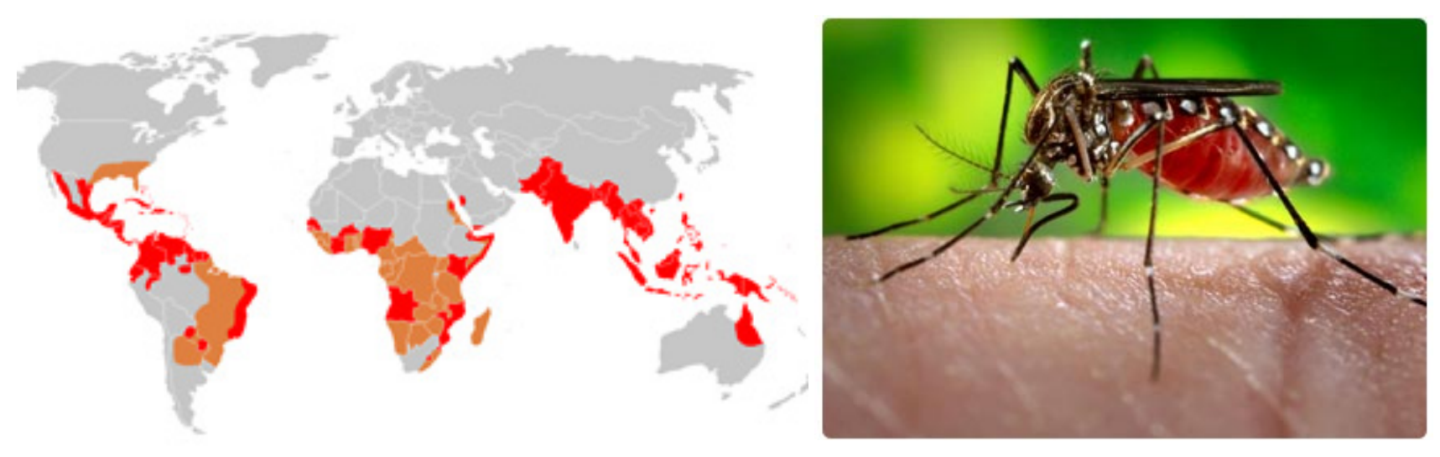
An Application of Information Theory in Predictive Modeling of Dengue Fever

Yang Liu (liux3204@umn.edu)[1], Matteo Convertino, Ph.D.[1, 2, 3, 4]
[1]Environmental Health Sciences Division, School of Public Health, University of Minnesota TC;
[2]Institute on the Environment, University of Minnesota TC;
[3]Institute for Engineering in Medicine, University of Minnesota TC; [4]Public Health Informatics Program, University of Minnesota TC

Background

Dengue Fever (DF) is a mosquito-borne infectious disease. The overall global burden of disease is 22,000 deaths and 390 million cases annually. The transmission of DF primarily relies on the female *Aedes aegypti*.


Source: Nature

Due to the dependencies between the natural environment and vector dynamics, a body of existing literature has examined the possibility of using environmental variables, such as temperature and precipitation, to predict DF incidence. Regardless of the modeling method, many of these studies share two features:

1. They examine one or two environmental variables and select their statistical features (e.g. maximum or mean) based on little empirical evidence.
2. They consider lag effects of the environment but lack a method to derive the optimal lag structure to predict DF incidence patterns.

Objective: This study tests multiple statistical features of four environmental variables: temperature, precipitation, humidity and Normalized Difference Vegetation Index (NDVI) to find the variables that are most relevant to predict DF incidence. Then the method aims to detect the lag window for each environmental variable that best relates to DF incidence in concert with all other variables. Such lag windows are applied to a generalized linear model (GLM) designed for prediction purpose.

Method

In information theory, the entropy of a random variable is:

$$H(X) = \sum_{x \in X} p(X) \log p(X)$$

where *p(X)* is the probability density function. *H(X)* is also referred to as the entropy of X. Extending from this concept, mutual information (MI) is frequently used to quantify the dependencies between two variables.

$$I(X,Y) = \sum_{x} \sum_{y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Since MI has no assumption on the linearity or the continuity of the random variables, it is considered a better measurement of association than correlation.

