Influenza Forecasting Framework based on Gaussian Processes

Christoph Zimmer [1] **Reza Yaesoubi** [2]

Abstract

The seasonal epidemic of influenza costs thousands of lives each year in the US. While influenza epidemics occur every year, timing and size of the epidemic vary strongly from season to season. This complicates the public health efforts to adequately respond to such epidemics. Forecasting techniques to predict the development of seasonal epidemics such as influenza, are of great help to public health decision making. Therefore, the US Center for Disease Control and Prevention (CDC) has initiated a yearly challenge to forecast influenza-like illness. Here, we propose a new framework based on Gaussian process (GP) for seasonal epidemics forecasting and demonstrate its capability on the CDC reference data on influenza like illness: our framework leads to accurate forecasts with small but reliable uncertainty estimation. We compare our framework to several state of the art benchmarks and show competitive performance. We, therefore, believe that our GP based framework for seasonal epidemics forecasting will play a key role for future influenza forecasting and, lead to further research in the area.

1. Introduction

The seasonal epidemic of influenza causes a tremendous burden on public health each year (Chretien et al., 2014). In the U.S. alone it is responsible for 9.2 to 35.6 million cases, 140,000 to 710,000 hospitalizations, and 12,000 to 56,000 deaths (Centers for Disease Control and Prevention, c). To allow for better public health policies and resource allocation, it is important to have reliable forecasts of future influenza development at hand. Therefore, the Center for Disease Control and Prevention (CDC) has invested effort into research on forecasting Influenza. This includes establishing a challenge to forecast influenza activities. In this challenge, various research groups utilize a variety of techniques and data sets to forecast difference influenza-related targets such as the number of influenza cases during future weeks.

Since the establishment of this challenge in 2013, progress has been made (Biggerstaff et al., 2016), within two directions: identifying and utilizing new data sources and developing new forecasting methodologies. The CDC influenza-like illness (ILI) data is published with a 1-3 weeks delay (Paul et al., 2014), so indirect online data has been used to obtain real-time ILI estimates. Google Flu trends (Ginsberg et al., 2009) uses the frequency of influenza related search terms such as ”cough” or ”fever” to estimate the number of ILI cases in real-time. Similarly, Twitter texts can be analyzed (Broniatowski et al., 2013) or the number of visits to Wikipedia pages (McIver & Brownstein, 2014).

With respect to methodological advances, a variety of statistical and mechanistic models have been developed to use both CDC ILI data and other real-time data sources to forecast influenza activities.(Adhikari et al., 2019; Brooks et al., 2015; Farrow, 2016; Osthus et al., 2017; Ray et al., 2017; Shaman et al., 2013; Yang et al., 2014; Zimmer et al., 2018a;b).

In this work, we develop a new framework for influenza forecasting that is based on Gaussian Processes regression. Applying this framework to the official CDC reference data set on influenza-like illness, we demonstrate the performance of this framework by retrospective forecasting on seven influenza seasons.

We compare our framework to several state of the art forecasting techniques and show that our results suggest clear and significant improvement for seasonal influenza forecasting in US compared to state of the art benchmarks.

Main contribution:

- A new framework based on Gaussian process regression for forecasting seasonal epidemics and quantifying the uncertainties in projections.

[1] Bosch Center for Artificial Intelligence, Renningen, Germany [2] Health Policy and Management, Yale School of Public Health, New Haven, USA. Correspondence to: Christoph Zimmer <christoph.zimmer@de.bosch.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

- Extensive benchmarking on the official CDC influenza data set compared with state-of-the- art forecasting methods.
- Statement of theoretical properties and its connection to algorithm specification.

The problem of seasonal epidemics forecasting can be briefly summarized as follows (details follow in section 3.2: for each year i and each week j, the CDC releases data on ILI, d^i_j. Now, denote the current week as j^* and the current year as i^*. The task of influenza forecasting is to provide estimates for $d^{i^*}_{j^*+t}$ for a $t \in \mathbb{N}$.

Section 2 will discuss related work, our novel framework will be introduced in Section 3 including some theoretical properties. Section 4 will demonstrate our framework's ability to provide accurate forecasts and state the results to benchmarks.

2. Related Work

Various methods have been developed for influenza forecasting. Many of them (Osthus et al., 2017; Shaman et al., 2013; Yang et al., 2014; Zimmer et al., 2018a;b) are based on physical models compartmentalizing the population in groups such as susceptible, infected and recovered. Defining transition between those groups allows to derive differential equations. Additional information like humidity dependence of infection rates can also be encoded (Shaman et al., 2013; Yang et al., 2014; Zimmer et al., 2018b). These approaches are different to our Gaussian Process (GP) based framework as they are based on physical knowledge/assumptions (e.g. differential equations). Their advantage is that they can also be applied to non-seasonal disease forecasting in which there is no previous seasons training data available. Additionally, they can also reveal insights into the key parameters driving the disease spread (Osthus et al., 2017; Yang et al., 2014; Zimmer et al., 2018a). On the other hand, they are based on physical assumptions and approximations which might lower their performance in forecasting seasonal influenza as our results indicate.

