Open Source Indicators for Modeling and Forecasting Water, Climate, and Environmental Stressors

Project Report

Naren Ramakrishnan

Executive Summary

Historically, water and climate shocks have been monitored and managed as isolated local hazards but the increasing interconnectivity and interdependency of global society suggests that there are significant gains to be had by developing a collective comprehensive model of how such hazards affect food, economic, and social systems, and thus influence human behavior and response. For example, during the 2007-2008 and 2010 global food crises, drastic responses such as food export control, increased subsidies, and hoarding were seen, which led to higher food price volatility and exacerbated existing tensions. The aim of this pilot project assesses whether publicly available data – including social media, news, as well as physical indicators - may be able to provide early indications and identify patterns of how water and climate shocks may contribute to disease outbreaks, civil unrest, and socio-political instability.

The focus of this pilot project is on local civil unrest activity in Latin American countries and on disease spread more globally, and quantifying the influence of water, climate, and environmental stressors on these classes of events. We use state-of-the-art algorithms from data mining, social media analytics, and multi-source forecasting to descriptively characterize and forecast (where possible) significant societal events and investigate whether climatic precursors can aid in better situational awareness and/or forecasting accuracy. For the most part, all analysis conducted here is done in languages native to the country (exceptions noted). All data analysis and processing was done on the commercial Amazon cloud using a streaming pipes-and-filters architecture.

For disease forecasting, we apply a spatio-temporal topic modeling algorithm over news articles coupled with physical indicators (temperature, weather, humidity) to forecast disease surveillance counts. Our approach allows the ablation of data sources, i.e., understanding the relative loss in performance by selecting removing data sources. This supports the investigation of incremental improvement obtained by incorporating climatic attributes in modeling. We focus on multiple countries (China, India, Singapore, United States, as well as the larger Americas) and multiple diseases (H7N9, Whooping Cough, Rabies, Salmonella, Dengue, Hand Foot and Mouth Disease, Malaria, and ILI). We demonstrate that for key (disease, country) combinations there are significant gains to be had by incorporating climatic attributes into the model. In particular, we observe significant improvement in forecasting for many (disease, country) combinations, such as (H7N9, China), and (Malaria, India), which exhibited a 30-40% increase after incorporating climate attributes. For other (disease, country) combinations such as (Dengue, India), (Dengue, Singapore) and (Salmonella, USA), we find no improvement in forecasting accuracy. This could be due to the inconsistent coverage of news in media as well as missing values in surveillance data. Improved data collection practices can lead to more uniform improvements.

For civil unrest modeling, we use a range of data analytics approaches to descriptively characterize the precursors to civil unrest. These approaches include text classification, storytelling, and dynamic query expansion, and are applied to a corpus of events, news articles, and/or tweets as appropriate. Text classification enables us to categorize news stories and documents as being climate-related or not. The storytelling algorithm identifies situations where precursors of civil unrest can be identified quite early in open source media, e.g., constructions or projects that negatively influence (or expected to influence) climatic considerations. Finally, the dynamic query expansion algorithm identifies an active citizenry that uses social media to voice concerns about ongoing developments. From a database of 25,352 civil unrest events across Latin America from January 2011 to March 2015, our approach identifies 974 events as climate-related events, which is equal to 3.84%. Although this overall % is small, there are specific countries and more importantly, specific (country, month) combinations where the % of climate-related protests exceeds 20%! Typical causes for civil unrest include droughts, floods, and heat waves. Typical rationale includes disruption of life and the government's inability to introduce countermeasures or responses in a timely manner.

There are three primary lessons from our study. First, our results indicate that open source indicators can provide a sufficiently location-specific early warning system for climatic events and precursors to disease outbreaks and civil unrest events. Although outside the scope of our study here, the analysis presented here can be readily prototyped into a continuous, online, cloud-based system that tracks activity in regions of interest and delivers analysis and/or forecasts of events as they develop. Such a system will become imperative in the near future for improved situational understanding and forecasting.

A second take-away from the study here is that massive data sources bring about the interconnectedness of societal events at a scale not possible before. Extending the methods studied here, we should be able to identify cascading and "network" effects among significant societal events. A policy maker should be able to use the resulting multimodal representation to pose "what-if", "why", and "why not" questions over scenarios, to understand relationships between specific strategies and outcomes. Most studies currently are retrospective in nature and being able to (conditionally) forecast would constitute a quantum leap in our capability to anticipate and mitigate severe climatic scenarios.

Finally, this study demonstrates that new research at the intersection of computer science, data analytics, environmental science, and policy planning is a blossoming area that must be fostered to make further inroads into this space. Data scientists are necessary to process the massive volumes of data being gathered on a human and societal scale but might lack awareness of the relevant scientific questions to pose over such data. Domain experts such as climate scientists and policy planners possess such awareness but are not necessarily experts in translating them to data analysis tasks. A synergistic collaboration will lead to relevant and practical research in mitigating water, climate, and environmental stressors.

Part 1: Modeling Climatic Stressors to Forecast Disease Outbreaks

Related Research

Prior studies [1-6] have indicated that sudden anomalies or deviation in climatic conditions (such as temperature and precipitation) are one of the major driving forces behind emergence and spread of infectious diseases, in different regions of the world. In general, climate change can be regarded as a prominent threat for public health and society. Past research studying the effects of climate change on disease outbreaks can be broadly classified into studies investigating waterborne diseases [7-9], foodborne diseases [10, 11], zoonotic diseases [12, 13] and emerging diseases [14].

In the area of waterborne diseases, researchers have specifically analyzed the effects of increased rainfall [7, 8] on outbreaks and the vulnerability of waterborne disease outbreaks caused by *E. coli O157:H7* and *Cryptosporidium* [9] to change in climatic factors such as excess precipitation, floods, high temperatures and drought. In [7], a Monte Carlo version of the Fisher exact test was used to analyze the relationship between precipitation and waterborne disease outbreaks in United States. It was found that 51% of waterborne disease outbreaks were preceded by precipitation events above the 90th percentile and 68% of outbreaks preceded by events above the 80th percentile.

For foodborne diseases such as Salmonellosis, researchers have specifically focused on examining the adverse effects of increased temperature [10, 11] on disease outbreaks. Akil et al. [10] found a strong correlation between rising temperatures and Salmonellosis outbreaks in three states (Mississippi, Tennessee and Alabama) of USA from 2002 to 2011. They used data mining approaches such as regression analysis and neural networks to understand the effects of temperature on Salmonella Infections. In [11], authors fitted log-linear models to examine the correlation between temperature and Salmonellosis notifications for 5 Australian cities over the period 1991 to 2001. They observed a positive association between monthly Salmonellosis notifications and average temperature of previous month.

In the area of zoonotic diseases such as Rabies, the focus of prior research has been mainly on investigating the role of increased temperature [12, 13] in outbreaks. Warmer temperatures allow infected host animals to survive winter season in large numbers, thus increasing the opportunity of transmission of zoonotic infection to humans [13]. Kim et al. [12] used a Poisson regression model to investigate Rabies incidence patterns with respect to climate factors and human exposure rate to Rabies in Alaska, USA. Recently Alexander et al. [14] investigated the role of climate change in triggering

outbreaks of an emerging disease, namely Ebola in different parts of West Africa during December 2013.

Methods

Our basic methodology to study disease outbreaks is a spatio-temporal topic model over the HealthMap [15, 16] (www.healthmap.org) corpus that can be coupled with climatic attributes. We assume that the input corpus of news articles is distributed over a fixed discretized time window. This input can therefore be transformed into a collection of tuples of the form **(word, location, timepoint): count**, where the count is defined as the total number a specific word was mentioned in all articles associated with the location and timepoint in the tuple. For example, a tuple ('dengue', ('India', 'Punjab'), 92): 17 means that the word "dengue" was mentioned 17 times in all the articles referring to the state of Punjab in India over the timepoint index 92 (here, timepoint index is simply an index into the database, and does not denote the actual time; in our studies counts are recorded either in terms of days or weeks, depending on the study/algorithm – details below) .

Following a similar approach to Jagarlamudi et al [17], we implemented a supervised spatiotemporal topic model [18] for topic and pattern discovery. We supervise the discovery process of each disease topic by providing a set of prior words (also called seed words). These seed words are user provided prior knowledge of each infectious disease and they encourage the topic model to find evidence of related disease topics in the HealthMap corpus. This supervised method helps in improving the discovery of word occurrences within each topic as the model tends to generate words that are related to the words in the seed set. Additionally, we model time and location jointly with the word occurrence patterns. This enables tracking of temporal and spatial patterns of such diseases in the news.

The second component of our method deals with forecasting disease outbreaks at a particular timepoint using news coverage (from the topic model) and climate attributes at different time-lags as features. Here overall news coverage for a particular disease in a country is defined by the topic model output which captures the prominence of a specific disease topic in the country at a given timepoint.

For geographically diverse countries such as China, India, and the United States disease outbreaks selectively manifest in certain locations and thus we need to incorporate climate changes over only those locations for targeted and robust forecasting. Typically, a disease outbreak has onset in a particular set of locations and then spreads to adjacent locations through human or animal contact. Thus changes in climatic factors over the onset locations have more influence in triggering the initial phase of the outbreak. At the same time, we also expect the local news media coverage to be higher at the onset locations than the adjacent locations where the disease outbreak activity is relatively low or zero. In other words, we postulate that the intensity of disease outbreak in a location is proportional to the extent of news coverage about that outbreak in that location.

To account for inconsistent time-lags of climate attributes for forecasting disease outbreaks across geographical regions, we implemented an ensemble approach where a single base regressor is trained using news coverage and values of a climate attribute at different time-lags as features. To generate the final forecast, we either fuse the forecasts resulting from all lags using a weighted average technique where the weights are proportional to the accuracies of the lags or we assign the final forecast to the forecast value of the lag having maximum accuracy.

Separate from infectious diseases, climatic attributes such as absolute humidity also play a crucial role in the propagation and spread of influenza-like-illness or ILI. We also study the problem ILI forecasting using a combination of sources [19, 20], including news (Healthmap), Twitter, weather, Google Search Trends, Google Flu Trends, and surveillance data (e.g., from the Pan American Health Organization or PAHO). See Figure 1. We used a matrix factorization based regression approach [19] using nearest neighbor embedding (MFN) to generate real-time ILI forecasts for 15 Latin American countries.

