Climate-based ensemble machine learning model to forecast dengue epidemics

Rochelle Schneider[1,2,3*], **Alessandro Sebastianelli**[1,4], **Dario Spiller**[1,5], **James Wheeler**[1,6], **Raquel Carmo**[1], **Artur Nowakowski**[1,7], **Manuel Garcia Herranz**[8], **Do-Hyung Kim**[8], **Hanoch Barlevi**[9], **Zoraya El Raiss Cordero**[9], **Silvia Liberata Ullo**[4], **Pierre-Philippe Mathieu**[1], **Rachel Lowe**[3,10,11]

1 European Space Agency, Φ-Lab, Frascati, 00044, Italy
2 European Centre for Medium-Range Weather Forecast, Forecast Department, Reading, RG2 9AX, United Kingdom
3 London School of Hygiene & Tropical Medicine, The Centre on Climate Change and Planetary Health, London, WC1H9SH, United Kingdom
4 University of Sannio, Engineering Department, Benevento, 82100, Italy
5 Italian Space Agency, Engineering and Future Space Systems Unit, Via del politecnico snc, Rome 00133, Italy
6 Geology and the Environment, University of Leicester, School of Geography, Leicester, LE1 7RH, United Kingdom
7 Warsaw University of Technology, Faculty of Geodesy and Cartography, Plac Politechniki 1, 00-661 Warsaw, Poland
8 UNICEF NYHQ, Office of Innovations, New York, 10017, United States
9 UNICEF Latin America and the Caribbean Office (LACRO), Climate, Environment, Energy & Disaster Resilience Section, Panama, Republic of Panama
10 London School of Hygiene & Tropical Medicine, Centre for Mathematical Modelling of Infectious Diseases, London, WC1H9SH, United Kingdom
11 London School of Hygiene & Tropical Medicine, Department of Infectious Disease Epidemiology, London, WC1H9SH, United Kingdom

ABSTRACT

Dengue fever is one of the most common and rapidly spreading arboviral diseases in the world, with major public health and economic consequences in tropical and sub-tropical regions. Countries such as Peru, 17.143 cases of dengue were reported in 2019, where 81.4% of cases concentrated in five of the 25 departments. When predicting infectious disease outbreaks, it is crucial to model the long-term dependency in time series data. However, this is challenging when performed on a countrywide level since dengue incidence varies across administrative areas. Therefore, this study developed and applied a climate-based ensemble model using multiple machine learning (ML) approaches to forecast dengue incidence rate (DIR) by department. The ensemble combined the outputs from Long Short-Term Memory (LSTM) recurrent neural network and Categorical Boosting (CatBoost) methods to predict DIR one month ahead for each department in Peru. Monthly dengue cases stratified by Peruvian departments were analysed in conjunction with associated demographic, geographic, and satellite-based meteorological data for the period January 2010–December 2019. The results demonstrated that the ensemble model was able to forecast DIR in low-transmission departments, while the model was less able to detect sudden DIR peaks in some departments. Air temperature and wind components demonstrated to be the significant predictors for DIR predictions. This dengue forecast model is timely and can help local governments to implement effective control measures and mitigate the effects of the disease. This study advances the state-of-the-art of climate services for the public health sector, by informing what are the key climate factors responsible for triggering dengue transmission. Finally, this project summarises how important it is to perform collaborative work with complementary expertise from intergovernmental organizations and public health universities to advance knowledge and address societal challenges.

1. Introduction

Dengue is a mosquito-borne viral infection mostly found in urban and peri-urban areas located in warm and tropical climate regions. The World Health Organization (WHO) reported that the global burden of dengue increased eight-fold over the last two decades[1]. This increase has been partly attributed to globalisation, climate change and urbanisation, but can also be explained by the improvement in local policies to record and report dengue cases to national (Ministry of Health) and international (WHO) related agencies[1]. However, most dengue infections are asymptomatic cases or misdiagnosed as other febrile illnesses, resulting in under-reporting[2]. Peru is a Latin American country particularly affected by dengue and according to the epidemiological alerts issued by the Peruvian Ministry of Health, all four dengue serotypes (DENV 1-4) are in circulation[3,4]. In 2019, 17.143 cases of dengue (37 deaths) were reported in Peru, representing a 243% increase compared with 2018[5]. The dengue incidence rate (DIR) by age group was reported as 34.18% for children and adolescents (0-17 years old), 24.88% for adults between 18 and 29-year-olds, 34.60% for 30 to 59-year-olds, and 6.68% adults over 60-years-old (6.68%)[5]. Since dengue is a climate-sensitive disease, temperature and rainfall variations influence the magnitude and seasonality of dengue transmission[6,7]. To date, there is no specific antiviral treatment for dengue, or a national mass dengue testing campaign to limit dengue transmission[8]. Therefore, determining the association between weather patterns and the surge of dengue cases is an important policy measure for an early response to future outbreaks.