Statistical time series modeling (Brooks et al., 2015; Ray et al., 2017) and crowd based approaches (see chapter 5.3.2 of (Farrow, 2016)) do not relay on the same set of assumptions as physical models and are, therefore, more similar to our GP based framework. Recently, also deep learning techniques have been applied (Adhikari et al., 2019). However, being Gaussian process based, our framework is novel and different from those previous techniques. (Senanayake et al., 2016) use a spatio temporal covariance and data from various states and all weeks of a year to model influenza-like illness forecasting. Our approach is different by training individual GP models for each forecasts based on a relatively small set of features of previous weeks, leading to small but reliable prediction intervals.

All forecasts are based on data that has been observed until the current time. The publication of the CDC official ILI data is usually 1-3 weeks delayed (Paul et al., 2014) as reports and tests from several units have to be collected. As this delay is obviously disadvantaguous for forecasting, there have been several attempts to use indirect web based data, in order to come up with timelier estimates of the CDC's ILI. These include Google flu trends (Ginsberg et al., 2009), Twitter (Broniatowski et al., 2013), or Wikipedia (McIver & Brownstein, 2014). These attempts are usually called *nowcasting* as they try to estimate the current situation. Our goal, *forecasting*, is different from *nowcasting* as it tries to forecast the future. This means that our framework is able to make use of those nowcasting approaches to receive a more timely data stream.

3. Methods

3.1. Gaussian Processes

We employ a Gaussian Process (GP) model to approximate the function $f : \mathcal{X} \subset \mathbb{R}^d \to \mathcal{Y} \subset \mathbb{R}$ (see (Rasmussen & Williams, 2006) for more details). Let us assume that we have so far collected n samples $X = (\boldsymbol{x}^1, \dots, \boldsymbol{x}^n)$ of d dimensional inputs $x^i = (x^i_1, \dots, x^i_d)$ and corresponding outputs $Y = (y^1, \dots, y^n)$. A GP is specified by its mean function $\mu(\boldsymbol{x})$ and kernel function $k(\boldsymbol{x}^i, \boldsymbol{x}^j)$. Given noisy observations, the GP posterior is given as

$$p(y|\boldsymbol{x}, X, Y) = \mathcal{N}(y|\mu(\boldsymbol{x}), \sigma(\boldsymbol{x})) , \qquad (1)$$

where the input $\boldsymbol{x}$ is a vector and consists of dimension d, i.e. $\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$. The output y contains the corresponding output measurement. Mean and variance are defined by

$$\begin{aligned} \mu(\boldsymbol{x}) &= \boldsymbol{k}(\boldsymbol{x}, X)^T(\boldsymbol{K}_n + \sigma^2\boldsymbol{I})^{-1}Y , \\ \sigma(\boldsymbol{x}) &= \boldsymbol{k}^{**}(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x}, X)^T(\boldsymbol{K}_n + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{k}(\boldsymbol{x}, X) , \end{aligned} \qquad (2)$$

The covariance matrix is represented by $\boldsymbol{K}_n \in \mathbb{R}^{n\times n}$. In this paper, we employ the Gaussian kernel as the covariance function, i.e. $k(\boldsymbol{x}^i, \boldsymbol{x}^j) = \sigma^2 \exp(-\frac{1}{2}(\boldsymbol{x}^i - \boldsymbol{x}^j)^T \boldsymbol{\Lambda}^2_f(\boldsymbol{x}^i - \boldsymbol{x}^j))$, which is parametrized by $\boldsymbol{\theta}_f = (\sigma^2_f, \boldsymbol{\Lambda}^2_f)$. Furthermore, we have an n-dimensional identity matrix $\boldsymbol{I}$, and σ^2 as output noise variance (see (Rasmussen & Williams, 2006)). $\boldsymbol{k}^{**}(\boldsymbol{x},\boldsymbol{x}) = \sigma_f^2 \in \mathbb{R}$ and the vector $\boldsymbol{k} \in \mathbb{R}^n$ contains kernel evaluations relating $\boldsymbol{x}$ to the previous n inputs.

3.2. Framework for Seasonal Epidemic Forecasting

In this section, we will describe the framework that we use to carry out weekly forecasts. As during the course of a seasonal epidemics forecasts need to be generated in multiple weeks in a row and for various targets, we note that we sequentially create individual forecasts for each week – based on the data being available from previous weeks. To this goal, we use available data from past seasons to train Gaussian process regression for each week, and generate forecasts based on the current years data.

We will first illustrate this with an example how to use the GP framework to seasonal epidemics forecasting, and then formalize the notation: let us assume to be in epidemic week (EW) – Sunday through Saturday – 50 of year 2010 and our target is the prediction of the next weeks data, namely EW 51. We use data from past years, e.g. the years 2003 until 2006, as training data. The input training data are values for EW40 (assuming that this is the start week of the seasonal epidemic) until EW50. The output training data are the value of EW51. Using this data, we can train a GP model describing the mapping from 'data seen in EW40 until EW50' to 'data in EW51'. We can employ this model to predict the value for EW51 in 2010 given the values of EW40 until EW50 in 2010.