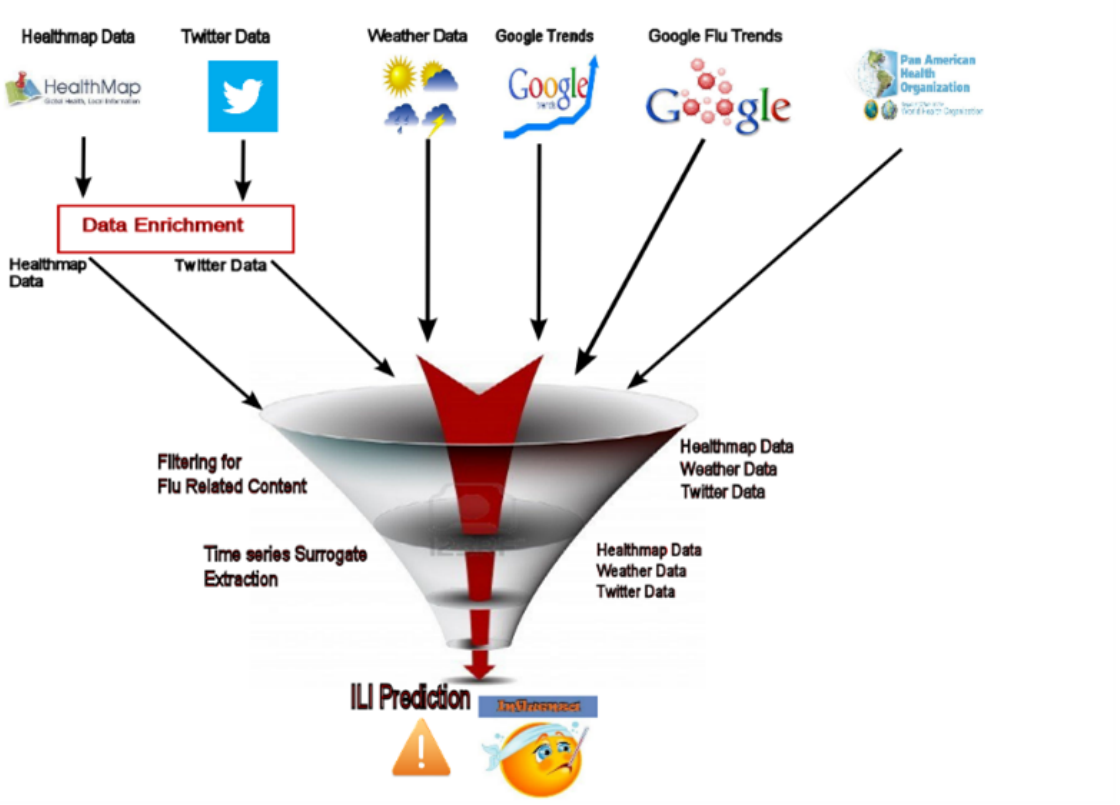


Figure 1: Streaming pipeline for forecasting ILI from Open Source Indicators

Regions and Periods of Study

Table 1. Disease-country combinations studied in this report.

Countries	Diseases	Time period of study	Evaluation period	Timepoint duration	Range of climatic attribute time-lags
China	1. H7N9 2. Dengue (vector-borne) 3. Hand, Foot and Mouth Disease (HFM)	Jan 2013 - Dec. 2014	Dec. 2013 - Dec. 2014	1 Month	0-5 months
United States	1. Whooping Cough 2. Rabies 3. Salmonella (Foodborne)	Jan 2011-Dec. 2013	Jan 2012-Dec 2013	1 epi week or CDC week (Sunday to Saturday)	0-11 epi weeks
Singapore	1. Dengue (vector-borne) 2. HFM	Jan 2013-Dec 2014	Jan 2014-Dec 2014	1 epi week or CDC week (Sunday to Saturday)	0-11 epi weeks
India	1. Foodborne Diseases 2. Dengue (vector-borne) 3. Malaria (vector-borne)	Jan 2013-Dec 2014	Jan 2014-Dec 2014	1 week (Monday to Sunday)	0-11 weeks
Latin American countries	1. Influenza-like-illness (ILI)	Jan 2010-August 2013	Jan 2013-August 2013	1 epi week or CDC week (Sunday to Saturday)	1-6 epi weeks

In Table 1, we depict the different (country, disease) combinations studied here. The choice of diseases in each country is influenced by the evidence of effects of climate change on these disease outbreaks in previous research studies [10, 12, 21, 22]. Along with the countries and diseases, we also show the time period of study, evaluation period and duration of a timepoint for each country. For each country, the evaluation period is chosen based on the timeline of major disease outbreaks in that country, e.g. for China, the evaluation timeline is chosen as Dec. 2013 to Dec. 2014 during which massive H7N9 (January 2014) and Dengue (October 2014) outbreaks have taken place. The timepoint duration is

chosen based on the timescale of the surveillance data available for that country, e.g. for China, surveillance data is available at the monthly level, so the timepoint duration is 1 month for this purpose. For India, United States and Singapore, surveillance data is available weekly, so each timepoint in our method represents one week. As a result, we can evaluate whether our method is able to forecast these specific outbreaks more accurately than a baseline approach. In Table 1, we also depict the range of climate attribute time-lags over which our ensemble technique operates to forecast disease outbreaks in that country.

Datasets

HealthMap: News articles corresponding to all diseases for each study country are provided by HealthMap, an online global disease alert system capturing local disease-related outbreak news articles from over 50000 electronic sources. Articles harvested mainly consisted of English words. However, for China, a good percentage of the articles are in the Chinese language. We translated them into English for ease of modeling. For each country, our first step is to filter those articles mentioning the names or related keywords of the study diseases. Then we enrich each filtered article using BASIS technologies' Rosette Language Processing (RLP) tools [23, 24] and then applied standard pre-processing techniques, such as stop-word removal and word-normalization. The words extracted from these filtered articles were found to contain general- (e.g. 'mosquito', 'contagious', 'nausea', 'foodborne', 'waterborne') as well as specific- (e.g. 'h7n9', 'dengue', 'malaria', 'pertussis') disease related words. Finally, we reverse-geocode (lat, long) coordinate of each filtered article to a location in the form of (country, state) pair, e.g. (8.384076, 77.011826): ("India", "Kerala"). In Table 2, we show the total number of articles downloaded from HealthMap for each country, total number of extracted words (after Basis Enrichment, stop-word removal and word-normalization) and total number of locations reverse-geocoded from the filtered HealthMap articles relevant to the study diseases in each country.

Table 2. Healthmap datasets utilized in this study.

Countries	#unique HealthMap articles	#unique disease-related keywords	#unique reverse-geocoded locations
China	22412	21879	30
India	3112	17160	30
United States	50787	45457	51
Singapore	290	4095	4

Climate Data: Climate attribute data for China, United States, India and Latin American countries is collected from the Global Data Assimilation System (GDAS) [25] and downloaded in 1 degree lat/long resolution from https://ladsweb.nascom.nasa.gov/ from 2010 to 2014. We downloaded 6 hourly readings of 5 climate attributes including absolute humidity, relative humidity, precipitation, specific humidity, and temperature. For Singapore, we downloaded the daily readings of average temperature, maximum temperature, minimum temperature and Precipitation at different weather stations from National Environment Agency (http://www.nea.gov.sg/) available from 2009-present. Total number of downloaded weather data for each country is given below:

Table 3. Downloaded weather data points for countries of interest.

Countries	#weather data points
China	9,074,179
United States	6,853,839
India	2,512,490
Singapore	138,646

For each country we filter those weather data points whose 'date' field falls within the study time periods mentioned in the section "Study Countries and Time Periods", e.g. for United States, we filter those weather data points whose 'date' field lies between January 2011 and December 2013.

The (latitude, longitude) values in the climate GDAS data can be different from the (latitude, longitude) values of the HealthMap articles where disease outbreaks are reported. The climate attributes at the (latitude, longitude) values of the weather data are interpolated to the (latitude, longitude) values of the HealthMap articles using an inverse distance weighting (IDW) [26, 27, 28] method for spatial interpolation.

For each type of climate attribute (CA), we calculate either its mean, maximum, minimum or cumulative sum for each coordinate over a timepoint to calculate the overall value of the climate attribute. If CA = Temperature, we take the mean, maximum and minimum of the temperature values over all the coordinates to obtain mean temperature, maximum temperature and minimum temperature in a location at a particular timepoint. If CA = Precipitation Rate, we calculate the cumulative sum of the rainfall values over the coordinates in location. If CA = Absolute Humidity or Relative Humidity, we calculate the average of the humidity values over the coordinates. Example movies depicting spatio-temporal variation of climatic attributes are given below.

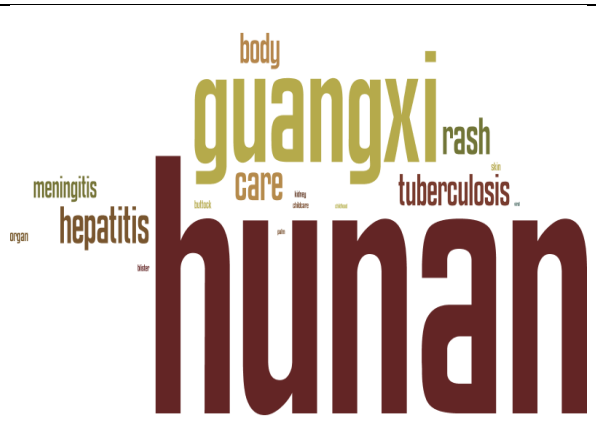
Results

We consider a baseline model where we use only news coverage as features in our forecasting model. To evaluate the effect of a climate attribute (Temperature, Rainfall, Absolute Humidity, Relative Humidity) on each disease in a region, we add the values of that climate attribute at different lags to the model in a view to assess its relative utility. For China, India and United States, we have data availability for Temperature, Rainfall, Relative Humidity and Absolute Humidity. For Singapore, the values of Relative Humidity and Absolute Humidity are unavailable. Accordingly, the baselines and possible variations explored are different.

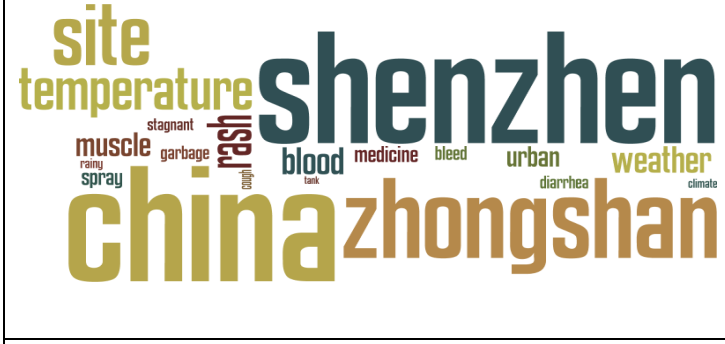
To evaluate the quality of our predictions, we adopt a key measure of performance known as the Quality Score (QS). The Quality Score (QS) is computed by comparing the predictions against official case counts for each timepoint. Following a similar approach as in [19, 20], QS is defined to be in the range [0,4] where a 4 indicates perfect forecasting accuracy.