In recent years, the adoption of machine learning (ML) techniques to perform Earth Observation (EO) classification and regression tasks[9,10] have substantially increased thanks to their ability to model any kind of predictor(s)-response association and to appropriately capture complex spatio-temporal relationships[11]. Several studies in the literature implemented different regression methods to forecast dengue cases across the globe[12–16], and a few used a single machine learning (ML) architecture[15–17] but none explored an ensemble model of multiple-ML approaches. The importance of exploring an ensemble multiple-ML models is because the dengue incidence rate (DIR) varies widely across different national administrative areas. Therefore, it becomes a difficult task for a single-ML model to fully capture the countrywide behaviour. For example, Figure A1 (in the Appendix) displays the DIR over 25 Peruvian departments. There is high DIR heterogeneity across the country, demonstrating departments with a typically low (e.g., Tacna and Moquegua), seasonal (e.g. Lambayeque and La Libertad), or year-round dengue incidence (e.g. Madre de Dios and Loreto).

The novelty of this work is threefold: (i) assess the benefits of new environmental data from satellite and satellite-based products to forecast DIR, (ii) introduce an innovative methodological ensemble approach to predict DIR one month ahead for each Peruvian department, and (iii) bring the most recent evidence on the climate factors that impact the dengue transmission in Peru, since this study found the last related publication covering up to the 2010 period[4]. The ensemble ML forecast model was designed, developed, and applied by the European Space Agency's scientific team according to the needs described by UNICEF Office of Innovations and UNICEF Latin America and Caribbean Regional Office. The outcome of this project is twofold: (i) provide a climate-based ML model to forecast DIR in Peru and (ii) develop a reproducible workflow, presented as a *cookbook* (*i.e.,* all processes are explained step-by-step, from data collection to model validation), where UNICEF in support of national governments can construct, test, validate and implement a working forecast model for other dengue-endemic countries.

2. Data and Methods

2.1 Study area and dengue data

Peru is a highly urbanized country situated on the central western coast of South America facing the Pacific Ocean, with a land size of about 1.28 million km^2. As of 2020, the country accommodates a population of around 32.6 million with about 34.8%, 54.3%, and 10.9% inhabitants between the age groups 0-19, 20-59, and 60-80+, respectively[18]. Peru has a diverse geography, comprising a western coastal plain, the Andes mountains, and the eastern tropical Amazon rainforest. The climate conditions range from moderate temperatures and low precipitation along the coast, heavy rainfall and high temperatures in the Amazon, and temperate up to cool temperatures in the Andes mountain range.

This study used the largest Peruvian administrative level (i.e. departments) to group monthly dengue cases from 2010 to 2019 provided by the Department of Epidemiology of Peru's Ministry of Health[19]. The number of cases by department was converted into dengue incidence rate (DIR) per 100.000 population. Population data by department for the same period was obtained from the National Institute of Statistics and Informatics of Peru (INEI)[18].

2.2 Satellite and satellite-based products

Three land products were obtained from EO satellites. Firstly, the monthly Normalised Difference Vegetation Index (NDVI) was used to indicate vegetation cover and health[20]. The NDVI (MOD09GA Version 6[21]) was calculated from the Moderate Imaging Spectroradiometer (MODIS) sensor onboard the Terra satellite. Global Forest Change[22] (version 1.8) was the second product explored, retrieved from Landsat 7 and Landsat 8 satellites, representing the percentage of annual forest loss. Finally, a digital elevation product obtained from the Shuttle Radar Topography Mission[23] (version 4) was used to obtain the mean altitude. Seven satellite-based meteorological variables were selected from the ERA 5 global reanalysis dataset provided by Copernicus Climate Change Service (C3S)[24]. Daily ERA 5 products collected for this study were: 2m height air temperature, dewpoint temperature, total precipitation, sea level pressure, surface pressure, 10m u- and v- components of wind. The grid cells from all variables were grouped by department through a geometric intersection tool. Monthly averages per department for each variable were computed from January 2010 to December 2019. The minimum and maximum values for 2m height air temperature were also computed, as well as latitude and longitude from each department's centroid to be used as a spatial proxy. All data was obtained from the Google Earth Engine[25] and the data was manipulated using Python[26].