Denote the weekly counts of epidemic cases as d_j^i where i the season and j the week of the season, e.g. d_0^{2007} is the value of week 40 in 2007 and d_{15}^{2007} is the value of week 3 in 2008 (again assuming that the season starts in EW 40). Now, suppose we are currently in week j^* of season i^*.

First, we need to assemble the training data: the choice of the training inputs depends on the week j^* and the season i^* and we denote it by $\mathcal{X}_{j^*}^{i^*}$. Then, $\mathcal{X}_{j^*}^{i^*} = (x^i | i \in I) = (d_j^i | j \in J\ ,\ i \in I)$ with $J \subseteq \{0 \le j \le j^*\}$ being previous weeks of a season and $I \subseteq \{\text{year}_{st}, \dots, i^* - 1\}$ being past seasons and year_{st} the first year of data recording. Let our target T be the prediction of $t = 1, \dots, T$ weeks ahead. Then, our training data outputs are the values of week $j^* + t$ and are defined as $\mathcal{Y}_{j^*}^{i^*} = (y_{t,j^*}^i | i \in I)$, with $y_{t,j^*}^i = d_{j^*+t}^i$. The GP is training on the set $\{\mathcal{X}_{j^*}^{i^*}, \mathcal{Y}_{j^*}^{i^*}\}$.

Next, this GP can be used to calculated predictive distribution given the data $x^{i^*} = (d_j^{i^*} | j \in J\ )$ so far observed this season. The t-week ahead prediction, $y_{t,j^*}^{i^*}$ is distributed as

$$p\left(y_{T,j^*}^{i^*} | x^{i^*}, \mathcal{X}_{j^*}^{i^*}, \mathcal{Y}_{j^*}^{i^*}\right) \qquad (3)$$

according to equation (1). The main steps are summarized in Algorithm 1.

Algorithm 1 Seasonal Epidemics Forecasting

```
Input: current week $j^*$, current year $i^*$, forecasting horizon $T$, one or more feature set of past weeks $J_1, \dots, J_N$ and seasons $I$, data recorded so far $d_j^i$ for $j \le j^*$ and $i \le i^*$.
for $t = 1\ T$ do
    for $l = 1\ N$ do
        Assemble target $T$ specific training data inputs: $\mathcal{X}_{j^*}^{i^*} = (d_j^i | j \in J_l\ ,\ i \in I)$
        and training data outputs $\mathcal{Y}_{j^*}^{i^*} = (y_{t,j^*}^i | i \in I)$
        Train a GP based on $\{\mathcal{X}_{j^*}^{i^*}, \mathcal{Y}_{j^*}^{i^*}\}$
        Forecast target according to equation 3, resulting in $\mu_l$ and $\sigma_l$
    end for
    Build ensemble forecast over $N$ members
end for
```

While this work focuses on t week ahead prediction, the framework can easily be extended to predicting the peak of the epidemics (target T1) by defining $y_{T1}^i = \max_{0 \le j \le 52} d_j^i$. Here, we assume the epidemic to be 52 weeks long. In case of influenza, this can be shorter, e.g. ~ 30 weeks. Further targets are the final epidemic size (T2), $y_{T2}^i = \sum_{j=1}^{52} d_j^i$, or also time targets such as the timing of the peak incidence (T3), $y_{T3}^i = \text{argmax}_{0 \le j \le 52} d_j^i$ or the onset of epidemics.

The scheme described above can be applied for any week of the season and any year to forecast any of the targets. This means that for each 30 week long season and K targets, we train $30 \times K$ GP models for our forecasting.

While it is possible to only use the optimum feature set of past weeks J, it is also possible to use an ensemble of the best N feature sets of past weeks, $J_1, \dots, J_n$. We will follow this approach as ensemble models have been shown to have strong and robust performance (Reich et al., 2019a).

The sets of features of past weeks J_l and seasons I can be optimized on a validation data set and we will explain in the numerical evaluation section how we choose those sets for real epidemic situations.

3.3. Theoretical Considerations

Each new epidemic season provides information to update the training of the function approximation for f. We will here show how this improves the forecasting precision in terms of decreasing predictive variance for each individual GP based forecast of our framework:

We denote the maximal information gain after observing n data points (epidemic seasons) as $\gamma_n = \max_{\{x^i\}_{i=1}^n \subset \mathcal{X}} I(\{y^i\}_{i=1}^n, \{f(x^i)\}_{i=1}^n)$. We know from (Srinivas et al., 2010) that this information gain asymptotically behaves as $\gamma_n = \mathcal{O}(log(n)^{d+1})$ where d is the dimension of the input space, so here $d = |J|$. Further, we know from (Zimmer et al., 2018c) (appendix), Lemma 4, that the sum of predictive variances can be bounded by the maximal information: $\sum_{i=1}^n \sigma_{i-1}(x^i) \leq CI(\{y^i\}_{i=1}^n, \{f(x^i)\}_{i=1}^n)$, where σ_{i-1} denotes the predictive variance based on GP trained with $i-1$ data points and C a constant depending only on the GP hyperparameters but not on n. Therefore, we can conclude that the sum of predictive variances has the following convergence rate:

Theorem 1. *Let $\{x^i\}_{i=1}^n$ be n arbitrary epidemic seasons within a compact and convex domain $\mathcal{X}$, and k be a kernel function such that $k(\cdot,\cdot) \leq 1$*

$$\frac{1}{n}\sum_{i=1}^n \sigma_{i-1}(x^i) = \mathcal{O}\left(\frac{1}{n} log(n)^{d+1}\right)$$

Proof.

