Big Data Science

Course: 18-788

Patrick McSharry
patrick@mcsharry.net
www.mcsharry.net
Twitter: @patrickmcsharry

Spring 2023

ICT Center of Excellence
Carnegie Mellon University

Big Data Science

WEEK 2A

Assignment 1

Feature \ Model	Linear	Quadratic	Cubic
Rainfall(t)			
Rainfall(t-k*)			
SMA(Rainfall, k')			

Objective:
Use information available in month t to forecast vegetation in month t

Feature construction and selection:
Optimize the delay k*
Optimize the moving average window k'

Model selection:
Identify the best model structure (R2, Adjusted R2 and MSE)
Consider cross-validation to avoid over-fitting.

Course outline

Week	Lecture A	Lecture B
1	Weather & agriculture	Climate change
2	Climate scenarios	Catastrophe models
3	Social trends	Finance
4	Sentiment analysis	Health
5	Telemedicine	Mobile data
6	Data4Dev	Socioeconomic status

Today’s Lecture

No.	Activity	Description	Time
1	Challenge	Economic impacts	10
2	Discussion	Weather dependence	10
3	Case study	Weather derivatives	10
4	Analysis	Tourism climatic index	20
5	Demo	Tourist scenarios	20
6	Q&A	Questions and feedback	10

Economic impacts

- An economic impact analysis (EIA) examines the effect of an event on the economy in a specified area, ranging from a single neighborhood to the entire globe.
- It usually measures changes in business revenue, business profits, personal wages, and/or jobs.

Weather dependence

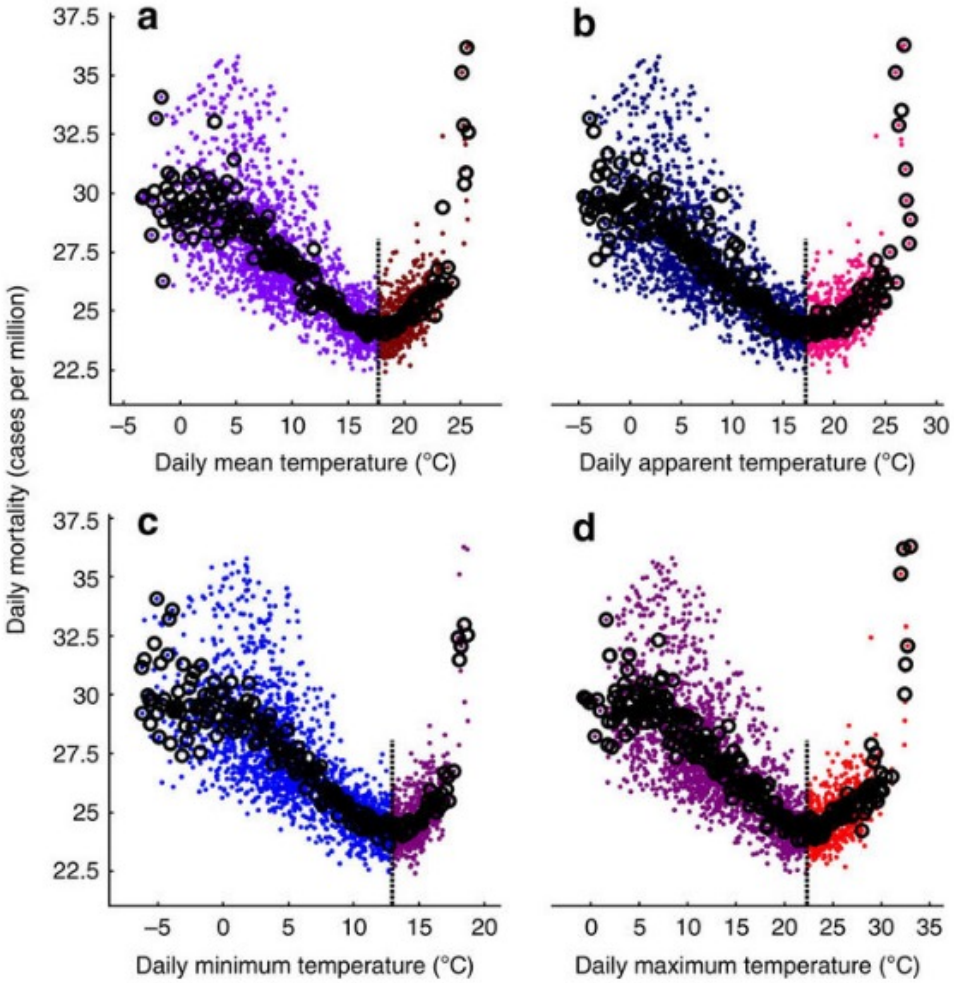
- Climate change scenarios can be generated to assess impact on a particular sector via:
- (1) **Data-driven** mapping (using machine learning) of weather data to economic activity
- (2) **Hybrid approach** – construct an index and relate this to economic activity
- (3) **Risk maps** may be determined from the index alone and used to make decisions

Mortality Quiz

- Which relationship best describes the dependence of mortality on temperature?
  a) Flat line
  b) Linear
  c) V-Shaped
  d) Inverted V-shaped

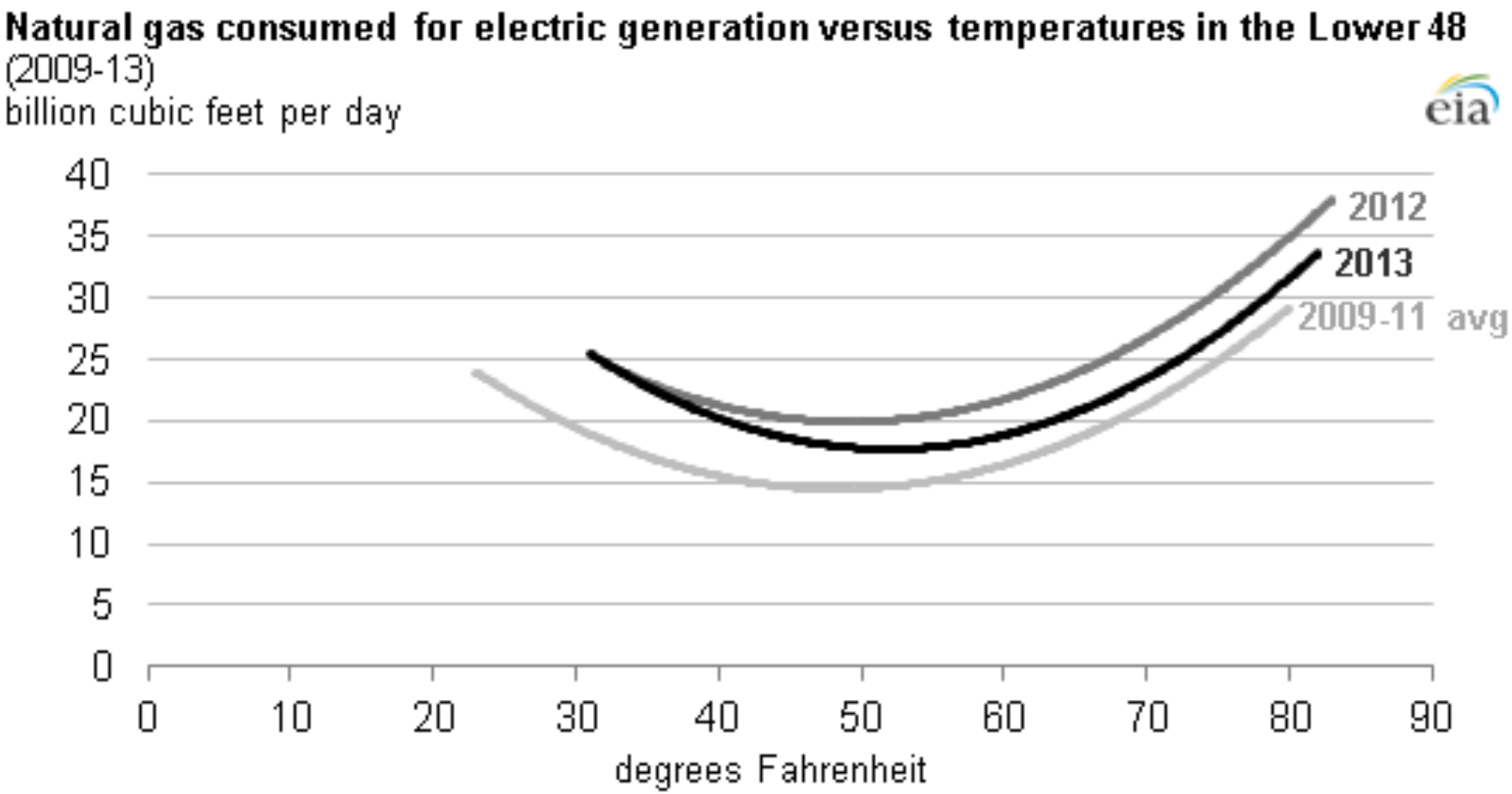
- **Slido.com #191830**

European mortality (1998-2003)

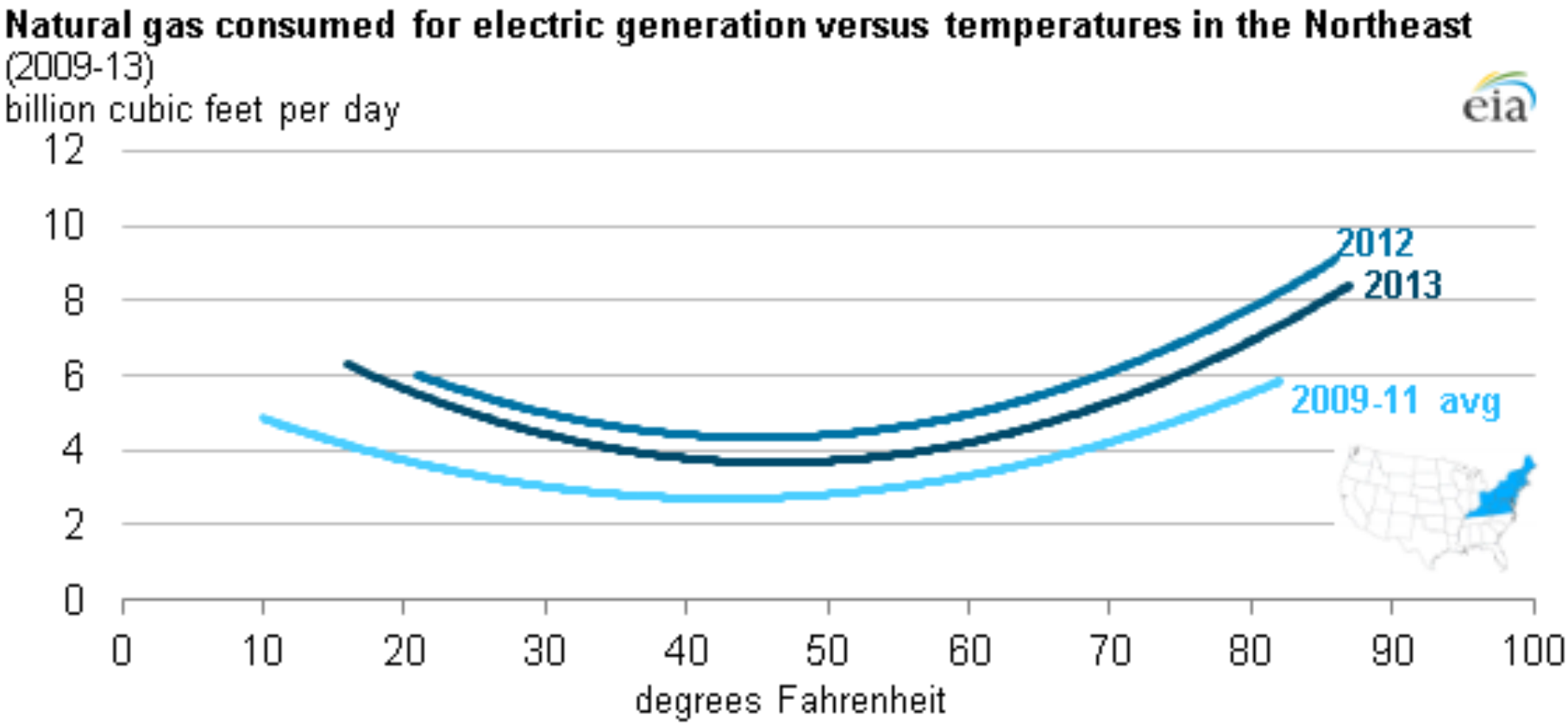


Ballester et al. (2011). Nature Communications, volume 2, 358.

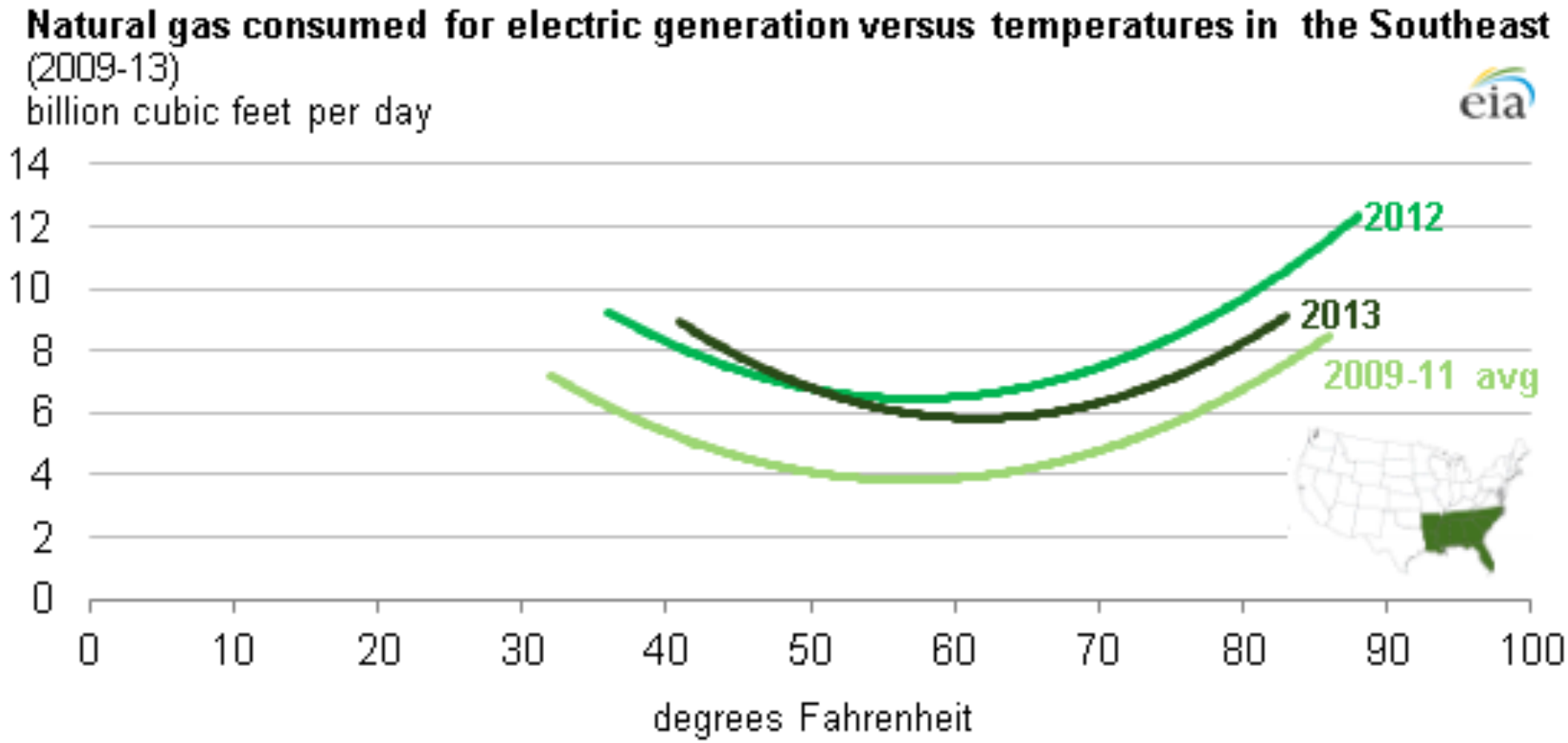
Natural Gas (lower 48 states)



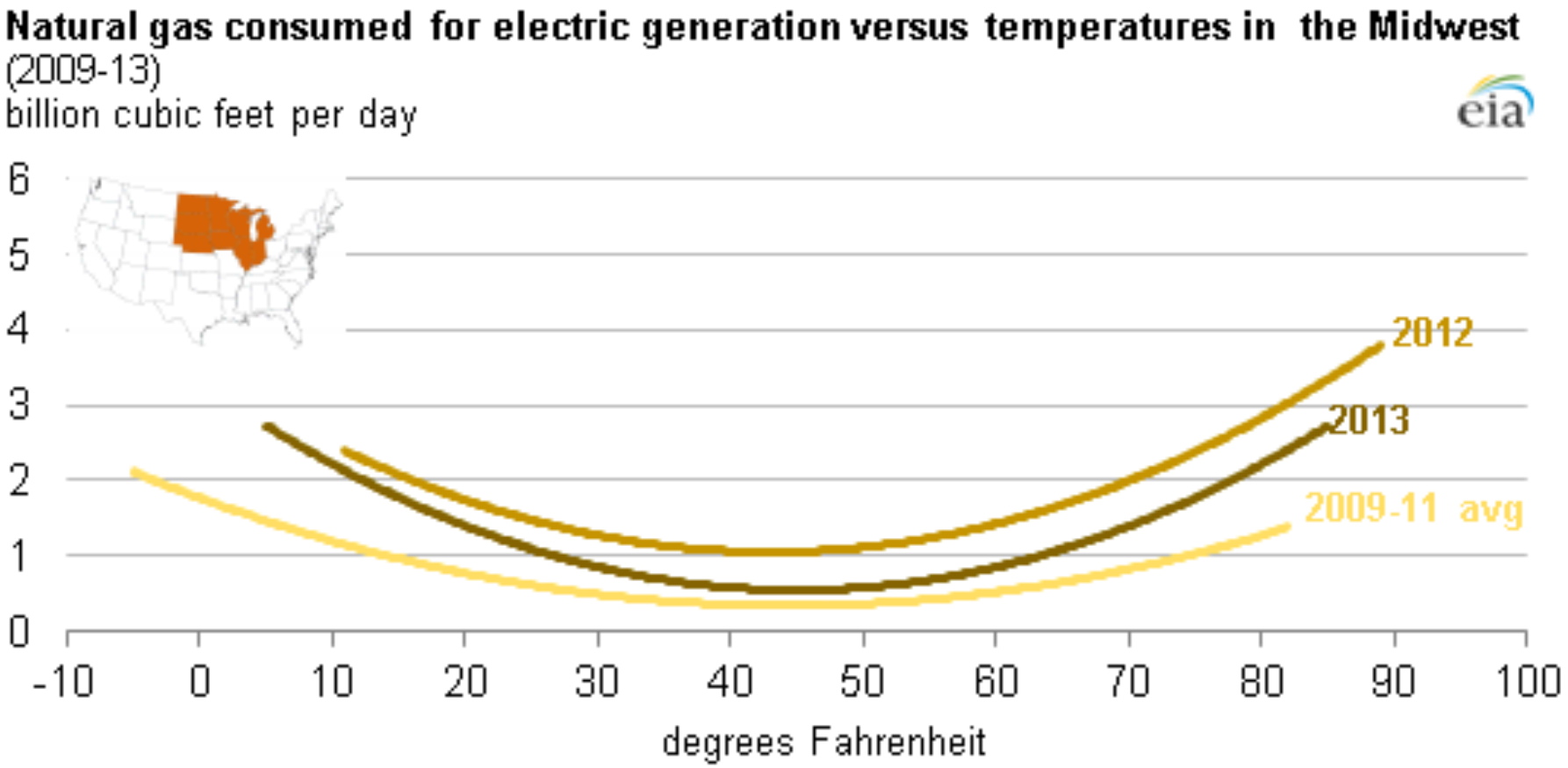
Natural Gas (NE)