The GLM used for DF prediction is defined as the following:

$$Y \sim \ NegBin(\mu, \theta)$$

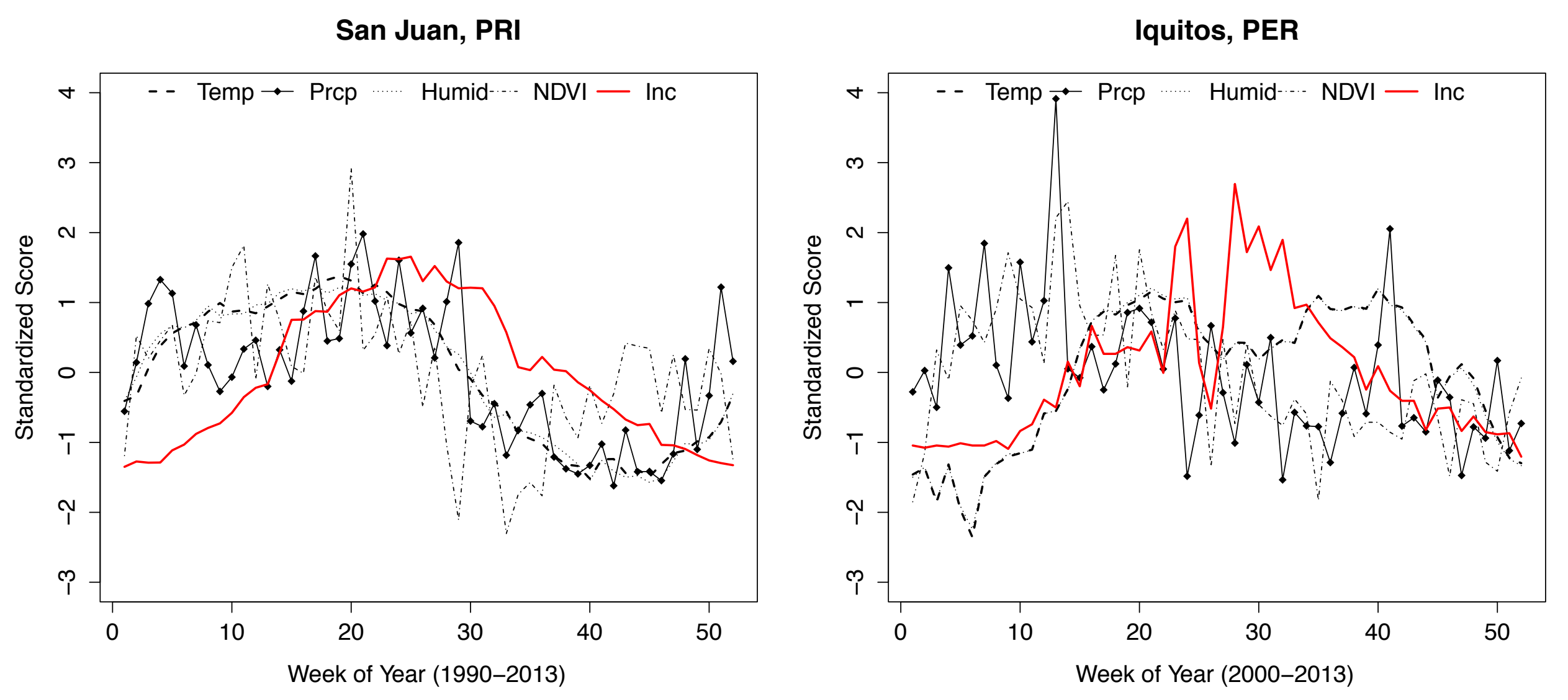
$$\log(\mu) = \beta_0 + \sum_{v=1}^{4} \sum_{l=n_i}^{m_i} ns(x_{vl}; \beta_{vl}) + \sum_{v=5}^{7} ns(x_v; \beta_v) + \log(pop)$$