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \sigma_{i-1}(x^i)\\ &\leq \frac{1}{n} CI(\{y^i\}_{i=1}^n, \{f(x^i)\}_{i=1}^n)\\ &\leq \frac{1}{n} C\gamma_n\\ &= \mathcal{O}\left(\frac{1}{n}\, log(n)^{d+1}\right)\end{aligned}$$

where the first bound is based on (Zimmer et al., 2018c) and the asymptotic property on (Srinivas et al., 2010). ■

Note, that this statement is similar to Theorem 2 in (Zimmer et al., 2018c) as it is the same convergence rate but also quite different as Theorem 2 in (Zimmer et al., 2018c) builds up on actively selected data points x^i while our Theorem 1 does not. Epidemic seasons x^i cannot be actively selected. Even though it is the same convergence rate, our statement is weaker as we can't say that another sequence of seasons would result in smaller variances (while Theorem 2 in (Zimmer et al., 2018c) could state that another sequence of points has smaller variance as the data points are actively selected). On the other hand, our statement is stronger as it holds for any sequence of seasons (not only actively selected inputs) which is key if data comes from seasonal epidemics.

Theorem 1 also shows that, while it might be tempting to use as many previous weeks as possible (large set J), it is worth to note that his adversely affects the speed of convergence. After we observed n_0 seasons and know how much information they contained, we can further bound the uncertainty of the following seasons:

Lemma 1. *Under the assumptions of Theorem 1*

$$\frac{1}{n}\sum_{i=n_0+1}^n \sigma_{i-1}(x^i) \leq \frac{1}{n} C\gamma_n - \frac{1}{n}\sum_{i=1}^{n_0} \sigma_{i-1}(x^i)$$

Proof. As seen in the proof of Theorem 1 it holds

$$\frac{1}{n}\sum_{i=1}^n \sigma_{i-1}(x^i) \leq \frac{1}{n} C\gamma_n$$

Splitting up the sum on the left hand side and substracting the first part (up to n_0) yields the desired relation. ■

Note, that while Theorem 1 states a convergence rate for the average predictive variance, this is only the predictive variance for the model. In practice, predictions would also contain observational noise and so this explains, why we would not end up with perfect forecasts.

4. Results – Benchmarking

We will use the official CDC data set on seasonal influenza forecasting to benchmark our framework to other state of the art methods.

4.1. The CDC Influenza Data

We use the official CDC data on influenza-like illness (ILI) as data source (Centers for Disease Control and Prevention, a). The CDC defines ILI as "fever (temperature of 100F [37.8C] or greater) and a cough and/or a sore throat without a known cause other than influenza" (Centers for Disease Control and Prevention, b). This data is openly available on a weekly basis since 1997. We treat different influenza seasons independently (as there is very little interaction between influenza seasons) and assume that a season starts at the epidemic week (EW) 40. While nowcasting approaches could be used to enrich the data set, this work uses the official CDC data in order to facilitate easier reproducability and comparability.

4.2. Benchmarking

We perform retrospective forecasting in order to benchmark the performance of several state of the art influenza forecasting techniques. Retrospective forecasting assumes to be in a certain week of a previous year and tries to forecast e.g. the next weeks cases, assuming to not have yet seen the following weeks. Then, the forecasting accuracy can be evaluated on the actual observed value. We use $k = 1, \ldots, K = 4$ as the forecasting targets as in the CDC challenge on influenza forecasting (Biggerstaff et al., 2016). We do retrospective forecasting for the seasons 2012/13 until 2018/19.

To evaluate the forecast, we use log-score – defined as the logarithm of the probability within a interval around the true value – as in the influenza forecasting challenge of the CDC (Biggerstaff et al., 2016). More precisely, log-score is the logarithmic probability of the forecast being in the correct bin (of size 0.1) or one of its five precedors or succesors. While this scoring rule is still being debated (Bracher, 2019) it remains the official CDC scoring rule(Reich et al., 2019b). An illustration of the log-score rule can be found in Figure A1 of (Zimmer et al., 2018b).

The reason to choose a probabilistic criterion is that probabilistic forecasts that allow for an accurate uncertainty estimation are of much more relevance for disease forecasting than an actual point forecast. Nevertheless, we also display results of point forecasts and root mean squared error (RMSE).

