Dengue fever: from extreme climates to outbreak prediction

Mai, S. T., Phi, H. T., Abubakar, A., Kilpatrick, P., Nguyen, H. Q. V., & Vandierendonck, H. (2023). Dengue fever: from extreme climates to outbreak prediction. In X. Zhu, S. Ranka, M. T. Thai, T. Washio, & X. Wu (Eds.), *2022 IEEE International Conference on Data Mining (ICDM): proceedings* (pp. 1083-1088). (IEEE International Conference on Data Mining (ICDM)). Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/ICDM54844.2022.00135

Published in:
2022 IEEE International Conference on Data Mining (ICDM): proceedings

Document Version:
Peer reviewed version

Dengue Fever: From Extreme Climates to Outbreak Prediction

Son T. Mai*, Ha T. Phi*, Abdullahi Abubakar*, Peter Kilpatrick*, Hung Q. V. Nguyen† and Hans Vandierendonck*
*Queen's University Belfast, UK
Email: {thaison.mai, p.kilpatrick, h.vandierendonck}@qub.ac.uk
†Griffith University, Australia
Email: quocviethung.nguyen@griffith.edu.au

***Abstract*—Dengue Fever (DF) is an emerging mosquito-borne infectious disease that affect hundred millions of people each year with considerable morbidity and mortality rates, especial on children. Together with global climate changes, it is continuously increasing in terms of number of cases and new locations. Thus, having effective early warning systems become an urgent need to improve disease controls and prevention. In this paper, we introduce a novel framework, called Proximity Time Ensemble, to predict DF outbreaks for multiple areas (provinces) and multiple time step ahead, and to study the effects of climate data on DF outbreaks. PT-Ensem consists of 6 key components: (1) an event-to-event probabilistic framework to study links among extreme climate events and DF outbreaks; (2) a proximity graph that connects similar provinces; (3) an ensemble prediction technique that combines many different advanced machine learning (ML) methods to predict outbreaks within t time steps in the future using extreme climate events as model inputs; (4) a data aggregate scheme to enrich training data for each provinces via its neighbors in the proximity graph; (5) a proximity propagation step that propagates predicted results among similar provinces via the proximity graph until maximal agreements are reached among provinces; and (6) a time propagation step to propagate results via different predicted time steps in each province. We use PT-Ensem to predict DF outbreaks for all provinces in Vietnam using data collected from 1997-2016. Experiments show that PT-Ensem acquires significant performance boost compared to many highly-rated ML models like XGBoost, LightGBM and Catboost in the outbreak prediction task. Compared to most recent deep learning approaches like LSTM-ATT, LSTM, CNN and Transformer for predicting DF incidence, PT-Ensem also dominates in both prediction accuracy and computation times.**

Keywords*—Dengue Fever prediction, Epidemic forecast, Outbreak prediction, Ensemble learning

I. Introduction

Dengue Fever (DF) is the most prevalent of arboviral and climate sensitive diseases transmitted by *Aedes* mosquitoes [1]. Patients with DF typically suffer from a wide range of symptoms, e.g., headache, vomiting, nausea, skin rash, and muscle and joint pain [2]. Severe DF can lead to more life-threatening problems, e.g., internal bleeding, respiratory distress, or organ failure [2]. Currently, there are no specific effective drugs for treating DF and no licensed vaccine available for it [1], [2]. Vector controls, which aim at using chemical, biological and environmental methods to target mosquitoes and their breeding sites, remain the primary means to prevent DF infections. Nevertheless, DF is continuously increasing rapidly compared to other communicable diseases. According to the World Health Organization (WHO), the number of reported DF infection cases has increased approximately 10 times during the last two decades (from 0.5 million in 2000 to 5.2 million reported cases in 2019). However, the number of real DF cases is actually much higher. E.g., Bhatt et al. [1] estimate around 390 million cases a year worldwide. The number of countries that have experienced DF epidemics has also increased from 8 in 1970 to more than 120 in 2020. Recent global climate changes have led to warmer and rainier weather conditions which support better the growth of mosquitoes and thus worsen future situations [3], [4]. Therefore, building early warning systems for DF has become an emerging need for more effective disease prevention and control and is continuously attracting many research efforts, e.g., [2], [5]–[15].

Identifying potential factors that affect DF incidence in specific areas is the foremost step towards effective early warning systems. Hence, previous research has extensively studied relationships between DF incidence and diverse ranges of potential factors e.g., human behaviors [16], [17], geographic information [18]–[20], socio-economic factors such as income [18], and especially meteorological factors [3], [21], [22]. Reported results can be local-specific. E.g., [23] found a positive correlation between rainfall and DF incidence with a lag time from 0-3 months in Hanoi, Vietnam, while [3] found no significant correlation. Based on these factors, a wide range of Machine Learning (ML) models have been employed to predict DF incidence rates/cases or outbreaks for many different areas e.g., Queensland in Australia [11], Guangzhou in China [8], Singapore [6], Honduras [6], Brazil [24], Bangkok in Thailand [18], Selangor in Malaysia [25], and Vietnam [13]. These models range from traditional to recent deep learning methods, e.g., Seasonal Autoregressive Integrated Moving Averaged (SARIMA) [11], Poisson regression [26], Support Vector Regression (SVR) [8], Gradient Boosting Machine (GBM) [7], [8], Generalized Additive Models (GAMs) [8], Generalized Linear Mixed Models (GLMMs) [13], Artificial Neural Networks (ANNs) [9], Back-propagation neural network (BPNNs) [7], Long-short term memory (LSTM) [7], Convolution Neural Networks (CNNs) [10], and Transfomer [10]. Inputs for these models also vary but climate data (e.g., rainfall and temperature) are frequently studied subjects [2].

Despite many research efforts, some problems remain. Most of the above mentioned studies focus on predicting DF incidence rates/cases weekly or monthly [7], [10]. Fewer works aim at DF outbreak prediction [10], [13], [24], [25]. Moreover, many of them actually do not predict outbreaks but infer them from forecast DF incidence [10], [13]. Most of these

techniques are heavily tailored to specific data (which are not publicly available) and specific areas e.g., [13], [18], which thus limits their applicability in wider contexts. Most techniques only predict DF for a single area (city or province) [6], [18]. Others study multiple areas, e.g., [7], [10], [13]. However, each area is typically treated independently. Their relationships are not properly exploited to improve performance.

Our contributions. In this paper, we propose a novel generic framework called Proximity Time Ensemble (PT-Ensem) for predicting DF outbreaks based on climate data for multiple areas that aims to address all the above problems. PT-Ensem fundamentally differs from all previous research as follows:

First, all existing works aim at predicting if an outbreak will occur at the time t exactly (month or week) in the future [10], [13], [24], [25]. However, we target a slightly different problem: predicting if an outbreak will happen *within* time t in the future. In this way, the predicted results at different time points can be used to independently verify each other due to their cumulative links, thus helping healthcare experts to make better decisions. It worth noting that our algorithm can be straightforwardly adapted to predict the outbreak at the exact time t while still retaining all of its other advantages.

Second, we propose a probabilistic model, called event-to-event relationships, to capture the links between DF outbreaks and extreme climate events for multiple areas. This model is used to study the relationship between different climate factors and DF incidence (c.f. Sections II and IV). Compared to traditional correlation analysis in previous works [3], [23], [27], it focuses only on major data points and thus is more robust to noise and small changes in the data. It is also used to model the similarity among different areas w.r.t. climate factors or DF incidence. The results are then exploited to adjust the prediction outcomes among all areas in a special scheme called the proximity propagation (c.f. Section III Phase D).

Third, while all existing works use raw climate data as inputs for their prediction models [8]–[11], [13], we propose to encode input data into sets of extreme events before feeding them to prediction models. By this way, years of climate data is compressed into a single point. Hence, the overall input size is reduced while retaining rich information. This helps to reduce model overfitting, thus enhancing both the accuracy and computation time. The range value differences of climate data in different areas are also flattened, thus creating an unified view for all areas based on extreme events. We then can aggregate data of similar provinces to train prediction models via a special constructed province proximity graph (c.f. Section III Phases B and C). It helps to increase the size and diversity of training data, thus reducing both over- and -underfitting problems and improving the final accuracy.

Fourth, we extensively study 20 different prediction models from traditional ones like Random Forest [28] and Extra Tree [29] to the most recent techniques like LightGBM [30] and Catboost [31] on our DF outbreak prediction problem. We choose the top 8 methods to combine them in an ensemble framework to enhance the performance. However, unlike all existing ensemble approaches [13], [24], [32], we aggregate outputs of all models into a probability of outbreak for each province, and feed all of them to a post-processing step, called the proximity propagation, to adjust the prediction result among provinces as described below.

Fifth, we propose two *unique* post-processing schema called proximity and time propagation (c.f. Section III Phases D and E) where prediction results are propagated among similar provinces (directed by a proximity graph) and across different prediction time periods to maximize prediction agreements among all studied areas. The intuition behind these two schemata is based on the well-known space and time transmission nature of DF which is also exploited by other research works [5], [9], [33], [34]. E.g., [9] incorporates human movement patterns into the input of prediction models to reflect the DF transmission. All of these works differ fundamentally to our approaches which are post-processing steps based on outputs of prediction models. We demonstrate in Section IV how they dramatically boost the prediction accuracy of PT-Ensem for all studied areas.