First, we evaluate the quality of our supervised spatio-temporal topic model in using seed words or prior words to extract relevant disease topics in specific countries. In Figure 2's panels below, we show phrases or terms discovered by the supervised spatio-temporal topic model for each disease topic in specific countries.

H7N9 disease topic, China

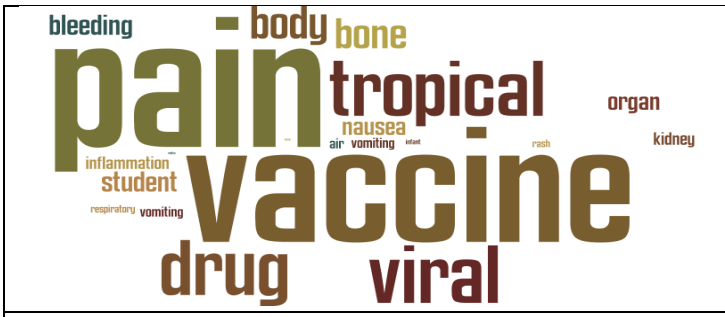


HFM disease topic, China



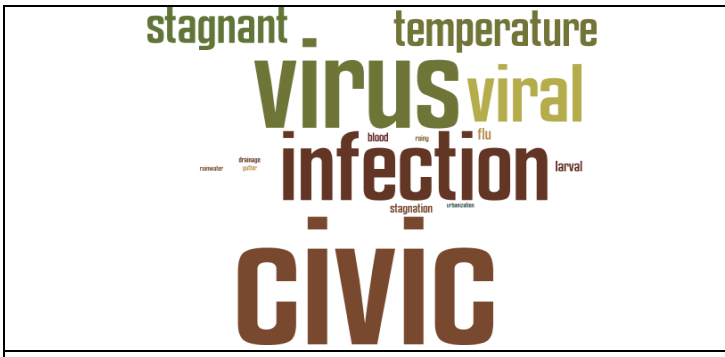
Dengue disease topic, China

Dengue disease topic, Singapore



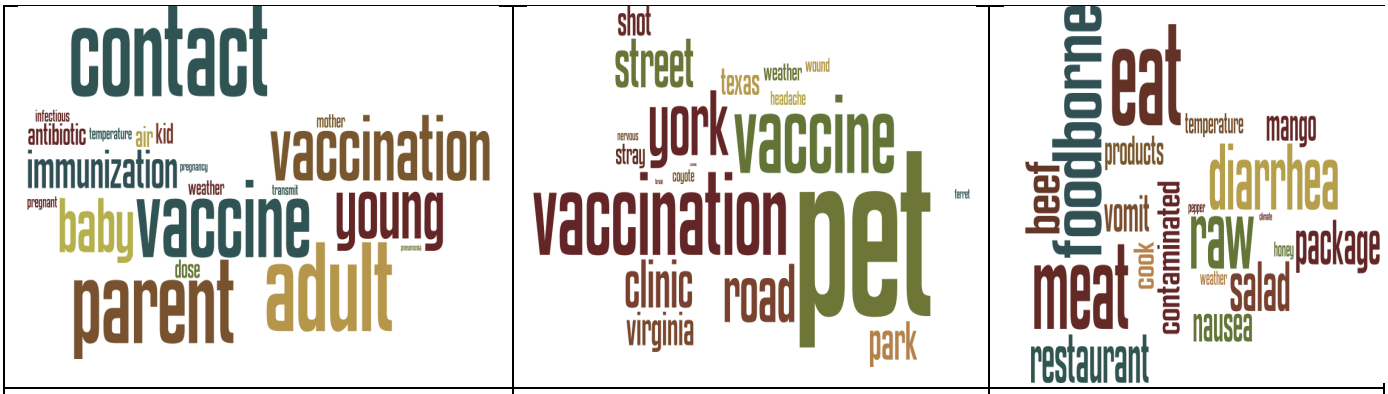
HFM disease topic, Singapore

Foodborne diseases, India



Dengue disease topic, India

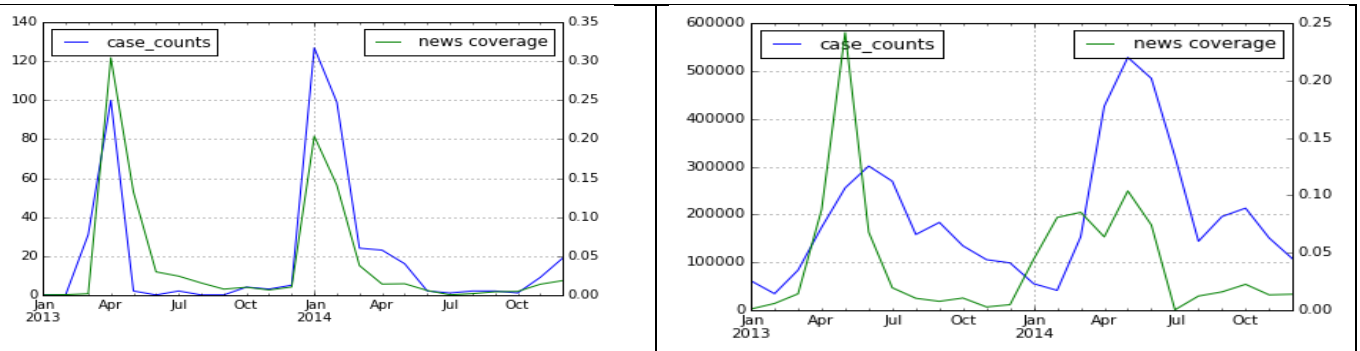
Malaria disease topic, India



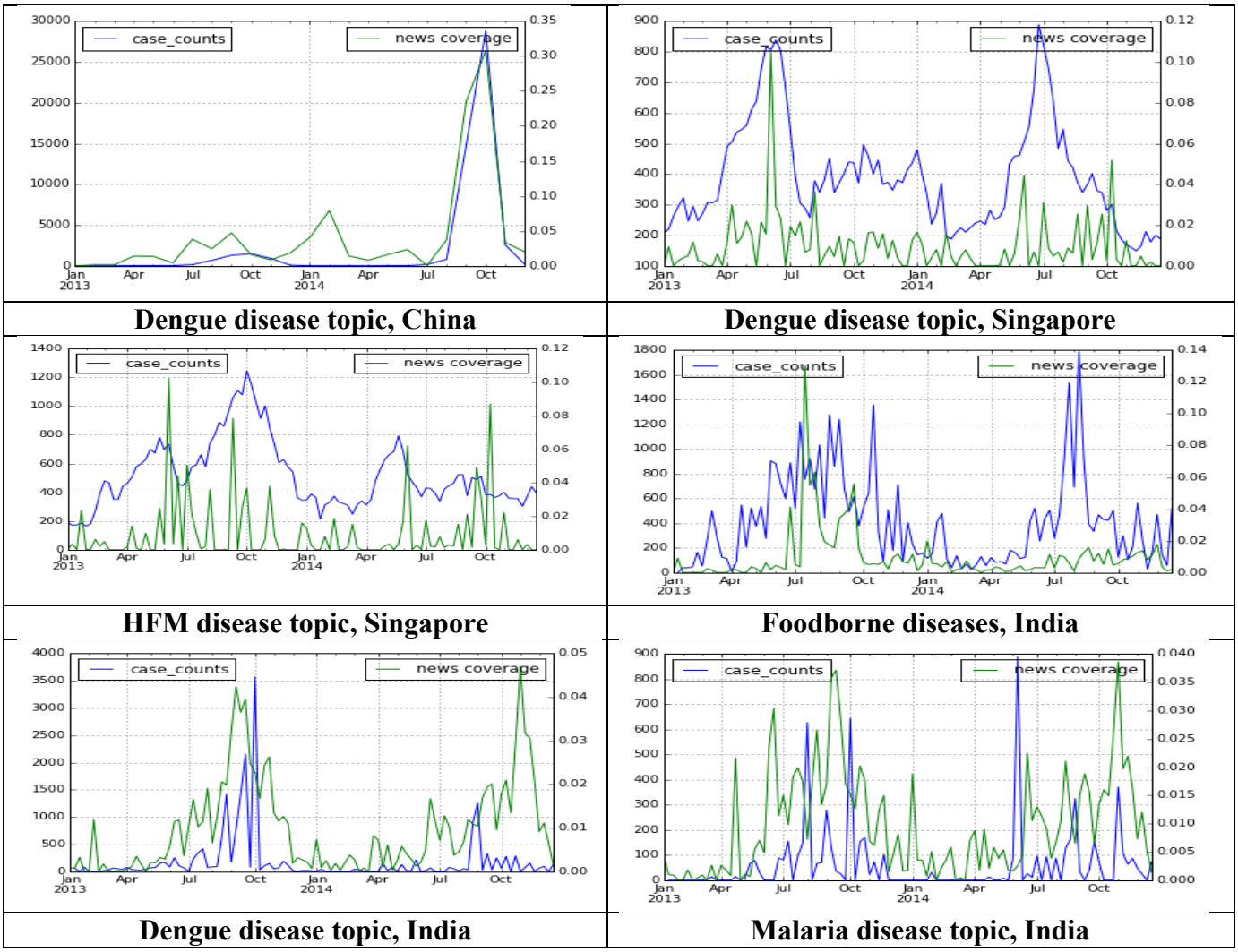
Whooping Cough disease topic, US	Rabies disease topic, US	Salmonella disease topic, US

Figure 2. Phrases or terms discovered by the supervised spatio-temporal topic model from the HealthMap corpus for each disease-country combination.

Next, we study the correlation (Figure 3 below) between temporal news coverage extracted from the HealthMap news corpus using our supervised topic model and the surveillance case counts for each (disease, country) combination. For most of the (disease, country) combinations, we observe a rise in temporal news coverage trends whenever there is an outbreak, and a fall in media coverage trends with low outbreak activity. However, there are several deviations, such as the Dengue outbreak in Singapore 2014, Salmonellosis outbreak in USA 2011 or HFM outbreak in China 2014 where temporal news trends are not able to capture the outbreak season. We posit that such deviations are a factor of multiple effects: (i) News media coverage during disease outbreaks is driven by interest. New coverage for certain diseases can be inconsistent over time and across regions. Specifically, for diseases with low public interest, the coverage can be low even though there is an ongoing disease outbreak. (ii) Articles in our database do not have the full text available. Insufficient information from such articles does not enable our framework to monitor disease progression appropriately during the relevant timepoints.



