SURROGATE ENSEMBLE FORECASTING FOR DYNAMIC CLIMATE IMPACT MODELS

Julian Kuehnert, Deborah McGlynn,[*] Sekou L. Remy & Aisha Walcott-Bryant
IBM Research Africa
Nairobi, Kenya
julian.kuehnert@ibm.com, {sekou,awalcott}@ke.ibm.com

Anne Jones
IBM Research Europe
Daresbury, United Kingdom
anne.jones@ibm.com

ABSTRACT

As acute climate change impacts weather and climate variability, there is increased demand for robust climate impact model predictions from which forecasts of the impacts can be derived. The quality of those predictions are limited by the climate drivers for the impact models which are nonlinear and highly variable in nature. One way to estimate the uncertainty of the model drivers is to assess the distribution of ensembles of climate forecasts. To capture the uncertainty in the impact model outputs associated with the distribution of the input climate forecasts, each individual forecast ensemble member has to be propagated through the physical model which can imply high computational costs. It is therefore desirable to train a surrogate model which allows predictions of the uncertainties of the output distribution in ensembles of climate drivers, thus reducing resource demands. This study considers a climate driven disease model, the Liverpool Malaria Model (LMM), which predicts the malaria transmission coefficient R0. Seasonal ensembles forecasts of temperature and precipitation with a 6-month horizon are propagated through the model, predicting the distribution of transmission time series. The input and output data is used to train surrogate models in the form of a Random Forest Quantile Regression (RFQR) model and a Bayesian Long Short-Term Memory (BLSTM) neural network. Comparing the predictive performance, the RFQR better predicts the time series of the individual ensemble member, while the BLSTM offers a direct way to construct a combined distribution for all ensemble members. An important element of the proposed methodology is that accounting for non-normal distributions of climate forecast ensembles can be captured naturally by a Bayesian formulation.

1 INTRODUCTION

The United Nations' Intergovernmental Panel on Climate Change (IPCC) sixth assessment report indicates that average global temperatures will rise by 1.5°C above preindustrial levels by 2040 (IPCC, 2021). This change brings increases in global sea levels, melting of glacial ice, and more extreme precipitation events (IPCC, 2021). This in turn is expected to bring increases in drought, wild fires, flooding, and variability in the regionality of diseases, and all with increased forecasting uncertainty.

Accurate forecasting of variables dependent on climate drivers remains a challenge. The underlying physical processes driving their variability are dynamic and nonlinear, making predictions through mathematical models difficult and resource expensive. Machine learning algorithms can help to gen-

[*] Currently at Virginia Tech, Department of Civil and Environmental Engineering, Blacksburg, VA 24061, USA, mcglyndf@vt.edu

erate predictions and associated uncertainties without explicitly knowing the underlying processes, thereby reducing computational costs.

In this work, we calculate the malaria parasite dynamics from daily temperature and precipitation data based on the Liverpool Malaria Model (LMM, Hoshen & Morse, 2004). The model has been used to understand malaria transmission dynamics, and predict malaria transmission over seasonal and climate change timescales (Jones & Morse, 2010; Caminade et al., 2014). Given its non-linear dependence on two climate variables (temperature and precipitation), this model serves as a useful example with which to explore the process of quantifying uncertainty in an endogenous variable.

Surrogate machine learning models are trained to predict the seasonal variability of the climate driven variables, using the ensembles of temperature and precipitation as inputs or feature variables, and the ensembles of transmission coefficients R_0 as output or the target variable. We implement two types of machine learning algorithms. For the first type, a Random Forest Quantile Regression (RFQR) model is used. This ensemble learning model was originally developed by Meinshausen & Ridgeway (2006) and offers a robust, non-linear, and non-parametric way to predict quantile ranges based on training observations. It has been shown to successully calibrate the dynamics of ensemble weather forecasts (Taillardat et al., 2016). For the second type, a Bayesian Long Short-Term Memory (BLSTM) model is used. Its nature of a recurrent Neural Network (RNN) is well suited for sequential data such as natural language processing and temporal data (Wang et al., 2019; Han et al., 2020). The LSTM model is a type of RNN that incorporates gated back propagation of data (Wang et al., 2020; Han et al., 2020) to mitigate the vanishing gradient problem (Khan, 2018). In our work, the LSTM is implemented as an approximation of a Bayesian NN to capture the uncertainty in the climate forecast ensembles.