We use the following methods for benchmarking: (A) Historical averages: While this is a naive benchmark, it is widely used (see e.g. (Biggerstaff et al., 2016)) as the threshold any new framework must pass. This benchmark simply uses the influenza counts of past seasons to build a kernel density estimate for predicting the target. (B) MSS is a recently published framework based on a humidity based SIRS model and a linear noise approximation (Zimmer et al., 2018a). (C) Linear regression uses linear models with different sets of past weeks as features. These features are selected on a validation data set (seasons 2010/11-2011/12). LinEns is a average ensemble over the three best linear models. (D) Sarima uses Seasonal auto regressive integrated moving average models as are also used in (Ray et al., 2017). Season is defined as one year and autogregressive and moving average terms are identified on the validation data set. (E) Epideep is a recently developed deep learning based influenza forecasting framework (Adhikari et al., 2019).

4.3. Implementation

We use Matlab 2016a and the GPML package (Rasmussen & Nickisch, 2019) for training Gaussian processes and predictions.

We use the seasons 2003/04 - 2007/08 as training data and the seasons 2010/11 and 2011/12 as validation data to determine appropriate feature sets I of previous weeks for the training data. We omit the 2008/09 - 2009/10 season as it was a pandemic season with unusual behavior. As we treat seasons independently by assumption, taking out two season does not change the procedure.

In order to determine the set of past weeks I used as input features of our framework, we use the 2010/11 - 2011/12 seasons as a validation data. We use various sets I (with up to five past weeks as features, $I \subset \{i^* - 4, i^* - 3, \ldots, i^*\}$) and choose the best three performing sets to become part of the ensemble as stated in Algorithm 1. This is done in the same way for the benchmarks linear regression to determine the input features as well as SARIMA to determine the autoregressive and moving average components.

The ensembles for the best set of features of past weeks for the GP based framework, and the best features for the Linear regression and the best autoregressive and moving average terms for Sarima are built by averaging the the densities of the predictive distributions of the individual methods' probabilistic forecasts. While this is a "naïve" ensemble and could be improved by e.g. training ensemble weights (ideally 4 target specific weights, and for each target 29 week specific weights) on additional validation seasons, we have not done this in the current work, as relatively few season are available so far. We anticipate further gain once this can be done in future research with additional data being available.

4.4. Performance of our GP based Framework on Seasonal Influenza forecasting

Our framework is able to do accurate point forecasting and to accurately quantify the uncertainties in projections. Figure 1 shows retrospective 1 week forecasts and their uncertainty estimation. We see for two seasons and two forecasting horizons that our framework performs well in point forecasts as well as a good estimation of the 95% prediction intervals. More seasons and targets can be found in the Appendix A.3. The appendix also contains a video visualizing influenza forecasting, see Figure A.10.

Next, we check whether our forecasts reach the desired coverage. We expect that on average 95% of the true values are within the 95% prediction interval. As this is a binomially distributed random variable, we can calculate its quantiles to see whether our framework is within this range. Figure 2 shows this range in green color and the actual values our framework achieves as a black line.

We note that the computational requirements for our framework allow for usage on personal computers. Retrospective forecasting for 7 seasons, 29 weeks per season and 4 targets is in the order of minutes with our framework.

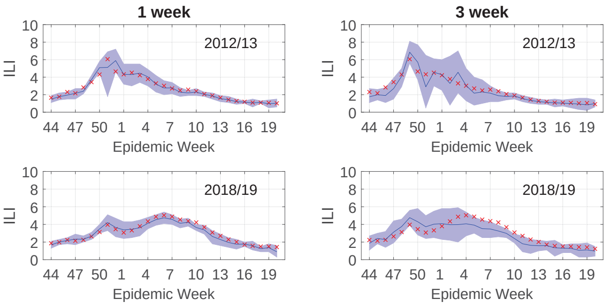

Figure 1. **Retrospective forecasts and their uncertainty:** One week retrospective influenza forecasting for two seasons and targets with our GP based framework for seasonal epidemics forecasting. Red x's are the actual observed values, and blue lines and shaded areas represent point forecasts and 95% prediction intervals.

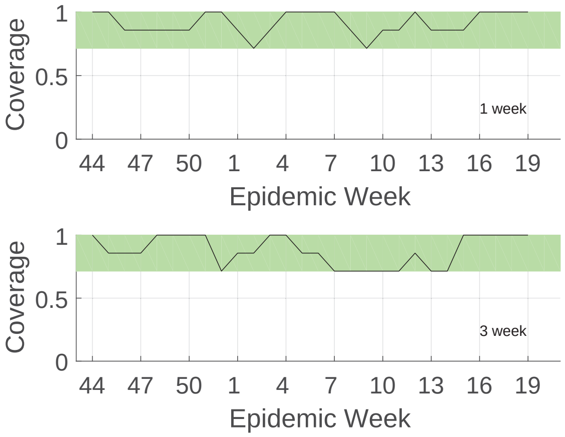

Figure 2. **Coverage of our framework:** Fraction of true values observed that are within the 95% prediction intervals (black line). As this is a binomially distributed random number, we can add its 95% confidence intervals (green shaded area) to check whether our framework yields reliable uncertainty estimation. Coverage for more forecasting targets in the Appendix figure A.7.