H7N9 disease topic, China	HFM disease topic, China



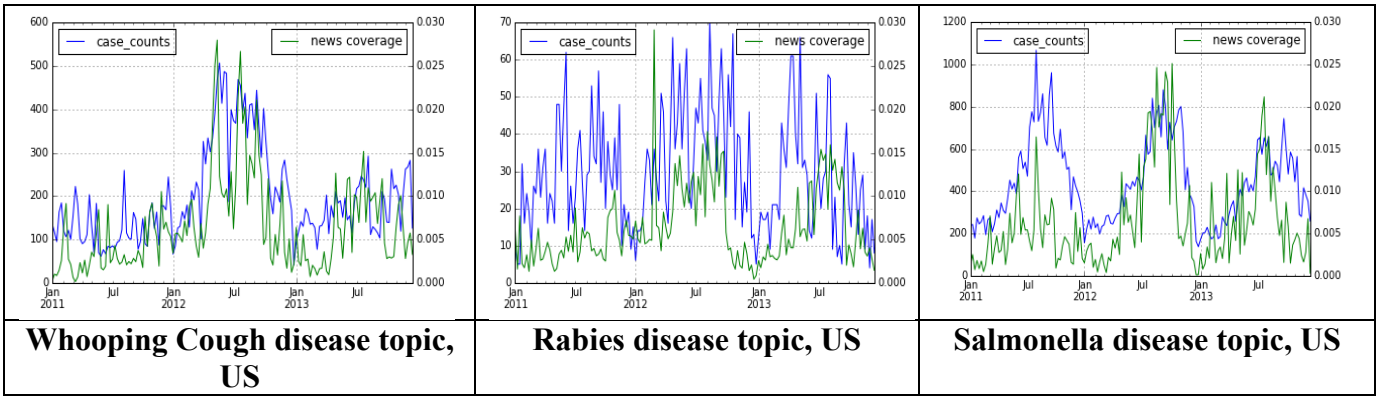


Figure 3. Correlation between temporal news coverage extracted from HealthMap news corpora and the surveillance case counts for each (disease, country) combination.

Finally, we evaluate the forecasting accuracy of our ensemble technique incorporating climate attributes against the baseline model which only uses news coverage as features. To evaluate the effect (positive or negative correlation) of each climate attribute on outbreaks of each disease in a region, we calculate the percentage change in overall QS of each climate attribute model over the baseline model.

Table 4. Percentage change (increase ↑ or decrease ↓) in overall accuracy (QS) of each climate attribute model over the baseline model for diseases in China.

Diseases	Baseline vs Temperature Model	Baseline vs Rainfall Model	Baseline vs Absolute Humidity Model	Baseline vs Relative Humidity Model
H7N9	**37%**↑	17.91%↑	34.68%↑	13.29%↓
Dengue	**40.63%**↑	15.63%↑	10.94%↑	14.06%↑
HFM	0%	0%	0%	0%

In Table 4, we see that forecasting performance for H7N9 and Dengue outbreaks are positively correlated with climatic attributes except Relative Humidity for H7N9 while HFM has no correlation with any of the climate attributes. For H7N9, Temperature is the most predictive climate attribute with 37% increase in forecasting accuracy followed by Absolute Humidity with 34.68% increase in accuracy. This is consistent with the observation in [29] that H7N9 viruses circulate more at higher levels in colder weather and at lower levels in warmer weather. For Dengue, Temperature shows the strongest correlation with 40.63% increase in forecasting accuracy.

Table 5. Percentage change (increase ↑ or decrease ↓) in overall accuracy (QS) of each climate attribute model over the baseline model for diseases in India.

Diseases	Baseline vs Temperature Model	Baseline vs Rainfall Model	Baseline vs Absolute Humidity Model	Baseline vs Relative Humidity Model
Malaria	**39.72%**↑	19.86%↑	16.31%↑	**39.72%**↑
Dengue	7.75%↓	**2.59%**↑	6.89%↓	12.93%↓
Foodborne Outbreaks	**1.14%**↑	6.29%↓	0%	0.57%↑

For India, we see that forecasting malaria outbreaks significantly benefits from the use of climate attributes. Temperature and Relative Humidity are the most predictive climatic factors for Malaria with 39.72% increase in forecasting accuracy. This is expected as rise in Temperature and Relative Humidity increases the lifespan of mosquitoes, giving them more opportunities to transmit malaria virus from one

person to another. Surprisingly, Dengue outbreaks are negatively correlated with most of the climate attributes except Rainfall where we observe a slight improvement (2.59%) in accuracy. This is inconsistent with our findings for China. Foodborne outbreaks show negligible correlation with the climate attributes.

Table 6. Percentage change (increase ↑ or decrease ↓) in overall accuracy (QS) of each climate attribute model over the baseline model for diseases in US.

Diseases	Baseline vs Temperature Model	Baseline vs Rainfall Model	Baseline vs Absolute Humidity Model	Baseline vs Relative Humidity Model
Rabies	3.15%↑	1.97%↑	3.15%↑	**3.54%**↑
Salmonella	**0.68%**↑	0.34%↑	0%	0.34%↓
Whooping Cough	1.65%↓	**0%**	1.32%↓	0.66%↓

In United States, forecasting Rabies outbreaks shows slight positive correlation with all of the climate attributes, with Relative Humidity being the most indicative attribute (3.54% increase in forecasting accuracy). However, for Salmonella and Whooping cough, we find negligible effect of the use of climate attributes in forecasting.

Table 7. Percentage change (increase ↑ or decrease ↓) in overall accuracy (QS) of each climate attribute model over the baseline model for diseases in Singapore.

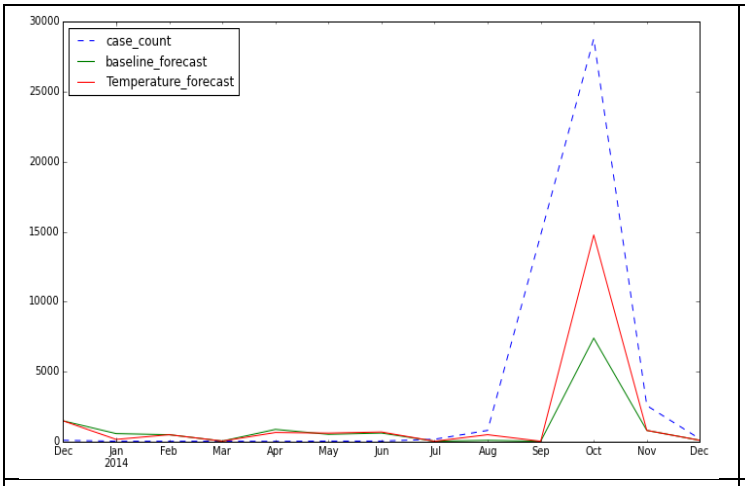
Diseases	Baseline vs Temperature Model	Baseline vs Rainfall Model
HFM	1.04%↑	**3.82%**↑
Dengue	0.35%↑	**0.71%**↑

In Singapore, HFM forecasting performance shows slight positive correlation with rainfall achieving 3.82% increase in forecasting accuracy. Again, contrary to our findings in China, Dengue in Singapore shows very negligible correlation with climate attributes.

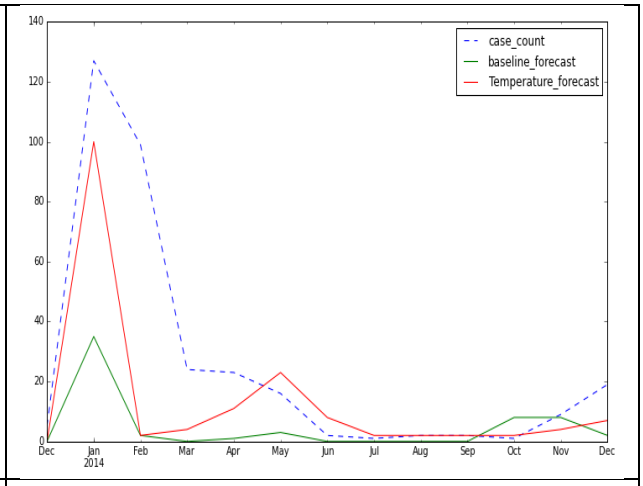
In overall, we observe significant improvement in forecasting for 3 (disease, country) combinations - (H7N9, China), (Dengue, China) and (Malaria, India) with 30-40% increase in predictive accuracy by incorporating climate attributes. In the figures below we plot the forecasts of our ensemble technique incorporating the best performing climate attribute (e.g. Temperature for H7N9) as features against the forecasts of the news-based baseline model for these 3 (disease, country) combinations. We also plot the actual surveillance case counts.

In the forecast plots for (Dengue, China) and (H7N9, China), we observe that our ensemble technique (red solid line) is able to forecast the massive outbreaks of H7N9 during January 2014 and Dengue during October 2014 more accurately with respect to the baseline model (green solid line). However, for Malaria in India, we observe that none of the models are able to forecast the outbreak spike in June, October and November 2014. However, the models have been able to forecast the outbreaks in July and September 2014. Also, note that the baseline model forecasts (green solid line) exhibit some false peaks, especially during Dec. 2014 and Jan-April 2014 even though the surveillance case counts are almost close to zero. However, our ensemble technique forecasts does not have these false peaks which led to the significant improvement in accuracy.

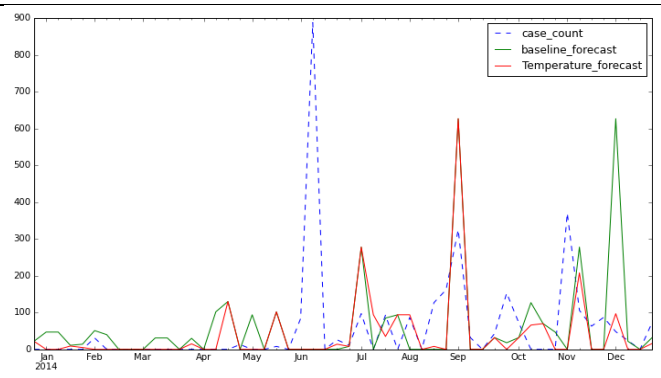
However, for certain (disease, country) combinations such as (Dengue, India), (Dengue, Singapore) and (Salmonella, USA), we find no improvement in forecasting accuracy. We posit that such deviations may be due to multiple reasons: i) our assumption that the intensity of disease outbreak in a region being proportional to the outbreak news coverage in the region can lead to errors in calculating the overall climate attribute value due to inconsistent nature of news media coverage across regions and time as discussed in section "News-Disease Correlation", ii) In Singapore, the climate data downloaded has a good number of missing values at multiple timepoints. We assume that the climate attribute value is zero at those timepoints. These missing values can hinder the performance of the model.



Forecast plot for Dengue, China



Forecast plot for H7N9, China



Forecast plot for Malaria, India

Figure 4. Forecast plots for 3 (disease, country) combinations where climate attributes had maximum impact on forecasting outbreaks.

For ILI, we compared the individual accuracy of sources shown in Figure 1 and found that climatic attributes are on average the most important source (see Figure 5).