2 Methodology

2.1 Climate and malaria transmission data

While the presented method can be applied to any variable that is highly correlated with climate and weather variability, we apply it to malaria forecasting as an example. Climate data was obtained from IBM®[1] PAIRS ECMWF seasonal forecasts (Lu et al., 2016; Klein et al., 2015; Johnson et al., 2019; Crawford et al., 2019) for the center of Nairobi, Kenya with coordinates (-1.286389, 36.817222). For the 5 year period from 2017 through 2021, 50 ensemble members of daily temperature and precipitation data were obtained in 6 month segments, reflecting the variability of seasonal forecasts. The starting dates of the forecasts are fixed on January 1 and July 1. This ensures that the seasonal peaks of malaria in Nairobi, which occur in April and November, are situated in the middle of the forecast period. The climate data was then propagated through the Liverpool Malaria Model (LMM) to calculate malaria transmission coefficient R0, using the simplified *steady state* model created by Jones (2007).

2.2 Surrogate model description

The RFQR was introduced by Meinshausen & Ridgeway (2006). It extends classical Random Forests, by allowing quantile ranges to be predicted and hence its own prediction uncertainty. This is realized by not only keeping the means of the target values in the leaves of the decision trees but storing all the samples of the target values. Quantiles can then be computed based on these distributions.

The BLSTM was implemented based on the work of Zhu & Laptev (2017). First, it combines a classical LSTM with an encoder-decoder in order to extract only representative features in the data, which improves predictive performance on unknown sample patterns. Then, following the approach by Gal & Ghahramani (2016), Monte Carlo dropout is introduced to randomly remove samples at each layer of the Neural Network. This is the process through which the Bayesian NN is approximated, and a posterior distribution of the prediction can be derived by repeating the experiments.

[1] IBM and the IBM logo are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on ibm.com/trademark.

2.3 Data preparation

Based on the above described data sources and Malaria Model, the datasets are constructed with following variables: {Date, Ensemble Member, Temperature [daily mean in degree Celsius], Precipitation (daily mean in mm), Malaria transmission R0}. Then, the following two training-test datasets are prepared:

- **Dataset 1** comprises 50 ensemble members for a single forecast period of a 6 month horizon from January 1 to June 30, 2021. The training dataset is built from 70 % (35) of the ensembles members, while the other 30 % (15) of the ensembles members are used for testing.
- **Dataset 2** comprises 50 ensemble members for eight forecast periods, each of a 6 month horizon starting at either January 1 or July 1 of 2017 through 2021. The dataset is then split into a training part comprising the first 4 years from 2017 through 2020, and a test part comprising the year 2021.

2.4 Model training and testing

The RFQR is set up with 1000 tree estimators, in which a minimum of 10 samples is required to split an internal node and the minimum number of samples per leaf is allowed to be 1. Loss is measured by the Mean Squared Error (MSE). The feature vector is constructed by {Day of Month, Month, Precipitation, Temperature} at time step t_n, while the target vector contains {R0} at time step t_n.

The BLSTM is set up as in Zhu & Laptev (2017), using the Mean Squared Error (MSE) as the loss function and learning with a rate of 0.01 during 200 epochs. For Dataset 1, a batch size of 1024 is used, while for the larger Dataset 2, the batch size is set to 4096. Each time series is split into sequences of 50 days while each 50^{th} day is predicted based on the previous 49, including the values of previous transmission coefficients R0. The feature vector is constructed by {Day of Month, Month, Precipitation, Temperature, R0} for time steps t_i to t_{i+49}, and the target is {R0} at time step t_{i+50}, where it is stepping through the length of the whole forecast period. 200 experiments with Monte Carlo dropout probability of 0.5 are performed to sample the posterior distribution.

2.5 Uncertainty estimation

The uncertainty of the model predictions is quantified by defining a quantile range from the lower $q_l = 15.87\,\%$ to the upper $q_u = 84.13\,\%$. We have chosen these values because they correspond to minus or plus the standard deviation in the case of a normal distribution.

The quantiles for predicted distribution of each test sample (i.e. for a given date t_i and a given ensemble member j) are inferred as follows:

1. RFQR: $q_l^j(t_i)$ and $q_u^j(t_i)$ are directly predicted.
2. BLSTM: compute $q_l^j(t_i)$ and $q_u^j(t_i)$ from the samples $s_k^j(t_i)$ of the M experiments with random Monte Carlo dropout.