To the best of our knowledge, there is no existing work that has all of the above mentioned features.

A case study in Vietnam. We employ our algorithm PT-Ensem for predicting outbreaks for all provinces in Vietnam using a wide range of climate data collected for 20 years from 1997 to 2016. Extensive results are shown in Section IV. PT-Ensem dramatically improves the prediction accuracy compared to other ML models like Random Forest [28], Extra Tree [29], LightGBM [30], and Catboost [31]. Compared to adapted recent deep learning approaches for DF incidence prediction in Vietnam like LSTM-ATT, CNN, and Transfomer [10], PT-Ensem clearly dominates them with 6.17 to 9.72 percentage points (per 100) improvement in the prediction accuracy and 24x runtime speed up. We additionally demonstrate how our event-to-event probabilistic framework can be used to study relationships between extreme climate events and DF outbreaks.

II. Background and Problem Formulation

Study area. Vietnam is a tropical country in Southeast Asia and has 3 main regions: Northern, Central, and Southern. Each region is further separated into subregions with different cultures, geography, and climate. The Northern provinces have the full 4 seasons with a humid tropical climate. Central provinces are hot and dry during summer but cool and rainy in winter. Southern provinces have constant warm temperatures and have only 2 main seasons: dry and rainy. According to The Global Climate Risk Index 2020, Vietnam is among the top countries that are vulnerable to climate change. The country has been suffering from increasing temperatures as well as frequent extreme weather events such as storms, droughts, or floods. By the end of this century, temperatures are predicted to rise by 1.8°C-2.5°C and rainfall to increase by 5-15mm per year [3]. These conditions are more suited for the virus transmission and thus will significantly worsen the DF incidence in the whole country [3], [23], [27]. Hence, effective early-warning systems for DF outbreaks in the whole country are urgently needed.

Data. Monthly DF incidence rates per 100,000 population for all provinces from 1997-2016 were provided by the National Institute of Hygiene and Epidemiology (NIHE), Vietnam. Figure 1 (top) shows the yearly incidence rates for all provinces. Central and Southern provinces, where the weather is warmer, rainier, and more humid, have higher incidence rates than Northern ones. Monthly climate data for each province from

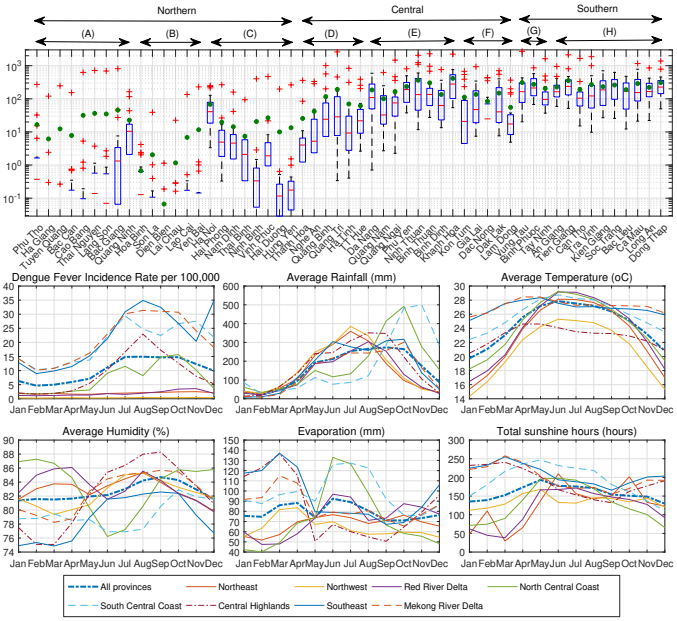

Fig. 1. (Top) Yearly incident rates per 100,000 population (in log scale) for all provinces from 1997 to 2016 using Whisker plots (green dots indicate mean values). Subregions: (A) Northeast, (B) Northwest, (C) Red River Delta, (D) North Central Coast, (E) South Central Coast, (F) Central Highlands, (G) Southeast, and (H) Mekong River Delta; (Bottom) The yearly averages of DF incidence rates and some climate factors from Jan to Dec for all subregions

1997-2016 are obtained from the Vietnam Institute of Meteorology, Hydrology and Environment (IMHEN) including: rainfalls (total rainfall (TRain) and highest rainfall (MaxDRain) in mm), number of rainy days (nRainDays) (days), temperatures (average temperature (AvgTemp), maximum average temperature (MaxAvgT), minimum average temperature (MinAvgT), absolute maximum temperature (MaxAbsT), and absolute minimum temperature (MinAbsT) in °C), humidity (average relative humidity (AvgHum) and minimum relative humidity (MinHum) in percentage), evaporation (TEva) (mm), and total sunshine hours (nSunHours) (hours). Figure 1 (bottom) shows the average values from Jan to Dec of DF incidence rates and several climate factors for all subregions compared to the whole country. The results show significant variations of peak times and value ranges for different subregions (and even their provinces). E.g., rainfalls in the north central coast provinces reach the peak in around Oct-Nov and range from 33.8 to 491.7 mm per year, while rainfall in the southeast provinces has peaks during June-Oct and ranges from 8.2 to 317.0 mm per year. These make it difficult to provide a unified model for effectively predicting DF for all provinces at once.

Problem formulation. Given a set of n province $\mathbb{P} = \{(P_i, C_i^1, \cdots, C_i^c, D_i)\}$ where P_i is province i, $C_i^j = (c_{i1}^j, \cdots, c_{im}^j)$ $(1 \leq j \leq c)$ are longitudinal climate factors of m months, and $D_i = (d_{i1}, \cdots, d_{im})$ is a longitudinal DF incidence rate of m months for province P_i, our target is to build an effective model to predict if DF outbreaks happen *within* a time frame of t months for all provinces, i.e., is there a DF outbreak *within* the next t months for each province P_i.

Dengue fever outbreak events. In public health, the notion of DF outbreak varies across different regions and countries [35]. While our method can be used with any of these notions, we employ the most common method of using mean and standard deviation of w recent years in this paper [13], [35]. If a DF rate exceeds k_d standard deviations from the mean value (default $k_d = 1$), we have an outbreak.

Definition 1 (DF outbreak event): For a province P_i with $D_i = (d_{i1}, \cdots, d_{im})$, a DF outbreak happens at time j if $d_{ij} \geq \mu_{ij} + k_d\sigma_{ij}$, where $\mu_{ij} = \sum_{t=j-12w}^{t=j} d_{it}/12w$ and $\sigma = \sqrt{\sum_{t=j-12w}^{t=j}(d_{it} - \mu_{it})^2/(12w-1)}$ are the mean and standard deviation of w recent year data of D_i.

Extreme weather events. In this paper, we approach the prediction task from a different viewpoint: *extreme (unusual) climate events*. The general idea is that a DF outbreak in a province has a link to significant climate change events in that province rather than small climate perturbations. Thus, instead of directly feeding raw climate data into prediction models like all existing works [6], [8], [11], [22], we discretize climate data into events of extreme (abnormal) weather conditions and use them as inputs for our prediction model. In this way, we also flatten the differences in climate value ranges of different provinces and thus create a unified view for all provinces based on climate events. This scheme allows us uniquely to aggregate data from different provinces to improve model training (c.f. Section III). Similar to DF outbreaks, we use mean and standard deviation of w recent years to identify extreme climate events. If a climate value is larger or smaller than k_c standard deviations from the mean value, it indicates an extreme climate event at that time.

Definition 2 (Extreme climate events): For a province P_i and a climate $C_i = (c_{i1}, \cdots, c_{im})$, an extreme climate event happens at time j if $c_{ij} \geq \mu_{ij} + k_c\sigma_{ij}$ or $c_{ij} \geq \mu_{ij} - k_c\sigma_{ij}$, where $\mu_{ij} = \sum_{t=j-12w}^{t=j} c_{it}/12w$ and $\sigma = \sqrt{\sum_{t=j-12w}^{t=j}(c_{it} - \mu_{it})^2/(12w-1)}$ are the mean and standard deviation of w recent years data of C_i.

For both Definitions 1 and 2, we use the default window $w = 5$ to avoid too long history effect like [13], [35].

Event-to-event relationships. W.l.o.g., let X_i and Y_j be climate factors (or DF incidences) of provinces P_i and P_j, respectively. Let $E(X_i) = \{x_1, \cdots, x_a\}$ and $E(Y_j) = \{y_1, \cdots, y_b\}$ be the set of extreme climate events (or DF outbreak events) that occur at time x_u $(1 \leq u \leq a)$ of X_i and y_v $(1 \leq v \leq b)$ of Y_i, respectively.