4.5. Results for Benchmarking with State of the Art

As described in the subsection benchmarking, we perform restrospective forecasts for the influenza season 2012/13 - 2018/19 with our GP based framework as well as several state of the art benchmarks. Each of the seasons starts at epidemic week (EW) 44 and ends 29 weeks later. This means 29 weeks of forecasting.

We perform 1-4 week forecasts and therefore, end up with a total of $7 \times 29 \times 4 = 812$ forecasts. For each of those, we can calculate the log-score. We note that the competitors show good performance as well as displayed in their retrospective forecast predicting interval figures A.4, A.5, A.6. However, we can see strong improvements by our new GP based framework: Figure 3 shows the reduction in log-score over those 203 forecasts as boxplots. We can see that our new GP based framework yields strong improvement in forecasting accuracy over all the other benchmarks for all forecasting targets. Wilcoxon signed rank test yields significance ($p < 0.001$) for all 1 week benchmarks (Hist, MSS, LinEns and SarimaEns). For 2-4 week targets, there is a growing correlation between the forecasting weeks and, therefore, Wilcoxon signed rank (assuming independence) should not be applied anymore on the set of forecasts from all season and weeks. However, Figure 3 shows that the improvement is similarly strong in terms of percentage of forecasts being improved by our framework.

Boxplots mainly focus on the median performance and it is possible that they do not detect rare outliers with very poor performance. Therefore, figure 4 shows the mean seasonal log-score gain for the four forecasting targets and four benchmarks and we do see an improvement of our framework compared to all benchmarks. Our advantage in terms of relative log-score gains is at least 20%-40% (LinEns) and up to 90% (Hist) and can be seen in figure A.1 in the appendix where we also show that our frameworks' predictions have smaller inter-quantile distances, figure A.2.

We also observe that the ensemble indeed performs better and more robust than the individuals as can be seen in the appendix figures A.8 and A.9.

The benchmark code for the Epideep (Adhikari et al., 2019) is currently only available for point estimates, so we include it in the Figure 5.

Even though less relevant in the context of seasonal epidemics forecasting, we also display squared error (SE) of point predictions and note that we still perform well. Figure 5 shows that we are better than MSS, Hist and Epideep for most forecasting targets. Sarima and LinEns perform similar or slightly better in terms of point forecasts which might be an effect of our framework being based on GPs which emphasize a probabilistic forecasts. Achieving this good probabilistic forecasts can come at the cost of slightly worse point predictions as GP build on normal distribution where an accurate uncertainty estimation might shift the mean towards a worse point prediction.

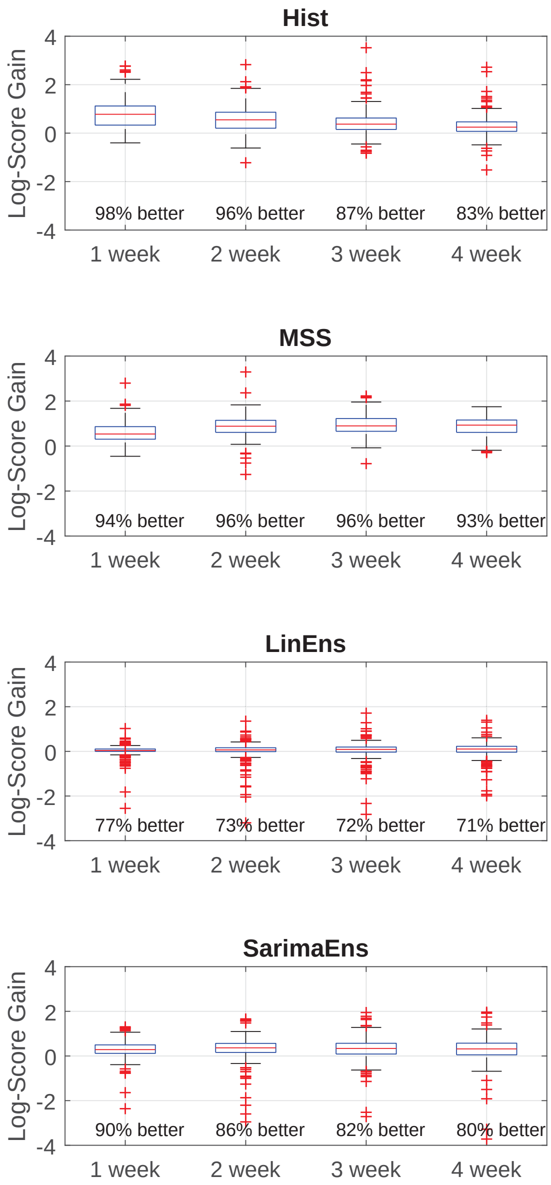

Figure 3. **Improvement in log-score by using our framework**: The four panels show four different state of the art benchmarks and each panel contains four forecasting targets. Each boxplot shows the improvement in log-score by using our novel framework compared to the respective state of the art method. Number below states percentage of log-scores improved by our new framework.