Figure 5. Comparison of ILI forecasting accuracy using MFN [19] using a single source at a time. Darker colors indicate higher accuracy.

To further analyze this effect we combined all the sources and performed an ablation test where we drop one source at a time and record the percentage decrease in accuracy. Comparative results of this ablation test is shown in Figure 6 where weather shows the most decrease in accuracy and hence once more indicating the importance of considering this physical source for endemic disease.

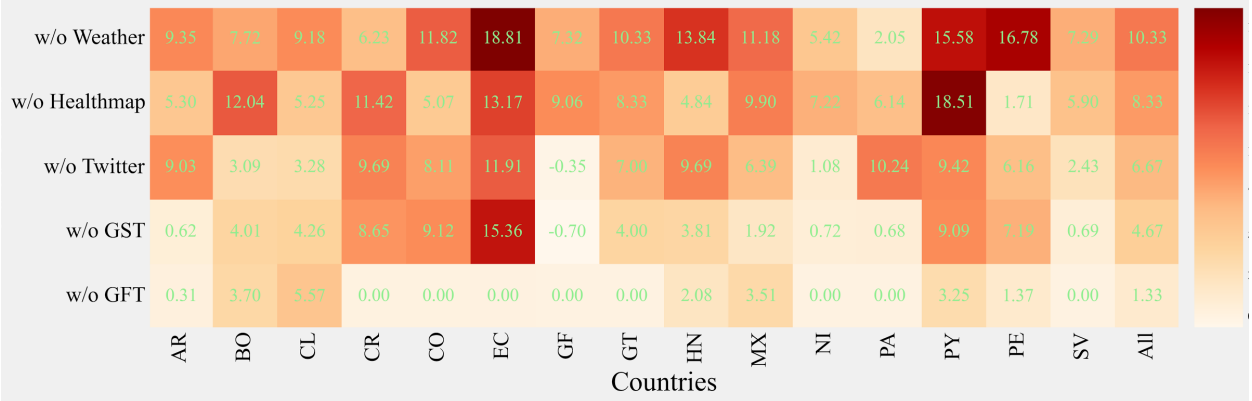


Figure 6. Comparison of decrease in ILI forecasting accuracy while ablating a single source. Darker colors indicate greater drops in accuracy; sources with most dark colors are thus the most important ones.

We extended our forecasting horizon using DPARX, a dynamic general linear model described in [20] and found similar results about climatic attributes. The forecasting accuracy for Latin American countries and US using DPARX is shown in Table 8.

Step	Method	AR	BO	CL	CO	CR	EC	GT	HN	MX	NI	PA	PE	PY	SV	US
1	ARX	2.94	2.51	3.10	2.90	2.21	2.81	2.83	2.96	2.25	2.18	2.78	2.51	2.84	2.83	3.51
	MFN	2.99	3.01	2.88	2.53	2.78	2.81	2.77	2.83	2.61	2.70	2.56	2.82	2.66	2.79	3.81
	DARX	3.09	2.84	3.17	2.84	2.57	2.94	2.83	2.89	2.91	2.77	2.72	2.67	2.79	2.72	3.71
	DPARX	2.98	2.84	3.07	3.01	2.70	2.97	2.87	2.93	2.84	2.86	2.82	2.78	2.86	2.77	3.72
2	ARX	2.56	2.05	2.63	2.71	1.61	2.56	2.63	2.76	1.15	1.36	2.56	2.05	2.62	2.64	3.21
	MFN	2.86	2.89	2.81	2.49	2.71	2.67	2.72	2.41	2.55	2.31	2.50	2.59	2.71	2.30	3.75
	DARX	2.98	2.69	3.00	2.69	2.63	2.79	2.72	2.81	2.66	2.28	2.55	2.49	2.68	2.66	3.60
	DPARX	2.67	2.73	2.86	2.83	2.66	2.79	2.78	2.78	2.62	2.49	2.71	2.63	2.64	2.68	3.61
3	ARX	2.25	1.65	2.21	2.50	1.06	2.30	2.39	2.59	0.60	0.94	2.42	1.72	2.39	2.46	2.92
	MFN	2.49	2.38	2.41	2.33	2.45	2.31	2.32	2.10	2.21	2.11	2.19	2.22	2.40	2.08	3.64
	DARX	2.68	2.32	2.68	2.57	2.52	2.72	2.50	2.65	2.47	2.00	2.52	2.32	2.54	2.53	3.41
	DPARX	2.33	2.44	2.63	2.70	2.58	2.66	2.59	2.61	2.36	2.31	2.75	2.44	2.51	2.55	3.42
4	ARX	1.98	1.37	1.73	2.31	0.72	2.07	2.22	2.41	0.39	0.83	2.21	1.46	2.21	2.30	2.56
	MFN	2.10	2.13	2.15	2.04	2.25	2.11	2.22	1.94	1.99	1.87	2.01	1.86	2.10	1.77	3.54
	DARX	2.42	2.12	2.39	2.49	2.34	2.52	2.42	2.51	2.17	1.74	2.38	2.27	2.30	2.42	3.18
	DPARX	2.10	2.23	2.32	2.64	2.38	2.52	2.55	2.45	2.06	2.15	2.72	2.38	2.27	2.53	3.20

Table 8. Comparison of multistep ILI forecasting accuracy using DPARX [20] using weather sources. Accuracy scores scaled to 0-4 with 4 indicating a perfect forecast.

Part 2: Climatic Precursors to Civil Unrest

Related Research

The sociopolitical effects of climate shocks, particularly ones leading to resource scarcities, such as riots, protests and even civil war are well documented. The drought in Syria that subsequently led to loss of agricultural output, one of the triggers for the civil war in 2011, is a classic example of the nexus between adverse climatic effects and social unrest [30, 31]. Recent quantitative analysis, that culls results from multiple disciplines like archaeology, economics, criminology, history, and political science, shows that the magnitude of climate's influence in human conflict is significant [32]. Studies show that there is a direct correlation between increased temperatures due to global warming [33], extreme rainfall (or lack thereof) [34] that can increase conflict by ~14%. Violent civil unrest outbreaks have been correlated with global and local climatic changes [36].

Although the above studies have established correlation, the scope of such analysis has been limited. While independently climatic changes may not directly cause social unrest, changes in the environment can alter the conditions for certain social interactions, negatively affect the economy, disrupt infrastructure projects, thus indirectly affecting the social stability of a region [33]. Wischnath et al. [35] categorizes the direct linkage between climatic change and human conflict into four categories viz. *escalating competition over resources, migration patterns leading to interracial tension, economic impacts due to increased levels of global drought, and individual responses to sudden climatic changes* [35]. While the causal relationship in the above scenarios is based on some fundamental analysis, intuition, and conjecture, prior research is not backed up by intensive data analysis. Most studies also study the effects of climate shocks and weather changes on the affected region, but little work demonstrated the cross-regional impacts of such effects.

Methods

We primarily utilize three main classes of techniques: *text classification, storytelling*, and *dynamic query expansion*. Text classification is the process of categorizing text documents (news articles, tweets) into one or more predefined categories based on their content. Typically this is done using a local or distributed vector representation of documents and techniques such as nearest neighbor modeling or naïve Bayes estimation to classify documents into classes. In nearest neighbor modeling, the algorithm ranks a document's "neighbors" (as determined via some similarity or distance measure) and distills the class labels of these neighbors (e.g., via a majority vote) into a final classification.

Storytelling is the process of "chaining" documents into a coherent sequence of time-evolving developments. The approach we take to conduct storytelling is to rely on weighted scores of similarities across news articles for three sets of features: textual features (related to keywords), spatial features (such as locations and geographical coordinates), and actors (such as person(s), and organizations mentioned in the articles).The chaining methodology is developed with the goal of identifying all documents related to a climate-related or civil unrest event and to keep track of the news story as new documents arrive. The algorithm operates in an incremental fashion wherein every new input article is analyzed as it arrives and is appended to already existing chain(s).

Finally, dynamic query expansion is a social network-based technique that utilizes heterogeneous features (e.g., containment, authorship, and replying etc.) extracted from Twitter data to expand a given seed query (or set of terms) – in our case, into climate and protest related keywords or hashtags. For more details, please see [23].

Where possible, results from all the above algorithms are compared against a manually curated set of events called the Gold Standard Report (GSR) described in [23], which contains news reports for civil unrest event from three major news sources in each country of interest.

Results

Our study encompassed 25,352 civil unrest events across Latin America from January 2011 to March 2015. These events were manually curated by subject matter experts (SMEs) at MITRE and available to all performers in the IARPA OSI program. Among all these 25,352 events, our text classifier identified 974 events as climate-related events, which is equal to 3.84%. Although this overall % is small, there are specific countries and more importantly, specific (country, month) combinations where the % of climate-related protests exceeds 20%. See Figures 7,8,9.