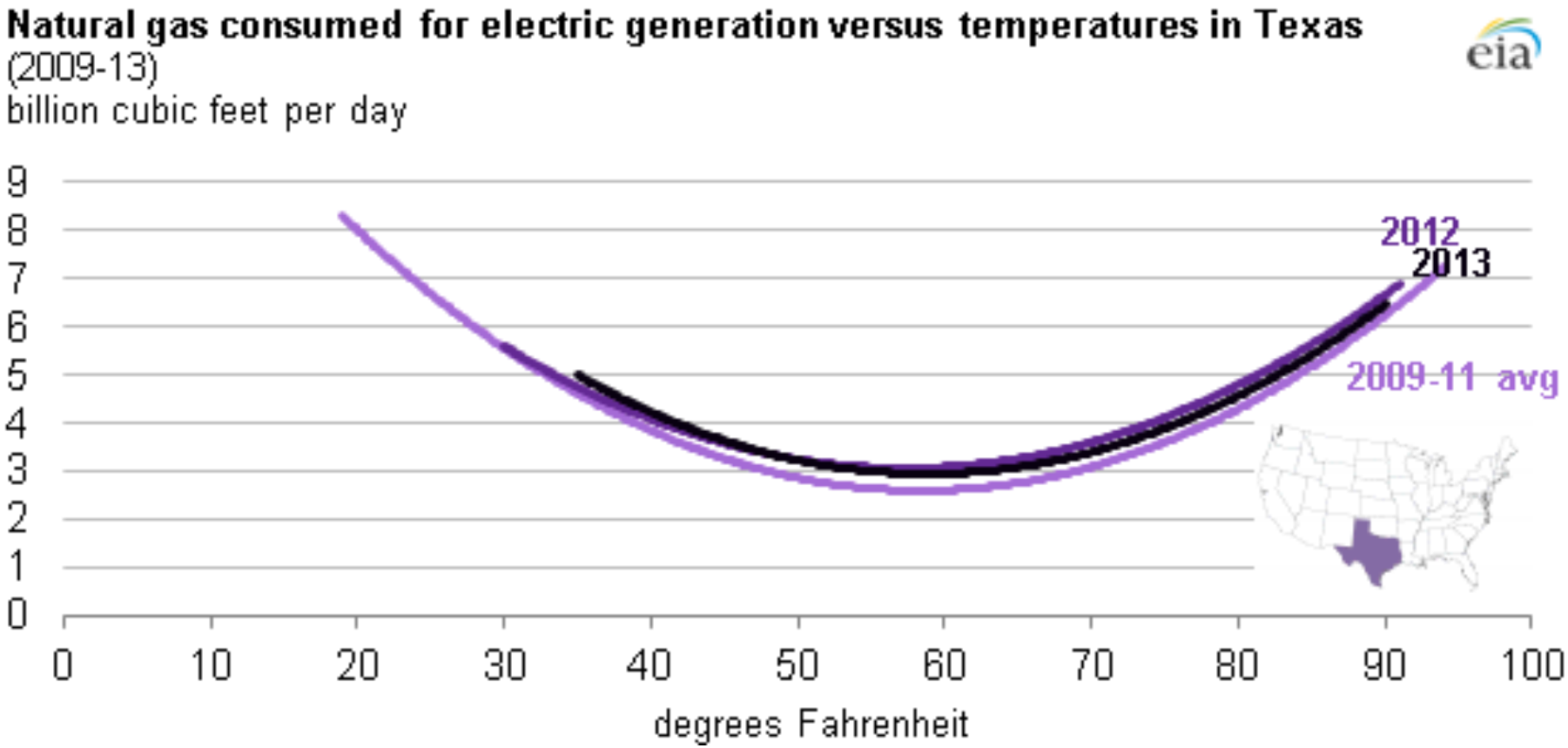
Natural Gas (SW)



Natural Gas (MW)

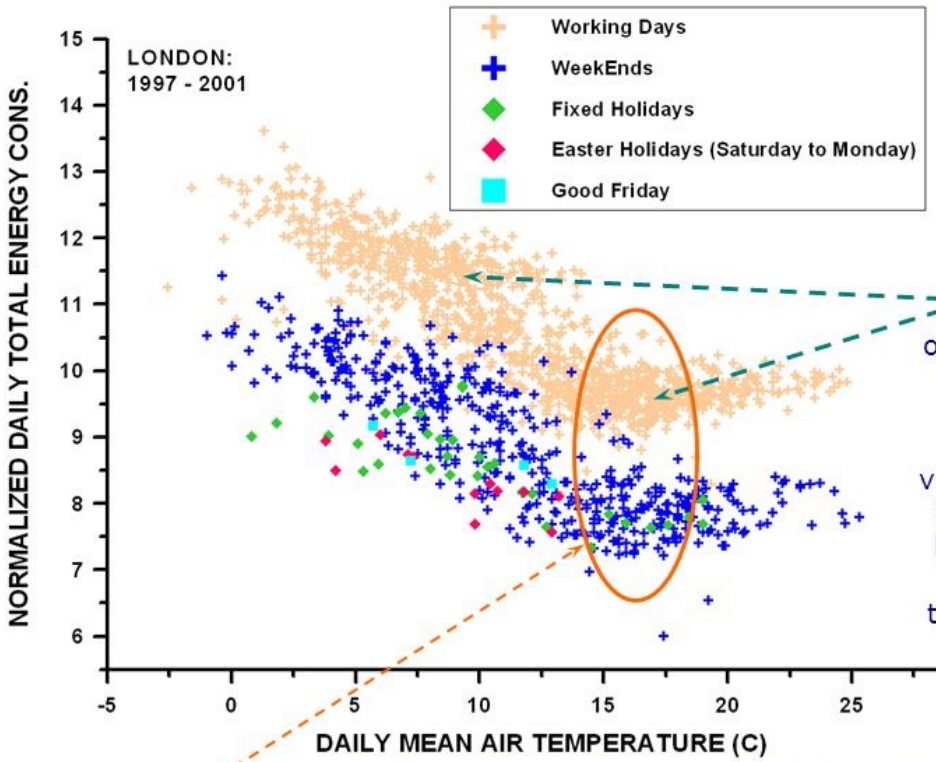


Natural Gas (Texas)



Temperature and Energy

Relationship between Energy Consumption and Air Temperature, for London, UK.

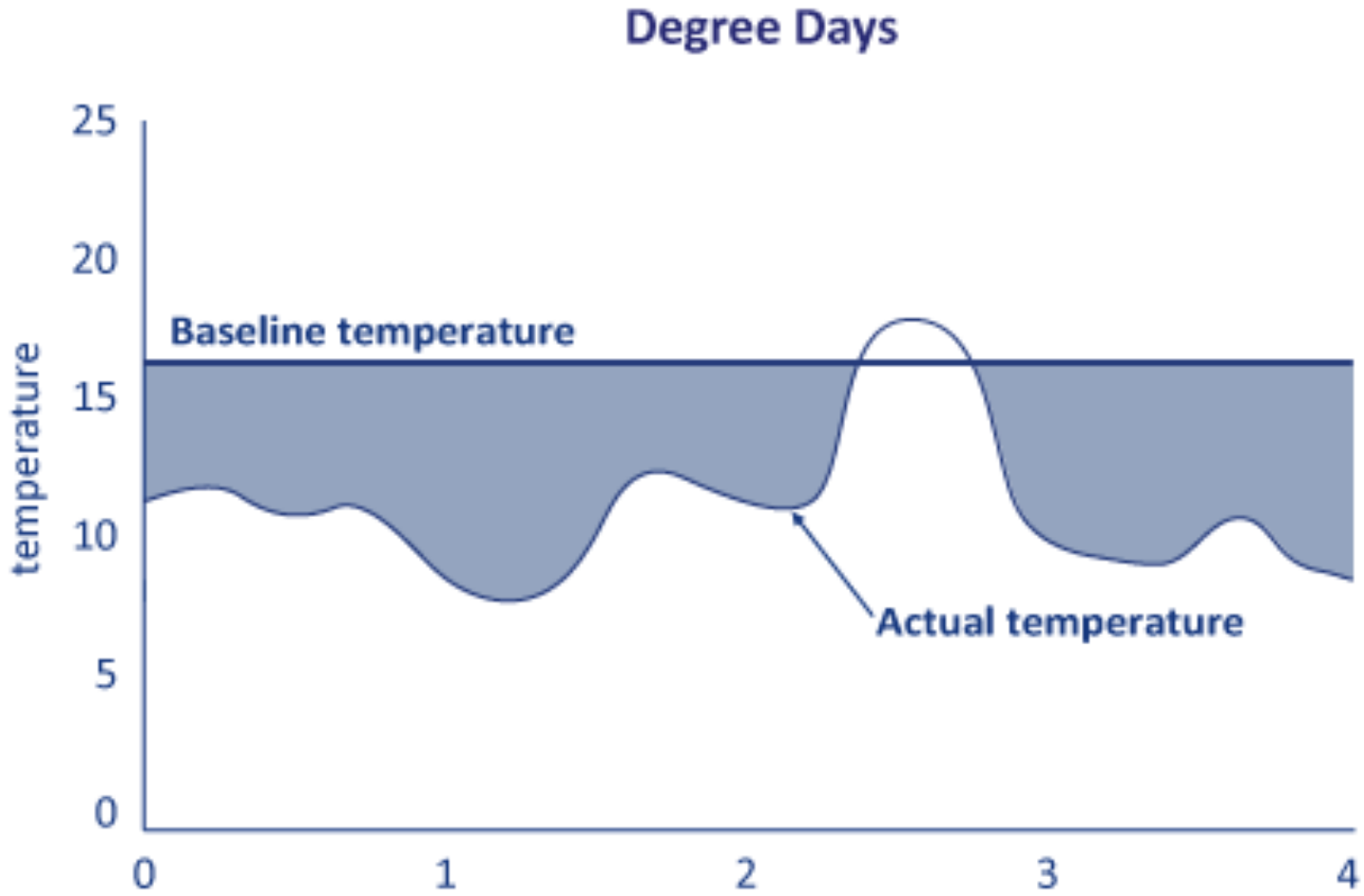


London, UK, also exhibits a NON linear relationship between air temperature and energy consumption BUT with only one minimum and one well defined maximum.

The minimum values of energy cons. appear to be around 16°C and this is the temper. for CDD calculations.

Above 16°C , energy levels for London tend to rise slightly but below this temperature energy levels increase significantly (space-heating needs).