Definition 3 (Forward/backward relationship): Given an event $x_u \in E(X_i)$, let $\overrightarrow{d}(x_u, Y_j)$ be the distance between x_u and its closest follow up events $y_v \in E(Y_j)$ (∞ if there is no follow up event), i.e., $\overrightarrow{d}(x_u, Y_j) = y_v - x_u$, where $v = argmin_z(y_z - x_u)$ s.t. $y_z \geq x_u$. Let $\overrightarrow{pdf}_{(X_i,Y_j)}(t)$ be the probability that an event in $E(X_i)$ has a follow up event in $E(Y_j)$ after exactly t months, i.e., $\overrightarrow{pdf}_{(X_i,Y_j)}(t) = \frac{1}{a}\sum_{u=1}^{a}\begin{cases}1 & \text{if } \overrightarrow{d}(x_u, Y_j) = t\\ 0 & \text{otherwise}\end{cases}$. Let $\overrightarrow{cdf}_{(X_i,Y_j)}(t)$ be the cumulative probability that an event in $E(X_i)$ has a follow up events in $E(Y_j)$ within t months, i.e., $\overrightarrow{cdf}_{(X_i,Y_j)}(t) = \sum_{z=0}^{t}\overrightarrow{pdf}_{(X_i,Y_j)}(z)$. The two functions $\overrightarrow{pdf}_{(X_i,Y_j)}(t)$ and $\overrightarrow{cdf}_{(X_i,Y_j)}(t)$ represent the forward relationships between X_i and Y_j, respectively, i.e., the probability of an event in $E(Y_j)$ occurs after an event in

$E(X_i)$ w.r.t. the time period t. Similarly, we use $\overleftarrow{pdf}_{(X_i,Y_j)}(t)$ and $\overleftarrow{cdf}_{(X_i,Y_j)}(t)$ to represent the backward relationship between X_i and Y_j, i.e., the probability that an event in $E(Y_j)$ occurs before an event in $E(X_i)$ w.r.t. the time period t. These functions can be straightforwardly defined following the forward cases by using the distance function $\overleftarrow{d}(x_u, Y_j) = x_u - y_v$ where y_v is the closet event in $E(Y_j)$ that occurs before $x_u \in E(X_i)$.

Assume that $E(X_i) = \{2, 5, 7, 9, 15\}$ and $E(Y_j) = \{1, 3, 7, 8, 10\}$. We have $\overrightarrow{d}(5, Y_j) = 7 - 5 = 2$, $\overrightarrow{d}(15, Y_j) = \infty$, $\overrightarrow{pdf}_{(X_i,Y_j)}(0) = 1/5 = 0.2$, $\overrightarrow{pdf}_{(X_i,Y_j)}(1) = 2/5 = 0.4$, and $\overrightarrow{cdf}_{(X_i,Y_j)}(3) = 0.2 + 0.4 + 0.2 + 0 = 0.8$. Similarly, we have $\overleftarrow{d}(9, Y_j) = 9 - 8 = 1$, $\overleftarrow{d}(15, Y_j) = 15 - 10 = 5$, $\overleftarrow{pdf}_{(X_i,Y_j)}(3) = 0$, $\overleftarrow{pdf}_{(X_i,Y_j)}(5) = 1/5 = 0.2$, and $\overleftarrow{cdf}_{(X_i,Y_j)}(1) = 0.2 + 0.4 = 0.6$.

The forward/backward probability functions provide a new unique view on the relationships among climate factors and DF incidences that capture large climate value changes, which is more robust to small data perturbations than traditional correlation analysis approaches like [3]. In Section IV, we demonstrate the uses of these functions to explicitly study/explain the impacts of different climate factors on DF outbreaks. These functions will also be used to capture the DF outbreak co-occurrence probability among provinces during the prediction propagation step in Section III Phase D.

III. The Proximity Time Ensemble Prediction

Our DF outbreak prediction model is built upon several key ideas: (i) converting raw climate data into events to have unified views; (ii) training prediction models using aggregated data from different provinces to avoid over- and underfitting problems; (iii) using many different prediction methods and aggregating their results; (iv) improving prediction results of a province by exploiting results of other provinces; and (v) propagating outcomes over time periods. Figure 2 illustrates our approach with 6 different main phases as follows.

Phase A: Data transformation. For each province P_i and a climate factor $C_i = (c_{i1}, \cdots, c_{im})$, we transform each continuous value c_{ij} into an event code $evn(c_{ij})$ representing its range w.r.t. extreme climate event thresholds k_c in Definition 2. By using different values of k_c, we can have finer transformations of c_{ij} by smaller ranges. In our experiments, we use two statistical common values $k_c = 1$ and $k_c = 2$ (representing 1 and 2 deviations from the mean value, respectively).

Definition 4 (Event transformation): Let $evn(c_{ij})$ be the event code value of c_{ij} wrt. mean μ_{ij} and standard deviation σ_{ij} at time j within w recent year. We have:

$$evn(c_{ij}) = \begin{cases} 2 & \text{if } c_{ij} \geq \mu_{ij} + 2\sigma_{ij} \\ 1 & \text{if } \mu_{ij} + 2\sigma_{ij} > c_{ij} \geq \mu_{ij} + \sigma_{ij} \\ 0 & \text{if } \mu_{ij} + \sigma_{ij} > c_{ij} \geq \mu_{ij} - \sigma_{ij} \\ -1 & \text{if } \mu_{ij} - \sigma_{ij} > c_{ij} \geq \mu_{ij} - 2\sigma_{ij} \\ -2 & \text{otherwise} \end{cases}$$

Figure 2 (A) illustrate the case of a single k_c threshold where we have 3 different ranges (and 3 code values): extreme (abnormal) low (-1), normal (0), and abnormal high (1). Besides providing a unified event view for different climate factors in different provinces as discussed in Section II, another major benefit of this data transformation is that each transformed data point consists of aggregate climate information from the past w years. Thus, this allows creation of effective prediction models without having to incorporate long tails of past data into their input (w year data to a single coded point). This helps to reduce model overfitting and computation time, thus improving overall performances as shown in Section IV.

Phase B: Province proximity graph construction. Let $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ be the graph, where $\mathbb{V} = \{P_1, \cdots, P_n\}$ be the set of all n provinces, and $\mathbb{E} = \{(P_i, P_j)\}$ be the set of edges that connect two provinces P_i and P_j if their similarity function $sim(P_i, P_j)$ exceeds a predefined threshold ϵ.

Definition 5 (Geographic province similarity): Given two provinces P_i and P_j, we have:

$$sim(P_i, P_j) = \begin{cases} 1 & \text{if } P_i \text{ and } P_j \text{ share a border} \\ 0 & \text{otherwise} \end{cases}$$

In this paper, we use the geography similarity function described in Definition 5 with the threshold $\epsilon = 0$, i.e., the graph $\mathbb{G}$ connects two provinces P_i and P_j if they are neighbors. The intuition behind it is that if two provinces are close, they are more likely to share similar climate patterns and DF incidences. Moreover, if a DF outbreak happens in P_i, it is more likely that it will affect other nearby provinces due to the transmission nature of DF as studied in [18], [20]. The proximity graph $\mathbb{G}$ is the backbone of our prediction method for controlling the learning process by aggregating training data and propagating prediction results among similar provinces in Phases C and D below, respectively.

Though Dengue Fever is climate sensitive, there are many other non-climate factors that can significantly affect it, e.g., human behaviors [17], education [36], environment [18], urbanization [1], and income [18]. This information can be incorporated into the similarity function besides the geography to better present the similarities among provinces and thus to better control the learning process. This approach is simpler than incorporating these factors directly into the regression models like existing techniques [5], [18]. But it is more flexible since these external factors can be explicitly controlled and different types of data, e.g., textual or non-longitudinal data can be easily incorporated via extended similarity functions. Due to the data types available in this study, we demonstrate our algorithm using only the geographic proximity graph.

Phase C: Dengue Fever outbreak prediction. Let τ be the longest time we want to predict into the future. For each province P_i and for each $1 \leq t \leq \tau$, we aim at predicting a binary output where 1 means that there will be an outbreak within the next t months for P_i and 0 otherwise.

C1: Create training data. For each province P_i, month j and time t, the prediction output $out_{ijt} = \begin{cases} 1 & \text{if } \overrightarrow{d}(j, D_i) \leq t \\ 0 & \text{otherwise} \end{cases}$, i.e., where $\overrightarrow{d}(j, D_i)$ is the time distance from j to the closest *strictly* follow up DF outbreaks (c.f. Definition 3 but with $y_z > x_u$, i.e., distance must be at least 1 month since we already use current DF outbreak/none-outbreak in the training data). The input data $in_{ij} = (evn(c^1_{ij}), \cdots, evn(c^c_{ij}), evn(d_{ij}))$ is a vector of encoded climate events and DF events. Besides the climate events, DF events are used as a disease historical factor to improve the prediction accuracy.