2.3 Ensemble machine learning approach

An ensemble ML approach was developed to forecast DIR one month ahead for each Peruvian department. The two supervised ML regression models involved are: (i) a deep learning model, called LSTM (Long Short-Term Memory)[27] and (ii) a tree-based ML design, called CatBoost (Categorical Boosting)[28]. LSTMs are well known for their ability to process the input data as a sequence of values with a short- and long-term memory of past inputs, while boosting models like CatBoost are well-known for their capacity to deal with learning problems based on heterogeneous features, noisy data, and complex dependencies. Figure 1 displays the ensemble framework, where each ML architecture runs in parallel using the same list of predictors (input X) and

the outcome variable (referred here as ground-truth *y*). The LSTM-based model is composed of three LSTM layers, with decreasing size of parameters and three fully connected layers. Each LSTM layer is responsible for the extraction of temporal features from the input data, while the fully connected part is responsible for the forecast. The CatBoost model is based on the gradient boosting approach, which is essentially a process of constructing an ensemble predictor as a combination of weaker models (base predictors) by performing gradient descent in a functional space. The proposed ensemble model of multiple-ML is based on the late data fusion paradigm, where the output of two models is averaged to calculate the final prediction. The LSTM and CatBoost were trained using *X* and *y* variables from all departments combined. The department ID was included as a predictor, enabling the models to provide a unique DIR forecast for each department.

The full dataset was divided into two sub-datasets: 2010-2017 training/testing set, referred to here as (X_{train}, y_{train}) and 2018-2019 validation set, referred as (X_{val}, y_{val}). Random X_{train} samples were used as feedforward input during the training. The internal weights of both models were adjusted based on mean absolute error (MAE) loss and its gradient, used to compare the model output (*i.e.*, forecasted DIR) with y_{train}(*i.e.* observed DIR). This study implemented a sliding window method, known as a backtesting strategy, where the train-test pair was composed of seven-one months. Therefore, the training/testing sets are now fully represented by $(X_{train}, y_{train}) \in R^{(M,N,P)}, R^{(M)}$ and the validation set, by $(X_{val}, y_{val}) \in R^{(K,N,P)}, R^{(K)}$; where *N* is the number of months (*i.e.* seven) in the *train* of train-test pair, *P* represents the number of predictors, *M*, and *K* are described in Equation 1:

$$M \text{ or } K = number\ of\ departments\ x\ (number\ of\ months\ [in\ the\ time\ interval] - N) \qquad \text{Equation 1}$$

where, *M* depends on the number of departments, the size in months of the training/testing (2010-2017) set, and *N*. *K* depends on the number of departments, the size in months of the validation (2018-2019) set, and *N*. Once trained, the LSTM and CatBoost algorithms were applied in the validation set, and a measure of performance was generated by computing the root mean square error (RMSE). To be able to compare the RMSE values between departments, a min-max normalization was applied at the department level. In this way, each variable of each department had a maximized dynamics ranging from 0 to 1. The normalization parameters have been saved to restore later the original range.

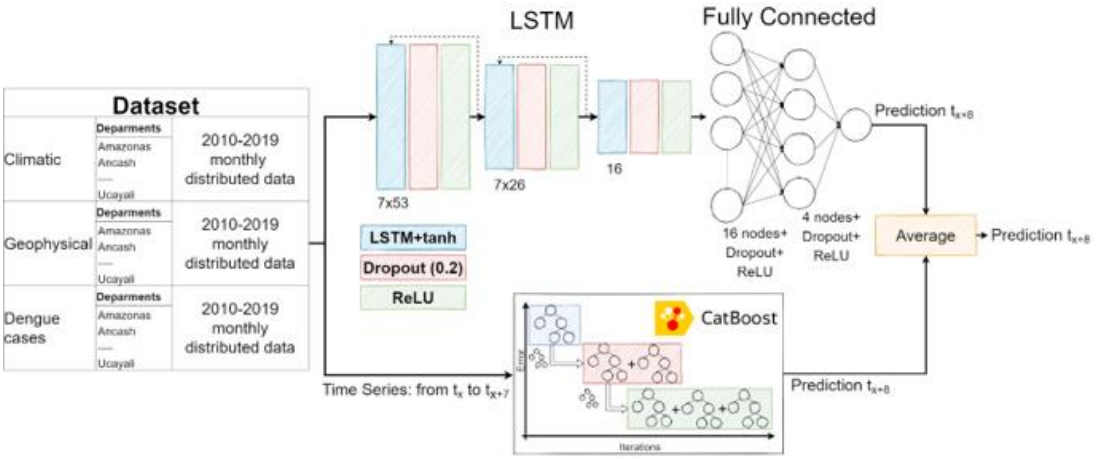

Figure 1. LSTM and CatBoost models framework.