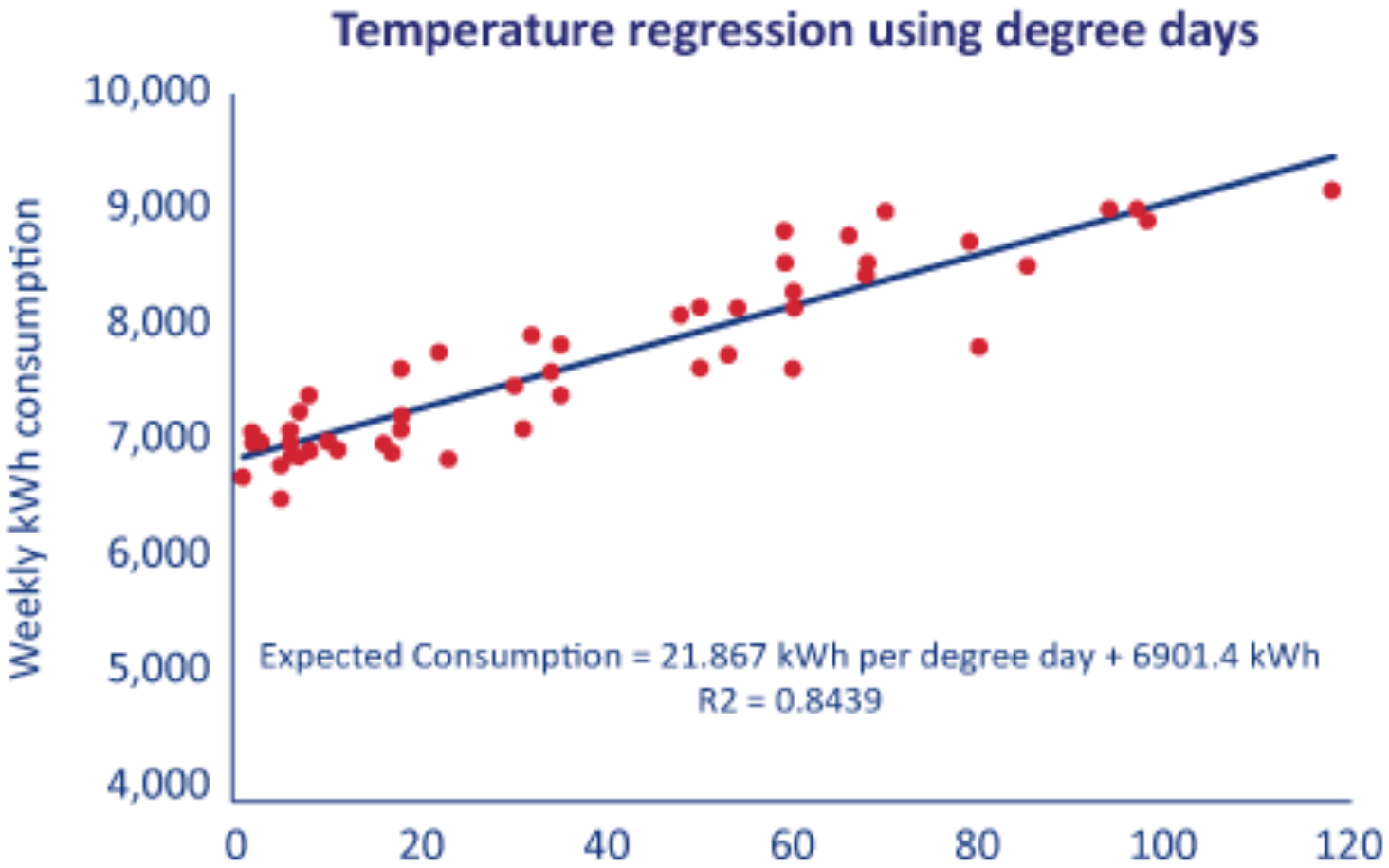
Degree Days



Baseline Temperature: 15.5°C

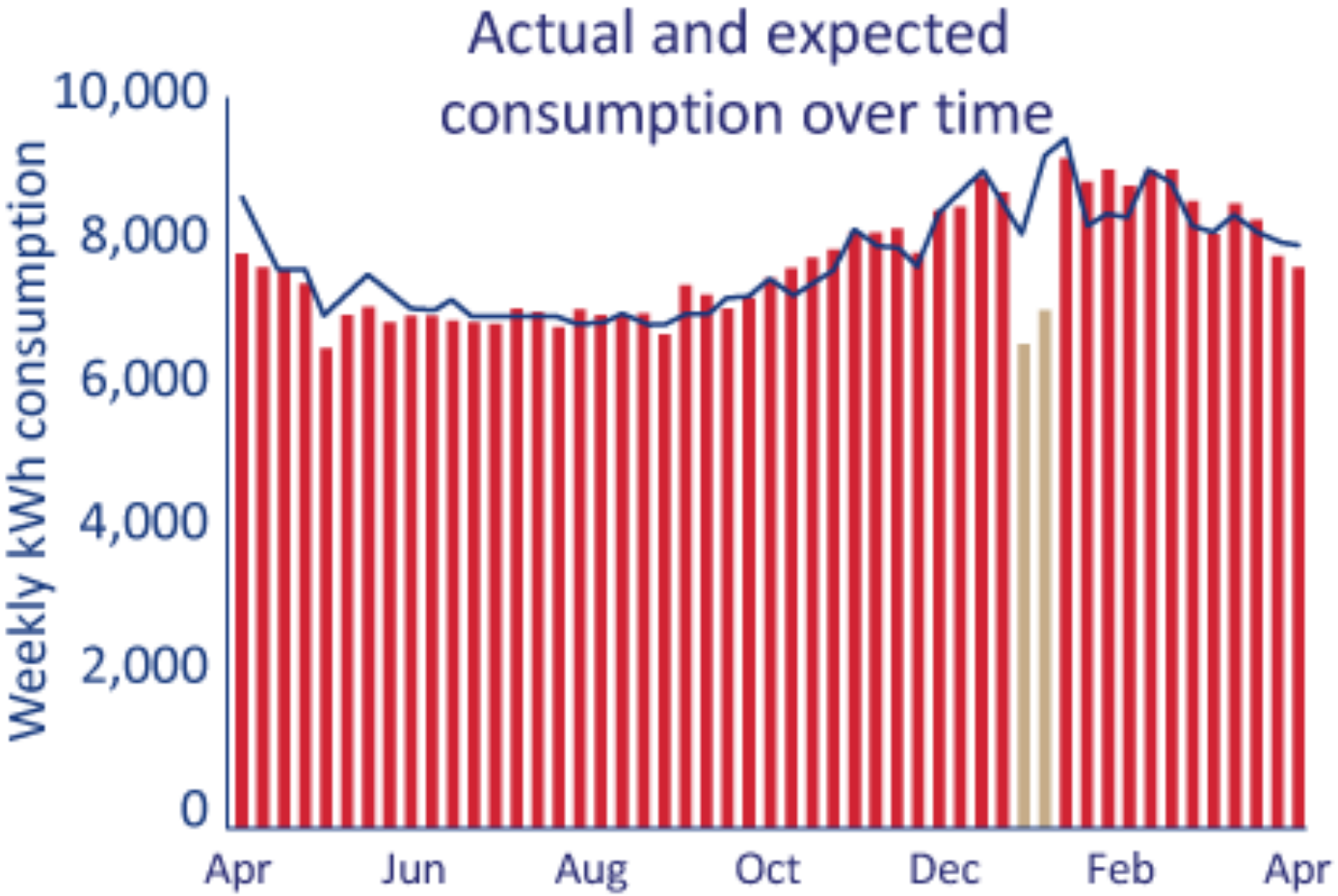
Source: www.eevs.co.uk/degreedays.html

Regression using Degree Days



Source: www.eevs.co.uk/degreedays.html

Forecasting consumption

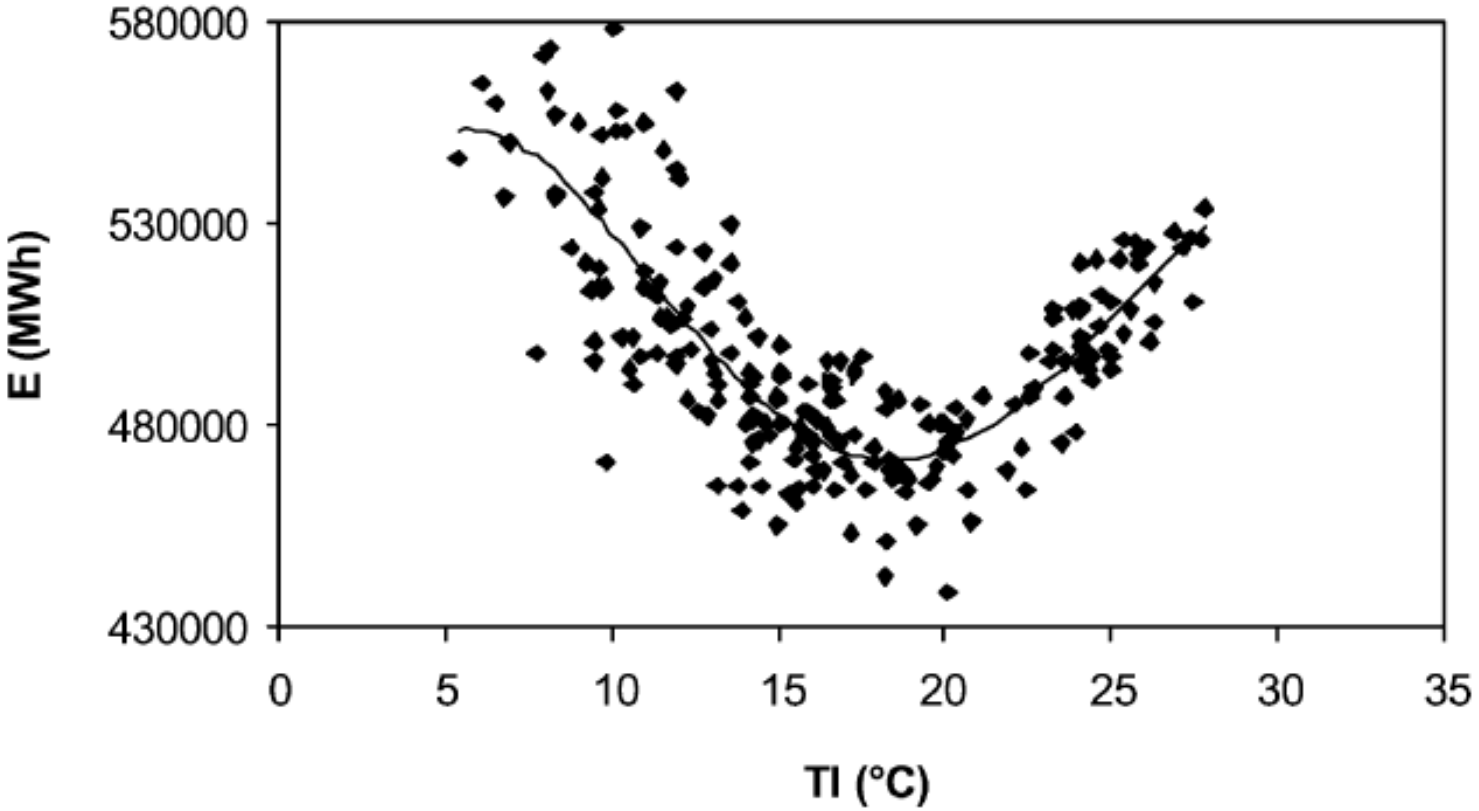


Source: www.eevs.co.uk/degreedays.html

Weather risk

- Fluctuations in the weather affect the revenues of many sectors (agriculture, energy, retail, transportation and construction)
- Weather risk tends to influence demand rather than price and adjustments of the latter are rarely adequate to compensate for lost revenues
- Insurance offers cover for low-probability extreme events but it does not provide compensation for the reduced demand that may result from slight variations in the weather such as the temperature being warmer or colder than expected.

Electricity load and temperature



- Pardo et al. (2002) analysed the dependence of electricity load on a population weighted temperature index in Spain

Electricity and temperature

- Electricity demand displays a non-linear V-shaped dependence on temperature
- Demand increases both for decreasing and increasing temperatures
- This results from the use of electric heating appliances in winter and air conditioners in summer
- There is a minimum around 18^{o}C where the demand is inelastic to temperature changes

Summer and winter regimes

- The nonlinearity in the response function prompted a separation of the effect into summer and winter regimes
- Typically 18^{o}C (or 65 ^{o}F) is used as a threshold variable to switch between summer and winter regimes
- Within each regime, the response is approximately linear which facilitates traditional linear time series analysis

Heating and cooling degree-days

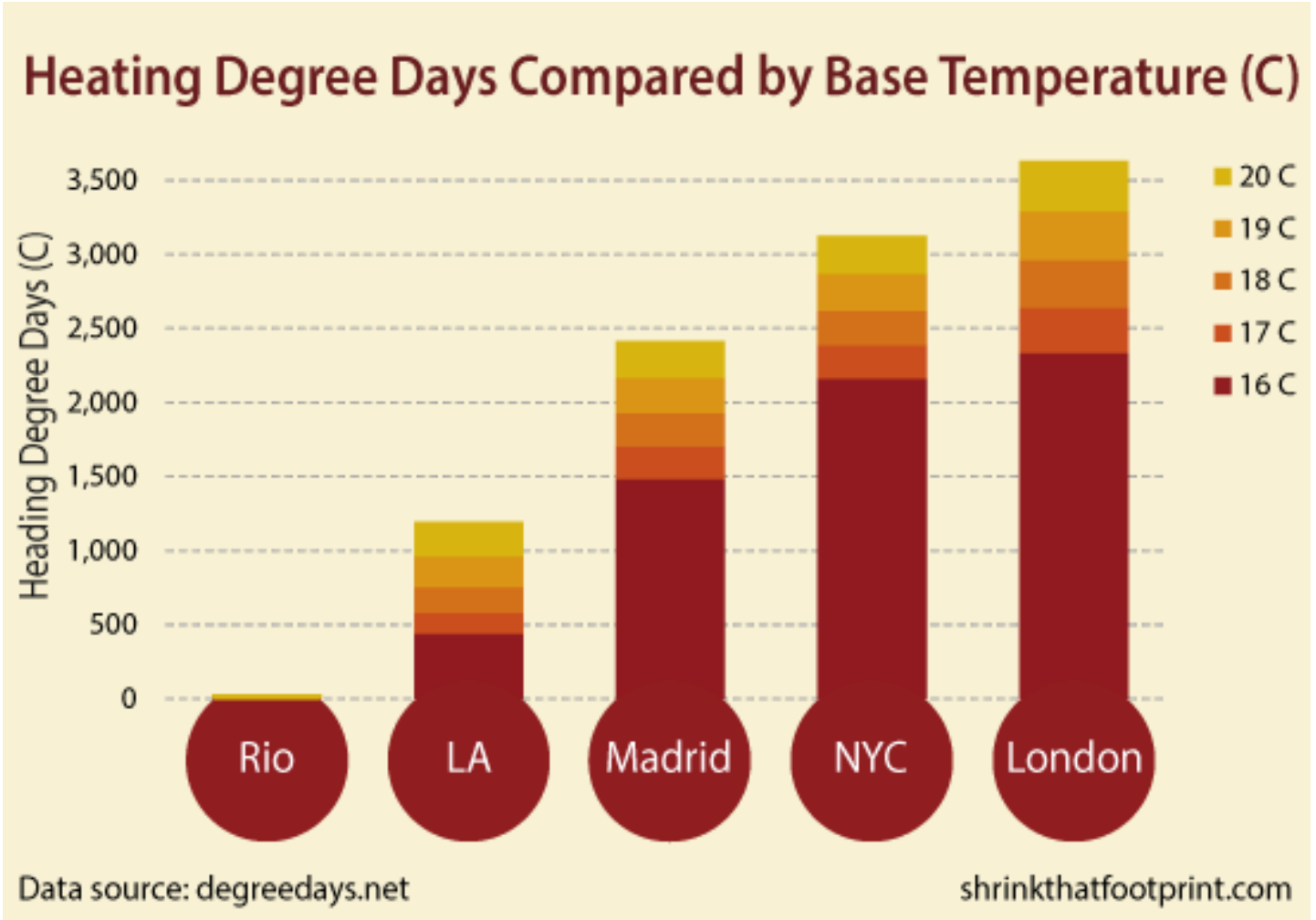
- Heating degree-day (HDD) and cooling degree-day (CDD) are indices designed to reflect the demand for energy needed to heat or cool:

$$\mathrm{HDD}_t = \max(18 - T_t, 0)$$

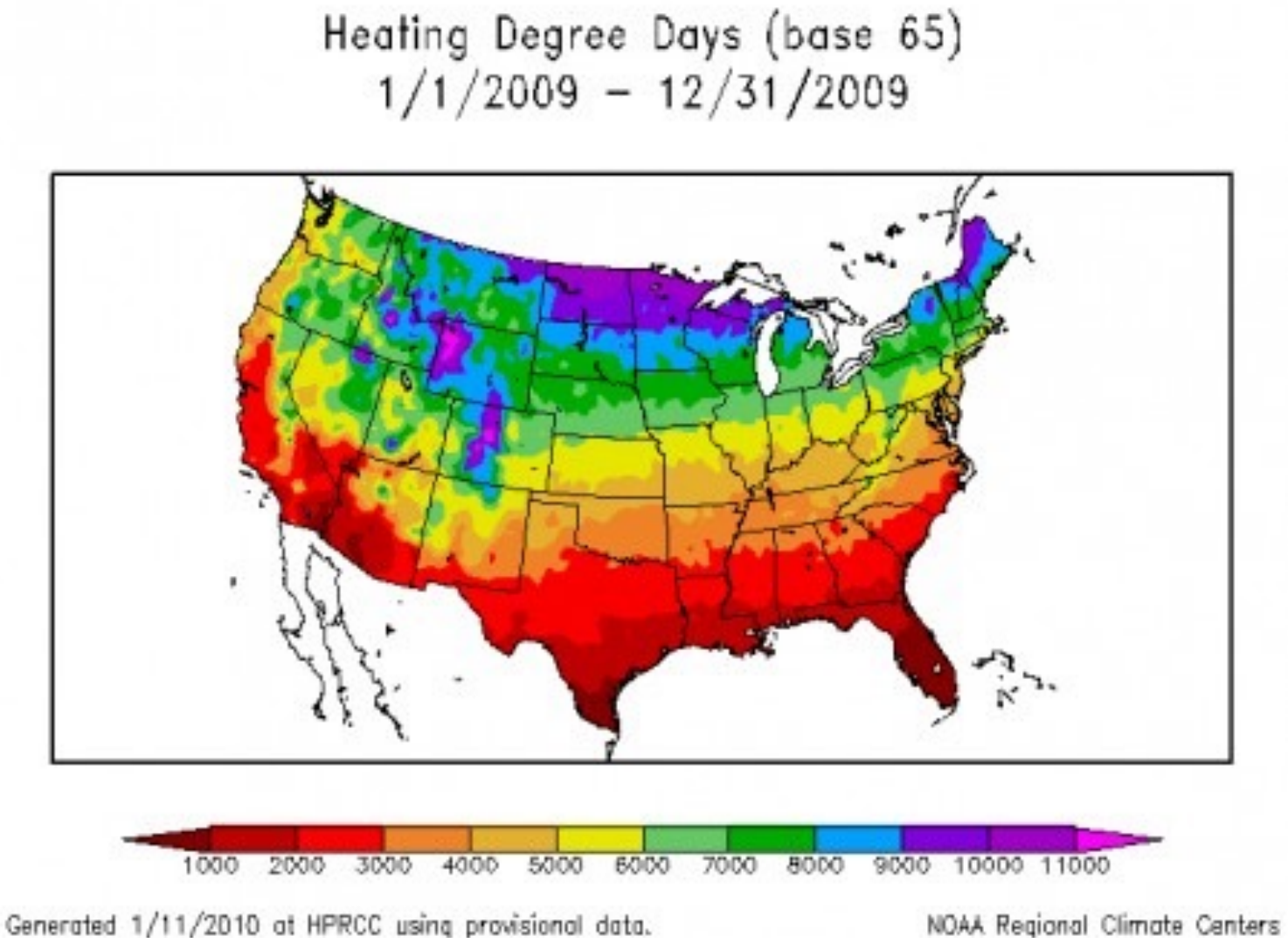
$$\mathrm{CDD}_t = \max(T_t - 18, 0)$$

- HDD measures the intensity and duration of cold in the winter
- CDD measures the intensity and duration of heat in the summer

Heating Degree Days



HDDs over the US



Source: http://snr.unl.edu/data/climate

Weather derivative quiz

- In contrast to insurance, weather derivatives protect against weather risk by covering:
  a) low-risk low-probability events
  b) low-risk high-probability events
  c) high-risk low-probability events
  d) high-risk high-probability events

- **Slido.com #191830**

Weather derivatives

- Weather derivatives are financial instruments that can form a risk management strategy to mitigate risk associated with adverse or unexpected weather conditions
- Unlike other derivatives the underlying asset (temperature, rainfall, wind, frost or snow) has no direct value
- Weather derivatives, in contrast to insurance, provide protection against low-risk high-probability events by treating the weather as a tradable commodity, similar to a stock price or interest rate

www.cme.com/trading/prd/weather/

Weather and revenues

- It is estimated that nearly 20% of the U.S. economy is directly affected by the weather, and the profitability and revenues of virtually every industry—agriculture, energy, entertainment, construction, travel, and others—depend to a great extent on the vagaries of temperature, rainfall, and storms.
- By creating a weather index that can be linked to the revenues of a particular organisation, it is then possible to trade the weather in order to mitigate the risk of adverse weather conditions
- Advantages of decreased earnings volatility are efficient use of equity, improving the company value to stakeholders and availability of lower debt costs and higher advance rates
- The Chicago Mercantile Exchange (CME) introduced exchange-traded weather derivatives in 1999 and now provides standardised contracts for 47 cities in the United States, Europe, Canada, Australia and Asia.

Weather and Financial Risk

Risk Holder	Weather Type	Risk
Energy Industry	Temperature	Lower sales during warm winters or cool summers
Energy Consumers	Temperature	Higher heating/cooling costs during cold winters and hot summers
Beverage Producers	Temperature	Lower sales during cool summers
Building Material Companies	Temperature/Snowfall	Lower sales during severe winters (construction sites shut down)
Construction Companies	Temperature/Snowfall	Delays in meeting schedules during periods of poor weather
Ski Resorts	Snowfall	Lower revenue during winters with below-average snowfall
Agricultural Industry	Temperature/Snowfall	Significant crop losses due to extreme temperatures or rainfall
Municipal Governments	Snowfall	Higher snow removal costs during winters with above-average snowfall
Road Salt Companies	Snowfall	Lower revenues during low snowfall winters
Hydro-electric power generation	Precipitation	Lower revenue during periods of drought

Source: www.climetrix.com

Derivative pricing

- Approaches employed for pricing weather derivatives vary greatly, reflecting different assumptions made about the variability of future weather conditions over the duration of the contract
- The disparity results from the availability of a range of atmospheric and statistical models
- Further complications arise from the inclusion of the effect of El Niño in the model.

Energy utility case study

- If an energy utility believes that November may be hotter than usual, possibly leading to reduced revenues, it could take out a HDD swap with a reference of 18^{o}C.
- Suppose the average temperature on each day was 19^{o}C, the utility would receive 30 (30 x 1) the agreed sum of money for each degree-day (compensation for the lost revenue due to a warm winter).
- Alternatively if the actual average temperature on each day was 16^{o}C, the utility would pay 60 (30 x 2) times the agreed sum of money for each degree-day (but would also benefit from the revenue generated by the cold winter).
- In this case, the utility is essentially selling the HDD index using a HDD swap or a HDD put option.

https://www.cmegroup.com/education/articles-and-reports/managing-climate-risk-with-cme-group-weather-futures-and-options.html

Restaurant case study

- In May 2002, Element Re announced a new weather derivative transaction with The Rock Garden, a London restaurant.
- The contract was designed to protect the restaurant against financial loss from colder-than-normal weather from March 1-June 30.
- The contract will pay out if the maximum daily temperature is less than a pre-agreed level for that month for more than a specified number of days (8.5^{o}C in March; 11^{o}C in April; 14.5^{o}C in May; 18^{o}C in June).

Agriculture

- Farmers are exposed to financial risk arising from the variability in crop yield that occurs due to different weather conditions.
- Crop yield and price are weather-dependent. In the case of corn, both temperature and precipitation provide significant explanatory variables.
- Drought in particular presents a substantial risk. Many farmers depend upon pre-financing against their future revenue streams for seed, fertilizer, and other agro chemical products.

Appetite for weather derivatives

- “I think it’s a perfect market. You can’t spook it, you can’t manipulate it. You can’t make people think it’s going to be 110 degrees in London next week,” he says. “And of course, weather is absolutely uncorrelated [to other asset classes].” - Peter Brewer, Cumulus Weather Fund
- No contract is too small. We’ve sold a weather derivative contract for a dollar” - David Friedberg, Weatherbill
- Weatherbill’s clients include car wash companies, hair salons and golf courses. “The other day one farmer rang me up and said his sows wouldn’t make a move to mate if the temperature went above 95 degrees Fahrenheit. He wanted to hedge against that.”

