes. The model predicts a gamma distribution although the true distribution is a normal distribution. But, the model correctly predicts the mean and variance. The PIT histogram is clearly not uniform because the distributions were not similar enough. Our example shows how shape is an important factor in model evaluation with the PIT. Although the ensemble models surely do have closer distribution shapes to their respective true distributions than the shapes in this example, since we are testing $10,500$ distributions for each model, it is unrealistic to assume our distributions are all very close, if not equal, to their respective unobservable true distributions. Therefore, it may be that the sheer number of shapes plays a role in the non-uniform trends we are seeing.

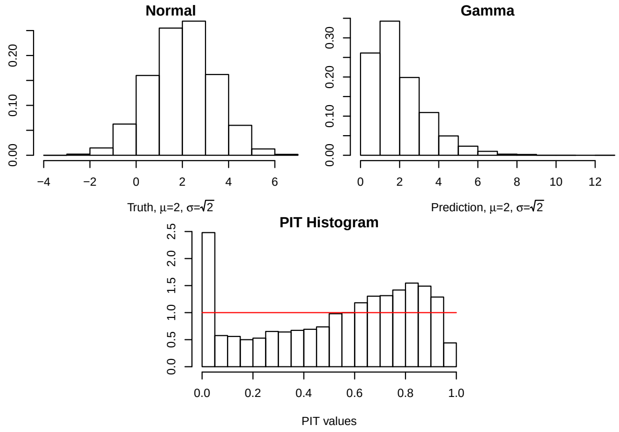


Figure 7: Different Shapes Example

Secondly, we ask ourselves: what exactly about the ensemble data might imply that a uniform histogram is not necessary to have a valid model? If we consider the models from a public health standpoint, it is unclear whether we want to see a uniform shape. A right skew shows us that in general, fewer observations fall in the right tail of their respective distributions. Thus, our distributions have a wider right tail distribution than needed. In the context of public health, large wILI percentages have a significance because they can indicate a more varied season or a large outbreak. Thus, it could be beneficial for our models to hedge against large outbreaks and in some way be prepared for them by placing more weight on observed values that fall above the median of each distribution than observed values that fall below. From a purely statistical view, low and high values have the same meaning but in our context, a level of precaution with high wILI values may be constructive for the CDC.

In conclusion, the PIT metric can tell us a lot about our models. Although it is evident that the PIT metric should be supported by measurements of bias and other evaluation metrics, it is useful in understanding the performance of models. In contrast to the log score obtained by the CDC, which only gives one value for each model, the PIT histogram gives insight into patterns of observed values in their respective distributions and can help improve a model by showing how its performance varies by target. Additionally, the visual aspect of the PIT method is unique, and the effectiveness of visualizing performance should not be overlooked; in many cases, something that is visually apparent might not be noticeable in data. Ultimately, the PIT metric can be a very informative metric but we must recognize it is not an all-encompassing metric to validate a model, especially when there are outside factors, such as the implications of an influenza outbreak, that affect the strength of a forecast.

References

Other relevant articles: [3], [4],[5] and github repository with all relevant code [7].

[1] Probability integral transform. `https://en.wikipedia.org/wiki/Probability_integral_transform`, Feb 2018.

[2] Francis Diebold, Todd Gunther, and Anthony Tay. Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 1997.

[3] Thomas M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560, 2001.

[4] Pablo Noceti, Jeremy Smith, and Stewart Hodges. An evaluation of tests of distributional forecasts. *Journal of Forecasting*, 22(6-7):447–455, 2003.

[5] Nicholas G Reich, Logan Brooks, and Sasikaran Kandula. Forecasting influenza in the u.s. with a collaborative ensemble from the flusight network, Nov 2017.

[6] Nick Reich. Building a collaborative ensemble to forecast influenza. `http://reichlab.io/2017/11/28/flusight-ensemble.html`, Nov 2017.

[7] Rebecca Silva. rsilva19/summer-ensembleforecast-pit. `https://github.com/rsilva19/Summer-EnsembleForecast-PIT/`.