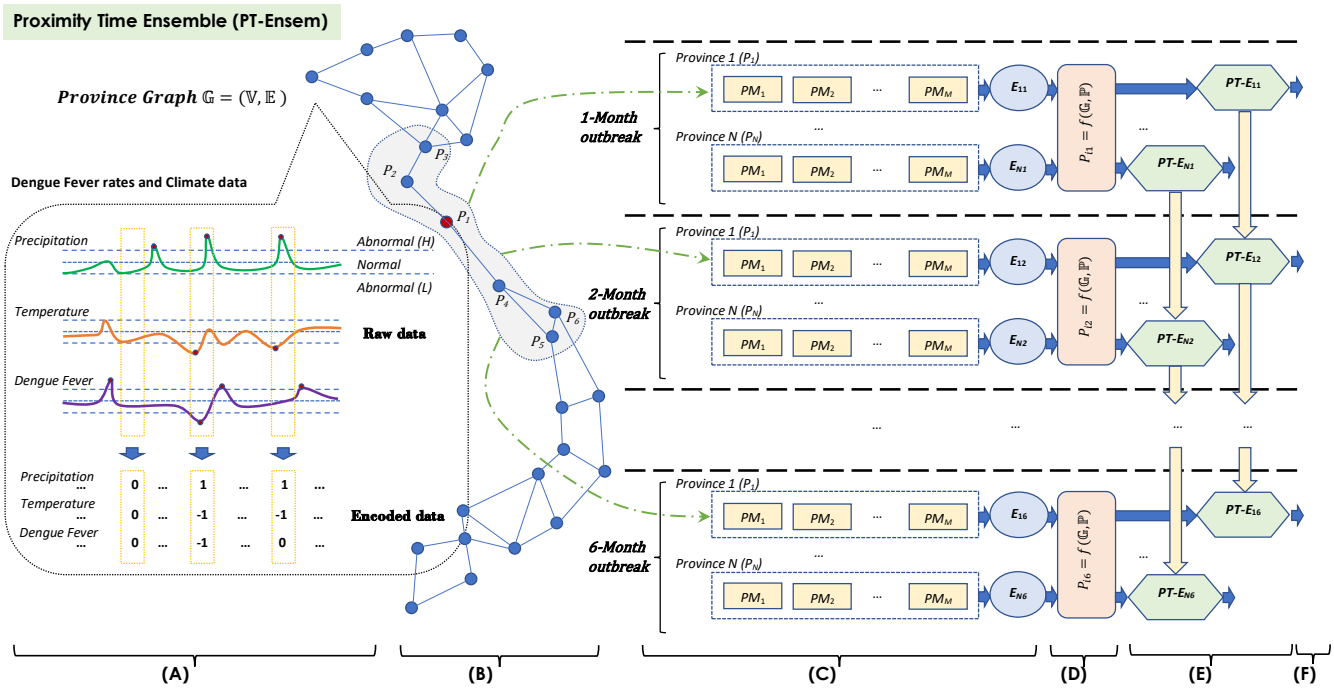

Fig. 2. The Proximity Time Ensemble (PT-E) approach for Dengue Fever outbreak prediction. The algorithm consists of 6 main phases: (A) climate and dengue fever data transformation for each province, (B) province proximity graph construction, (C) ensemble DF outbreak prediction, (D) proximity-based prediction adjustment, (E) time-based prediction propagation, and (F) final decisions

C2: Data aggregation. Input data for all provinces are represented in a unified view of events, thus flattening the differences in data value ranges. Hence, we propose to enrich training data for each province by additionally including data from other similar provinces. This helps to increase the size and diversity of training data, thus reducing both over- and under-fitting problems to significantly improve the performance as we will demonstrate in Section IV. Data aggregation is a key point of our algorithm and it is conducted via the proximity graph $\mathbb{G}$ created in Phase B. For each province P_i, let κ-$hub(P_i)$ be the set of provinces that are up to k edges away from P_i in the graph $\mathbb{G}$, i.e., provinces that are similar but not too far away from P_i. E.g., 2-$hub(P_1) = \{P_1, P_2, P_3, P_4, P_5, P_6\}$ in Figure 2. Let $train(P_{it})$ be the training data of P_i for t-months ahead prediction. The κ-hub extended training data κ-$train(P_{it}) = \cup_{P_j \in \kappa\text{-}hub(P_i)} train(P_{jt})$ is a union set of training data of all provinces P_j that are κ edges away from P_i (default $\kappa = 2$) on the graph $\mathbb{G}$.

C3: Ensemble prediction. Ensemble is a common method to improve learning performance and has been widely applied in DF prediction, e.g., [32]. In this paper, we use M different prediction models PM_1 to PM_M to predict the outbreak for each province P_i at each time period t separately as shown in Figure 2 using the training data κ-$train(P_i)$ described above. However, instead of returning the final binary prediction outcomes, e.g. via majority votes, we return $e_{it} = \frac{1}{M}\sum_{j=1}^{M} out(PM_j, t)$, where $out(PM_j)$ is the prediction output of the model PM_j at time t, i.e., e_{it} is the probability of DF outbreaks during the next t months. The final results are only being decided after the proximity and time adjustment processes in Phase D and E below.

Phase D: Proximity-based prediction propagation. Since DF is a transmissible disease, an outbreak in a province may affect or be affected by other provinces, e.g., via human mobility [34] or extreme climate events across nearby provinces [21]. Hence, in this phase, we propagate the prediction results among provinces via the proximity graph $\mathbb{G}$ to improve the overall prediction accuracy. The intuition is simple, e.g., if a nearby province P_j of P_i is predicted to have outbreaks, it is more likely that P_i will also have an outbreak.

Let $p(D_i) = |E(D_i)|/m$ be the probability of an DF outbreak in province P_i. Let $p(D_j|D_i)$ be the probability that we have an outbreak in P_j given an outbreak in P_i, i.e., $p(D_j|D_i) = \overrightarrow{pdf}_{(D_i,D_j)}(0)$ (note $p(D_i|D_i) = 1$). If an outbreak will happen in P_j, the probability it will happen in P_i, denoted as w_{ij}, can be calculated via Bayes's theorem as $w_{ij} = \frac{\overrightarrow{pdf}_{(D_i,D_j)}(0)\cdot|E(D_i)|}{|E(D_j)|}\alpha$, where $\alpha \in [0, 1]$ (default $\alpha = 0.1$) is a predefined propagation constant to control the influence intensity among different provinces (note $w_{ii} = 1$), i.e., to limit the effects of wrong DF outbreak prediction in one province on others. For each province P_i, we use w_{ij} as a weight to propagate prediction results from its proximity similar province $P_j \in \lambda$-$hub(P_i)$ (default $\lambda = 2$) to P_i.

For each province P_i and prediction time t, let $err(P_i) = \sum_{P_j \in \lambda\text{-}hub(P_i)} w_{ij}(e_{it} - e_{jt})^2$ be a weighted sum of square prediction disagreements between P_i and its λ-hub neighbors P_j. Our target is to update e_{it} to minimize $err(P_i)$. Taking the derivative of $err(P_i)$ wrt. e_{it}, we can update e_{it} as $\hat{e}_{it} = \frac{\sum_{P_j \in \lambda\text{-}hub(P_i)} e_{jt} w_{ij}}{\sum_{P_j \in \lambda\text{-}hub(P_i)} w_{ij}}$. The updating process is iteratively performed for all provinces in each round until the average of changes for all provinces $diff(\mathbb{P}) = \frac{1}{n}\sum_{i=1}^{n} |e_{it} - \hat{e}_{it}|$ converges, i.e, $diff(\mathbb{P})$ is smaller than a predefined threshold

θ (default $\theta = 0.001$) or the number of iterations reaches a predefined threshold ϕ (default $\phi = 1000$).

Phase E: Time-based prediction propagation. Due to the time spanning nature of the DF disease, i.e., if an outbreak occurs in a month, there is high chance that the next month will also have an outbreak. Also, we predict outbreaks happen within a time frame of t months which cover $t-1$ periods. These cumulative relationships can be exploited to improve the prediction accuracy. Therefore, for each province P_i and prediction time t, we propose to propagate the results from $t-1$ to t as follows: $e_{it} = e_{it} + \beta e_{it-1}$, where $\beta \in [0, 1]$ (default $\beta = 0.3$) is a predefined propagation constant to control the influence between two consecutive prediction time points, i.e., to limit the effect of wrong outbreak predictions being propagated into the future.

Phase F: Final decision. Since the outcome e_{it} of P_i at time t is a continuous value, we need to find a cut-off threshold γ_{it} to determine an outbreak (i.e., $e_{it} \geq \gamma_{it}$). Here we exploit the training data to automatically find an optimal cut-off threshold γ_{it} for P_i. First, for each time t, we get the ensemble results for the training set $train(P_{it})$. Second, we repeat the proximity-based and time-based propagation in Phases D and E on the training outcomes. Lastly, we use a grid-search with a step of 0.01 from 0 to 1 to find γ_{it} that maximizes a predefined classification scoring function, e.g., accuracy, balanced accuracy or F1-score. In this work, we employ the balanced accuracy to balance both the outbreak and non-outbreak prediction. This fits well with our problem where the non-outbreak events dominate the outbreak ones, leading to data imbalance. The identified value γ_{it} will be used as a cut-off threshold to detect DF outbreaks.

IV. Experiments

Experiments are conducted on a workstation with 4.0Ghz CPU and 32GB RAM using Python 3. All codes and data will be publicly available upon request.

Parameters. Unless otherwise stated, we use the default parameters (c.f. Section III) for PT-Ensem. For other prediction models, we use grid-searches to find suitable parameters for them to ensure fair comparisons.