Source: FT, 20 June, 2007

Word Cloud

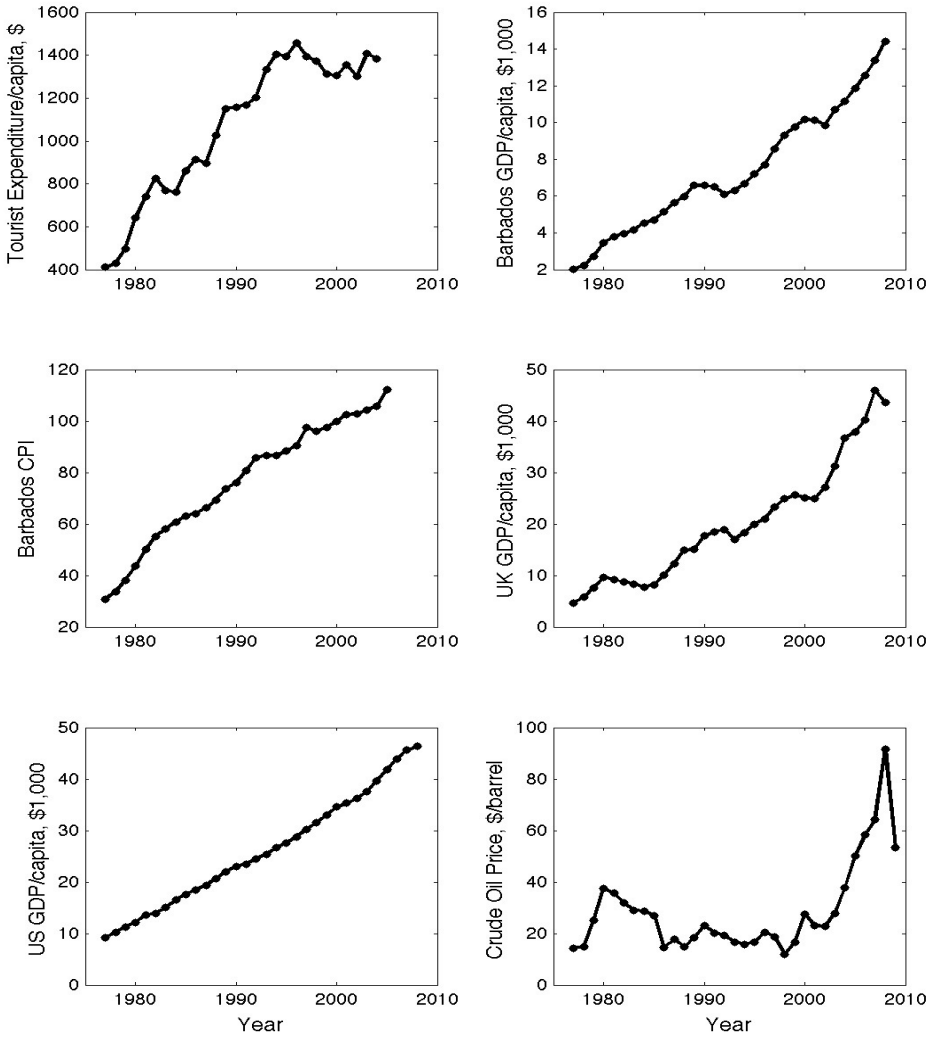
- Name a weather-dependent economic activity

- **Slido.com #191830**

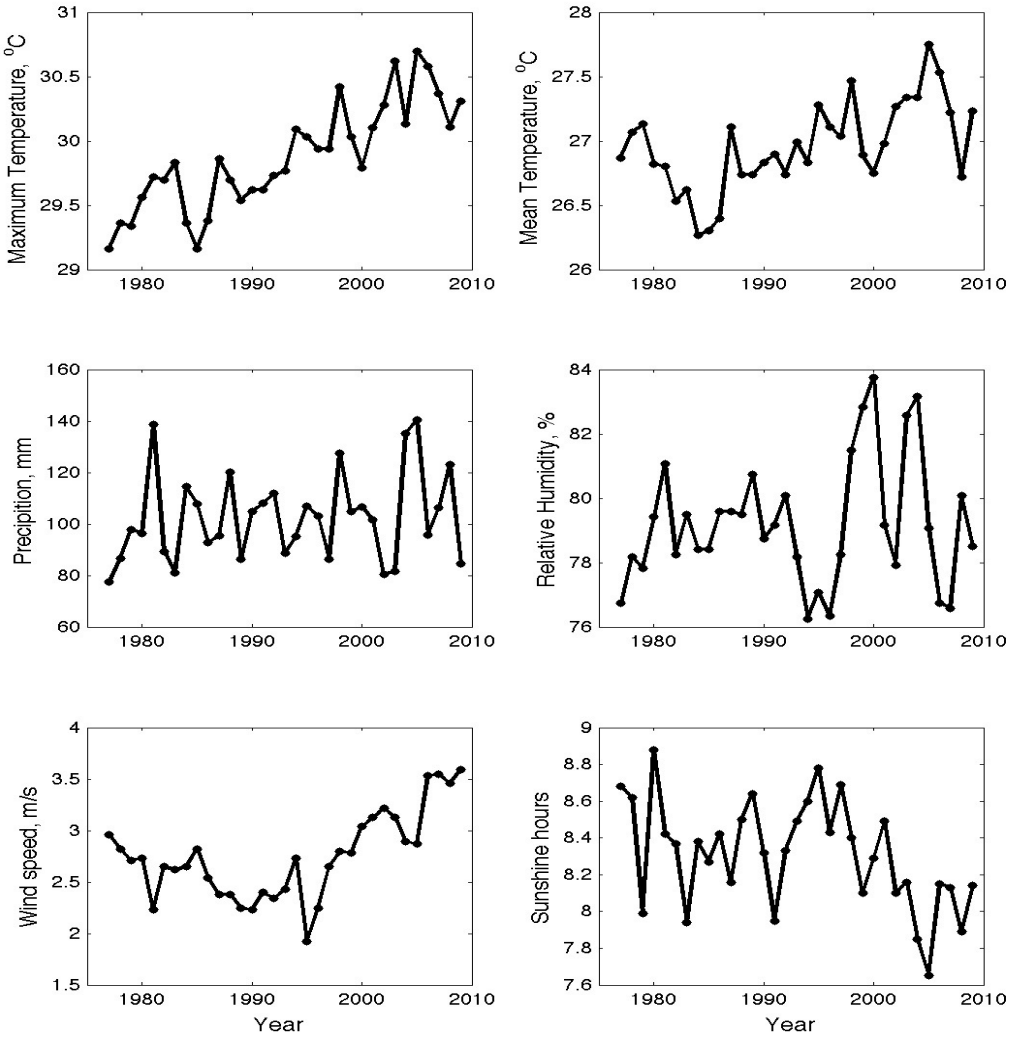
Climate and Tourism

- Impact of climate change on tourism
- Influence of weather variables on tourist arrivals
- Annual analysis is not sufficient since tourists plan trips on a monthly time scale
- Analysis of monthly weather data is a requirement
- Complex nonlinear dependencies between weather and tourism
- Lack of sufficient data to derive relationship between weather and tourism demand

Socio-economic time series



Climate variables



Methodology

- Occam's razor (principle of parsimony): seek the simplest model that explains the data in order to avoid over-fitting problems
- Use the tourism climate index (TCI) to encode the tourism monthly dependency on weather
- Construct a smoothed TCI since tourism demand will adjust to climate change over a few years
- The last five years were used to estimate a linear model in order to identify the climate change signal

Data analysis

- Insufficient evidence to use macro-economic variables, price of crude oil
- Nonstationarity and existence of structural changes present substanital challenges
- Future analysis would benefit from additional datasets with monthly resolution and across a group of similar countries

Tourism Climatic Index

- TCI requires observations of seven monthly weather variables:
- (1) maximum daily temperature (°C);
- (2) mean daily temperature (°C);
- (3) minimum daily relative humidity (%);
- (4) daily relative humidity (%);
- (5) precipitation (mm);
- (6) daily duration of sunshine (hours); and
- (7) wind speed (m/s).

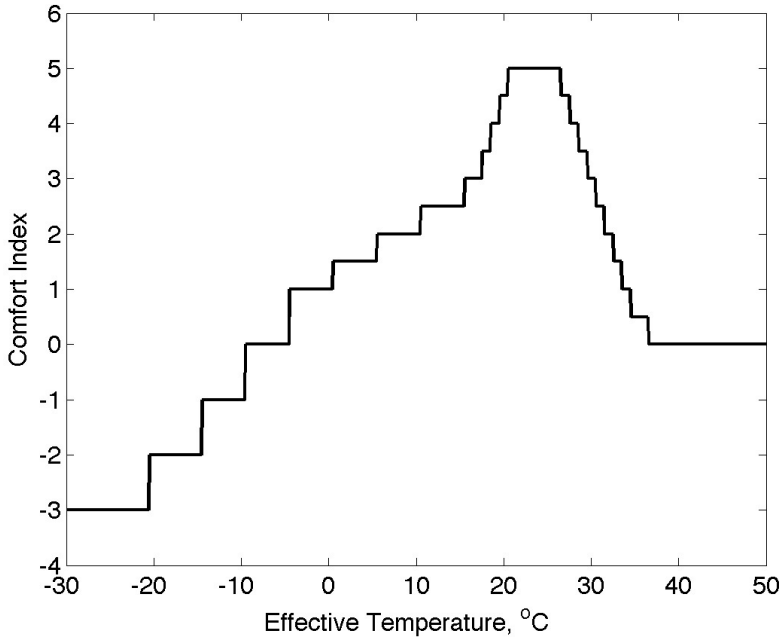
Source: Mieczkowski (1985)

Tourism Climatic Index

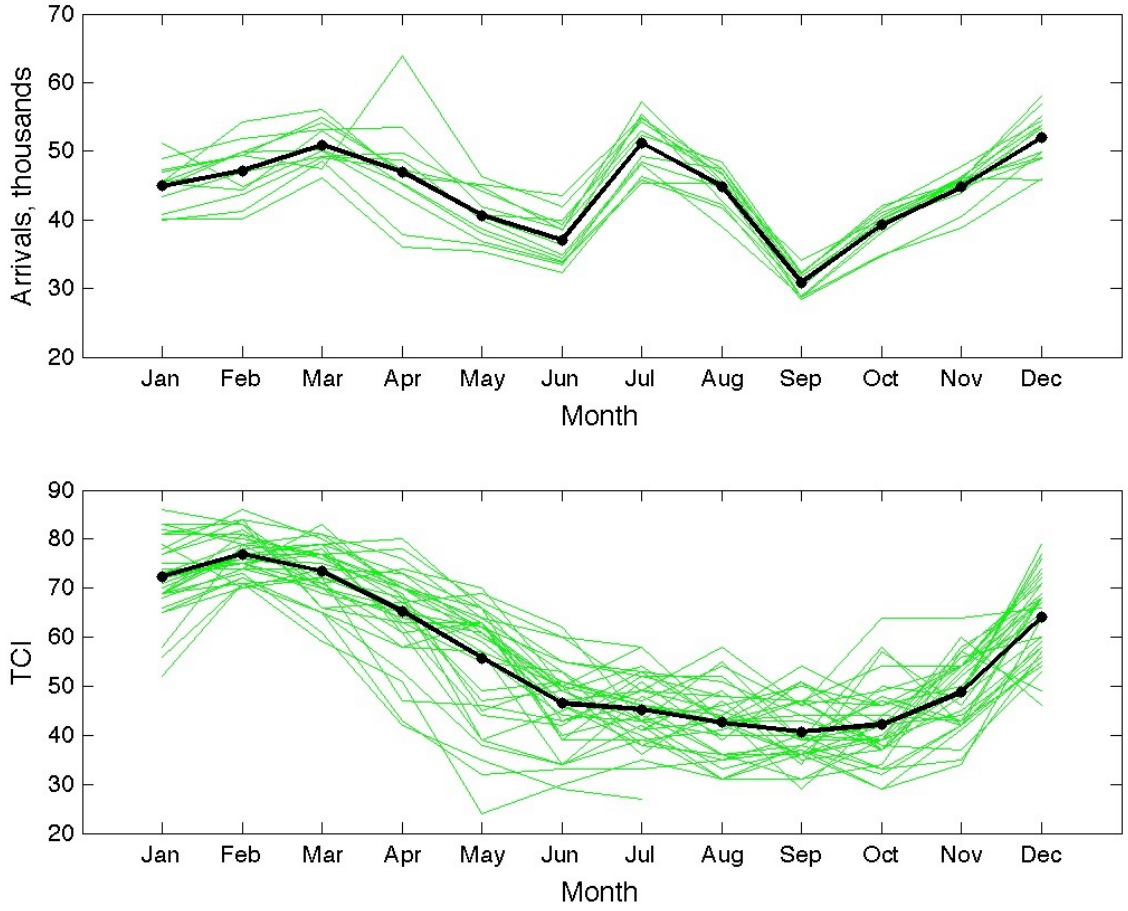
- The TCI is constructed from five sub-indices:

  $$TCI = 8CID + 2CIA + 4R + 4S + 2W,$$

- where
- CID = daytime comfort index,
- CIA = daily comfort index,
- R = precipitation,
- S = sunshine, and
- W = wind speed.



Tourist arrivals and TCI



Selected model

- Linear trend in monthly tourist arrivals

$$\frac{A_t^n}{A_0^n} = \alpha + \beta\left(\frac{STCI_t^t}{STCI_0^n}\right) + \varepsilon_t^n$$

- Estimation of only two parameters; both significant at $p<0.001$
- Background levels A_0^n and $STCIA_0^n$ employed data up to 2000

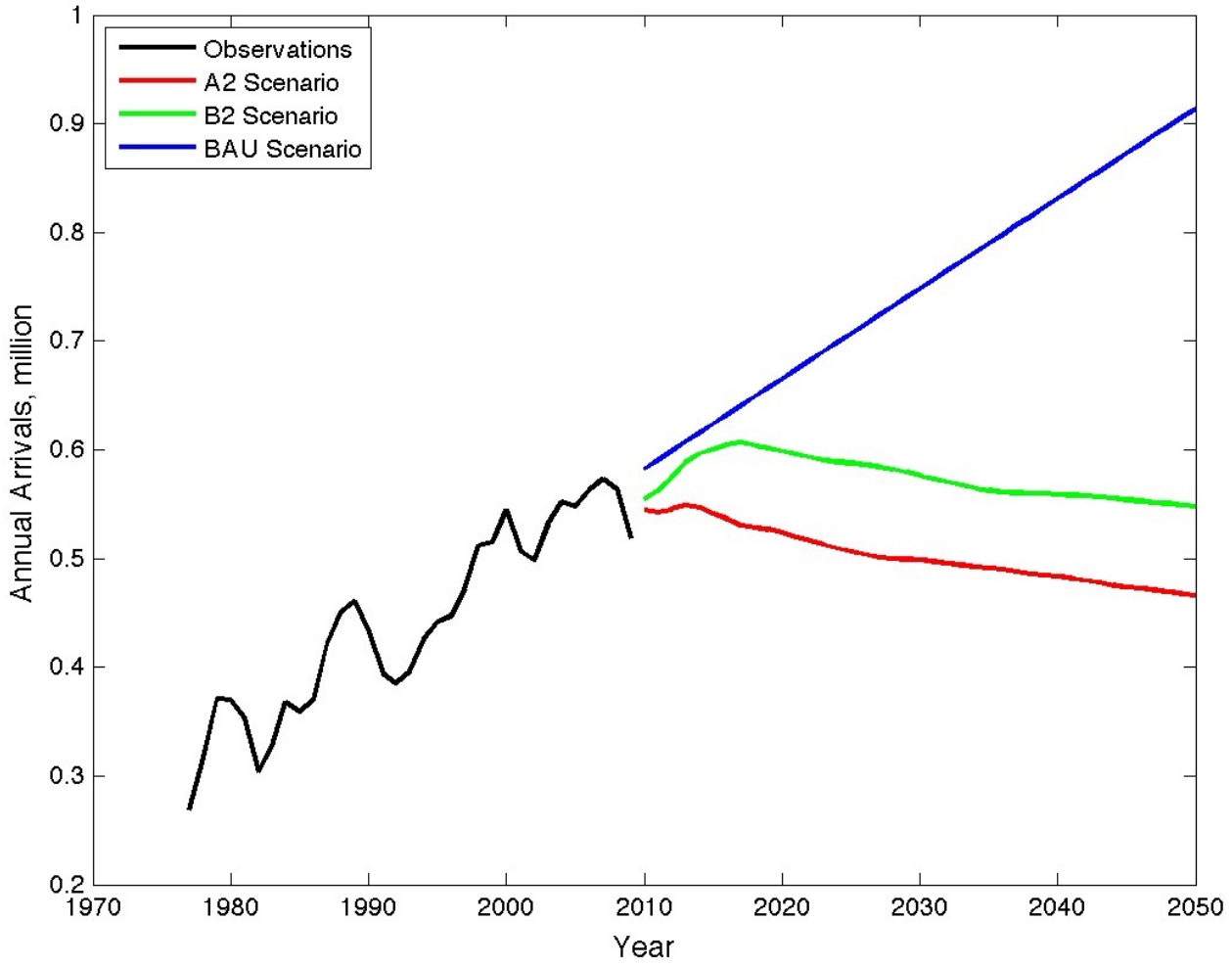
SRES Climate Scenarios

Variable \ Scenario	A2 Scenario	B2 Scenario
Temperature (□C)	+ 2.3	+ 1.84
Precipitation (mm)	- 24	- 6.4
Relative Humidity (%)	+ 0.6	- 0.4
Sunshine (hours)	+ 0.7	+ 0.6
Wind speed (m/s)	+ 0.6	- 0.12
Sea level (m)	+ 0.3	+ 0.1

•The business as usual (BAU) scenario was formed by forecasting future tourist arrivals using linear extrapolation

•The BAU serves as a benchmark for what would happen were climate change to have no impacts on tourism