5. Discussion and Conclusion

Our novel GP based framework for influenza forecasting has shown strong performance in leading to accurate probabilistic forecasts for seasonal influenza Figure 1 and 2. In addition, it has shown significant improvements compared to state of the art competitors.

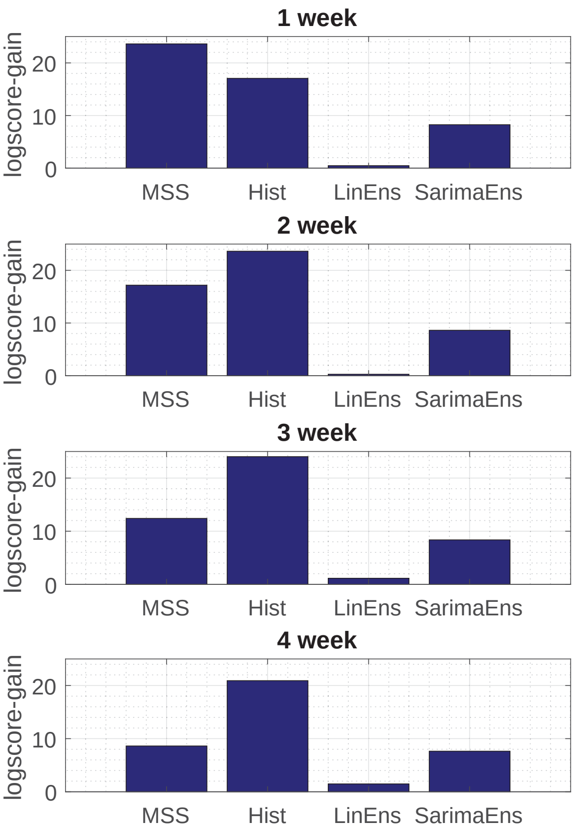

Figure 4. **Average log-score gain by using our framework:** The four panels are four different forecasting horizons and each panel contains results with respect to four different state of the art competitors. Each bar shows the mean seasonal logscore gain.

We note that this work is focused on probabilistic forecasts as they are more relevant in the context of seasonal epidemics forecasting than point predictions. While we do see a strong improvement in terms of probabilistic forecasts (Figure 3), we note that our GP based framework for seasonal epidemics forecasting is built for probabilistic forecasts and, therefore, still beats some state of the art competitors in point forecasting (Figure 5) but not all of them. This does not negatively impact its relevance and public health impact due to its strong performance on probabilistic forecast.

While our ensembles already show stronger performance than the individuals (Figure A.8 and A.9), we note that there are other approaches for building ensembles than taking the average, e.g. one could optimize the weights of each individual contributor. Furthermore, one could determine different weights for different targets and different epidemic weeks. While this should in general be superior, we note that it also requires a decently seized validation data set. If determine weights individually for each target and epidemic

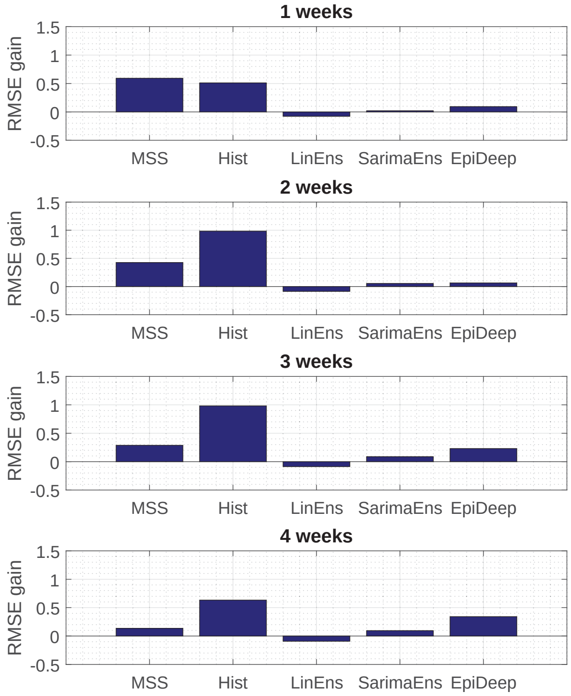

Figure 5. **Comparison of RMSE:** Even though less relevant in influenza forecasting, we also display RMSE. Each panel corresponds to one forecasting target and each panel contains 5 different state of the art benchmarks. Each bar represents the improvement (positive value) or decrease (negative value) in RMSE over all seasons and EWs compared to our framework.

week, each validation season gives us only one validation data point, so we would need several seasons to determine reliable weights. As influenza data has only been recorded year round since 2003, we postpone this idea to future research once more seasons have become available.

In summary, we developed a new Gaussian process based framework for seasonal epidemics forecasting. We demonstrate its capability by forecasting seasonal influenza and show that it is capable to produce precise point forecasts as well as accurate uncertainty quantification. We benchmark against several state of the art competitors and show significant improvement in results. Therefore, we believe that our Gaussian process based framework is a key contribution to improving influenza forecasting.

References

Adhikari, B., Xu, X., Ramakrishnan, N., and Prakash, B. A. Epideep: Exploiting embeddings for epidemic forecasting. *KDD*, 2019. doi: https://doi.org/10.1145/3292500.3330917.