Prediction evaluations. We split the data into a training set from 1997-2013 and a test set from 2014-2016. To evaluate the performance of the binary outbreak prediction, we use 3 common measures: balanced accuracy, specificity and sensitivity. Specificity is the ratio of number of correct predicted normal months and total number of normal months. Sensitivity is the ratio of number of correct predicted outbreak months and total number of outbreak months. Balanced accuracy, which is the average of specificity and sensitivity, is used as the main evaluator due to its effectiveness on binary classification problem, especially when the label is imbalanced [37].

Performance of different methods. Figure 3 shows the performance of 19 different prediction models (including recent highly-rated models e.g., CatBoost [31] or LightGBM [30]) on all provinces using their own training data. Due to the class imbalance between outbreak (1) and none-outbreak (0), we use class weights as ratios between outbreaks and non-outbreaks to improve the performance. In terms of overall performance

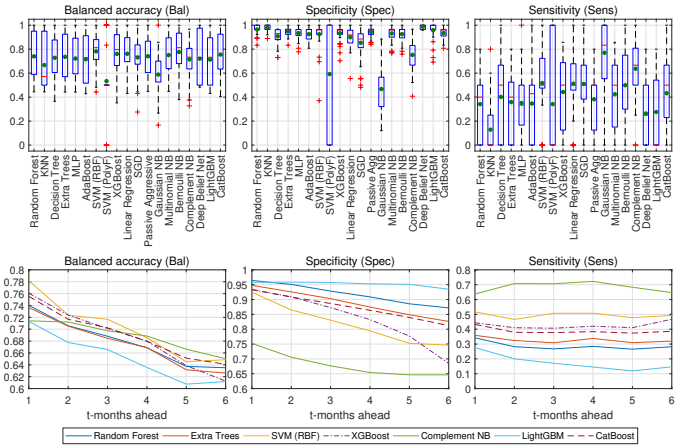

Fig. 3. Performance of 19 different prediction models without data aggregation: (top) performance metrics for all provinces at 1-month ahead prediction (Whisker boxplots with green dots being mean values); and (bottom) performance metrics for 1-6 months ahead prediction of some models

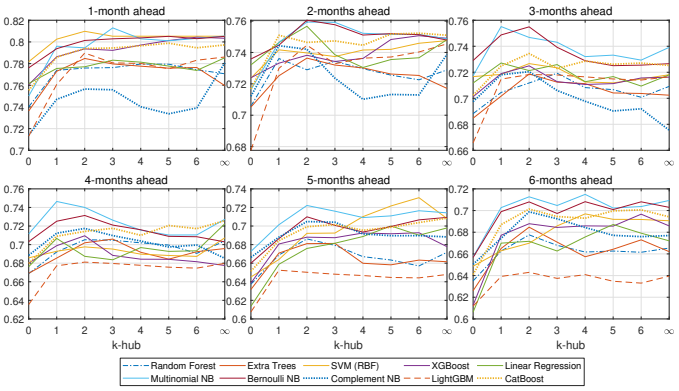

Fig. 4. Performance (balanced accuracy) of different models w.r.t. different values of κ in our data aggregation scheme ($\kappa = 0$ and $\kappa = \infty$ mean we do not use data aggregations and aggregate all provinces together, respectively)

(balanced accuracy), SVM with Radial Basis functions (SVM (RBF)), XGBoost and Catboost are among those with frequent top results, while KNN, SVM with polynomial function (SVM(PolyF)) and Gaussian Naive Bayes (Gaussian NB) constantly have worst outcomes. Most methods show very high specificity and very low sensitivity, meaning that they fail to predict most outbreaks and tend to put everything as normal. Moreover, the longer we predict into the future, the worse the overall performance obviously due to larger uncertainty.

Effect of training data aggregations. Figure 4 presents the performance of different top models (the rest is omitted for clarity) using the data aggregation approach described in Phase C (c.f. Section III). When data of neighbor provinces are started to be aggregated to train prediction models ($\kappa = 1$), the prediction accuracy increases significantly in all cases as expected (richer and more diverse data help to reduce both over- and under-fitting problems). In major cases, when κ is large enough, the performance starts to decrease since data from too differing provinces may be mixed, thus decreasing the training data quality. However, the overall performance very rarely drops below the non-aggregation ones ($\kappa = 0$). Peak results are typically acquired with κ from 1 to 5.

Performance of our PT-Ensem approach. Based on the

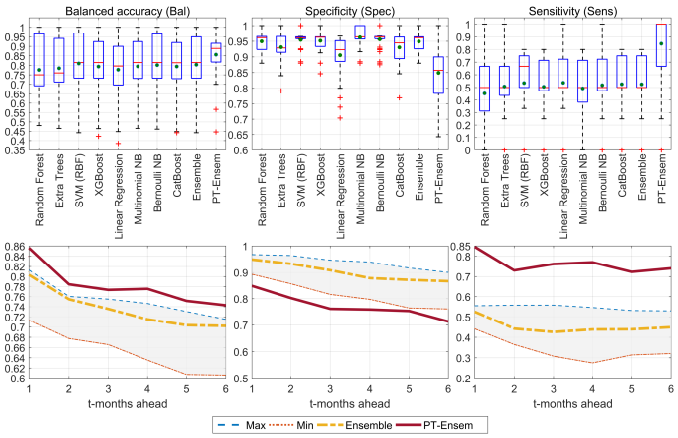

Fig. 5. Performance of PT-Ensem compared to the best 8 prediction models and the traditional ensemble method (majority votes) for all provinces. (Top) Performance metrics for 1-month prediction of all provinces in Whisker boxplot with green dots as mean values. (Bottom) Averaged balanced accuracies for 1-6 month prediction on all provinces. The Min and Max lines represent the highest and lowest balanced accuracy of all 8 selected models with different data aggregation values ($\kappa = (1-6, \infty)$). We only use default parameters for Ensemble and PT-Ensem

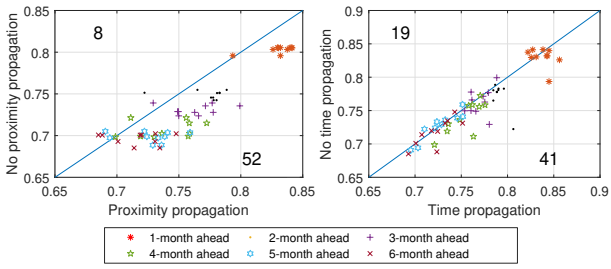

Fig. 6. Averaged balanced accuracies of PT-Ensem with/without proximity and time propagations for all provinces with different parameters κ = 1-5 and λ = 1-2. A point below the crossline means using proximity and time propagations are better than not using them

above studies, we choose the 8 best prediction models to use with PT-Ensem (c.f. Figure 5). For 1-month ahead prediction (Figure 5 (top)), PT-Ensem significantly outperforms all other methods with the averaged balanced accuracy of 0.856 over all provinces compared to the best result of 0.809 of SVM (RBF). It also outperforms state-of-the-art majority vote ensemble approach with a score of 0.803. While its averaged specificity is slightly lower than others (0.849), the averaged sensitivity is significantly different to the second best models (0.845 vs. 0.537, respectively), indicating a dramatic improvement in DF outbreak prediction capability without raising many false alarms. Figure 5 (bottom) further compares PT-Ensem with the best possible results of all its member models for 1 to 6 months prediction. While the traditional Ensemble approach helps to increase the performance in most cases, it is still lying below the max line. However, PT-Ensem significantly boosts the overall performance over the limitations of its components in both long and short-term predictions. These performance improvements come from the proximity and time propagation in Phases D and E (c.f. Section III) as analyzed below.

Effects of the proximity and time propagation. Figure 6 compares the averaged balanced accuracy for all provinces with and without the proximity and time propagation. Over 60 studied cases, using proximity helps to improve the prediction

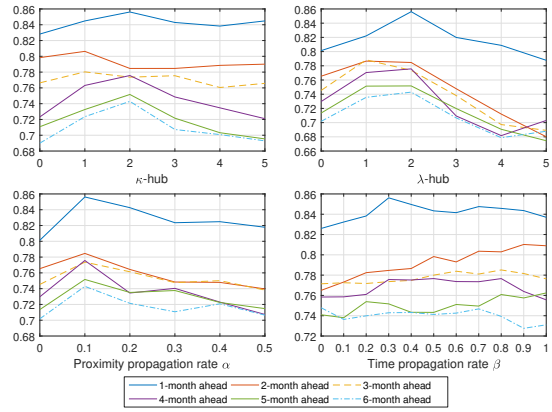

Fig. 7. Averaged balanced accuracy for PT-Ensem over all provinces wrt. different parameters

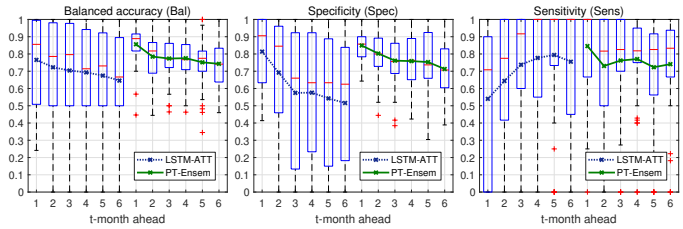

Fig. 8. Performance of PT-Ensem compared to LSTM-ATT (the most recent DF incidence rate/case prediction method for Vietnam) using Whisker boxplots

accuracy in 52 cases (86.6%). While using time-propagation helps to improve the results in 41 cases (68.3%). The proximity propagation has stronger effects than time-propagation (points move further down from the crossline). The further the prediction, the weaker the propagation effect (more points above the crossline). This is quite obvious since more wrong results will be propagated among provinces and across prediction periods.