3 Results and discussion

The dataset included 281.264 dengue cases reported between 2010 and 2019 in 25 departments in Peru (Figure A1, in the Appendix). The national DIR reported in 2019 was 52.66 cases per 100.000 population, which was 90 times greater in Madre de Dios department and 8, 5, and 2 times greater in Tumbes, Loreto, and Ucayali respectively. Madre de Dios is the third-largest department in Peru but the least populated, located almost entirely in the low-lying Amazon rainforest. Excluding Madre de Dios, all departments within the top five highest DIR were found in the north, where the minimum T_{mean} does not go below 17°C (Figure A2). This study explored the most relevant environmental factors reported in the literature[12,29,30] that influence dengue transmission. Comparing Figure A1 with Figure A3 showed that high altitude is a clear environmental factor responsible for lowering the temperature and consequently reducing DIR in departments located along the Andes mountain range. On the contrary, departments located in the Amazon rainforest (*i.e.*, Loreto, Amazonas, San Martin, Ucayali, and Madre de Dios) had a high dengue persistence year-round with punctuated dengue outbreaks during the rainy season (November-April). Therefore, this study grouped the departments by three transmission types: (i) low (*e.g.*, Lima, Apurimac, and Tacna), (ii) seasonal (*e.g.*, Lambayeque, La Libertad, and Ancash), and (iii) year-round (*e.g.*, Loreto, San Martin, and Madre de Dios). Chowell et al.[4] also grouped Peruvian provinces based on a similar stratification, which they named as jungle, coastal, and mountain regions. They aimed to investigate the relationship between dengue incidence and climate factors during 1994-2008 and their results showed

that the jungle region (in this study named as a year-round group) is responsible for multiple dengue introductions into coastal areas (*i.e.,* seasonal group) since their favourable environmental conditions promote constant mosquito breeding sites.

During the training of the CatBoost model, the importance of each predictor was assessed. This process identified what were the most influential variables to predict DIR for each train-test pair. As expected for a forecast model, the most important variable was DIR from the previous months. Following the DIR, CatBoost model ranked maximum, minimum, and mean air temperature as well as the wind components (speed and direction) as the most important variables. Air temperature is a well-know highly ranked variable but wind component has been recently reported by the literature[31]. The high variation in DIR across departments motivated this study to propose an ensemble-based forecast model since a single-learner would not be able to explain this complex heterogeneity in spatial-temporal dynamics of dengue transmission. Table 1 displays the results of LSTM, CatBoost, and the ensemble models for three departments that represent each group (full table in the appendix, Table A1). Some departments (e.g., Lambayeque) demonstrated a coherence between LSTM and CatBoost forecast performance; however, the benefit in applying a hybrid approach is also to control overfitting. For example, for Lima, by taking the average of both models' results to compute the final forecasted DIR, the ensemble penalised any potential single-model overfit. Note, the ensemble did not self-adjust to provide good performance only to a specific transmission type. Higher RMSE results found in some departments (e.g. Tumbes) might be attributable to additional characteristics of dengue transmission that were not accounted in the models, entomological surveillance data (geographical distribution and density of the vector population)[32] and socio-economic data (e.g. proportion of dwellings with electric lighting, running water, and hygienic services)[2,30]. Figure A4 displays three plots showing monthly DIR predictions during 2018-2019 generated by the three models (LSTM, CatBoost, and ensemble) and *y* (ground truth) for three department types (low, seasonal, and endemic). In departments with low DIR, the models were able to forecast well without overestimating the dengue incidence. Regarding seasonal and endemic departments, the plots showed a good ability to generalize and forecast in many situations.

This study also faced some limitations that must be acknowledged, for example, the very high peaks in specific months for Madre de Dios' DIR were not well captured by the models since these unique outbreak values were not recorded by any other department. This phenomenon was expected since the models did not have sufficient samples to learn this type of dynamics and therefore the prediction turned out to be lower. Therefore, future directions for this project are pointing to the exploration of weekly data to mitigate high peaks of DIR to improve the modelling performance.

Table 1. LSTM, CatBoost, and ensemble models' performance for three departments. The RMSE metric describes the models' error, expressed by the normalised DIR to allow comparison between departments.