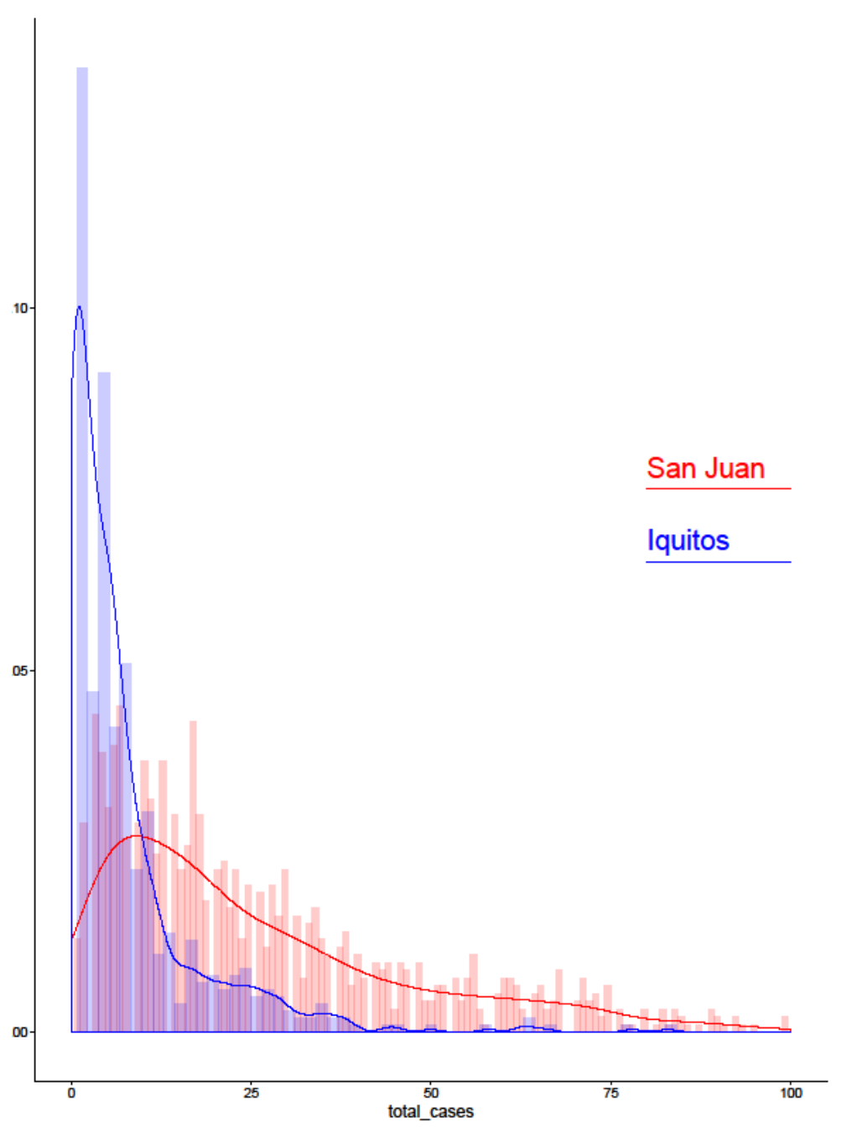
μ, θ: features of the distribution;
n, m: lower and upper boundary of the lag window l;
v: independent variables;
pop: population.