Parameter analysis. The effects of parameters κ, λ, α and β on the averaged balanced accuracy of PT-Ensem for all provinces are shown in Figure 7. When the influence areas are expanded (i.e., κ and λ are increased), the prediction accuracy increases and reaches peak performance at κ = 1-2 and λ = 1-2 before dropping (especially for λ). The main reason is that the more provinces involved in the propagation, the more chance wrong prediction results are propagated to others, leading to performance degradation. We suggest to keep κ = 1-4 and λ = 1-2. The two parameters α and β are used to control the influences among provinces and time periods, i.e., propagating *good* predictions and restricting *bad* predictions towards others. Overall, when $\alpha \neq 0$ and $\beta \neq 0$ (i.e., use propagation), the performance increases compared to not using them (the same results can be seen in Figure 6). The peak result is around $\alpha = 0.1$ while varying significantly from 0.2 to 1.0 for β. So, we suggest to keep $\alpha \in [0.1, 0.2]$ and $\beta \in [0.2, 1]$.

Comparisons with deep learning approaches. In [10], various deep learning methods (LSTM, CNN, Transformer, and LSTM-ATT) and most common traditional methods (SARIMA, Poisson Regression, XGBoost, and SVR with Radial Basis and Linear kernels) are employed to predict DF incidence rates using climate data for 20 provinces of Vietnam. Thus, we adapt these methods to our problem by using predicted incidence rates from 1 to 6 months to calculate

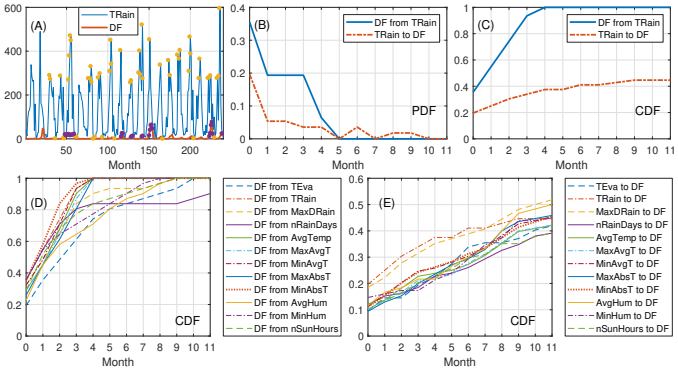

Fig. 9. Links between climate factors and DF in Ha Noi (capital of Vietnam). (A) Monthly rainfalls and DF incidence rates from 1997-2016 (yellow and purple circles indicate extreme rain events and DF outbreaks, respectively). (B) The backward probability that a DF outbreak has a previous rainfall event at a specific time (DF from TRain) and the forward probability that a rainfall event has a follow up DF outbreak at a specific time (TRain to DF). (C) Similar to (B) but with cumulative probabilities. (D) The backward cumulative probability between DF and all climate factors. (E) The forward cumulative probability between DF and all climate factors

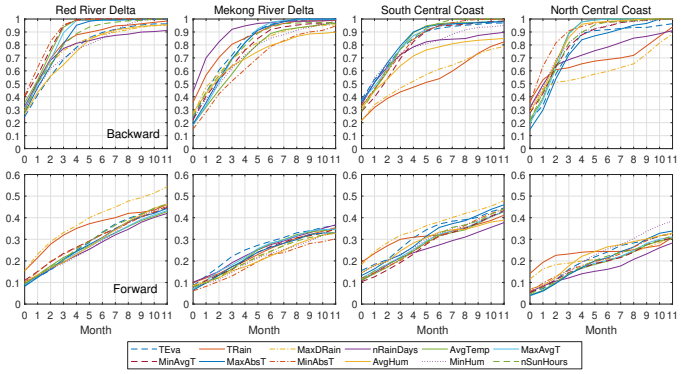

Fig. 10. Forward/backward relationships between climate and DF for some subprovinces

the outbreaks as described in Section II. Note that this is also the most common way to predict DF outbreaks in the literature (i.e., not predict outbreak explicitly but via the incidence rates/cases, e.g., [8], [10]). Similar to [10], we use grid-searches to find the best parameters for these methods to ensuring fair comparison with our method.

Figure 8 shows the averaged balanced accuracy, specificity and sensitivity of LSTM-ATT, the best method reported in [10], and PT-Ensem over all provinces. PT-Ensem completely dominate LSTM-ATT in all prediction periods. E.g., the LSTM-ATT/PT-Ensem balanced accuracy results for 1 to 6 month ahead predictions are 0.765/0.856, 0.722/0.784, 0.704/0.773, 0.693/0.775, 0.674/0.751, 0.645/0.743, respectively. Overall, the performance differences are from 6.17 to 9.72 percentage points over 100, a significant improvement. The Whisker boxplots also show that PT-Ensem provides stable (compact) high quality results for all provinces, while the results of LSTM-ATT vary significantly on different provinces.

Runtime comparisons. Another notable advantage of PT-Ensem, compared to LSTM-ATT, is its much lighter computational cost. While it takes PT-Ensem less than 30 minutes to predict all provinces (default parameters), LSTM-ATT needs more than 12 hours to complete (with optimal found parameters) (i.e., around 24x slower than PT-Ensem).

Exploring Extreme climate and DF outbreaks. In this part, we illustrate how to explore the links between climate and DF using the forward/backward relationship functions (c.f. Definition 3). Due to space constraints, we present only some findings here, leaving the rest for a longer version of this paper.

Figure 9 illustrates the relationships between climate factors and DF outbreaks for Ha Noi. There are 31 (12.9%) DF outbreaks and 62 (25.8%) extreme rainfall events during 20 years. The backward functions $\overleftarrow{pdf}_{(DF,TRain)}(t)$ and $\overleftarrow{cdf}_{(DF,TRain)}(t)$ in (B and C) shows the probability that a DF outbreak has an extreme rainfall occurrence exactly t month or within t months before it ($t = 0$ means co-occurrence). E.g., there is a 19.3% chance that a DF outbreak has a previous rainfall event at exactly 3 month. Within 4 months, the cumulative chance is 100%, i.e., whenever we have a DF outbreak, we know that extreme rainfall events occur within 4 months before. These match well with reported lag times of around 3 months between a peak rainfall and a peak in DF incidence for Hanoi [3], [23]. However, while peak analysis can miss important unusual outbreaks due to its averaged scheme, this backward function shows a more detailed story. On the other hand, the forward functions $\overrightarrow{pdf}_{(TRain,DF)}(t)$ and $\overrightarrow{cdf}_{(TRain,DF)}(t)$ in (B and C) tell us the probability of a DF outbreak after an extreme rainfall event, e.g., 3.5% after exact 3rd month and 33.9% after 3 months. These indicate that extreme rainfall events are not a sole cause of DF outbreaks. This is not a surprise since there are many other known factors that can affect DF incidence, e.g., geography, urban/rural areas or socio-ecology factors [5], [14], [34]. (D) and (E) compare all our 12 climate factors on DF outbreaks (the higher the values, the stronger the relationships). DF outbreaks are more likely to be associated to temperature and rainfall events than evaporation and humidity events (D). And, rainfall events are more likely to lead to DF outbreaks than all other factors (E). These match with previous analyses for Hanoi [23] (and other areas, e.g., Mexico [12]) but contradict [3], where evaporation and humidity are found to be correlated to DF incidence but not rainfall and temperature in Hanoi from 2008-2015. One reason for the disagreement is that they use much less data than us (8 vs 20 years). Moreover, correlation measurements (on raw DF incidence data) are very sensitive to small fluctuations in data, (possibly) leading to missing relationships. In contrast, our approach focuses on big changes (major points) and thus is less sensitive to these small fluctuations.

Different provinces show (significant) different relationship patterns between DF and climate. However, neighbor provinces are more likely to have more similar patterns. E.g., Ha Noi and its neighbor provinces Hai Phong, Nam Dinh, Thai Binh, Ninh Binh, Vinh Phuc, and Hung Yen have rainfall as the most influential forward factor, except Hai Duong, where it is hard to distinguish effects of different climate factors. This again confirms the intuition behind our data aggregation scheme as shown above. In Figure 10, we show the mean forward and backward relationships of all provinces in some subregions to have a bigger view from region perspectives. In Red River Delta, DF and climate have stronger links than all other areas (indicated by higher forward/backward values). Rainfall is the

strongest forward factor and temperature is the strongest backward one. Most of its provinces show similar behavior. Many provinces in Mekong River Delta tend to behave differently. E.g., AvgHum is the least important factor for Dong Thap and TEval is the strongest factor, while MinAbsT is the weakest and nRainDays is the strongest factor for Ca Mau and Bac Lieu. nRainDays and MinAbsT are the strongest and weakest ones overall but their differences to others are not very clear. Northern and Southern central coastal regions seem to have very similar behavior. Rainfall is the weakest backward factor but is among the strongest forward factors. nRainDays is the weakest forward but among average backward factors. Similar to Mekong River Delta, the relationships between climate and DF are quite diverse. For Southeast subregions MinAbsT is the weakest factor and TEval is the strongest one. For other subregions, no clear dominant climate factors are found.