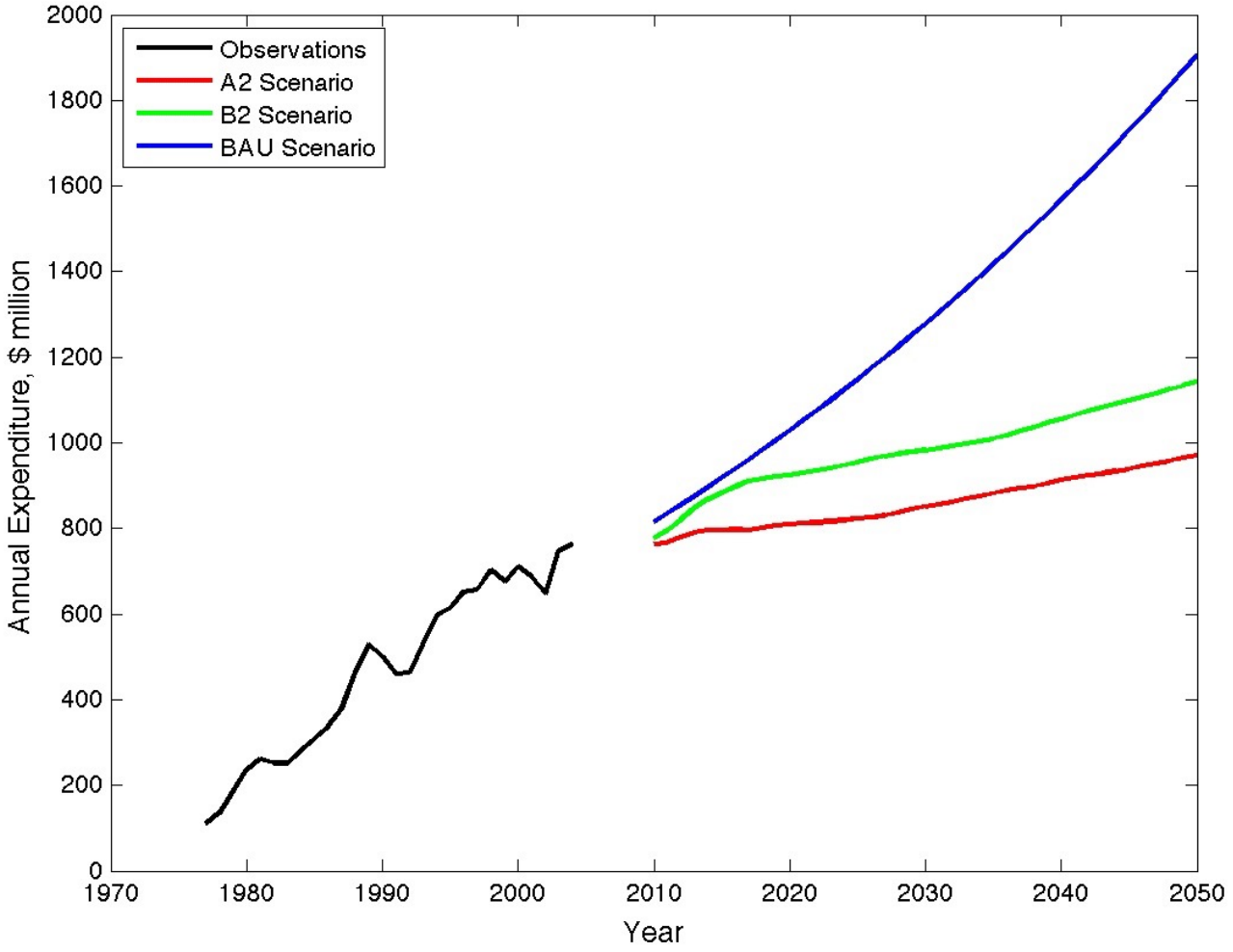
Tourist arrival scenarios



Summary of Scenarios

- B2 scenario suggests that annual arrivals will continue to increase until 2017, after which they decrease due to the adverse effect of climate change on the TCI values
- By comparison, the A2 scenario produces substantially greater changes to the climate variables, thereby reducing the number of tourist arrivals from 2013 onwards.

Tourist arrival expenditure



Assuming expenditure per tourist is $1,400 and an annual rate of inflation of 1%

Data Analytics

WEEK 2B

Course outline

Week	Lecture A	Lecture B
1	Weather & agriculture	Climate change
2	Climate scenarios	Catastrophe models
3	Social trends	Finance
4	Sentiment analysis	Health
5	Telemedicine	Mobile data
6	Data4Dev	Socioeconomic status

Today's Lecture

No.	Activity	Description	Time
1	Challenge	Financial losses	10
2	Discussion	CAT model	10
3	Case study	Hurricanes	10
4	Analysis	Exceedance probability	20
5	Demo	EP curves	20
6	Q&A	Questions and feedback	10

Catastrophe

- Catastrophe [Oxford English Dictionary]:
- An event causing great and usually sudden damage or suffering;
- a disaster:
- an environmental catastrophe
- Usage: inaction will only bring us closer to catastrophe

Insurance Poll

- List some catastrophes (known as perils) that the insurance industry might cover?

- **Slido.com #820813**

Perils

- Flood
- Windstorms (hurricanes, typhoons)
- Earthquakes
- Wildfires
- Landslides
- Hailstorm
- Terrorism

Flood risk assessment

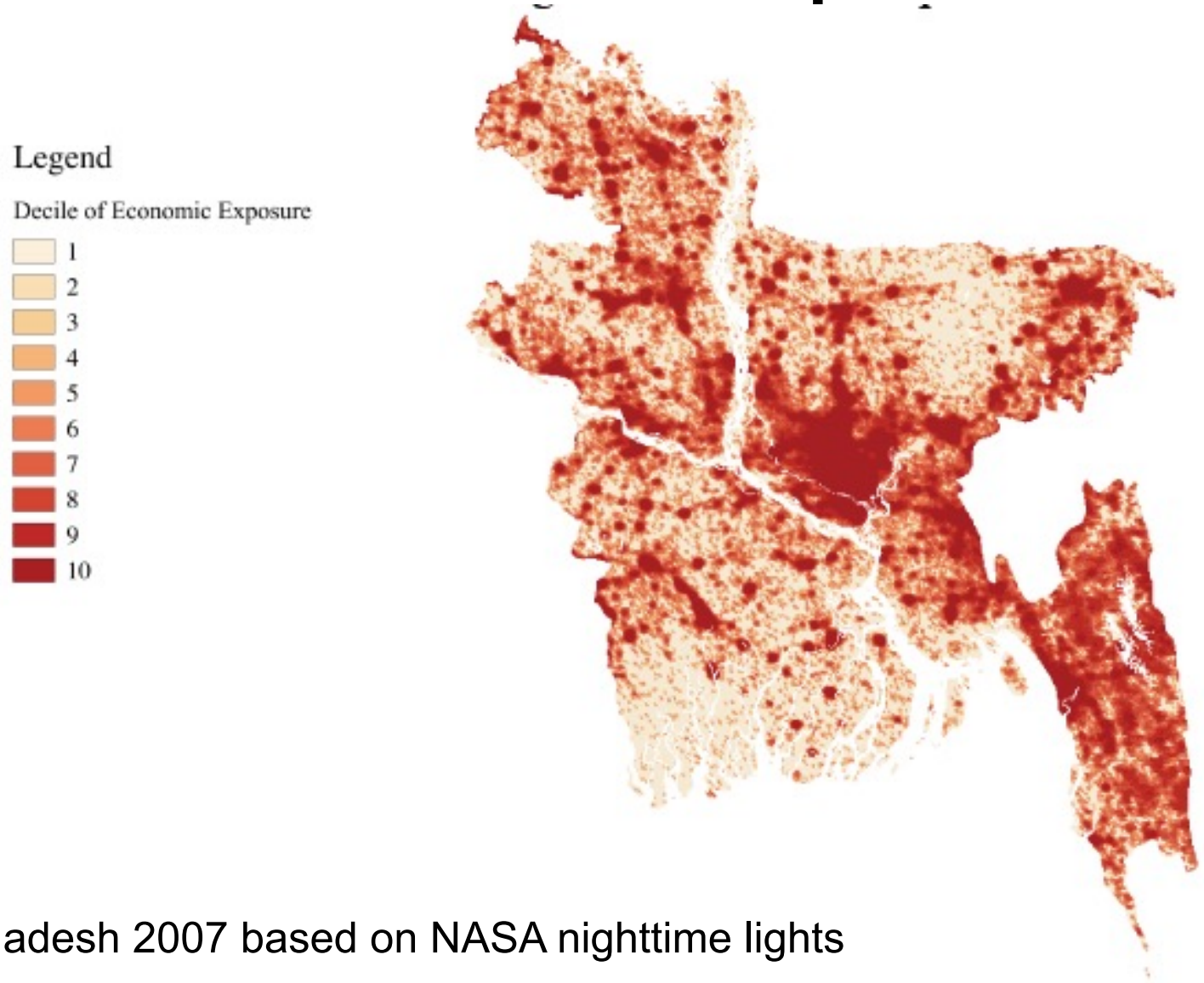
- High-resolution satellite datasets measuring nighttime lights, estimated population, and recorded surface water
- Estimates of economic and human risk
- New metric of human vulnerability using nighttime light intensity per capita
- Estimates of monetary losses of economic activity due to disasters close to official values (within 95% confidence interval)
- Case study: 2007 floods in Bangladesh

Collaboration with Tom Swartz

Satellite Data

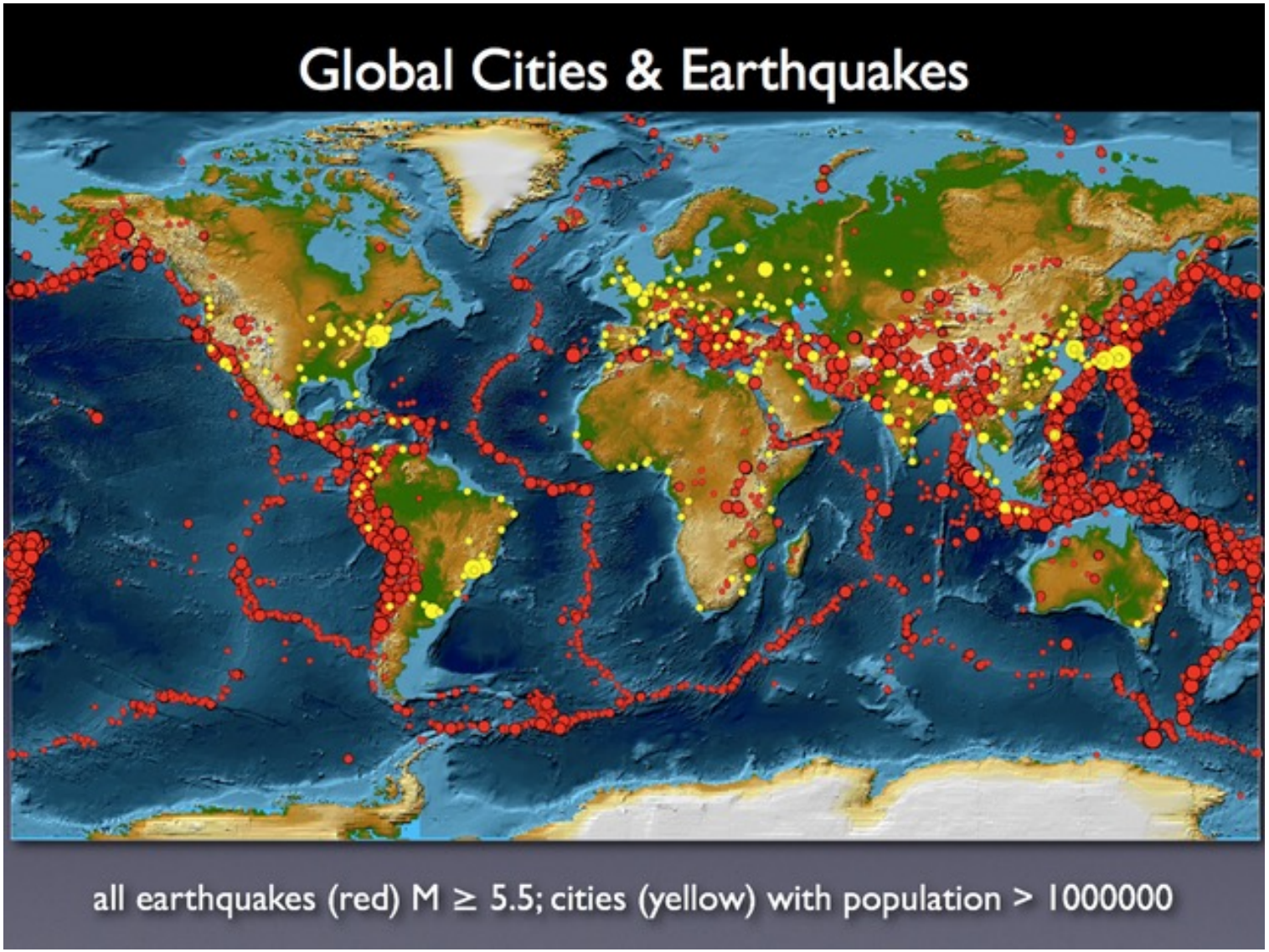
Source: NASA

Economic Exposure



Bangladesh 2007 based on NASA nighttime lights

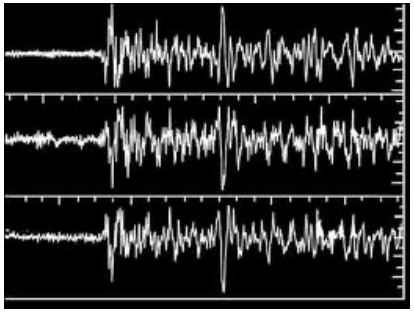
Earthquakes

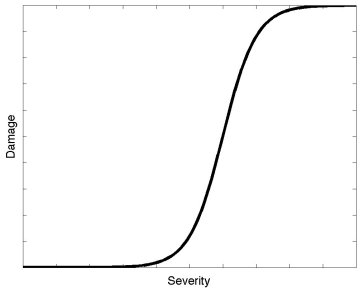


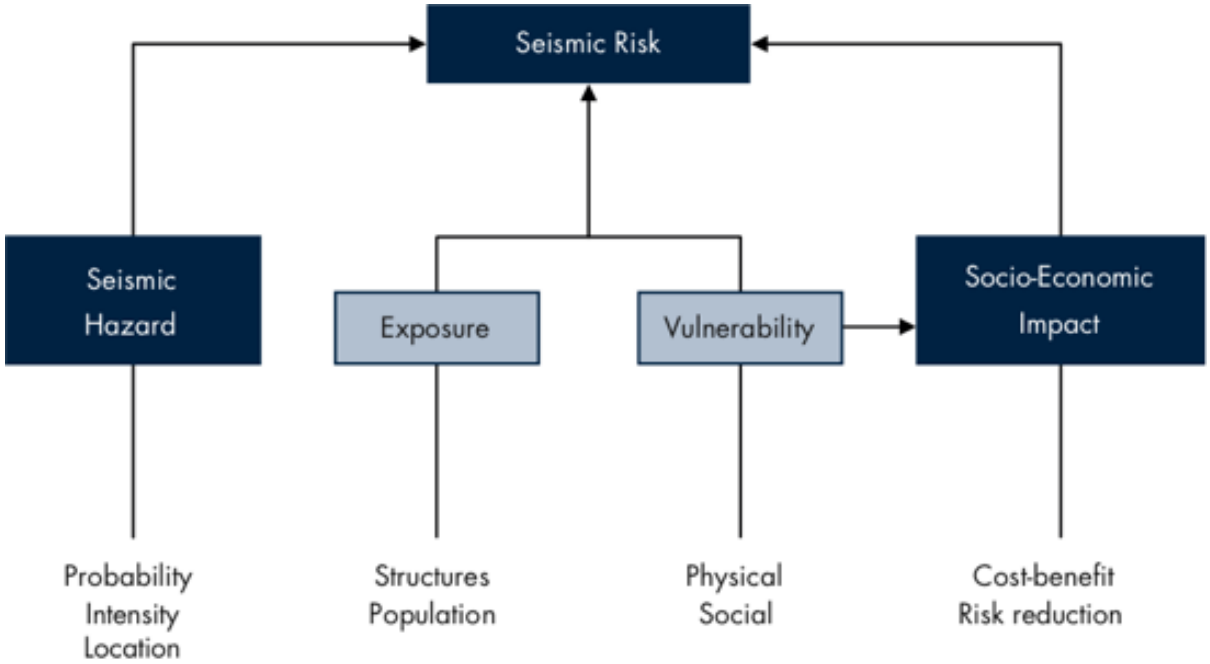
Source: Rick Allmendinger, Cornell University

Catastrophe modelling

Risk = Hazard x Exposure x Vulnerability







Source: Global Earthquake Model

Socio-economic impact

- **Direct impacts**: mortality, morbidity, damage and destruction of assets;
- **Indirect impacts**: loss of income, production and services;
- **Tertiary impacts**: degradation of social capital and institutional capacity.

Model Risk

- Numerous major losses which were not adequately modelled. For example in 2011:
  - Japan ($360bn); NZ quake ($40bn); Thai flood ($46.5bn)
- Vendor model changes
  - Modelled versus real world & uncertainty
- Regulation associated with Solvency II and increased focus on 1 in 200 year events
  - Compliance issues relating to model uncertainty and risk assessment

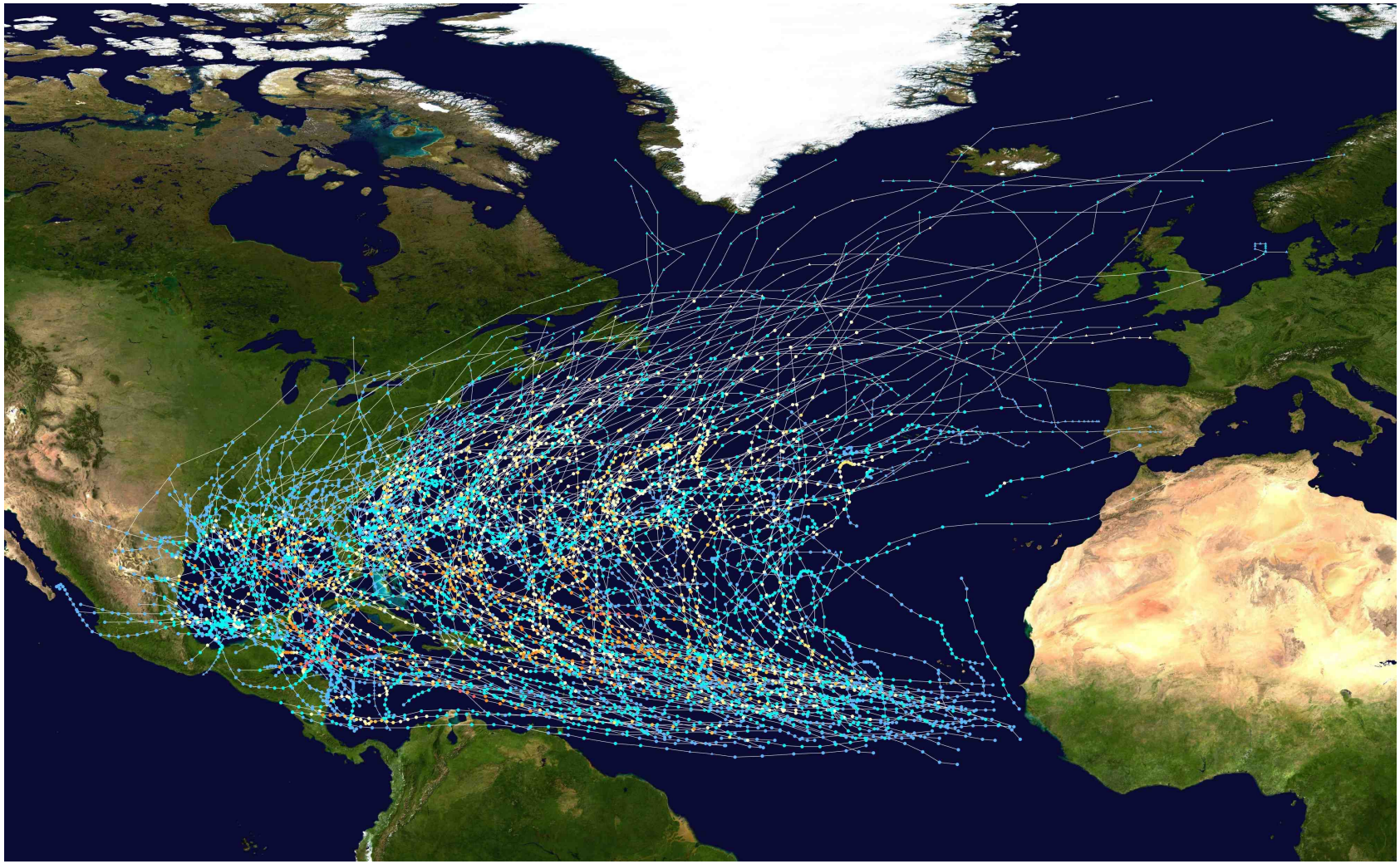
Systemic Risk

- Exposure is more interconnected than ever
  - Sovereign risk, Arab Spring, Spatiotemporal extremes
- Supply chain risk (outsourcing, lean manufacturing, just in time inventory)
- Local and regional models are inadequate for assessing potential global losses
- Thailand flooding was an unexpected loss
  - Initial estimate of $10 bn (actual $46.5 bn)
  - Contingent business interruption
  - Semiconductors, car manufacturing