When considering all test ensemble members (i.e. for a given date across all N ensemble members), the combined ensembles are inferred as follows:

1. RFQR: compute the combined lower and upper quantile Q_l and Q_u by averaging individual quantiles $Q_l(t_i) = 1/N \sum_j^N q_l^j(t_i)$ and $Q_u(t_i) = 1/N \sum_j^N q_u^j(t_i)$.
2. BLSTM: derive Q_l and Q_u as for the RFQR. Additionally, we derive a lower and upper quantile range Qd_l and Qd_u directly from the samples $s_k^j(t_i)$ across all experiments $k = 1, ..., M$ and all ensembles $j = 1, ..., N$.

3 Results and discussion

3.1 Predictions for individual ensemble members

For an initial evaluation of the dynamic behavior of the predictions, we randomly pick three individual ensemble members and compare the dynamic quantile range with the true R0 timeseries.

The comparison is shown in Figure 1 for ensemble members $\{40, 45, 50\}$ and for RFQR (left) and BLSTM (right). Qualitatively evaluated, the dynamic behavior of the quantile range predicted by the RFQR captures well the time series of the target R0 values. Even if the predictions from the BLSTM reflect some of the dynamic characteristics, the predicted quantile range does not succeed in capturing all data points.

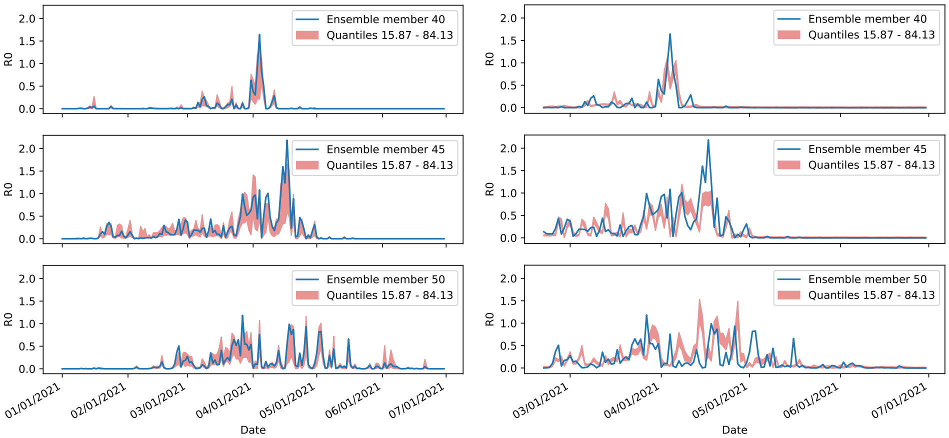

Figure 1: Target time series (blue) and predicted quantile range (red) from the RFQR (left) and from the BLSTM (right) for three individual ensemble members.

3.2 Predictive performance comparison

The predictive performance of the two surrogate models is now compared quantitatively for all test ensemble members. Predictions of the defined quantile range for the 15 test ensemble members in Dataset 1 are shown in Figure 2, using the RFQR (left) and the BLSTM (right). The samples are sorted in ascending order of the target values (*True R0*) which are plotted as blue dots. In both models the prediction uncertainty increases with increasing R0. This is possibly based on two factors. One, there is a class imbalance in that within the dataset there are more days in the year where the values of R0 are small (smaller than the median value). Two, the R0 time series contain larger fluctuations during periods of peak transmissions (i.e. when R0 is high). This means that when R0 should be high, there is more variability in the underlying data which possibly increases the prediction uncertainty.

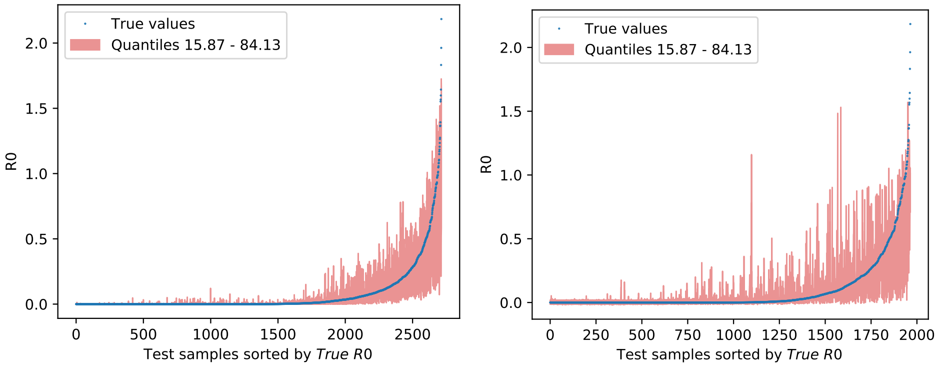

Figure 2: Target values and predicted confidence intervals from RFQR (left) and from BLSTM (right) for all daily samples of the 15 test ensemble members, sorted by the target values.