V. Related Works and Discussion

Dengue fever and climate. Climate factors are known to have direct impacts on mosquito development and behaviors [4], [15], [17], [38], thus affecting DF incidence. E.g., the development rates of *Ae.aeypti* mosquito increases when temperature is from 12°C to 30°C but drops quickly after 40°C [4]. Mosquito bite rate increases with temperature in Thailand [38]. Barrera et al. [17] report that high rainfall increases mosquito density in Puerto Rico. However, too high rainfall can flush out their breeding sites, thus reducing their numbers [15].

Other researches attempt to directly elucidate the lag correlations between different climate factors and DF incidence rates/cases [3], [21], [22]. These can be used to design effective warning systems by choosing highly correlated factors as predictors or designing look-back parameters in regression models [11], [21]. However, the results vary considerably depending on studied areas and time periods. E.g., Do et al. [23] report positive correlations between rainfall and temperature and DF incidence for Ha Noi, Vietnam with lag times 0-3 months and 0-2.5 months between peaks of rainfall, temperature, and DF, respectively. However, Tran et al. [3] report no correlation between rainfall and temperature on DF incidence for the same city but with different wards and time periods. Minimum temperature is reported to be correlated with DF incidence with 1-2 months lag for Mexico [12] and Guangzhou, China [15]. Many other papers report positive/negative correlations between DF incidence and other climate factors such as sunshine hours [3], [39], wind speed [15], humidity [3], [15], [39], evaporation [3], and El Niño [12]. Though the finding relationships can be location-specific, the frequent reported correlations make climate the most widely used predictor for building early warning systems for DF [2]. However, these also make building a general prediction model that can work well for different areas a challenging task. PT-Ensem does not rely on any specific lag regression models like [12], [13], thus it can be used with other climate data in other areas straightforwardly without any expert knowledge.

Our forward/backward relationship functions provide a different view based on significant data points on the links between climate and DF. Compared to traditional correlation analysis, it is less sensitive to small or noisy data values, thus can reflect better the relationship between them. Our study also shows varying relationships between climate and DF across different provinces and regions.

Dengue fever and other factors. Despite being a climate sensitive disease, many other non-climate factors can affect the DF incidence. E.g., Stoddard et al. [16] report that local human movement can lead to the spread of dengue virus. Barrera et al. [17] point out significant correlations between human behaviors such as water storage or garbage collection to DF incidence in Puerto Rico. Many other factors have been reported including (but not limited to): education [36], degree of urbanization [1], and socio-economic covariates [18]. These factors have been used as predictors for many proposed models [5], [9], [27]. E.g., Liu et al. [9] use spatial interactions of human movements among regions as input features to predict DF incidence in Guangzhou, China. The neighborhood graph of PT-Ensem also implicitly captures the movement of people across nearby provinces. Compared to climate data, many of these non-climate data are very hard to collect and historical data may not be fully available to put into prediction models, thus limiting their usability. In this work, we only have climate data to study but exploring the performance of PT-Ensem with non-climate data is an interesting task in the future.

Dengue fever predictions. Most of existing works for DF incidence/outbreak prediction are built upon linear regression models, e.g., SARIMA [11], Poisson regression [12], [26], GAMs [7], [8], and GLMMs [12]. These models heavily rely on expert knowledge to analyse the lag relationships between input data and DF incidence to choose suitable regression windows as well as to design different model components. Thus, they lack flexibility and can hardly be used with other data and other areas without enormous redesign efforts, a non-trivial task. Hence, we hardly see comparisons among these models in the literature. Among these models, spatial-temporal Bayes approaches like [12], [13] are also based on the same idea of linking nearby provinces together to capture the spatial aspects of DF transmissions [16] like PT-Ensem. They all combine features from other provinces as additional components in their prediction models. This is fundamentally different to the data aggregation scheme via a proximity graph of PT-Ensem (Phase D) where we do not extend the prediction models with additional features but only combine data from other provinces to extend training sets. Moreover, PT-Ensem has two other *unique* extra schemata where the predicted results are propagated among provinces guided by a proximity graph and among different predicted time periods. To the best of our knowledge, none of the works has these propagation features. In [7], transfer learning is combined with LSTM for DF incidence prediction. The model is first trained on Guangzhou data. The pretrained model is then used as a starting point to train other cities. This approach somehow resembles the aggregation scheme of PT-Ensem. However, when facing provinces with very different climate patterns and value ranges like Vietnam (c.f. Figure 1), it can bring up negative effects as seen in Figure 4, especially when raw data is used. PT-Ensem has an underlying proximity graph to prevent excessively different data being used in the prediction models and an event encoding scheme to flatten the differences in value ranges. In [5], [34], [40], graphs that represent human movements are also constructed and are used as input features for prediction models. In contrast, PT-Ensem only uses its proximity graph to guide the aggregation and propagation

phases. In [10], various deep learning methods (LSTM-ATT, LSTM, CNN, and Transformer) are employed to predict DF incidence for 20 provinces in Vietnam using only climate data. Comparison among PT-Ensem and these methods are studied in Section IV where PT-Ensem significantly outperforms them in terms of prediction accuracy and runtimes. Superensem [13], a GLMM model, also predicts outbreaks for all provinces in Vietnam like PT-Ensem. Comparison between them is thus interesting. However, Superensem is heavily tailored using many specific data which is unfortunately out of our reach while adapting it to work with our data is a non-trivial task as mentioned above. Superensem and others [8], [24] also employ ensemble techniques like Phase C of PT-Ensem. However, there are only a few prediction models involved (only one in [12], [24] but with different parameters). They also limit to the use of majority voting scheme in most cases.

VI. Conclusion

In this paper, we introduce for the first time a novel method, called PT-Ensem, for predicting DF outbreaks for multiple areas at the same time. Compared to all existing works, PT-Ensem is *unique* in the way it compresses input data into sets of extreme climate events before using them to predict DF outbreaks and aggregate training data and propagate predicted results among similar provinces with guidance from a proximity graph. We employ PT-Ensem to predict the DF outbreaks from 1 to 6 month ahead for all provinces in Vietnam as a case study. Extensive experiments have been conducted using PT-Ensem and a wide range of 24 different competitors from traditional methods like Random Forest, Extra Tree to recent approaches like LightGMB, Catboost and deep learning techniques like LSTM-ATT, LSTM, CNN and Transformer. PT-Ensem significantly outperforms others in terms of both long-and short-term prediction accuracy. Detailed analysis shows that the performance improvement of PT-Ensem comes from all three major parts: aggregation, proximity propagation, and time propagation. In terms of runtime, PT-Ensem is orders of magnitude faster than deep learning approaches. Moreover, PT-Ensem is not overly reliant on expert knowledge. Thus, it can be easily applied to other areas and climate data. We also demonstrate how the forward/backward probabilistic functions of PT-Ensem can be used to study the effects of different climate factors on DF incidence in Vietnam and provide a different view to common correlation analysis approaches used in other works. Our future works aim at exploring other non-climate factors into PT-Ensem to improve the performance, using PT-Ensem to study DF outbreaks in other areas, and studying other transmissible diseases like flu or diarrhea.

Acknowledgment

We special thank Anh Le for helping us building the province graph and Hanh Tran for data collection and discussion on the project. We also special thank anonymous reviewers for valuable comments that help to significantly enhance the paper quality.

References

[1] S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh *et al.*, "The global distribution and burden of dengue," *Nature*, vol. 496, no. 7446, pp. 504–507, 2013.

[2] P. Siriyasatien, S. Chadsuthi, K. Jampachaisri, and K. Kesorn, "Dengue epidemics prediction: a survey of the state-of-the-art based on data science processes," *IEEE Access*, vol. 6, pp. 53 757–53 795, 2018.

[3] T. Thi Tuyet-Hanh, N. Nhat Cam, L. Thi Thanh Huong, T. Khanh Long, T. Mai Kien, D. Thi Kim Hanh, N. Huu Quyen, T. Nu Quy Linh, J. Rocklöv, M. Quam *et al.*, "Climate variability and dengue hemorrhagic fever in hanoi, viet nam, during 2008 to 2015," *Asia Pac J Public Health*, vol. 30, no. 6, pp. 532–541, 2018.

[4] L. Eisen, A. J. Monaghan, S. Lozano-Fuentes, D. F. Steinhoff, M. H. Hayden, and P. E. Bieringer, "The impact of temperature on the bionomics of aedes (stegomyia) aegypti, with special reference to the cool geographic range margins," *J Med Entomol*, vol. 51, no. 3, pp. 496–516, 2014.

[5] D. Salami, C. A. Sousa, M. d. R. O. Martins, and C. Capinha, "Predicting dengue importation into europe, using machine learning and model-agnostic methods," *Sci Rep*, vol. 10, no. 1, pp. 1–13, 2020.

[6] T. K. Martheswaran, H. Hamdi, A. Al-Barty, A. A. Zaid, and B. Das, "Prediction of dengue fever outbreaks using climate variability and markov chain monte carlo techniques in a stochastic susceptible-infected-removed model," *Sci Rep*, vol. 12, no. 1, pp. 1–17, 2022.