Biggerstaff, M., Alper, D., Dredze, M., Fox, S., Fung, I. C.-H., Hickmann, K. S., Lewis, B., Rosenfeld, R., Shaman, J., Tsou, M.-H., Velardi, P., Vespignani, A., and Finelli, L. Results from the centers for disease control and prevention's predict the 2013 – 2014 influenza season challenge. *BMC Infectious Diseases*, 16, 2016. doi: 10.1186/s12879-016-1669-x.

Bracher, J. On the multibin logarithmic score used in the flusight competitions. *Proceedings of the National Academy of Sciences*, 116(42):20809–20810, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1912147116. URL https://www.pnas.org/content/116/42/20809.

Broniatowski, D. A., Paul, M. J., and Dredze, M. National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. *Plos ONE*, 8, 2013. doi: 10.1371/journal.pone.0083672.

Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J., and Rosenfeld, R. Flexible modeling of epidemics with an empirical bayes framework. *Plos Computational Biology*, 11, 2015. doi: doi:10.1371/journal.pcbi.1004382.

Centers for Disease Control and Prevention. Fluview interactive, a. URL https://www.cdc.gov/flu/weekly/fluviewinteractive.htm.

Centers for Disease Control and Prevention. Overview of influenza surveillance in the united states, b. URL https://www.cdc.gov/flu/weekly/overview.htm#Outpatient.

Centers for Disease Control and Prevention. Disease burden of influenza, c. URL https://www.cdc.gov/flu/about/disease/burden.htm.

Chretien, J. P., George, D., Shaman, J., Chitale, R. A., and McKenzie, F. E. Influenza forecasting in human populations: a scoping review. *PLoS ONE*, 9(4):e94130, 2014.

Farrow, D. *Modeling the past, present, and future of influenza*. Doctoral dissertation, Carnegie Mellon University, 2016.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature Letters*, 457, 2009. doi: 10.1038/nature07634.

McIver, D. J. and Brownstein, J. S. Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *Plos Computational Biology*, 10, 2014. doi: 10.1371/journal.pcbi.1003581.

Osthus, D., Hickmann, K., Caragea, P., Higdon, D., and Valle, S. D. Forecasting seasonal influenza with a state-space SIR model. *Annals of Applied Statistics*, 11, 2017.

Paul, M. J., Dredze, M., and Broniatowski, D. Twitter improves influenza forecasting. *Plos Current Outbreaks*, 1, 2014. doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.

Rasmussen, C. E. and Nickisch, H. Gaussian processes for machine learning (gpml) toolbox. *Journal of Machine Learning Research*, 11, 2019.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Ray, E., Sakrejda, K., Lauer, S., Johansson, M., and Reich, N. Infectious disease prediction with kernel conditional density estimation. *Stat. Med.*, 36, 2017.

Reich, N. G., Brooks, L. C., Fox, S. J., Kandula, S., McGowan, C. J., Moore, E., Osthus, D., Ray, E. L., Tushar, A., Yamana, T. K., Biggerstaff, M., Johansson, M. A., Rosenfeld, R., and Shaman, J. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154, 2019a. ISSN 0027-8424. doi: 10.1073/pnas.1812594116. URL https://www.pnas.org/content/116/8/3146.

Reich, N. G., Osthus, D., Ray, E. L., Yamana, T. K., Biggerstaff, M., Johansson, M. A., Rosenfeld, R., and Shaman, J. Reply to bracher: Scoring probabilistic forecasts to maximize public health interpretability. *Proceedings of the National Academy of Sciences*, 116(42): 20811–20812, 2019b. ISSN 0027-8424. doi: 10.1073/pnas.1912694116. URL https://www.pnas.org/content/116/42/20811.

Senanayake, R., O'Callaghan, S., and Ramos, F. Predicting spatio–temporal propagation of seasonal influenza using variational gaussian process regression. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016.

Shaman, J., Karspeck, A., Yang, W., Tamerius, J., and Lipsitch, M. Real-time influenza forecasts during the 2012-2013 season. *Nat Commun*, 4:2837, 2013.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. *Proceedings of the 27 th International Conference on Machine Learning, Haifa, Israel, 2010*, 2010.

Yang, W., Karspeck, A., and Shaman, J. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLOS Computational Biology*, 10:e1003583, 2014.

Zimmer, C., Leuba, S. I., Cohen, T., and Yaesoubi, R. Accurate quantification of uncertainty in epidemic parameter estimates and predictions using stochastic compartmental models. *Statistical Methods in Medical Research*, 28, 2018a.

Zimmer, C., Leuba, S. I., Yaesoubi, R., and Cohen, T. Use of daily Internet search query data improves real-time projections of influenza epidemics. *Journal of the Royal Society Interface*, 15, 2018b. doi: https://doi.org/10.1098/rsif.2018.0220.

Zimmer, C., Meister, M., and Nguyen-Tuong, D. Safe active learning for time-series modeling with gaussian processes. In *32nd Conference on Neural Information Processing Systems*, 2018c.