Challenges for CAT models

- Extreme events often cluster in time and space
- Independent analysis of risks may be misleading
- Using a Poisson assumption for a long memory process will underestimate the risk
- Correlated risks refer to a combination of hazards such as wind and rain
- Earthquake and tsunami, Japan (2011) cost $360 billion (World Bank) due to multi-hazards, nuclear crisis and loss of economic activity

Hurricanes Tracks



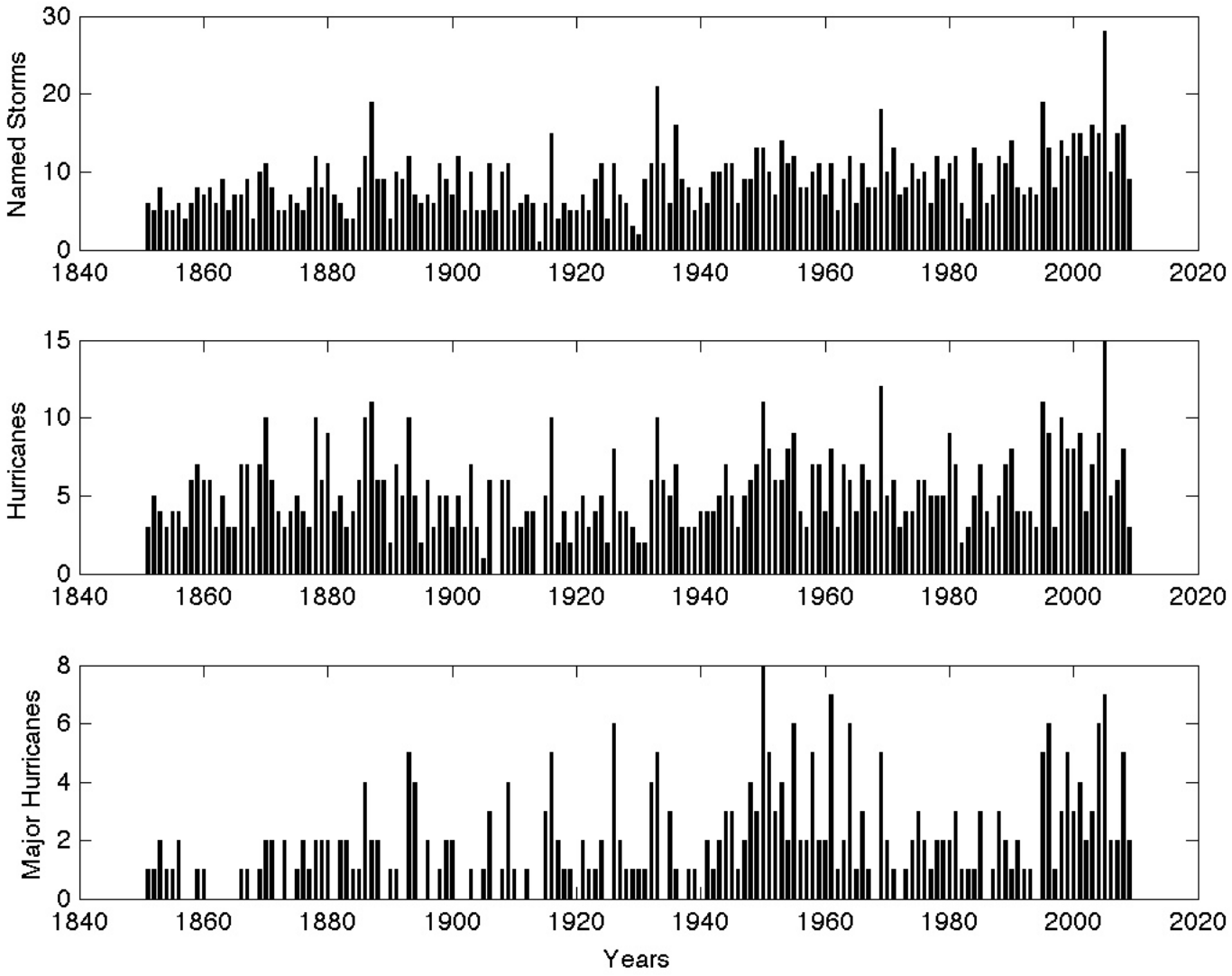
Saffir-Simpson Hurricane Scale

Category	Wind Speed			Storm Surge		
	mph	m/s	km/h	kts	ft	m
Five	≥156	≥ 70	≥ 250	≥ 136	> 18	> 5.5
Four	131 - 155	59 - 69	210 - 249	114 - 135	13 - 18	4.0 – 5.5
Three	111 - 130	50 - 58	178 – 209	96 - 113	9 - 12	2.7 – 3.7
Two	96 - 110	43 - 49	154 - 177	83 - 95	6 - 8	1.8 – 2.4
One	74 - 95	33 - 42	119 - 153	64 - 82	4 - 5	1.2 – 1.5

Classification:

- Named Storms: Tropical Storms, Hurricanes and Subtropical Storms
- Hurricanes: Saffir-Simpson Hurricane Scale 1 to 5
- Major Hurricanes: Saffir-Simpson Hurricane Scale 3, 4, or 5

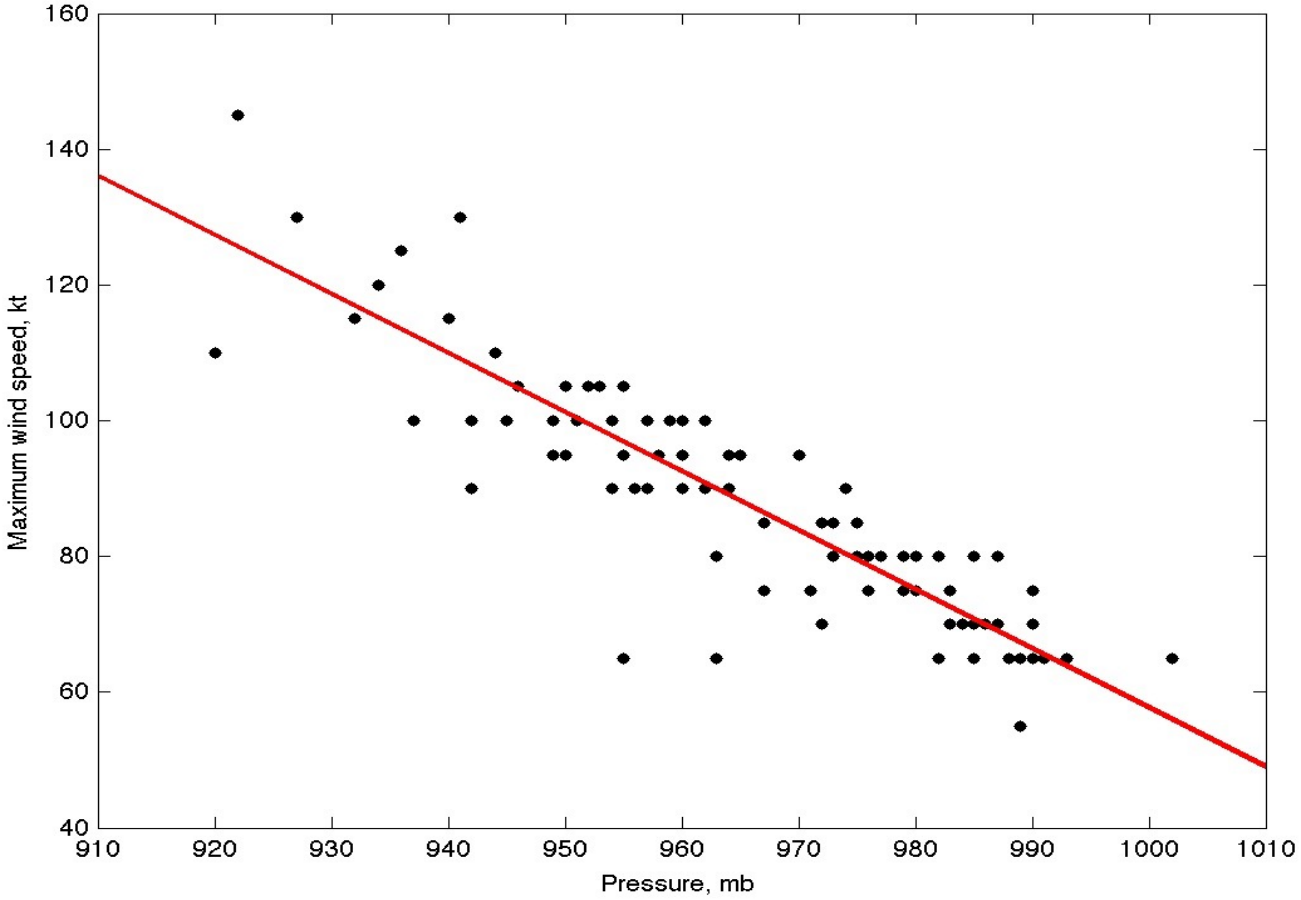
Hurricane counts



Destructive Hurricanes

Date	Name	States	Category	Lowest Pressure, mbar	Max Wind Speed, mph	Estiamted Loss, $bn
29 Aug 2005	Katrina	LA, MS	3	902	175	90.1
24 Aug 1992	Andrew	FL, LA	5-3	922	175	41.1
13 Sep 2008	Ike	FL, LA, TX	4	935	145	38.0
24 Oct 2005	Wilma	FL	3	882	185	32.3
13 Aug 2004	Charley	FL	4	941	150	18.7
16 Sep 2004	Ivan	FL	3	910	165	20.7
23 Sep 2005	Rita	TX	3	895	180	11.1
02 Sep 2004	Frances	FL	2	935	145	13.8
22 Sep 1989	Hugo	SC	4	918	160	17.5
26 Sep 2004	Jeanne	FL	3	950	120	8.0
01 Sep 2008	Gustav	FL, LA	4	941	155	6.7
03 Oct 2002	Lili	LA	4	938	145	1.1
26 Sep 2002	Isidore	LA	3	934	125	1.6

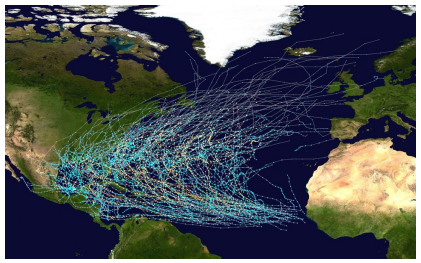
Max wind speed versus pressure

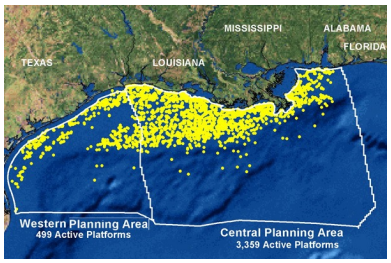


Maximum wind speed versus central pressure of hurricanes making US landfall between 1900 and 2008. Regression has $R^2 = 0.8$. Data Source: HURDAT, NOAA.

Hurricanes & Oil

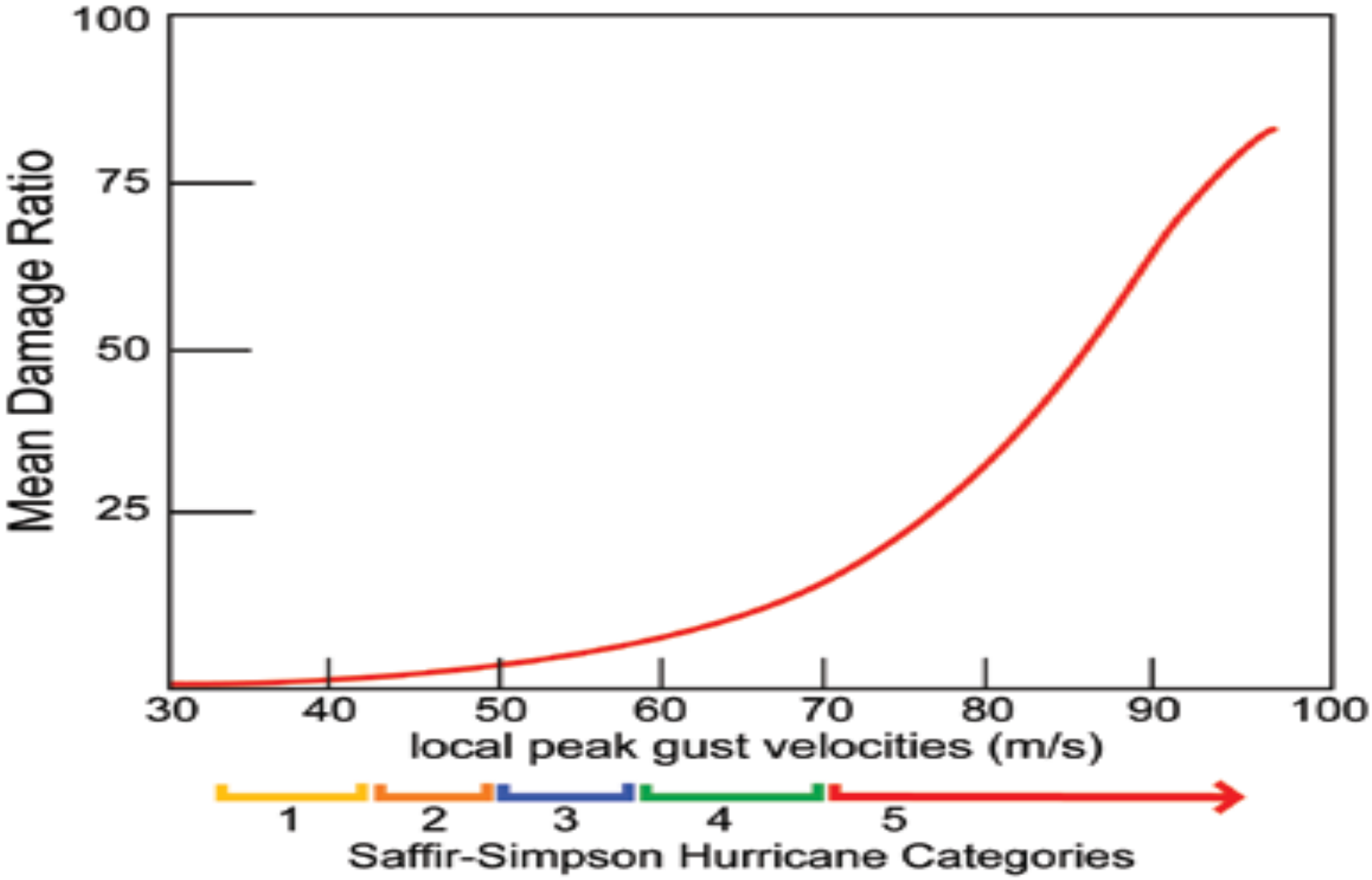
Risk = Hazard x Exposure x Vulnerability





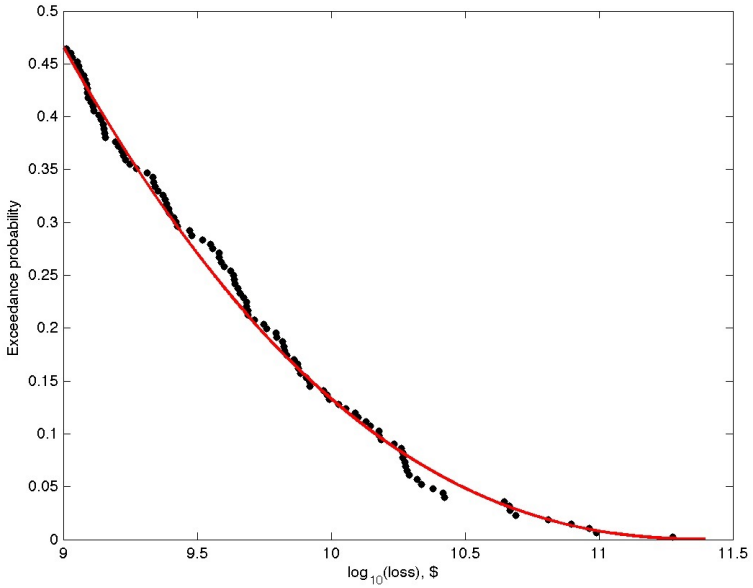
Hazard: How are hurricanes generated in the Atlantic ocean?
Exposure: Will the hurricane path coincide with the location of your assets?
Vulnerability: How likely is it that the hurricane will cause damage?