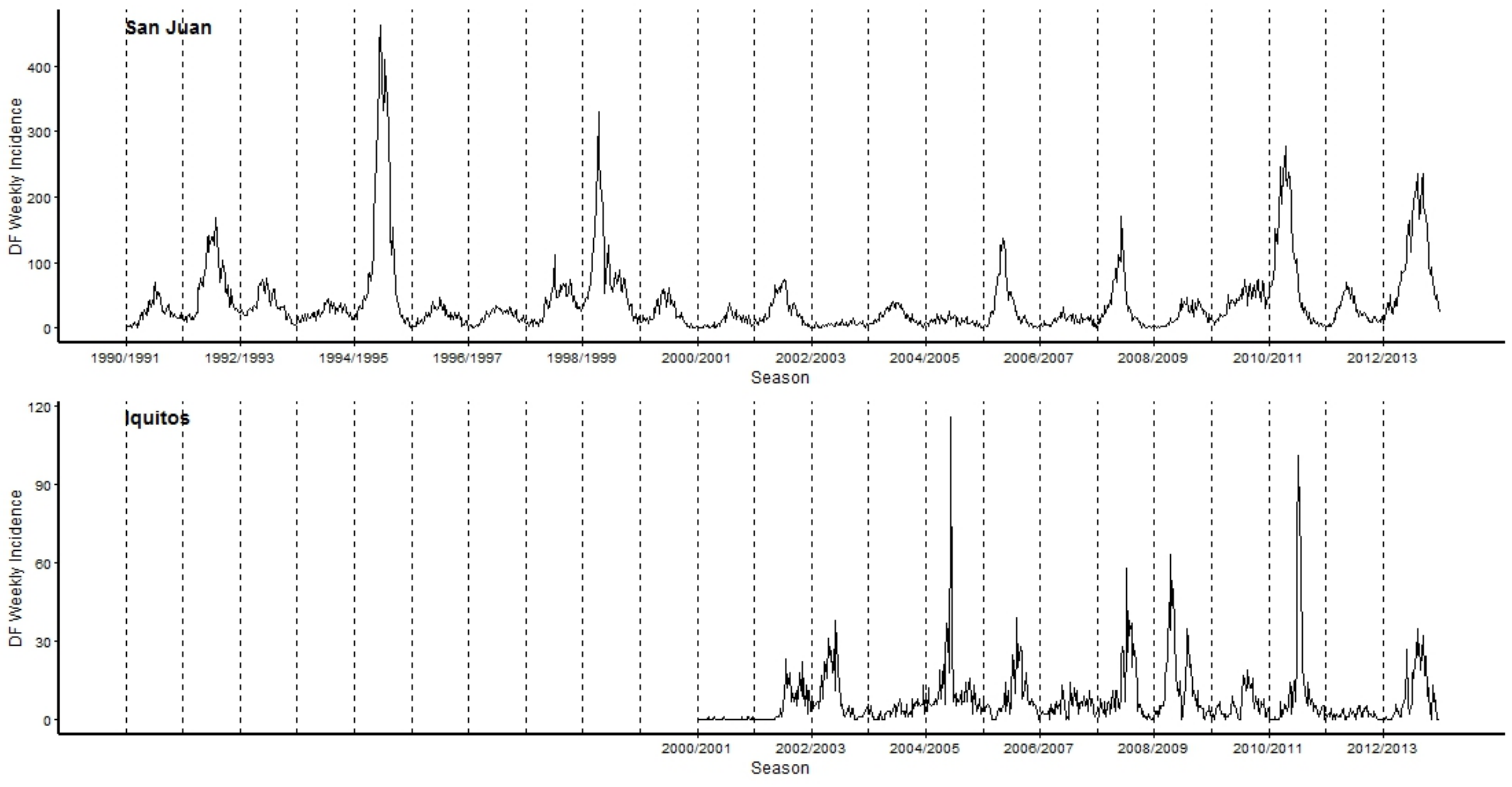
Data

- The study sites are San Juan, the coastal capital city in northern Puerto Rico, and Iquitos, an inland city surrounded by the tributaries of the Amazon River in northern Peru.