Dengue transmission type	Department	LSTM RMSE	CatBoost RMSE	Ensemble RMSE
Low transmission	Lima	0.243	0.357	0.298
Seasonal transmission	Lambayeque	0.244	0.241	0.242
Year-round transmission	Madre de Dios	0.218	0.208	0.212

4 Conclusion

Dengue control and prevention is a challenging task and must be performed via multiple levels of coordinated response (i.e., global and local). This study explored artificial intelligence technologies together with the latest advances from EO satellites and satellite-based products to forecast one-month DIR across 25 departments in Peru, chosen as a pilot country for the analysis. This study had a truly public health purpose since the project was a cooperation between ESA and UNICEF with the mission to provide knowledge and awareness about the impact of climate factors on dengue transmission that could be basis for dengue prevention and mitigation policies. The study identified that departments located at higher altitudes kept the weather less favourable to dengue transmission, while proximity to the Amazon rainforest has a significant influence to keep some departments in an endemic scenario. The findings also showed that there is a strong heterogeneity between departments in Peru; therefore, an ensemble-ML approach allows us to capture the transmission dynamics of dengue across three groups: low, seasonal, and year-round transmissions. The ensemble model demonstrated good accuracy across all groups, but some departments still need information beyond environmental factors, such as entomological surveillance data and socio-economic data to better describe the spatio-temporal DIR patterns. The increase of temperature due to climate change will make more departments prone to dengue outbreaks[33] which might increase the number of endemic departments responsible for importing cases to the other regions. Therefore, policy measures should strengthen the government response actions to adapt and control the increasing risk of recurrent outbreaks in new areas.

Finally, UNICEF's interest is to generate evidence on the correlation between climate change and dengue with the intention of influencing the improvement of health services and systems and children's health levels. Therefore, the next steps of this project is to subset the methodological approach to focus on children (*i.e.,* 0-19 years old) since they are more vulnerable to suffer from severe dengue cases[34,35].

Acknowledgements

The authors would like to thank ESA for providing financial support and technical resources to develop this project. We are grateful to Noelle Cremer, former trainee at the ESA Φ-Lab, for collaborating during the initial weeks of the project. We also acknowledge useful discussions and feedback of this work with other members of UNICEF: Desiree Raquel Narvaez (UNICEF NYHQ, Health Section - Programme Division), Karina Cantizano (UNICEF LACRO – Survive & Thrive Section), and Carlos Calderón Bonilla (UNICEF Emergency Officer Peru Country Office). This work was generated using Copernicus Climate Change Service (C3S) information [2010-2019]. The authors would like to thank the European Centre for Medium-Range Weather Forecasts (ECMWF) that implements the C3S on behalf of the European Union.

References

1. Dengue and severe dengue. https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue.
2. Buczak, A. L., Koshute, P. T., Babin, S. M., Feighner, B. H. & Lewis, S. H. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. *BMC Med Inform Decis Mak* **12**, 124 (2012).
3. Alertas epidemiológicas y documentos normativos – CDC MINSA. https://www.dge.gob.pe/portalnuevo/vigilancia-epidemiologica/subsistema-de-vigilancia/dengue/alertas-epidemiologicas-y-documentos-normativos-dengue/.
4. Chowell, G. *et al.* Spatial and temporal dynamics of dengue fever in Peru: 1994–2006. *Epidemiol. Infect.* **136**, 1667–1677 (2008).
5. Epidemiological Update: Dengue - 7 February 2020 - PAHO/WHO | Pan American Health Organization. https://www.paho.org/en/documents/epidemiological-update-dengue-7-february-2020.
6. Lowe, R. *et al.* Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study. *The Lancet Planetary Health* **5**, e209–e219 (2021).
7. Butterworth, M. K., Morin, C. W. & Comrie, A. C. An Analysis of the Potential Impact of Climate Change on Dengue Transmission in the Southeastern United States. *Environmental Health Perspectives* **125**, 579–585 (2017).
8. Bhatt, S. *et al.* The global distribution and burden of dengue. *Nature* **496**, 504–507 (2013).
9. Sebastianelli, A. *et al.* AIRSENSE-TO-ACT: A Concept Paper for COVID-19 Countermeasures Based on Artificial Intelligence Algorithms and Multi-Source Data Processing. *IJGI* **10**, 34 (2021).
10. Del Rosso, M. P., Sebastianelli, A. & Ullo, S. L. *Artificial Intelligence Applied to Satellite-based Remote Sensing Data for Earth Observation*. (The Institution of Engineering and Technology, forthcoming).
11. Schneider, R. *et al.* A Satellite-Based Spatio-Temporal Machine Learning Model to Reconstruct Daily PM2.5 Concentrations across Great Britain. *Remote Sensing* **12**, 3803 (2020).
12. Lowe, R. *et al.* Climate services for health: predicting the evolution of the 2016 dengue season in Machala, Ecuador. *The Lancet Planetary Health* **1**, e142–e151 (2017).
13. Colón-González, F. J. *et al.* Probabilistic seasonal dengue forecasting in Vietnam: A modelling study using superensembles. *PLoS Med* **18**, e1003542 (2021).
14. Lowe, R. *et al.* The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in Southeast Brazil. *Statist. Med.* **32**, 864–883 (2013).
15. Carvajal, T. M. *et al.* Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. *BMC Infect Dis* **18**, 183 (2018).
16. Polwiang, S. The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003-2017). *BMC Infect Dis* **20**, 208 (2020).
17. Ong, J. *et al.* Mapping dengue risk in Singapore using Random Forest. *PLoS Negl Trop Dis* **12**, e0006587 (2018).
18. PERÚ Instituto Nacional de Estadística e Informática. https://www.inei.gob.pe/estadisticas/censos/.
19. Reporte de Tabla de casos notificados por causas. 2010-2020. | Centro Nacional de Epidemiología, Prevención y Control de Enfermedades. CDC - Perú. https://www.dge.gob.pe/salasituacional/sala/index/1_TablaCasosSE/82.
20. Deering, D. W. Rangeland reflectance characteristics measured by aircraft and spacecraft sensors. (Texas A&M University. Libraries, 1978).
21. Vermote, Eric & Wolfe, Robert. MOD09GA MODIS/Terra Surface Reflectance Daily L2G Global 1kmand 500m SIN Grid V006. (2015) doi:10.5067/MODIS/MOD09GA.006.
22. Hansen, M. C. *et al.* High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* **342**, 850–853 (2013).
23. Jarvis, A., Guevara, E., Reuter, H. I. & Nelson, A. D. Hole-filled SRTM for the globe : version 4 : data grid. (2008).
24. Hersbach, H. *et al.* The ERA5 global reanalysis. *Q.J.R. Meteorol. Soc.* **146**, 1999–2049 (2020).
25. Gorelick, N. *et al.* Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* **202**, 18–27 (2017).
26. Welcome to Python.org. *Python.org* https://www.python.org/.
27. Sherstinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network.