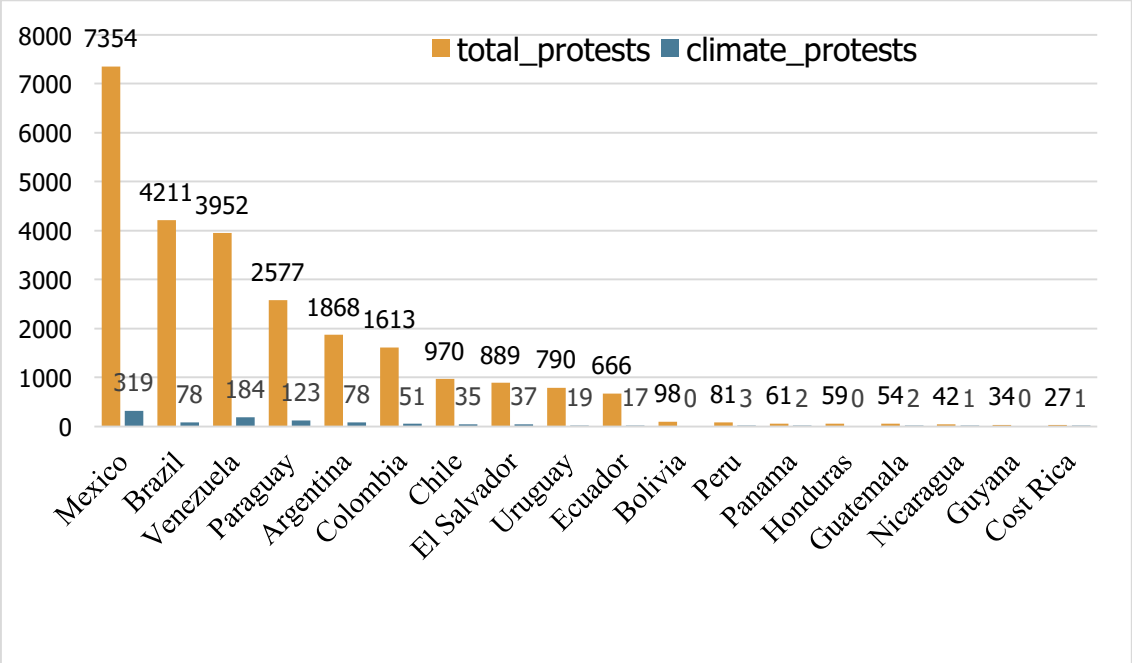


Figure 7 GSR protest event distributions

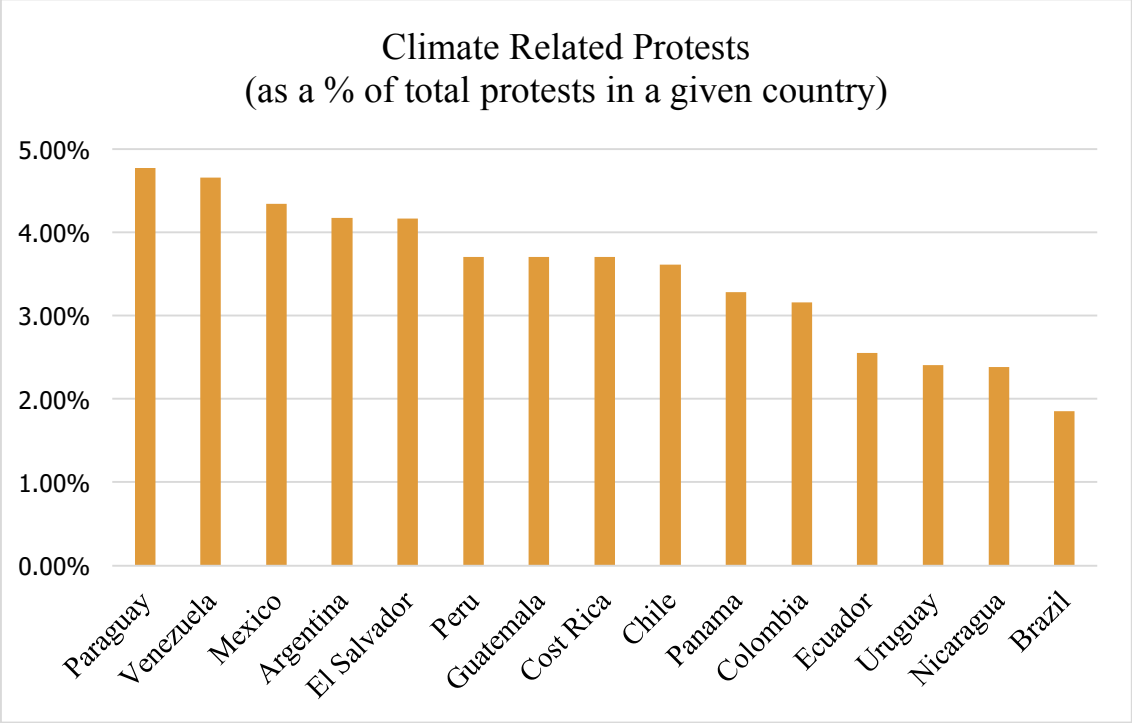


Figure 8 Climate-related protest events as a % of total protests in a country.

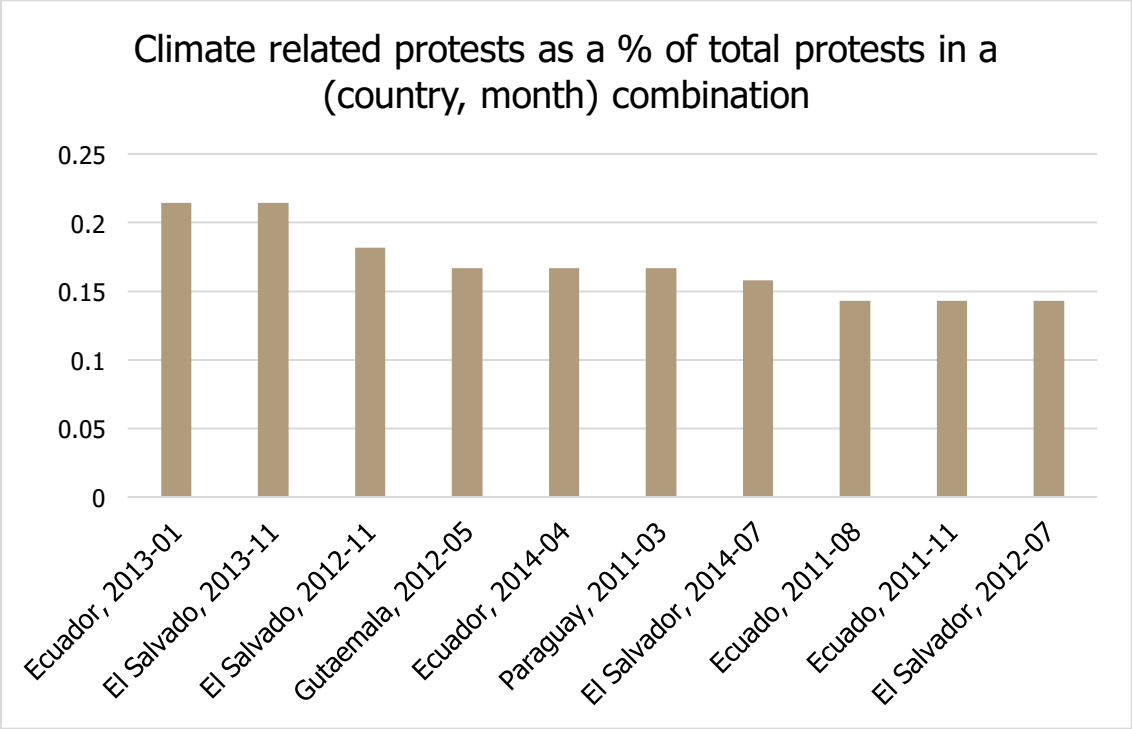


Figure 9 Top 10 Climate related protest events (%) by country/month, from Jan 2011 to Mar 2015

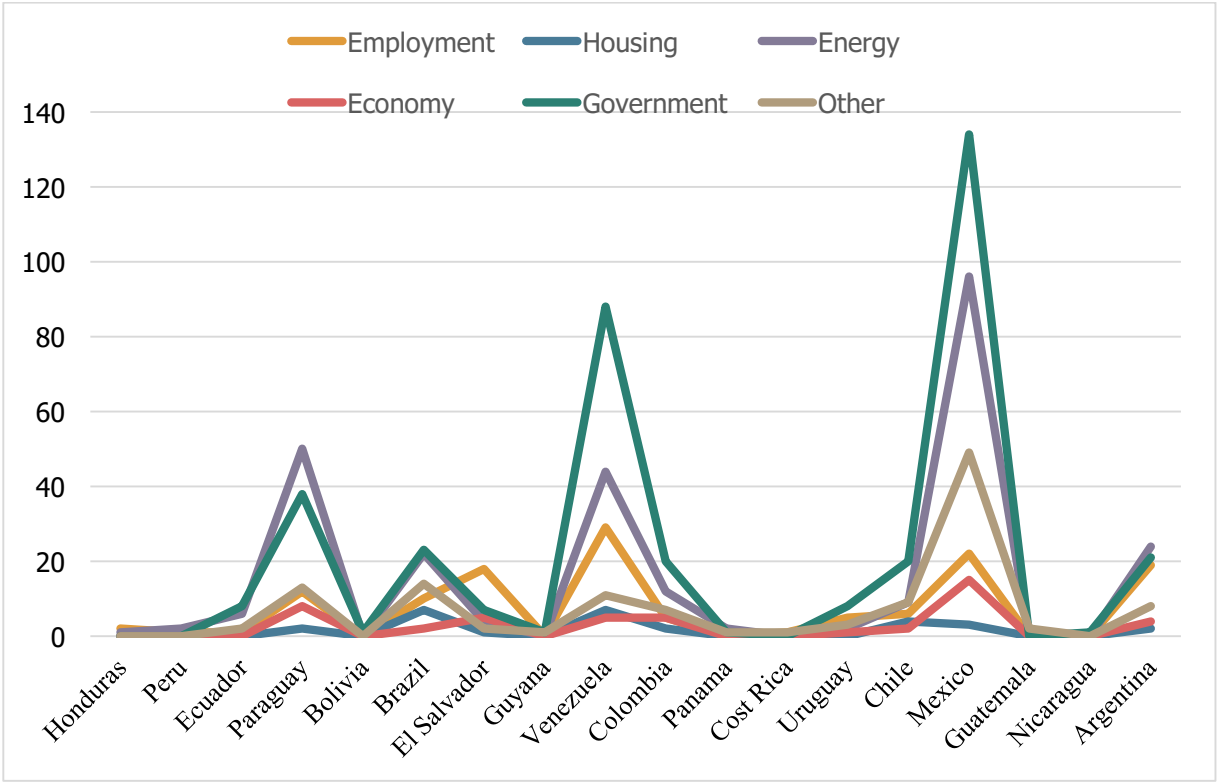


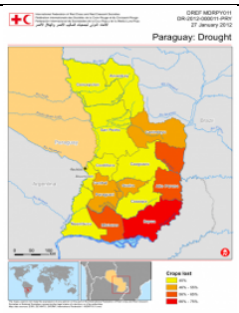
Figure 10 Distribution of climate-related civil unrest events in Latin America.

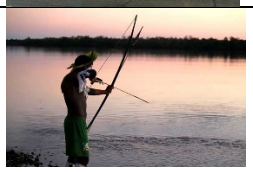
Mexico has both the largest absolute number of protest events and climate-related protest events. However, in terms of percentages, Costa Rica is the highest (8%) and Bolivia has the lowest (1%) number of climate-related civil unrest events. We also found that February, March, and May have more climate-related protest events than other months n overall, while November has the least. A breakdown of causes of protests is given in Fig 10. As can be seen, most protests (implicitly or otherwise) are targeted against government policies.

We now turn to providing specific examples of climate related protests identified by our text classification algorithm (see Table 9).

Table 9. Sample protests with climatic themes.

Country & Date	Photos (if available)	Description
Brazil Jan 2015	 	Climate scientists suggest that Brazil's weather patterns have been disrupted by the loss of the Amazon rainforests. In this picture, a worker from the São Paulo state company that provides water and sewage services inspects cracked ground near Jaguary dam, Braganca Paulista.
Mexico Sep 2013	 	On September 15, 2013, Mexico experienced a period of most intense rainfall with the onset of Hurricane Manuel. Within 24 hours, Hurricane Ingrid also struck Mexico's gulf coast. These two hurricanes left over 150 people dead.
Brazil Nov 2013		Students of the Federal University of Pernambuco (UFPE), Recife, protest the heat in classrooms. Discomfort caused by the heat weave caused eight thousand workers of the Rio Grande Shipyard to strike in the city of Rio Grande.