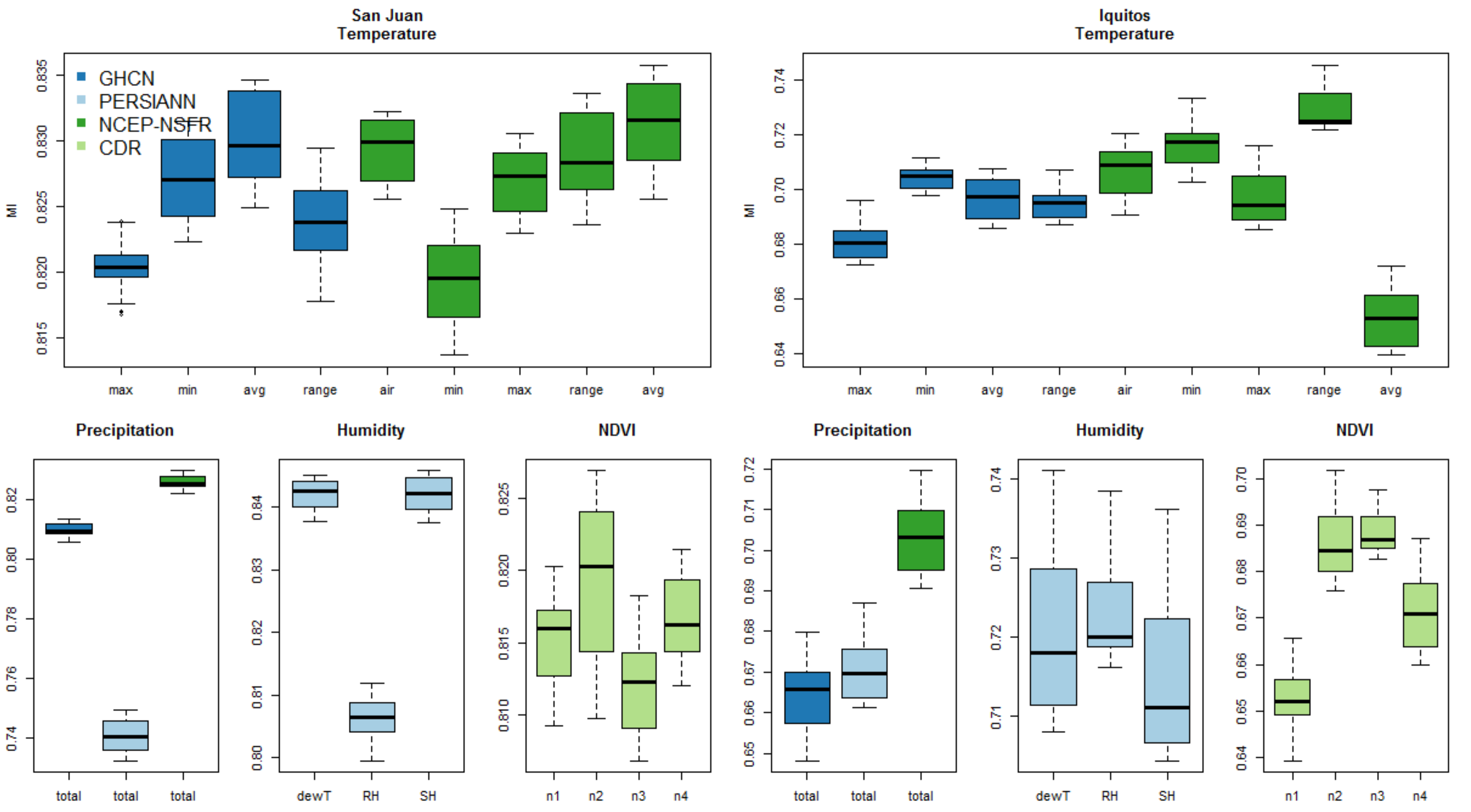
- DF shows strong seasonality at both locations with clear epidemic and non-epidemic periods every year. It has always been present in San Juan. In Iquitos, there had not been cases until 2000.
- The climate of the two sites and their synchrony with the DF season are quite different. DF season starts in May in San Juan and in July in Iquitos. Rainfall is concentrated in a few months in San Juan and scattered across the year in Iquitos.
- The probability distribution of DF weekly incidence in San Juan is relatively flatter than in Iquitos. Iquitos has more days with a small number of weekly incidence. San Juan has more days with a larger number of weekly incidence. These are power-law distributions with exponential decay for small events.