Physica D: Nonlinear Phenomena **404**, 132306 (2020).
28. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. *arXiv:1706.09516 [cs]* (2019).
29. Xu, J. *et al.* Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method. *IJERPH* **17**, 453 (2020).
30. Parselia, E. *et al.* Satellite Earth Observation Data in Epidemiological Modeling of Malaria, Dengue and West Nile Virus: A Scoping Review. *Remote Sensing* **11**, 1862 (2019).
31. Chumpu, R., Khamsemanan, N. & Nattee, C. The association between dengue incidences and provincial-level weather variables in Thailand from 2001 to 2014. *PLoS ONE* **14**, e0226945 (2019).
32. WHO | Vector surveillance and control at ports, airports, and ground crossings. *WHO* http://www.who.int/ihr/publications/9789241549592/en/.
33. Lowe, R. *et al.* Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study. *PLoS Med* **15**, e1002613 (2018).
34. Alvarado-Castro, V. M. *et al.* Clinical profile of dengue and predictive severity variables among children at a secondary care hospital of Chilpancingo, Guerrero, Mexico: case series. *Boletín Médico Del Hospital Infantil de México (English Edition)* **73**, 237–242 (2016).
35. Hernández-Suárez, C. M. & Mendoza-Cano, O. Empirical evidence of the effect of school gathering on the dynamics of dengue epidemics. *Global Health Action* **9**, 28026 (2016).