Paraguay Sep 2013		Severe water shortage and lack of rain for an extended period of time coupled with inattention of the government to farmers' plights caused a massive farmers protest on 09/25/2013.
Venezuela Mar 2013	 	Outer bands impact caused by Hurricane Isaac (Aug 2012) was in general followed by the government's lack of post-disaster reconstruction mechanisms. Many of the flood's victims were still living in ruins, causing a general population protest on 03/25/2013.
Argentina Mar 2013		Urban waterlogging caused by severe storm systems and the government's failure in taking emergency measures for civilians caused general protests on 03/23/2013 and 03/25/2013.
Chile Mar 2013	 	Events were held to commemorate the tsunami and earthquake of February 27, 2010. This event turned into a violent riot on 03/14/2013 at Concepción where more than 60 participants clashed with police and burned several police vehicles.
Chile Aug 2010		Twenty neighbors of a housing complex commune protested with banners to denounce that there are five buildings that are not repaired after the earthquake in February 2010.
Brazil 2012-2013	 	Severe water shortage and lack of rains, coupled with the government's inability to acknowledge the crisis, led to a general protest on 12/14/2012, followed by a farmers protest on 01/14/2013 (Pedra Branca, Brazil).

Honduras Oct 2012	 	A massive storm lasted several days, caused flooding, and threatened lives and property. A general protest towards government was called on 10/15/2012 for its indifference and disregard.
Chile Jan 2014		A group of residents protested against a project to destroy an area devoted to green areas in the Conchali.
Brazil Jun 2012	 Sacrificing the Amazon and its Peoples for Dirty Energy	About 300 indigenous people and environmentalists protested against the construction of the Belo Monte hydroelectric dam. The demonstration occurred about 50 kilometers from the county seat, in the area known as a cofferdam, a kind of earth dam built on the Xingu River Consortium Builder Belo Monte.
Argentina Jan 2013		Tired of repeated blackouts, the residents of Villa Crespo cut yesterday afternoon Scalabrini Ortiz street, at the height of Murillo. They claimed that since November last year they have suffered interruptions in electrical service.

While the above provide specific anectodal examples of climate related protests, the storytelling algorithm places such news stories in context. The storytelling algorithm is employed to track the evolution of an event, from its beginning to end, with perhaps a development into a civil unrest event. To avoid confusion, we build story chains for each country separately. Each news article is modeled as a bag of terms (involving named entities such as people, locations, organizations) and these terms are tracked over time in a story chain. Story chains involving climatic keywords and that end in protests are especially interesting to help suggest precursors. See examples in Figures 11,12, and 13.

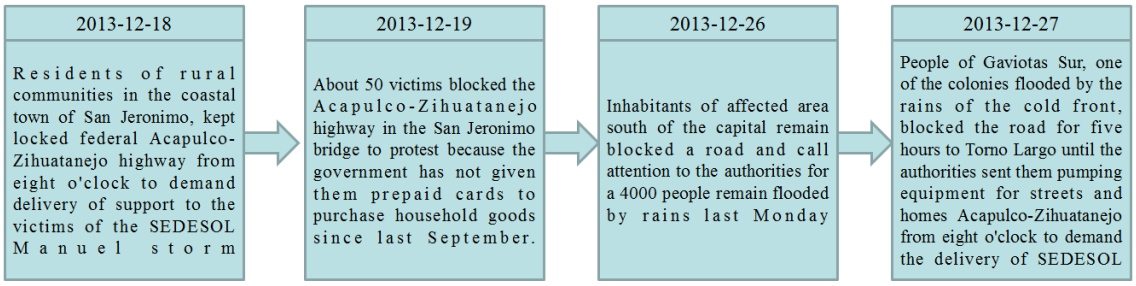


Figure 11. Mexico hurricane protest story chain.

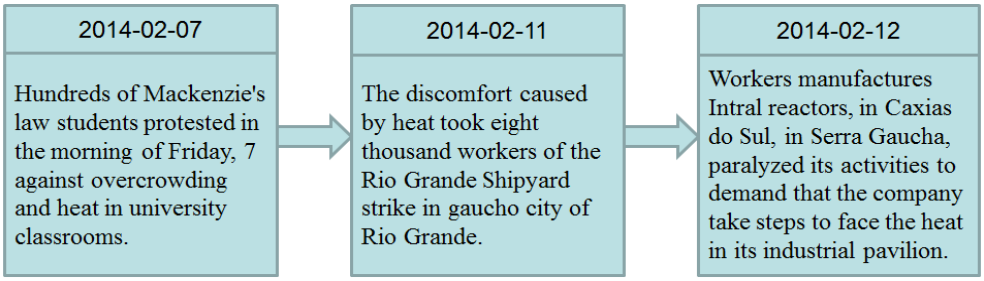


Figure 12 Brazil heat protest story chain.

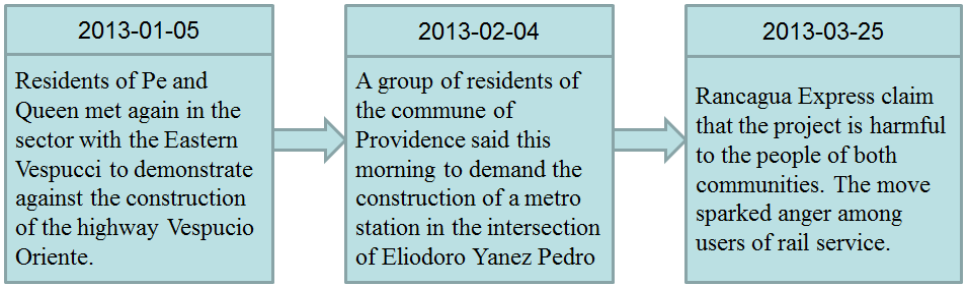


Figure 13 Chile construction protest story chain.

Finally, we describe results from our dynamic query expansion algorithm. The number and distribution of events for each country extracted from social media can serve as an important precursor or indicator of ongoing happenings. For instance, Fig. 14 shows that during the Mexican hurricane events in Sep 2013 (Manuel/Ingrid), there were significant upticks that cause anomalous activity relative to other, more "normal", days.

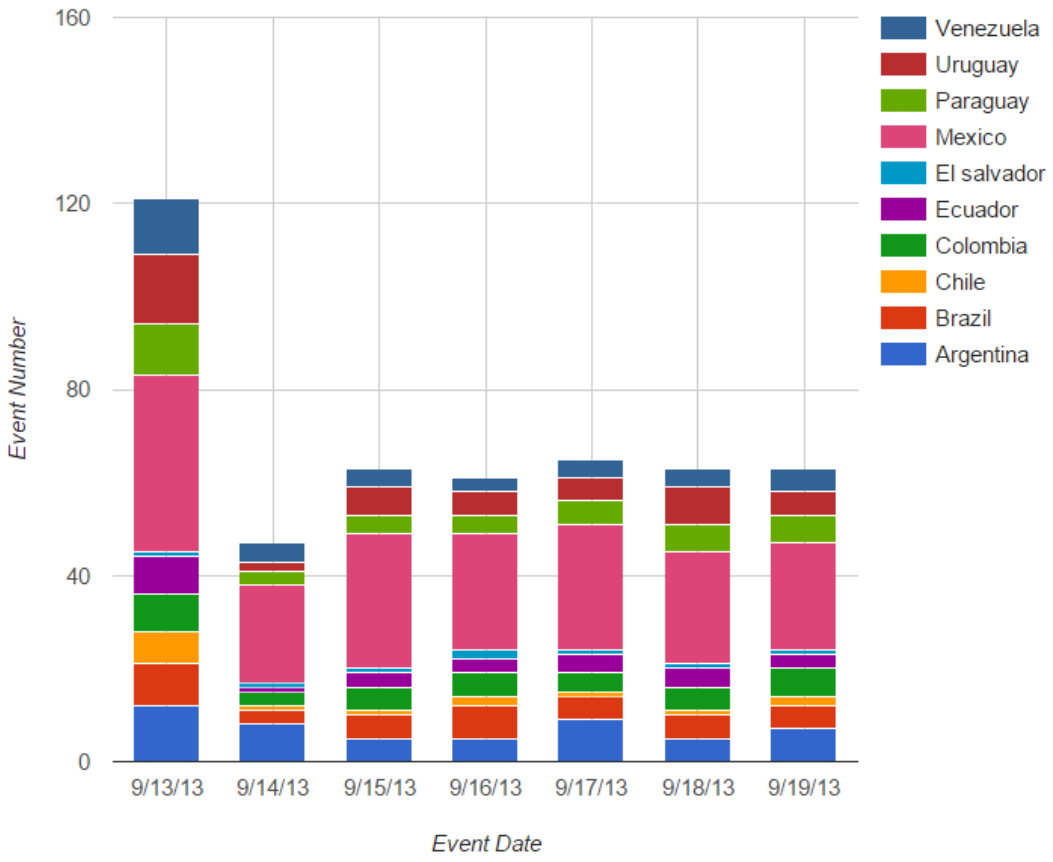


Figure 14 Protest related events extracted using the DQE algorithm in Sep 2013.

Studying these events in greater detail gives us greater insight into the functioning of DQE, which we proceed to do next.

Case Study 1: Mexico Hurricane, Sep 2013

In Sep. 2013, Mexico suffered from two Hurricanes: Ingrid and Manuel. On September 15, 2013, Mexico experienced the most intense rainfall in the country's history with the onset of Hurricane Manuel that battered the northwest. Within 24 hours, Hurricane Ingrid also struck Mexico's gulf coast. Hurricane Manuel then made its second landfall. These two hurricanes left over 150 people dead and caused about six billion dollars in damage. After the hurricane, related protests and other civil unrest events broke out dozens of times, and lasted for more than 2 years because the government's response had been woefully inadequate.

Figure 15. Illustration of DQE on tweets from Mexico (Sep 2013)

Fig 15 illustrates the updation of keywords list at each iteration leading to a convergence identifying specific storm-related keywords (e.g., Manuel). A more detailed view of the iteration is given in Fig. 16.

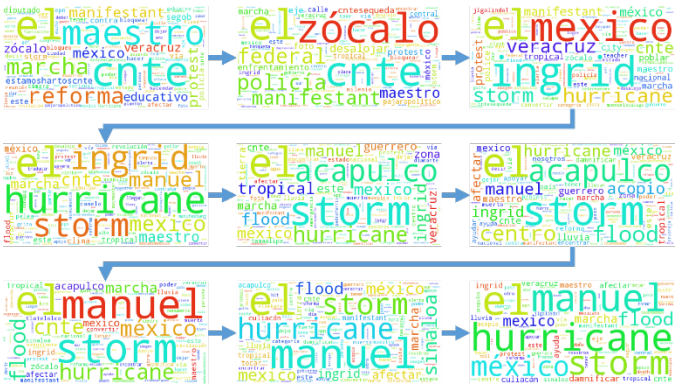


Figure 16 Word cloud of tweets related to the Mexico hurricanes Ingrid and Manuel from Sep. 12th to Sep. 20th 2013.