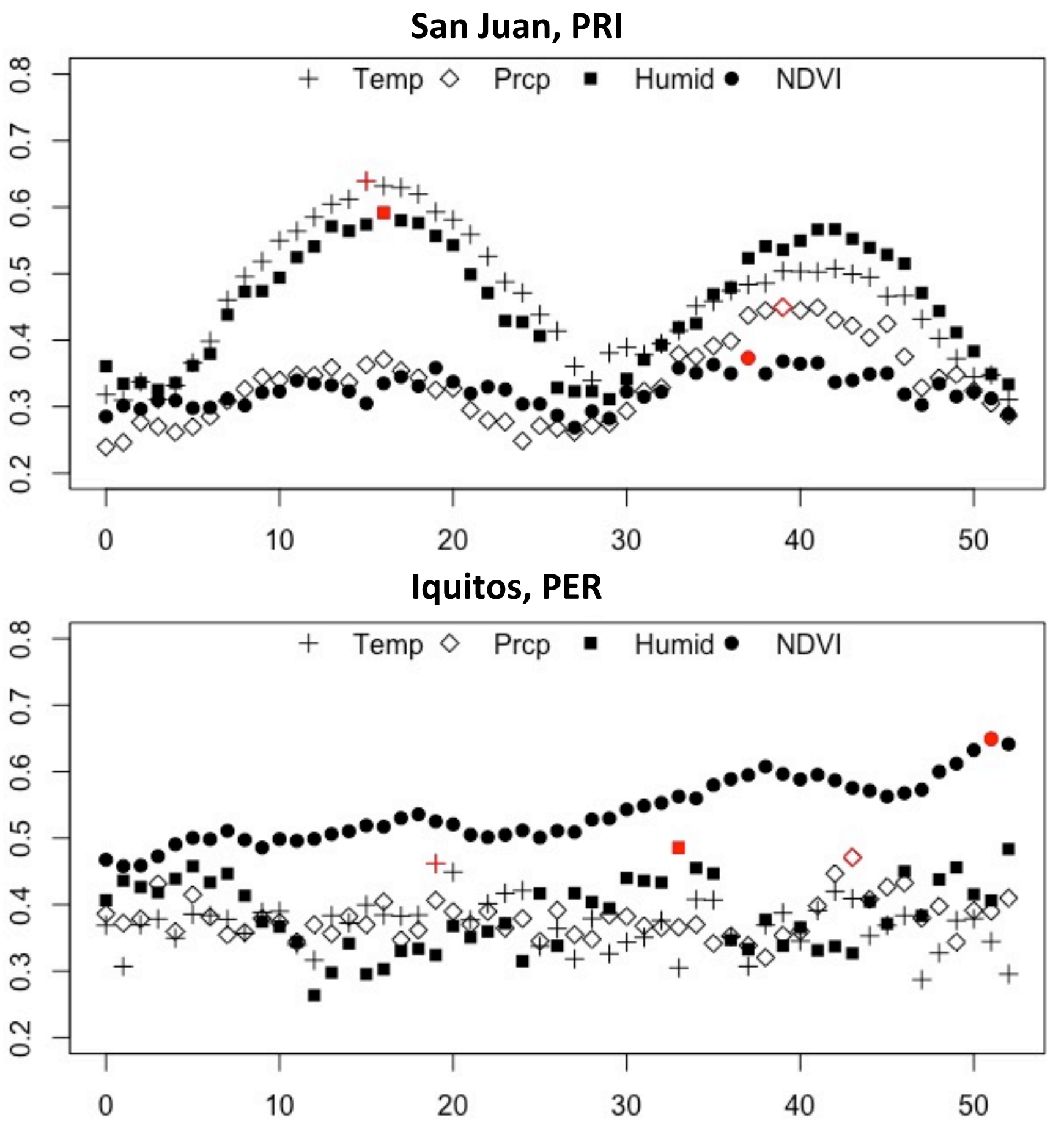


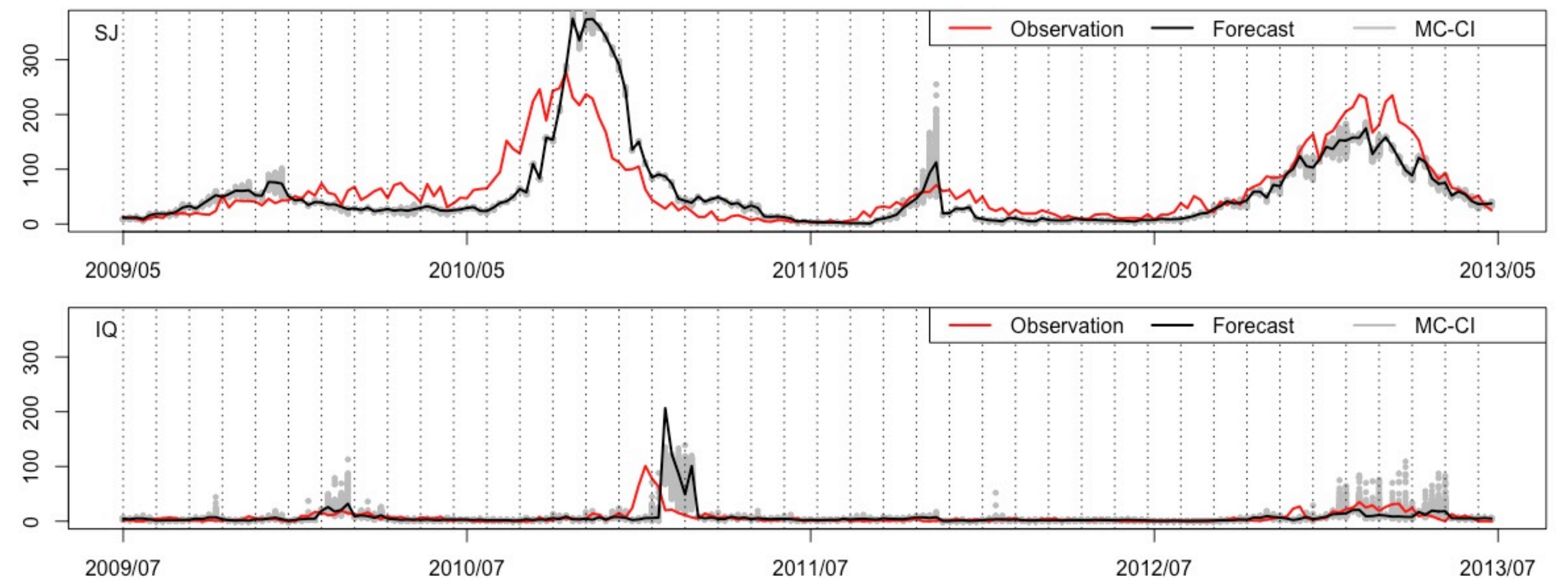


Results



- 4 environmental variables and a total of 19 environmental statistical features were evaluated for their association with DF weekly incidence. We selected one feature with the highest MI for every environmental variable as a potential candidate of the GLM. Other variables included were seasonality, long-term trend and an auto-regressive component.
- A 4-8 week long window of importance was selected for each environmental variables. Multiple model selection criteria were used.
- Variance inflation factor (VIF) (threshold = 4) was used to measure severity of multi-collinearity. As a result, humidity related variables were dropped for both the San Juan and Iquitos.





Conclusions

- This study has shown the potential of the application of information theory in predictive model design; in particular for the variable selection and the lag-structure formation phase.
- Results from the variable selection phase show that different statistical features measuring the same environmental variables will lead to different magnitudes of associations with DF incidence. In fact, the same features of the same variables derived from different methods may lead to different results. This study used four data sources including station data, satellite data and re-analysis data. In public health modeling, the choice of environmental information can be critical and should be handled with caution.
- The lag-structure is based on the best predictive power. The lag may suggest a physical connection between the variables and outcome. Lag-structure is not consistent for the two study sites.
- Through the application of mutual information, we show that temperature and rainfall are the most important environmental variables affecting DF. This is consistent with the current understanding of disease dynamics.