Appendix

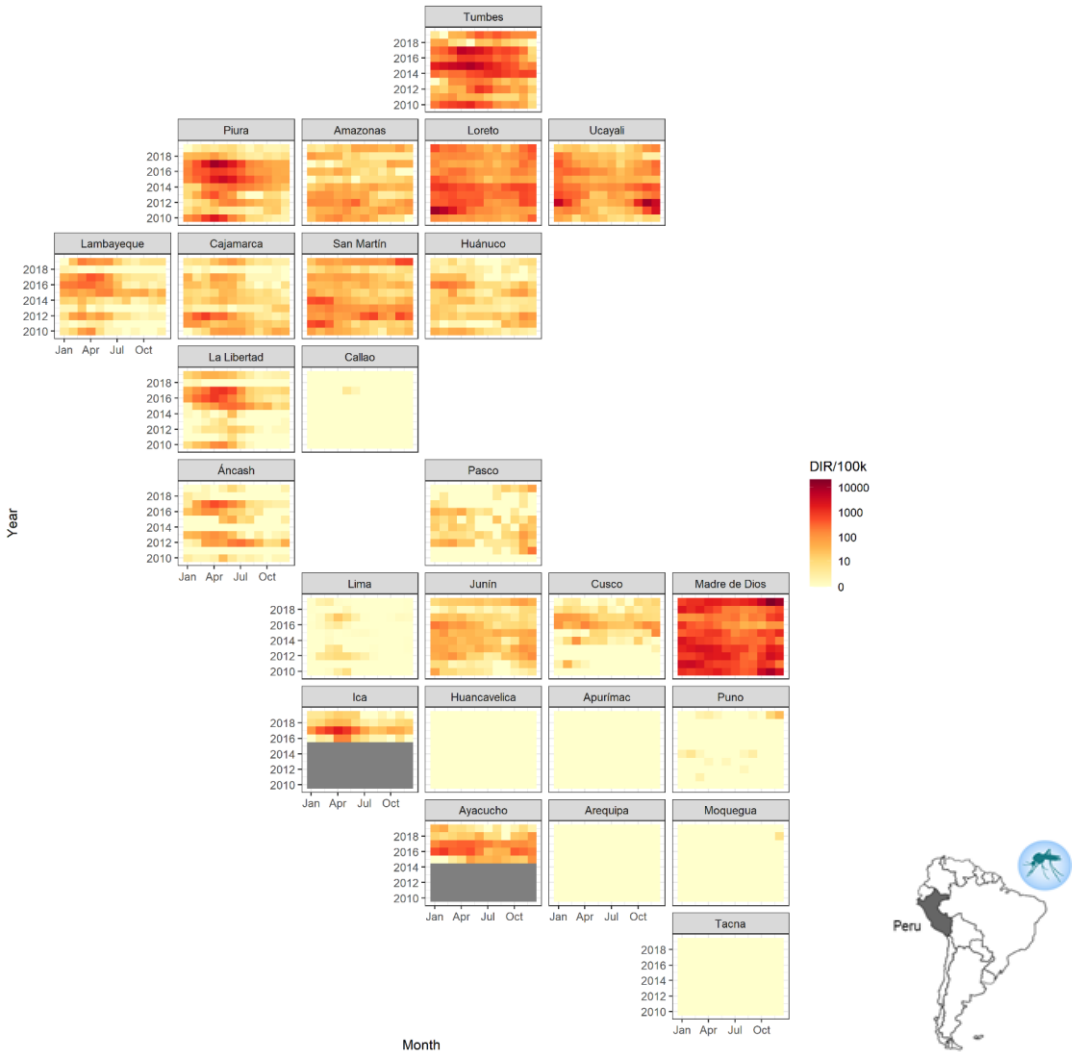

Figure A1. Spatial and temporal variation in dengue incidence rate (DIR) (per 100.000 people) in Peru by department. Dengue records for Ayacucho and Ica departments were only available from 2014 and 2015, respectively.

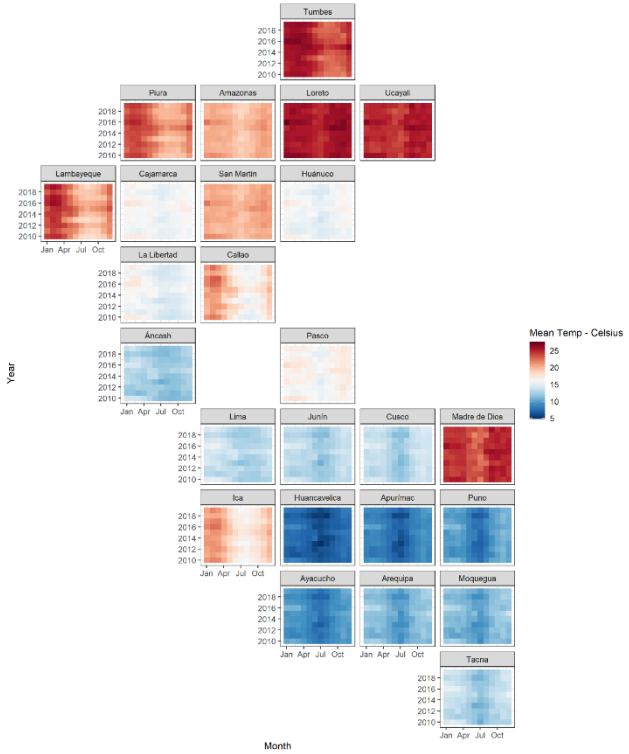

Figure A2. Monthly mean 2m height air temperature (2010-2019) by department.