[7] J. Xu, K. Xu, Z. Li, F. Meng, T. Tu, L. Xu, and Q. Liu, "Forecast of dengue cases in 20 chinese cities based on the deep learning method," *Int J Environ Res Public Health*, vol. 17, no. 2, p. 453, 2020.

[8] P. Guo, T. Liu, Q. Zhang, L. Wang, J. Xiao, Q. Zhang, G. Luo, Z. Li, J. He, Y. Zhang *et al.*, "Developing a dengue forecast model using machine learning: A case study in china," *PLoS Negl Trop Dis*, vol. 11, no. 10, p. e0005973, 2017.

[9] K. Liu, M. Zhang, G. Xi, A. Deng, T. Song, Q. Li, M. Kang, and L. Yin, "Enhancing fine-grained intra-urban dengue forecasting by integrating spatial interactions of human movements between urban regions," *PLoS Negl Trop Dis*, vol. 14, no. 12, p. e0008924, 2020.

[10] V.-H. Nguyen, T. T. Tuyet-Hanh, J. Mulhall, H. V. Minh, T. Q. Duong, N. V. Chien, N. T. T. Nhung, V. H. Lan, H. B. Minh, D. Cuong, N. N. Bich, N. H. Quyen, T. N. Q. Linh, N. T. Tho, N. D. Nghia, L. V. Q. Anh, D. T. M. Phan, N. Q. V. Hung, and M. T. Son, "Deep learning models for forecasting dengue fever based on climate data in vietnam," *PLOS Neglected Tropical Diseases*, vol. 16, no. 6, pp. 1–22, 06 2022.

[11] W. Hu, A. Clements, G. Williams, and S. Tong, "Dengue fever and el nino/southern oscillation in queensland, australia: a time series predictive model," *Occup Environ Med*, vol. 67, no. 5, pp. 307–311, 2010.

[12] F. J. Colón-González, I. R. Lake, and G. Bentham, "Climate variability and dengue fever in warm and humid mexico," *Am J Trop Med Hyg*, vol. 84, no. 5, 2011.

[13] F. J. Colón-González, L. Soares Bastos, B. Hofmann, A. Hopkin, Q. Harpham, T. Crocker, R. Amato, I. Ferrario, F. Moschini, S. James *et al.*, "Probabilistic seasonal dengue forecasting in vietnam: A modelling study using superensembles," *PLoS Med*, vol. 18, no. 3, p. e1003542, 2021.

[14] A. Aswi, S. Cramb, P. Moraga, and K. Mengersen, "Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review," *Epidemiol Infect*, vol. 147, 2019.

[15] C. Wang, B. Jiang, J. Fan, F. Wang, and Q. Liu, "A study of the dengue epidemic and meteorological factors in guangzhou, china, by using a zero-inflated poisson regression model," *Asia Pac J Public Health*, vol. 26, no. 1, pp. 48–57, 2014.

[16] S. T. Stoddard, B. M. Forshey, A. C. Morrison, V. A. Paz-Soldan, G. M. Vazquez-Prokopec, H. Astete, R. C. Reiner, S. Vilcarromero, J. P. Elder, E. S. Halsey *et al.*, "House-to-house human movement drives dengue virus transmission," *PNAS*, vol. 110, no. 3, pp. 994–999, 2013.

[17] R. Barrera, M. Amador, and A. J. MacKay, "Population dynamics of aedes aegypti and dengue as influenced by weather and human behavior in san juan, puerto rico," *PLoS Negl Trop Dis*, vol. 5, no. 12, p. e1378, 2011.

[18] R. Jain, S. Sontisirikit, S. Iamsirithaworn, and H. Prendinger, "Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data," *BMC Infect Dis*, vol. 19, no. 1, pp. 1–16, 2019.

[19] M. Kikuti, G. M. Cunha, I. A. Paploski, A. M. Kasper, M. M. Silva, A. S. Tavares, J. S. Cruz, T. L. Queiroz, M. S. Rodrigues, P. M. Santana

et al., "Spatial distribution of dengue in a brazilian urban slum setting: role of socioeconomic gradient in disease risk," *PLoS Negl Trop Dis*, vol. 9, no. 7, p. e0003937, 2015.

[20] CDC, "Dengue hemorrhagic fever–us-mexico border, 2005," *Morb Mortal Wkly Rep*, vol. 56, no. 31, pp. 785–789, 2007.

[21] Z. Liu, Z. Zhang, Z. Lai, T. Zhou, Z. Jia, J. Gu, K. Wu, and X.-G. Chen, "Temperature increase enhances aedes albopictus competence to transmit dengue virus," *Front Microbiol*, vol. 8, p. 2337, 2017.

[22] M. A. Johansson, D. A. Cummings, and G. E. Glass, "Multiyear climate variability and dengue?el nino southern oscillation, weather, and dengue incidence in puerto rico, mexico, and thailand: a longitudinal data analysis," *PLoS Med*, vol. 6, no. 11, p. e1000168, 2009.

[23] T. T. T. Do, P. Martens, N. H. Luu, P. Wright, and M. Choisy, "Climatic-driven seasonality of emerging dengue fever in hanoi, vietnam," *BMC Public Health*, vol. 14, no. 1, pp. 1–10, 2014.

[24] S. F. McGough, L. Clemente, J. N. Kutz, and M. Santillana, "A dynamic, ensemble learning approach to forecast dengue fever epidemic years in brazil using weather and population susceptibility cycles," *J R Soc Interface*, vol. 18, no. 179, p. 20201006, 2021.

[25] N. A. M. Salim, Y. B. Wah, C. Reeves, M. Smith, W. F. W. Yaacob, R. N. Mudin, R. Dapari, N. N. F. F. Sapri, and U. Haque, "Prediction of dengue outbreak in selangor malaysia using machine learning techniques," *Sci Rep*, vol. 11, no. 1, pp. 1–9, 2021.

[26] P. Siriyasatien, A. Phumee, P. Ongruk, K. Jampachaisri, and K. Kesorn, "Analysis of significant factors for dengue fever incidence prediction," *BMC Bioinform*, vol. 17, no. 1, pp. 1–9, 2016.

[27] B. Bett, D. Grace, H. S. Lee, J. Lindahl, H. Nguyen-Viet, P.-D. Phuc, N. H. Quyen, T. A. Tu, T. D. Phu, D. Q. Tan *et al.*, "Spatiotemporal analysis of historical records (2001–2012) on dengue fever in vietnam and development of a statistical model for forecasting risk," *PLoS One*, vol. 14, no. 11, p. e0224353, 2019.

[28] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001.

[29] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach Learn*, vol. 63, no. 1, pp. 3–42, 2006.

[30] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *NIPS*, vol. 30, 2017.

[31] A. V. Dorogush, V. Ershov, and A. Gulin, "Catboost: gradient boosting with categorical features support," *arXiv preprint arXiv:1810.11363*, 2018.

[32] A. L. Buczak, B. Baugher, L. J. Moniz, T. Bagley, S. M. Babin, and E. Guven, "Ensemble method for dengue prediction," *PLoS One*, vol. 13, no. 1, p. e0189988, 2018.

[33] M. A. Johansson, F. Dominici, and G. E. Glass, "Local and global effects of climate on dengue transmission in puerto rico," *PLoS Negl Trop Dis*, vol. 3, no. 2.

[34] R. Bomfim, S. Pei, J. Shaman, T. Yamana, H. A. Makse, J. S. Andrade Jr, A. S. Lima Neto, and V. Furtado, "Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas," *J R Soc Interface*, vol. 17, no. 171, p. 20200691, 2020.

[35] O. J. Brady, D. L. Smith, T. W. Scott, and S. I. Hay, "Dengue disease outbreak definitions are implicitly variable," *Epidemics*, vol. 11, pp. 92–102, 2015.

[36] N. Zhao, K. Charland, M. Carabali, E. O. Nsoesie, M. Maheu-Giroux, E. Rees, M. Yuan, C. Garcia Balaguera, G. Jaramillo Ramirez, and K. Zinszer, "Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in colombia," *PLoS Negl Trop Dis*, vol. 14, no. 9, p. e0008056, 2020.

[37] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *ICPR*, 2010, pp. 3121–3124.

[38] T. W. Scott, P. H. Amerasinghe, A. C. Morrison, L. H. Lorenz, G. G. Clark, D. Strickman, P. Kittayapong, and J. D. Edman, "Longitudinal studies of aedes aegypti (diptera: Culicidae) in thailand and puerto rico: blood feeding frequency," *J Med Entomol*, vol. 37, no. 1, pp. 89–101, 2000.

[39] H. V. Pham, H. T. Doan, T. T. Phan, and N. N. T. Minh, "Ecological factors associated with dengue fever in a central highlands province, vietnam," *BMC Infect Dis*, vol. 11, no. 1, pp. 1–6, 2011.

[40] J. Liu-Helmersson, H. Stenlund, A. Wilder-Smith, and J. Rocklöv, "Vectorial capacity of aedes aegypti: effects of temperature and implications for global dengue epidemic potential," *PLoS One*, vol. 9, no. 3, p. e89783, 2014.