- For DF, the final forecasts were generated 5 weeks ahead of time for four consecutive years after the training period. The results were evaluated using statistical measurements such as the Mean Absolute Error (MAE), as well as seasonal features such as season onset and epidemic peak size.
- In order to assess the accuracy of this model compared with other approaches, we established a baseline model using a seasonal autoregressive integrated moving average (SARIMA) model framework. The accuracy improvement is most significant for season peak size and peak timing. This model is 25-200% more likely to generate accurate predictions. Accuracies are particularly high with 1-2 months lead time.

End Notes

References:

Hii, Y. L., Rocklov, J., Wall, S., Ng, L. C., Tang, C. S., & Ng, N. (2012). Optimal Lead Time for Dengue Forecast. PLoS Neglected Tropical Diseases, 6(10).
Hii, Y. L., Zhu, H., Ng, N., Ng, L. C., & Rocklov, J. (2012). Forecast of Dengue Incidence Using Temperature and Rainfall. PLoS Neglected Tropical Diseases, 6(11).
Liu Y., Convertino M. The Application of Information Theory in Designing Dengue Fever Prediction Model. (Work In Progress.)

Acknowledgement:
Y.L. thanks the support from the Discovery Grant (PI: Convertino). Y.L. and M.C. thank the support from the U.S. Centers for Disease Control and prevention (CDC) for providing the public health data and the National Oceanic and Atmospheric Administration for providing environmental data (GHCN, PERSIANN, NCEP-CFSR and NDVI CDR). They jointly hosted the Dengue Fever Forecast Exercise, from which this study evolved from.

Software:
All analysis in this study is performed using R (3.2.4), packages used include 'infotheo', 'forecast', 'glmulti', 'usdm', 'ggplot2', 'ggmap'.