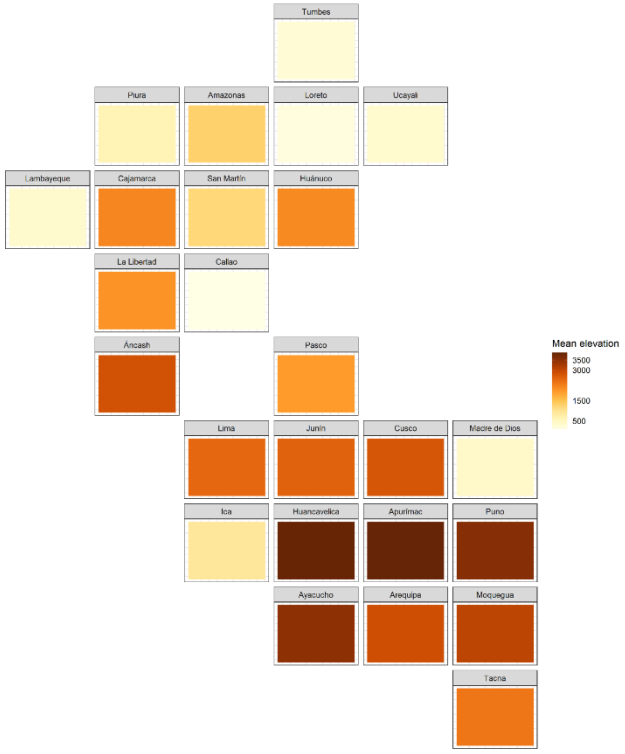

Figure A3. Mean altitude by department.

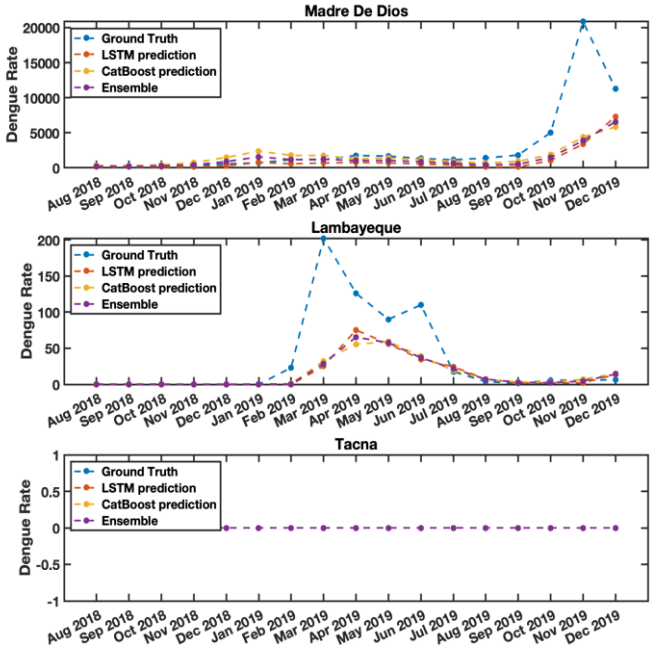

Figure A4. LSTM, CatBoost, and the ensemble models predictions for a year-round (top), seasonal (middle), and a low transmission (bottom) department type.

Table A1. LSTM, CatBoost, and ensemble models' performance for all departments. The RMSE metric describes the models' error, expressed by the normalised DIR to allow comparison between departments.

Department	LSTM RMSE	CatBoost RMSE	Ensemble RMSE
Amazonas	0.264	0.274	0.269
Ancash	0.243	0.234	0.239
Apurimac	0.033	0.006	0.018
Arequipa	0.014	0.004	0.008
Ayacucho	0.256	0.219	0.238
Cajamarca	0.219	0.238	0.229
Callao	0.028	0.024	0.020
Cusco	0.299	0.322	0.307
Huancavelica	0.009	0.005	0.006
Huanuco	0.297	0.304	0.299
Ica	0.144	0.098	0.116
Junin	0.214	0.229	0.219
La Libertad	0.274	0.286	0.299
Lambayeque	0.244	0.241	0.242
Lima	0.243	0.357	0.298
Loreto	0.299	0.291	0.294
Madre De Dios	0.218	0.208	0.212
Moquegua	0.249	0.246	0.247
Pasco	0.194	0.207	0.200
Piura	0.066	0.051	0.051
Puno	0.206	0.214	0.210
San Martin	0.249	0.259	0.253
Tacna	0.002	0.004	0.002
Tumbes	0.333	0.359	0.344
Ucayali	0.158	0.134	0.144
Overall	0.190	0.193	0.190