Comparing the two models RFQR and BLSTM, more uncertainty is observed in the BLSTM predictions. However, not only is the RFQR quantile range narrower, it also outperforms the BLSTM in capturing the target values in the quantile range. This is reflected in Table 1, which gives the true data points laying within the quantile range from 15.87 % to 84.13 % as a percentage of the total number of samples.

Table 1: Rates of target values laying within the confidence intervals predicted from RFQR and BLSTM, respectively, for different datasets and differet quantile ranges.

	RFQR	**BLSTM**
Dataset 1, Range $[q_l, q_u]$	82.9 %	61.8 %
Dataset 2, Range $[q_l, q_u]$	80.6 %	61.1 %
Dataset 1, Range $[Q_l, Q_u]$	15.5 %	38.9 %
Dataset 2, Range $[Q_l, Q_u]$	11.8 %	34.5 %
Dataset 1, Range $[Qd_l, Qd_u]$	na	69.1 %
Dataset 2, Range $[Qd_l, Qd_u]$	na	73.9 %

Similar results can be seen from the model predictions with Dataset 2. Figure 3 shows the predicted quantile range for the test ensemble in year 2021 using the RFQR (left) and BLSTM (right), while the target R0 values are plotted as blue dots. Again, the uncertainty of the RFQR is generally lower compared to the BLSTM, while a high percentage of the target values are in the quantile range of the RFQR, outperforming the BLSTM (see Table 1).

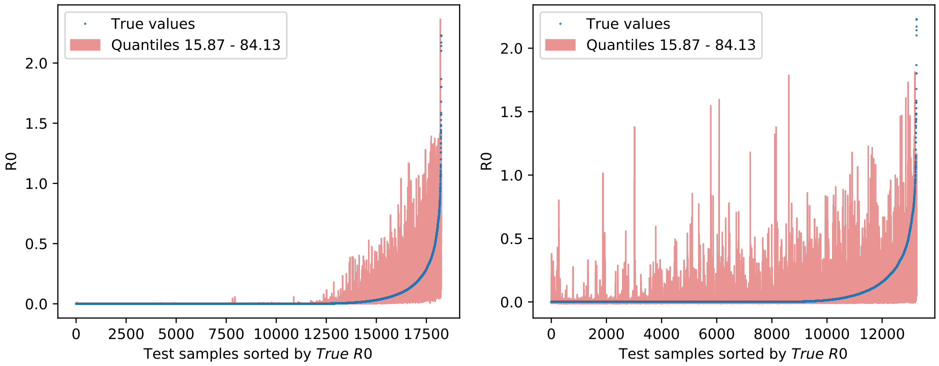

Figure 3: Target values and predicted confidence intervals from RFQR (left) and from BLSTM (right) for all daily samples and all 50 ensemble members of the two test forecast periods in 2021, sorted by the target values.

3.3 Ensemble prediction uncertainty

We now assemble the individual model predictions across all ensemble members on a daily basis. First we will calculate $[Q_l, Q_u]$ based on the individual member quantile ranges $[q_l, q_u]$. To assess how well the combined quantile range $[Q_l, Q_u]$ captures the variation of the target values R0, the percentage of samples within the range is calculated. The corresponding values are listed in Table 1 for both Dataset 1 and Dataset 2. It can be observed that upon averaging the lower and upper quantiles, the combined range is only able to capture a subset of the previously captured samples. The process of averaging hence works like a low-pass filter, removing the fluctuations which are necessary to keep the target values within the range. Of note, the BLSTM is now able to capture more data points. This is because the underlying uncertainty had more variations.

To capture more of the target values, we now calculate the ensemble quantile range $[Qd_l, Qd_u]$ directly from the samples of the model instead of from the previously inferred quantiles. In the case of the BLSTM, these samples are generated during repeated predicting with Monte Carlo dropout. Analysing the hit rates, see Table 1, the percentages of captured data points are again now comparable with the percentages from analysing the individual member samples. For visual inspection, figures of the resulting dynamic range are included in Appendix A.