Hurricane Damage



Catastrophe Modelling

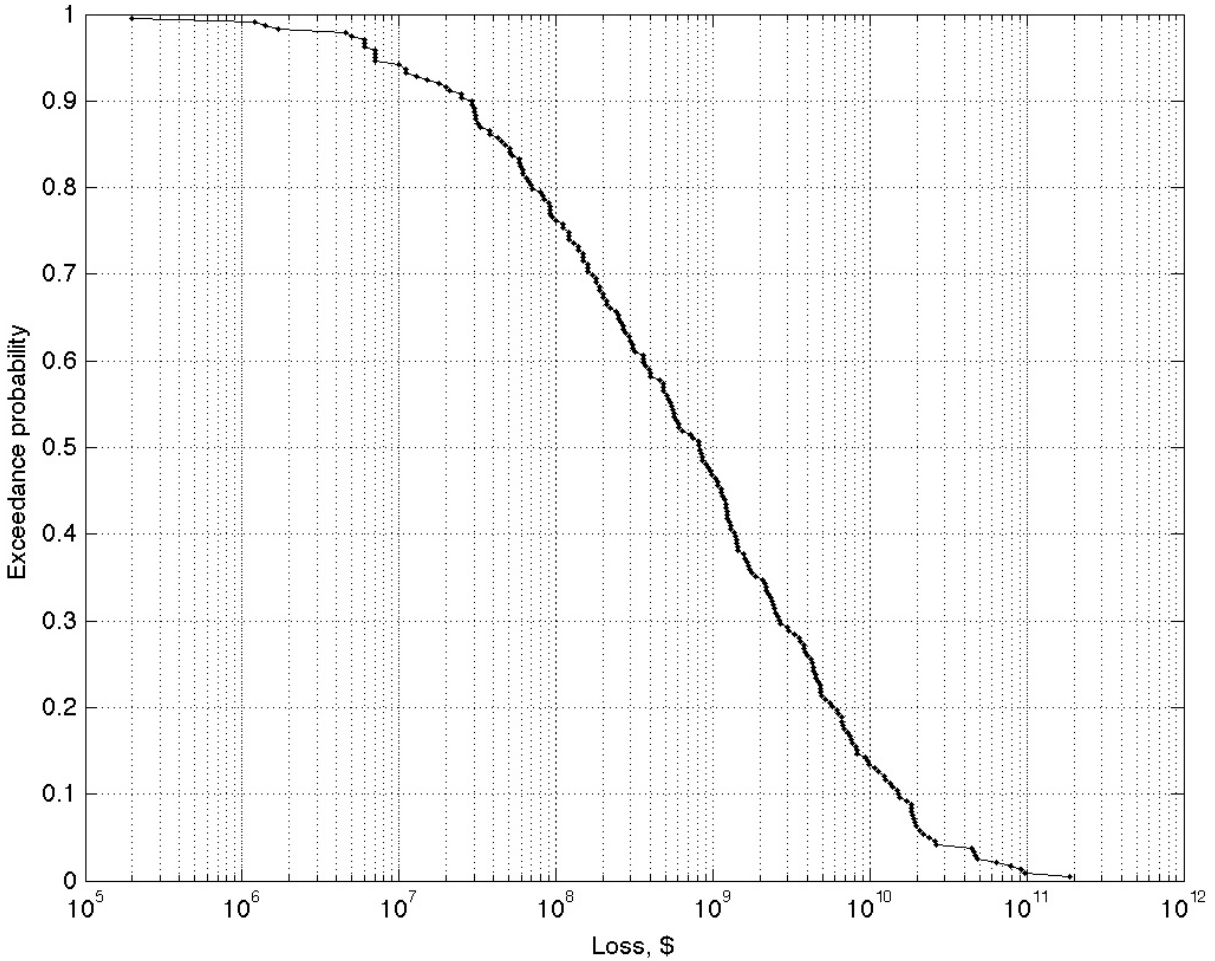
Risk = Hazard x Exposure x Vulnerability



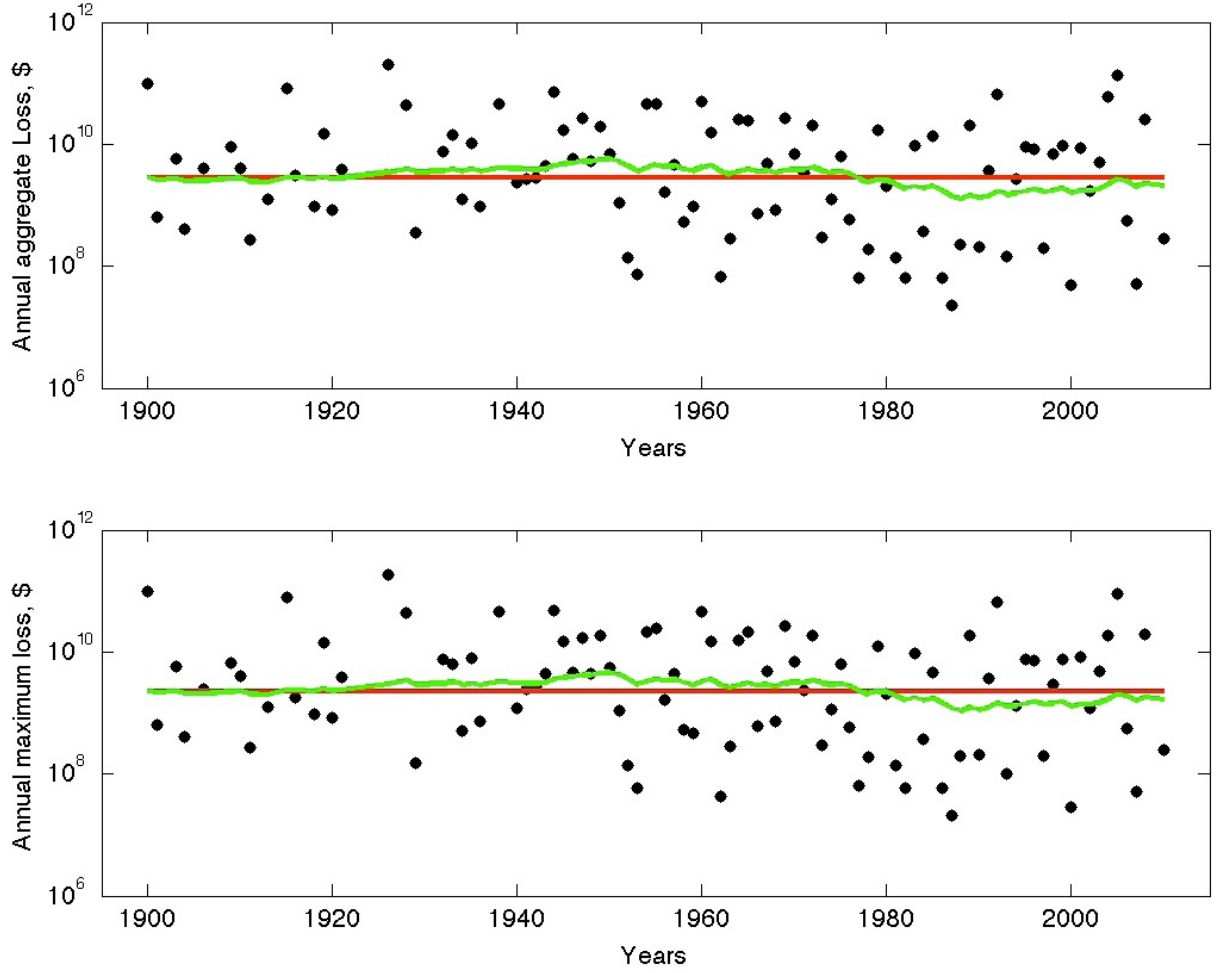
Under certain assumptions, extreme value theory or simulation can be used to perform extrapolation in order to assess the probability of extreme losses.

Loss from a 1 in 200 year event is estimated at $152 billion. Loss from Katrina was $90 billion.

Exceedance probability



Losses over time



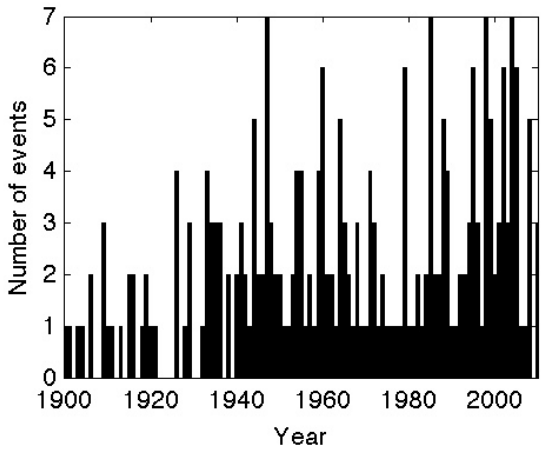
Recent decline in losses may be a result of a combination of decreased hazard (low hurricane activity) and improved engineering standards along the coast of the US.

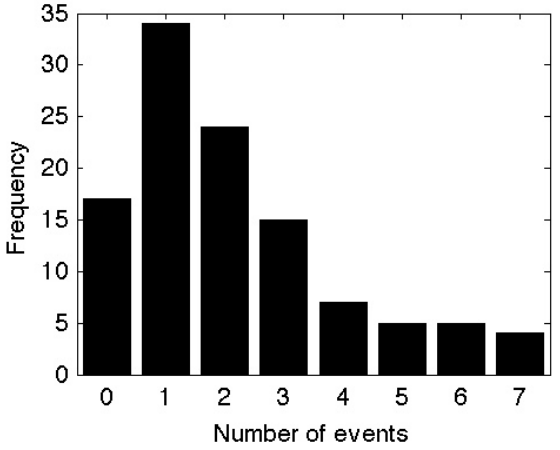
The exact explanation requires a separation of the effects due to the hazard, exposure and vulnerability.

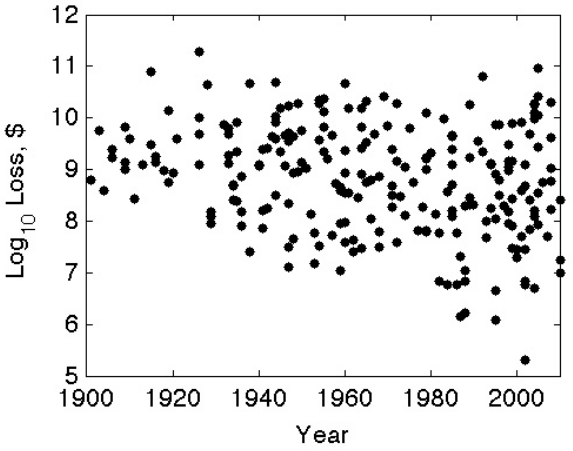
While engineering standards may have reduced the exposure, it is likely that there has been an increase in the exposure along the coast.

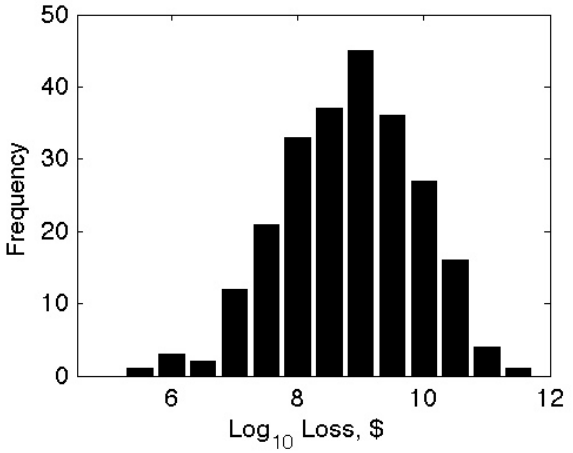
Risk = hazard x exposure x vulnerability

Loss summary

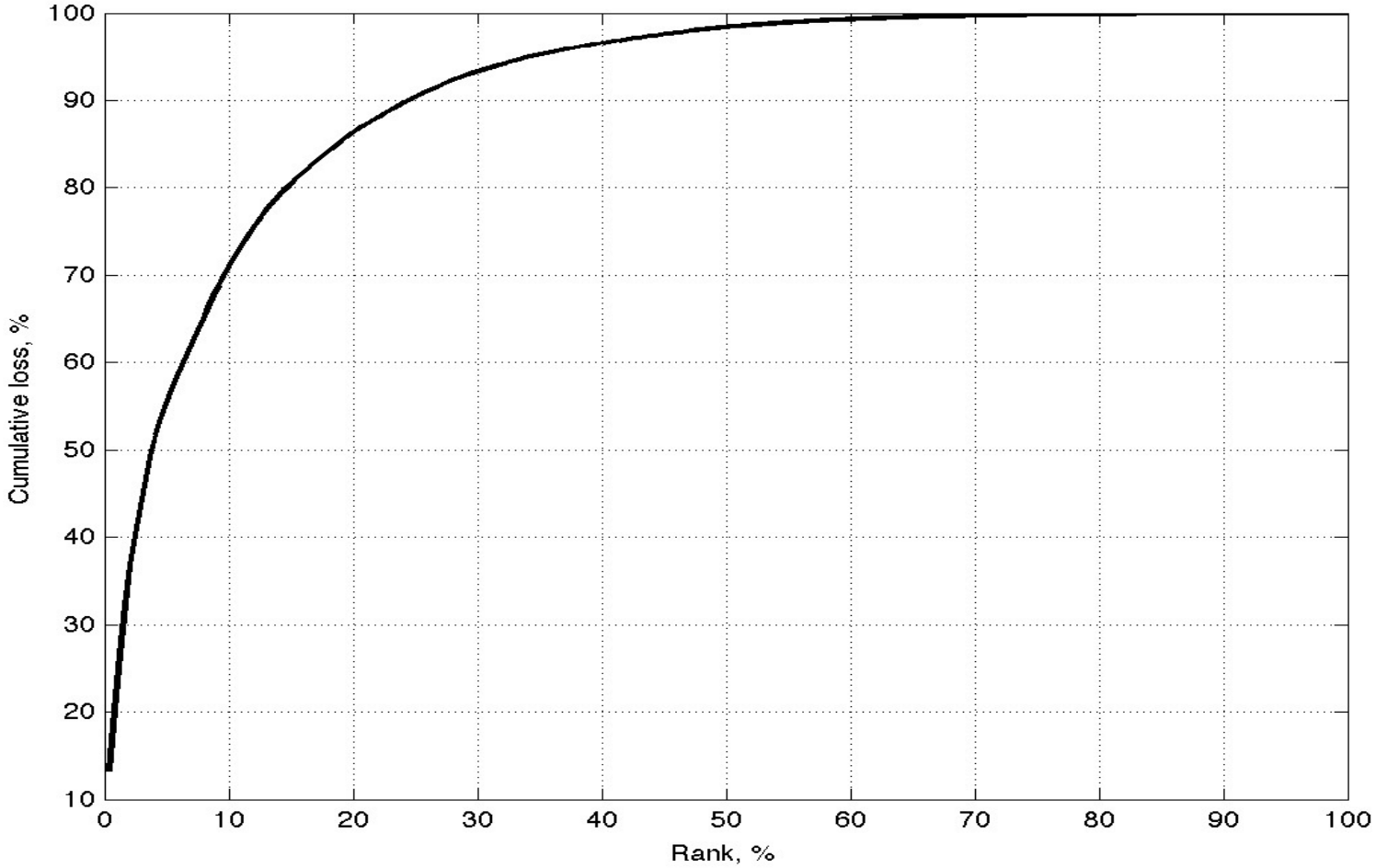








Cumulative Hurricane Loss



Cumulative loss versus rank showing that the largest 20% of the events account for 86% of the total losses arising from hurricanes.

Hurricane losses

- Which power law exponent would you expect to describe how financial losses from hurricanes scale with wind speed?

  a) One
  b) Two
  c) Three
  d) Four

- **Slido.com #820813**

Power-law

- The relationship between damage and maximum wind speed of the hurricane has been studied by a number of researchers.
- We would expect the damage to grow monotonically with wind speed.
- Appealing to physics, from first principles we might expect damage to be proportional to the amount of energy or power delivered by the hurricane.

What power-law coefficient?

- The kinetic energy of the hurricane scales with the square of the wind speed.
- The power of the hurricane suggests that damage should be proportional to the cubic power of the wind speed (as for power generation from wind turbines).
- An analysis of damage normalised by GDP was found to depend on the ninth power of the maximum wind speed (Nordhaus, 2010).

Estimating the power-law

- We considered the power-law relationship

  $L(w)=L_0 w^{\delta}$.

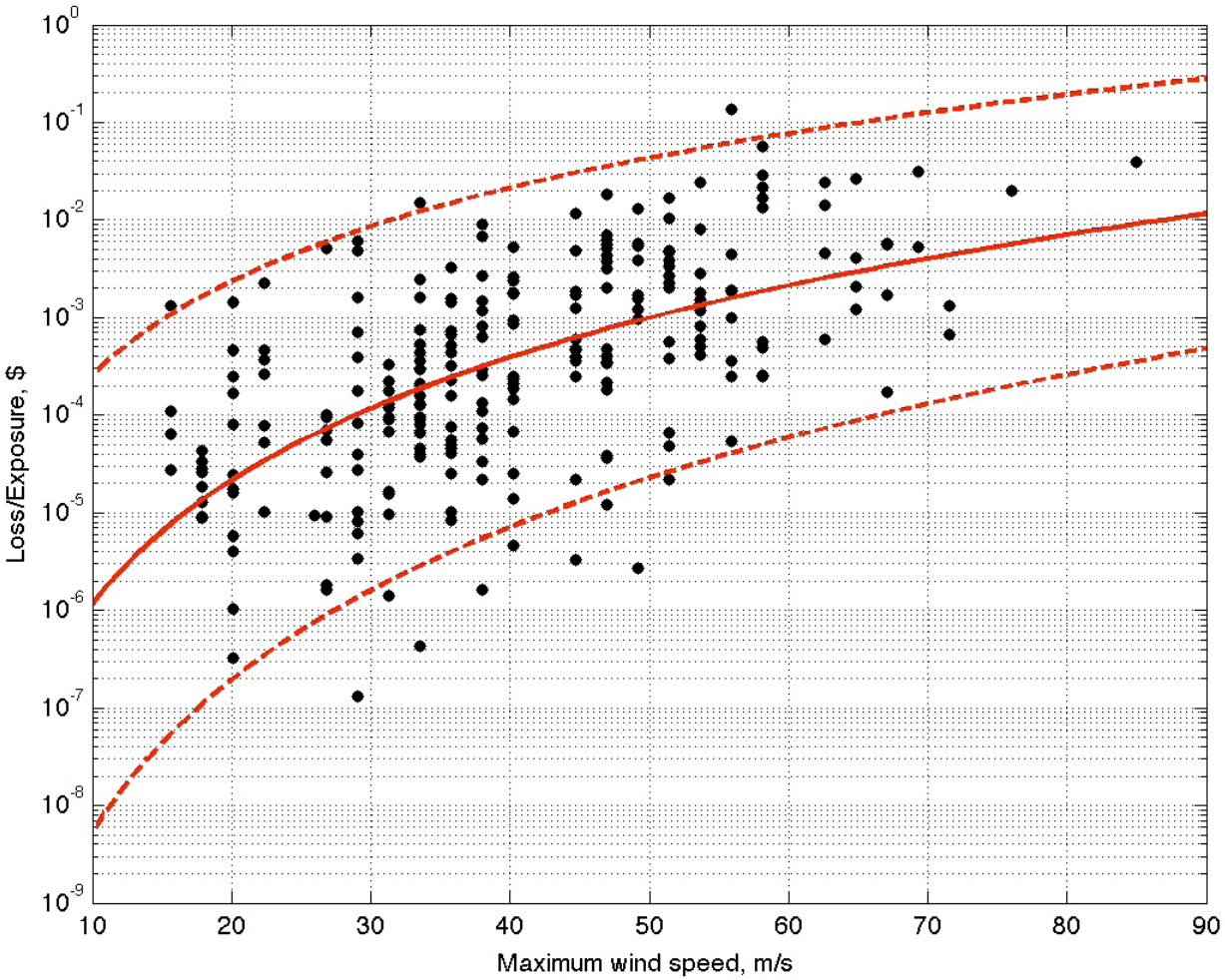
- Fitting the linear relationship

  $\log L(w) = \log L_0 + \delta \log w$

- yielded

  $\delta = 3.87$ and $R^2 = 0.32$.