Figure 17 illustrates a similar progression of word clouds pertaining to civil unrest events after Hurricane Ingrid and Manuel. On the top left, Figure 17a illustrates a protest in Acapulco, a costal city of Mexico seriously damaged by the two hurricanes. Figure 17c, on the bottom left, is a protest demanding more help from the Mexican government due to the cold winter of 2013. Other graphs shown in Figure 17b and 17d on the right side are civil unrest events demanding more help under similar circumstances. In fact, dozens of civil unrest events broke out in the next two years, with the latest protest occurring as late as June 2015.

Figure 17 Word Cloud of Civil Unrest Events after Mexico Hurricane Ingrid and Manuel in Sep. 2013

Case Study 2: Extreme Weather, Culiacán, Mexico, Sep. 2013

Culiacán, an important and big city in northwestern Mexico suffered from severe drought for an extended period and was also affected by storm and floods on Sep. 18th, 2013. In Fig 18, we see coordinated protests across multiple cities, with similar motivations. In particular, a storm hit the city of Zihuatanejo, which is the fourth-largest city in the Mexican state of Guerrero, leading to subsequent protests. Acapulco de Juárez, commonly called Acapulco, was inundated by floods caused by the hurricanes mentioned earlier, disrupting life and stoking civil unrest. These floods also affected the city of San Luis Río Colorado on Sep 18, 2013, again leading to strikes.

Figure 18 Mexico climatic protest (Sep 2013).

Case Study 3: Brazil Drought, May 2012

A severe drought condition in Brazil in 2012 gave rise to sustained protests (Fig. 19). This drought was considered the worst drought in Brazil within the past 50 years.

Figure 19 Brazil drought related protests studied using dynamic query expansion.

Part 3: Lessons Learned

Our disease forecasting study has shown that there are specific, tangible, benefits to be had in incorporating climatic attributes into our modeling of several infectious diseases. By evaluating our approach broadly across a range of countries and diseases, we have demonstrated that when news coverage is regular and data quality is uniform, incorporation of climatic attributes improves the quality score more significantly than other surrogate data sources. Our study is also the first to enumerate climate-themed and climate-related protests on a massive scale in multiple Latin American countries. The % of such protests varies from 8% to 20% depending on the specific country or month under consideration.

Open source indicators can thus provide a sufficiently location-specific early warning system for climatic events and precursors to disease outbreaks and civil unrest events. Although outside the scope of our study here, the analysis presented here can be readily prototyped into a continuous, online, cloud-based system that tracks activity in regions of interest and delivers analysis and/or forecasts of events as they develop.

A second take-away from the study here is that massive data sources bring about the interconnectedness of societal events at a scale not possible before. Extending the methods studied here, we should be able to identify cascading and "network" effects among significant societal events. For instance, what are the most common "pathways" by which a climatic event leads to a protest? Can we forecast not just the imminent occurrence of such events but also societal responses and reactions at a broader scale? If we can combine the lessons learnt here with a representation of policy and countermeasures, we can use the resulting multimodal representation to pose "what-if", "why", and "why not" questions over scenarios, to understand relationships between specific strategies and outcomes. Most studies currently are retrospective in nature and being able to (conditionally) forecast would constitute a quantum leap in our capability to anticipate and mitigate severe climatic scenarios.

Finally, this study demonstrates that new research at the intersection of computer science, data analytics, environmental science, and policy planning is a blossoming area that must be fostered to make further inroads into this space. Data scientists are necessary to process the massive volumes of data being gathered on a human and societal scale but might lack awareness of the relevant scientific questions to pose over such data. Domain experts such as climate scientists and policy planners possess such awareness but are not necessarily experts in translating them to data analysis tasks. A synergistic collaboration will lead to relevant and practical research in mitigating water, climate, and environmental stressors.

Additional areas to be explored include the integration of network models of large-scale societal phenomena with data-driven approaches. This will permit the integration of model-based and data-driven approaches, leveraging their respective strengths. For instance, with modern high performance computing and data capacity, it is possible to create a network model of an entire city (or region) upon which we can impose multiple behaviors, scenarios, and countermeasures. Such a model can then be coupled with long-term climate forecasts to provide greater lead time in anticipating disruptive events.

References

1. Patz, JA et al. "Climate change and infectious diseases." Climate change and human health: risks and responses (2003): 103-37.
2. Patz, Jonathan A et al. "Global climate change and emerging infectious diseases." Jama 275.3 (1996): 217-223.
3. Mirski, Tomasz, Michał Bartoszcze, and Agata Bielawska-Drózd. "Impact of climate change on infectious diseases." Pol J Environ Stud 3 (2012): 525-532.
4. Semenza, Jan C, and Bettina Menne. "Climate change and infectious diseases in Europe." The Lancet infectious diseases 9.6 (2009): 365-375.
5. Morand, Serge et al. "Climate variability and outbreaks of infectious diseases in Europe." Scientific reports 3 (2013).
6. Lipp, Erin K, Anwar Huq, and Rita R Colwell. "Effects of global climate on infectious disease: the cholera model." Clinical microbiology reviews 15.4 (2002): 757-770.
7. Curriero, Frank C et al. "The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948-1994." American journal of public health 91.8 (2001): 1194-1199.
8. Hunter, PR. "Climate change and waterborne and vector-borne disease." Journal of applied microbiology 94.s1 (2003): 37-46.
9. Charron, Dominique F et al. "Vulnerability of waterborne diseases to climate change in Canada: a review." Journal of Toxicology and Environmental Health, Part A 67.20-22 (2004): 1667-1677.
10. Akil, Luma, H Anwar Ahmad, and Remata S Reddy. "Effects of Climate Change on Salmonella Infections." Foodborne pathogens and disease 11.12 (2014): 974-980.
11. D'Souza, Rennie M et al. "Does ambient temperature affect foodborne disease?." Epidemiology 15.1 (2004): 86-92.
12. Kim, BI et al. "A conceptual model for the impact of climate change on fox rabies in Alaska, 1980–2010." Zoonoses and public health 61.1 (2014): 72-80.
13. Hueffer, Karsten et al. "Zoonotic infections in Alaska: disease prevalence, potential impact of climate change and recommended actions for earlier disease detection, research, prevention and control." International journal of circumpolar health 72 (2013).
14. Alexander, KA et al. "What factors might have led to the emergence of Ebola in West Africa?." PLOS Neglected Tropical Diseases, to appear. http://blogs.plos.org/speakingofmedicine/files/2014/11/Alexanderetal.pdf (2014).
15. Freifeld, Clark C et al. "HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports." Journal of the American Medical Informatics Association 15.2 (2008): 150-157.
16. Brownstein, JS, and CC Freifeld. "HealthMap: the development of automated real-time internet surveillance for epidemic intelligence." Euro Surveill 12.11 (2007): E071129.
17. Jagarlamudi, Jagadeesh, Hal Daumé III, and Raghavendra Udupa. "Incorporating lexical priors into topic models." Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics 23 Apr. 2012: 204-213.
18. Rekatsinas, Theodoros, Saurav Ghosh, Lise Getoor, Naren Ramakrishnan et al. "SourceSeer: Forecasting Rare Disease Outbreaks Using Multiple Data Sources." Proceedings of the SIAM International Conference on Data Mining (SDM'15), Vancouver, BC, Apr-May 2015.

19. Chakraborty, Prithwish, Naren Ramakrishnan et al. "Forecasting a moving target: Ensemble models for ILI case count predictions." Proceedings of the 2014 SIAM International Conference on Data Mining. Proceedings. Society for Industrial and Applied Mathematics 2014: 262-270.
20. Wang, Zheng, Prithwish Chakraborty, Naren Ramakrishnan et al. "Dynamic Poisson Autoregression for Influenza-Like-Illness Case Count Prediction." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 10 Aug. 2015: 1285-1294.
21. Fu, Xiuju et al. "Time-series infectious disease data analysis using SVM and genetic algorithm." Evolutionary Computation, 2007. CEC 2007. IEEE Congress on 25 Sep. 2007: 1276-1280.
22. Fuller, Trevon et al. "Identifying areas with a high risk of human infection with the avian influenza A (H7N9) virus in East Asia." Journal of Infection 69.2 (2014): 174-181.
23. Ramakrishnan, Naren et al. "'Beating the news' with EMBERS: forecasting civil unrest using open source indicators." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining 24 Aug. 2014: 1799-1808.
24. Doyle, Andy et al. "Forecasting Significant Societal Events Using The Embers Streaming Predictive Analytics System." Big data 2.4 (2014): 185-195.
25. Kleist, Daryl T et al. "Introduction of the GSI into the NCEP global data assimilation system." Weather and Forecasting 24.6 (2009): 1691-1705.
26. Apaydin, Halit, F Kemal Sonmez, and Y Ersoy Yildirim. "Spatial interpolation techniques for climate data in the GAP region in Turkey." Climate Research 28.1 (2004): 31-40.
27. Ozelkan, Emre et al. "Spatial interpolation of climatic variables using land surface temperature and modified inverse distance weighting." International Journal of Remote Sensing 36.4 (2015): 1000-1025.
28. Lloyd, CD. "Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain." Journal of Hydrology 308.1 (2005): 128-150.
29. Avian Influenza A (H7N9) Virus | Avian Influenza (Flu) , Centers for Disease Control and Prevention (www.cdc.gov).
30. Gleick, Peter H. "Water, drought, climate change, and conflict in Syria." Weather, Climate, and Society 6.3 (2014): 331-340.
31. Kelley, Colin P et al. "Climate change in the Fertile Crescent and implications of the recent Syrian drought." Proceedings of the National Academy of Sciences 112.11 (2015): 3241-3246
32. Hsiang, Solomon M, Marshall Burke, and Edward Miguel. "Quantifying the influence of climate on human conflict." Science 341.6151 (2013): 1235367.
33. Burke, Marshall, Solomon M Hsiang, and Edward Miguel. "Climate and Conflict." 16 Oct. 2014.
34. Devlin, Colleen, and Cullen S Hendrix. "Trends and triggers redux: Climate change, rainfall, and interstate conflict." Political Geography 43 (2014): 27-39.
35. Wischnath, Gerdis, and Halvard Buhaug. "On climate variability and civil war in Asia." Climatic Change 122 (2014): 709-721.
36. Hsiang SM, Meng KC, Cane MA (2011) Civil conflicts are associated with global climate. Nature 476:438-441.