4 Conclusion and future work

In this work, we use a Random Forest Quantile Regression (RFQR) model and a Bayesian Long Short-Term Memory (BLSTM) neural network as surrogates for a dynamic climate impact model to predict dynamic quantile ranges from climate ensemble forecasts. In the application with the Liverpool Malaria Model (LMM) for the prediction of transmission coefficient R0, higher uncertainty was predicted on high R0 values. This is attributed to both a class imbalance in R0 in the dataset and increased fluctuation in larger values of R0. Even though the BLSTM is designed to predict time series data, the performance of RFQR is better, likely because it can model the dynamics of the underlying physical processes. In order to derive a combined ensemble quantile range, it is observed that averaging the predicted quantile ranges of the individual ensemble members acts as a strong low-pass filter because it only captures a small portion of the target values. Computing the ensemble quantile range directly from the underlying predictions samples, instead of the predicted quantiles, allows a high percentage of the target value variations to be captured. Nonetheless, other methods may need to be considered to capture the extremes of the distribution, such as for example Extreme Value Theory or Autoregressive Conditional Heteroscedasticity.

References

Cyril Caminade, Sari Kovats, Joacim Rocklov, Adrian M. Tompkins, Andrew P. Morse, Felipe J. Colón-González, Hans Stenlund, Pim Martens, and Simon J. Lloyd. Impact of climate change on global malaria distribution. *Proceedings of the National Academy of Sciences*, 111(9):3286–3291, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1302089111. URL https://www.pnas.org/content/111/9/3286.

Todd Crawford, James Belanger, Michael Ventrice, and John Williams. Calibrated Probabilistic Seasonal Forecasts at IBM / The Weather Company : Business Applications. (October):22–24, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yang Han, Jacqueline C.K. Lam, Victor OK Li, and Qi Zhang. A Domain-Specific Bayesian Deep-learning Approach for Air Pollution Forecast. *IEEE Transactions on Big Data*, pp. 1–1, 2020. ISSN 2332-7790. doi: 10.1109/TBDATA.2020.3005368. URL https://ieeexplore.ieee.org/document/9127775/.

Moshe B. Hoshen and Andrew P. Morse. A weather-driven model of malaria transmission. *Malaria Journal*, 3(1):32, 2004. doi: 10.1186/1475-2875-3-32. URL https://doi.org/10.1186/1475-2875-3-32.

IPCC. Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. 2021. ISSN 0373-6245.

S. J. Johnson, T. N. Stockdale, L. Ferranti, M. A. Balmaseda, F. Molteni, L. Magnusson, S. Tietsche, D. Decremer, A. Weisheimer, G. Balsamo, S. P. E. Keeley, K. Mogensen, H. Zuo, and B. M. Monge-Sanz. Seas5: the new ecmwf seasonal forecast system. *Geoscientific Model Development*, 12(3):1087–1117, 2019. doi: 10.5194/gmd-12-1087-2019. URL https://gmd.copernicus.org/articles/12/1087/2019/.

Anne E Jones and Andrew P Morse. Application and validation of a seasonal ensemble prediction system using a dynamic malaria model. *Journal of Climate*, 23(15):4202–4215, 2010.

Anne Elizabeth Jones. *Seasonal ensemble prediction of malaria in Africa*. PhD thesis, University of Liverpool, 2007.

G. M. Khan. Artificial neural network (ANNs). *Studies in Computational Intelligence*, 725:39–55, 2018. ISSN 1860949X. doi: 10.1007/978-3-319-67466-7_4.

Levente J. Klein, Fernando J. Marianno, Conrad M. Albrecht, Marcus Freitag, Siyuan Lu, Nigel Hinds, Xiaoyan Shao, Sergio Bermudez Rodriguez, and Hendrik F. Hamann. PAIRS: A scalable geo-spatial data analytics platform. *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, pp. 1290–1298, 2015. doi: 10.1109/BigData.2015.7363884.

Siyuan Lu, Xiaoyan Shao, Marcus Freitag, Levente J. Klein, Jason Renwick, Fernando J. Marianno, Conrad Albrecht, and Hendrik F. Hamann. IBM PAIRS curated big data service for accelerated geospatial data analytics and discovery. *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, pp. 2672–2675, 2016. doi: 10.1109/BigData.2016.7840910.

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of Machine Learning Research*, 7(6), 2006.

Maxime Taillardat, Olivier Mestre, Michaël Zamo, and Philippe Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6):2375–2393, 2016.

Bin Wang, Tianrui Li, Zheng Yan, Guangquan Zhang, and Jie Lu. DeepPIPE: A distribution-free uncertainty quantification approach for time series forecasting. *Neurocomputing*, 397:11–19, 2020. ISSN 18728286. doi: 10.1016/j.neucom.2020.01.111.

Mengyang Wang, Hui Wang, Jiao Wang, Hongwei Liu, Rui Lu, Tongqing Duan, Xiaowen Gong, Siyuan Feng, Yuanyuan Liu, Zhuang Cui, Changping Li, and Jun Ma. A novel model for malaria prediction based on ensemble algorithms. *PLOS ONE*, 14(12):e0226910, dec 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0226910. URL `https://dx.plos.org/10.1371/journal.pone.0226910`.

Lingxue Zhu and Nikolay Laptev. Deep and Confident Prediction for Time Series at Uber. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2017-November:103–110, 2017. ISSN 23759259. doi: 10.1109/ICDMW.2017.19.

A Appendix: Ensemble prediction uncertainty

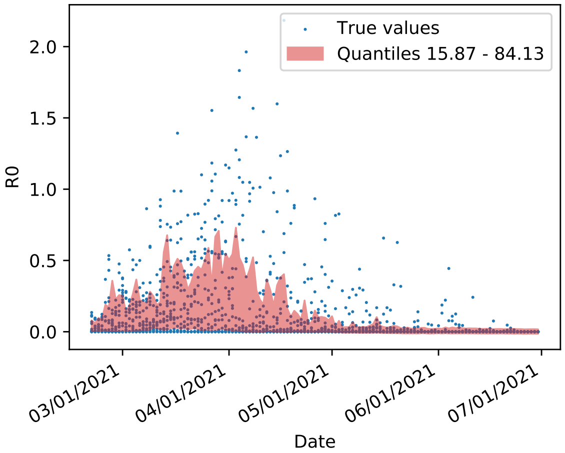

Figure 4: Uncertainty prediction using the BLSTM for the 15 test ensembles members for the first half of 2021 in Dataset 1. The model was trained on the other 35 ensemble members for the same forecast period from January 1 to July 1.

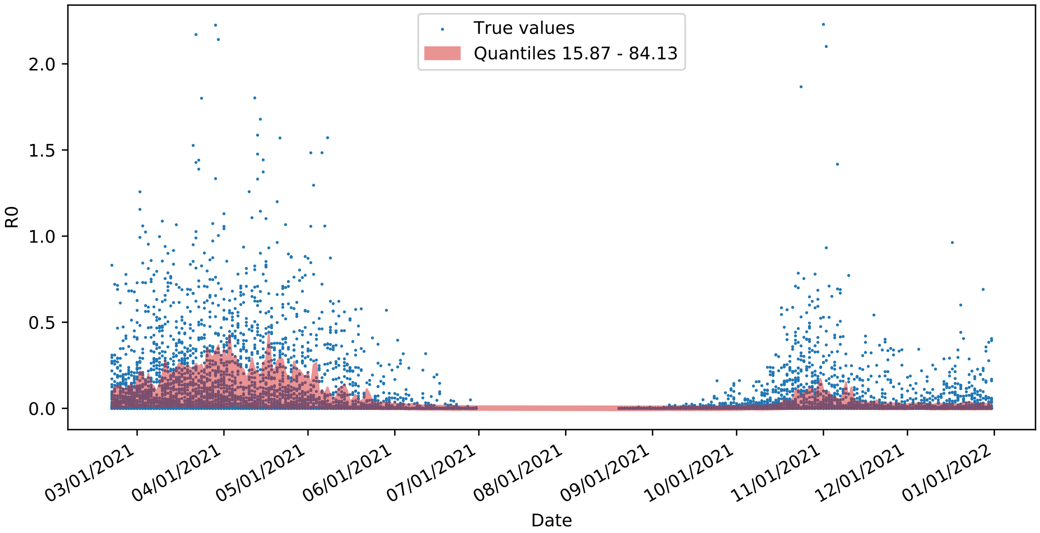

Figure 5: Uncertainty prediction using the BLSTM for all 50 test ensemble members in both forecast periods of 2021 in Dataset 2. The model was trained on all 50 ensemble members for 8 six-month forecast periods from 2017 to 2020.