Loss versus wind speed

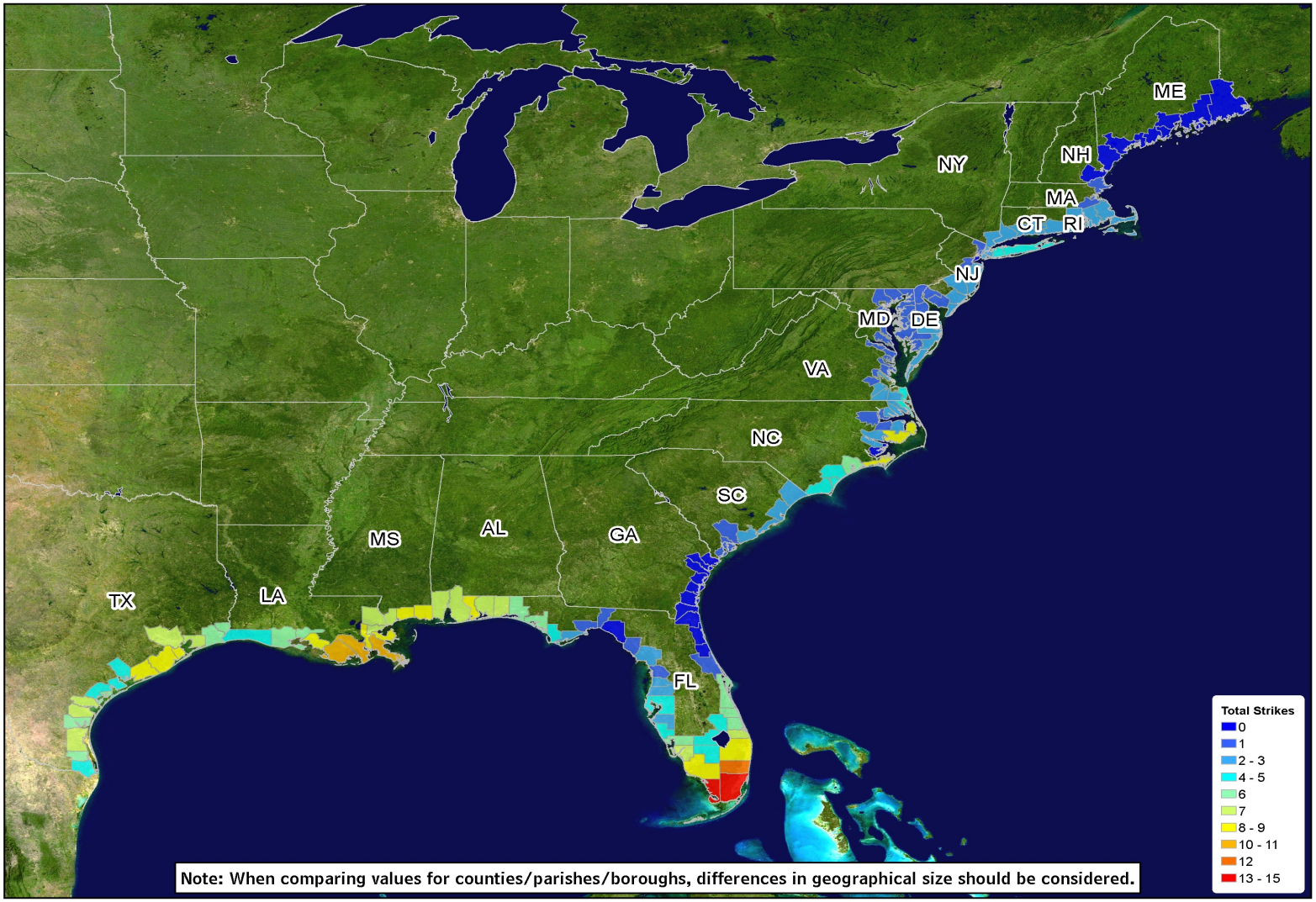


Quiz

- Which US state has been affected by the largest number of hurricanes?
  a) Florida
  b) Louisiana
  c) New York
  d) Texas

- **Slido.com #820813**

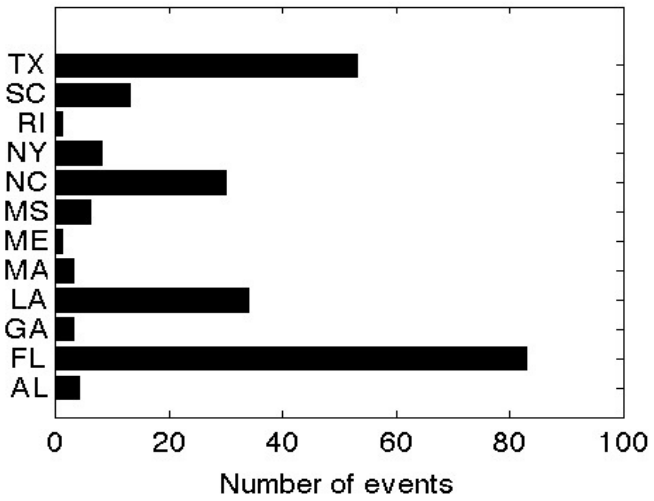
Hurricane Landfall

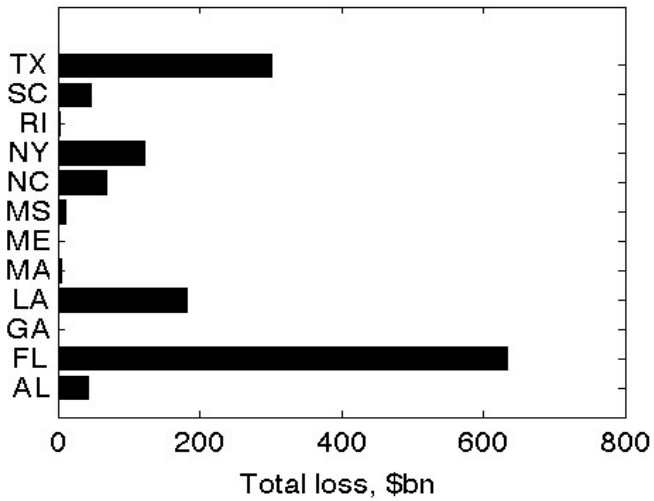


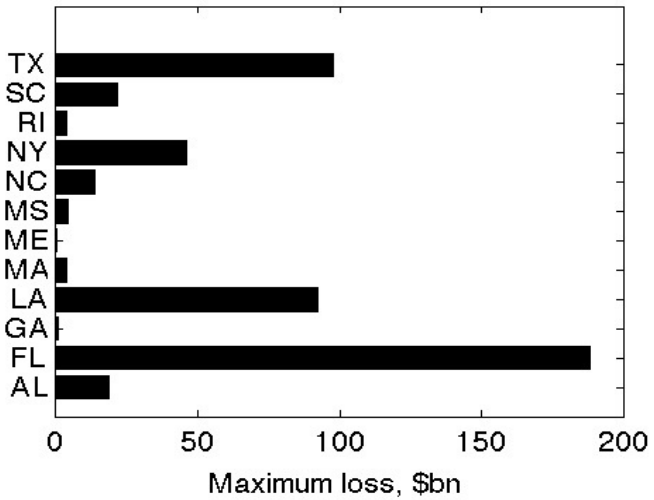
Total number of major hurricane strikes by counties/parishes/boroughs, 1900-2010

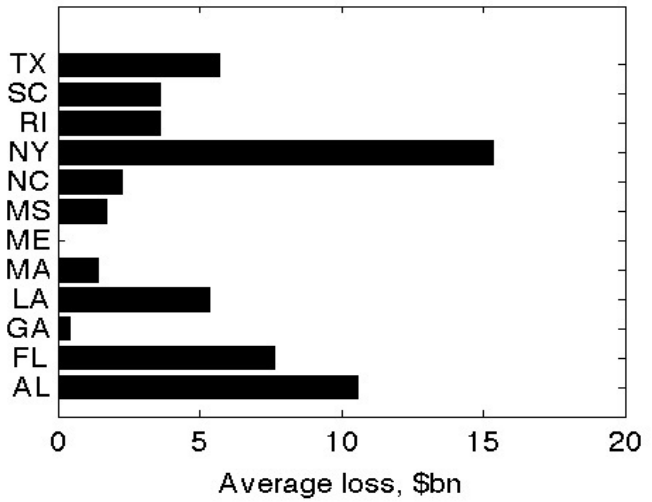
Data from NWS NHC 46: Hurricane Experience Levels of Coastal County Populations from Texas to Maine. Jerry D. Jarrell, Paul J. Hebert, and Max Mayfield. August, 1992, with updates.

Hurricanes by State

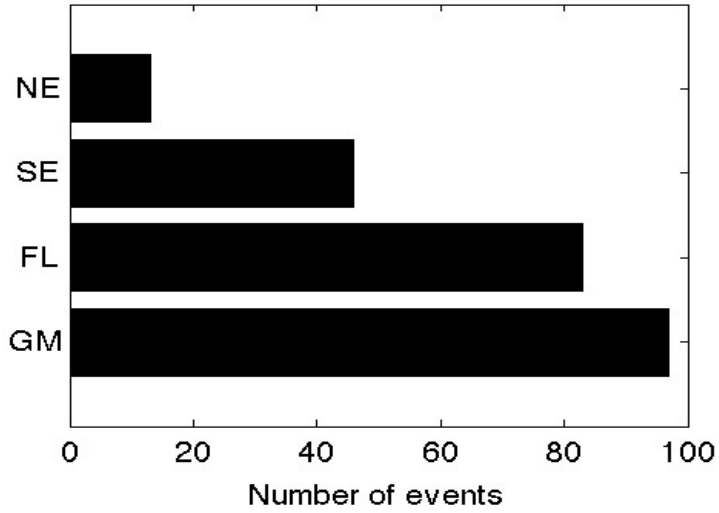


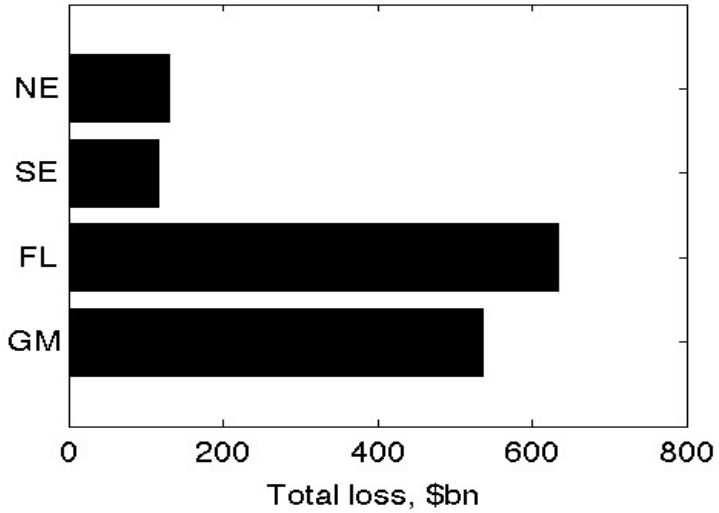


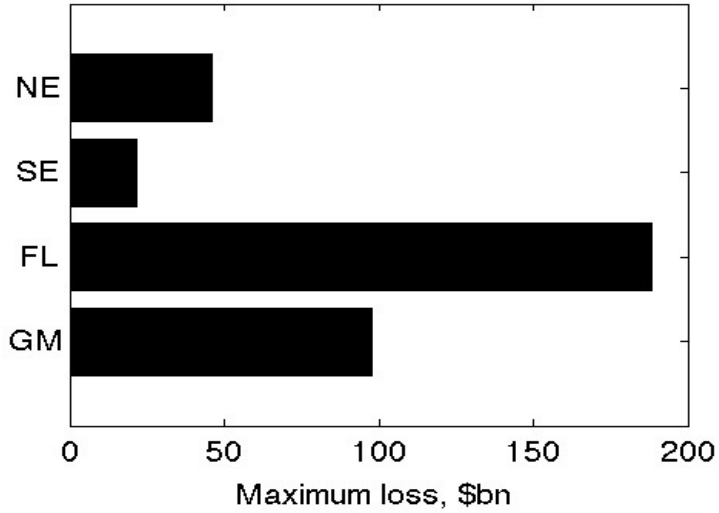


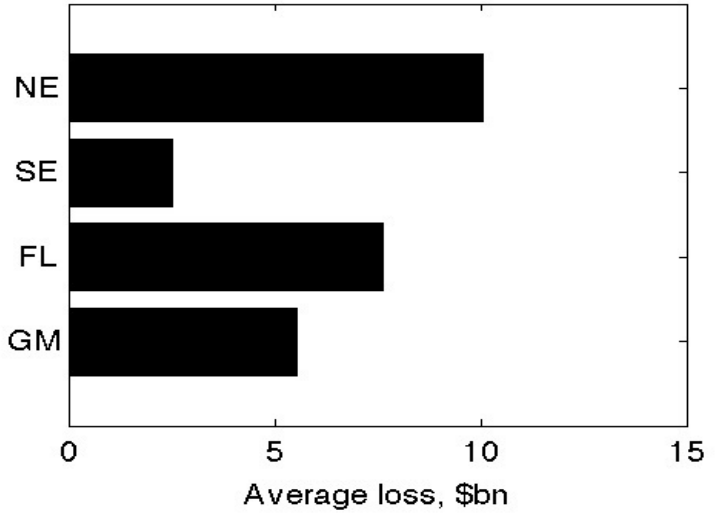


Hurricanes by Region



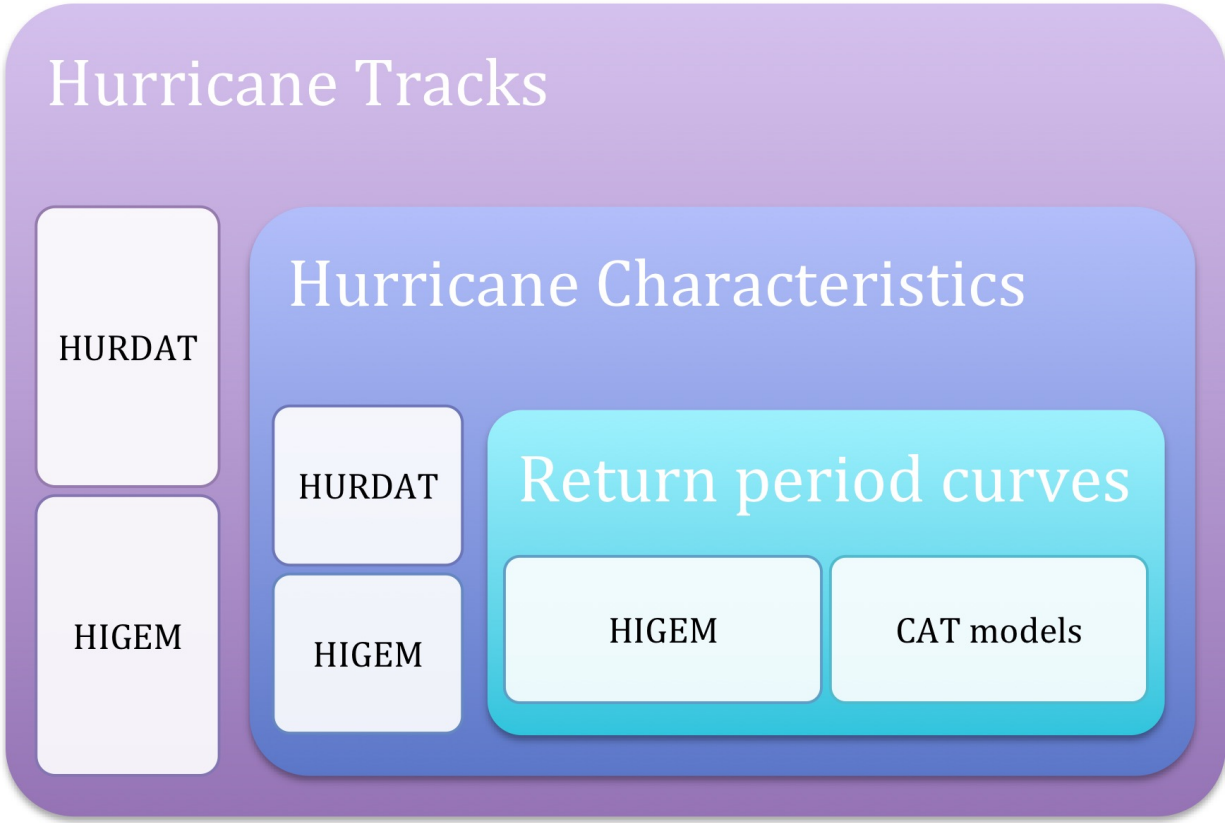






Risk forecasting using dynamical models

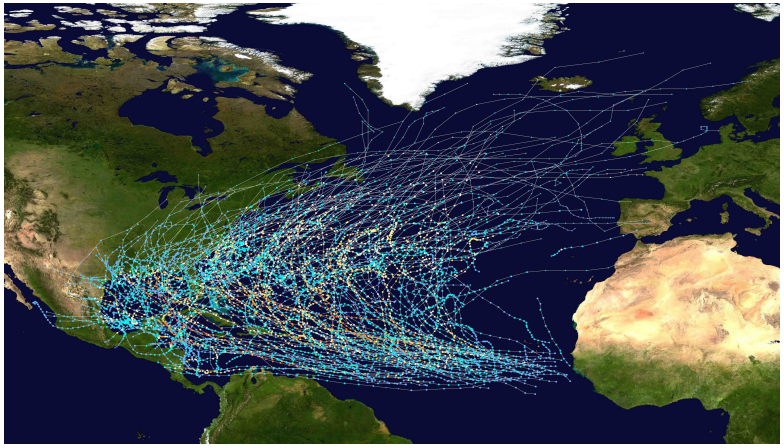
- Limited observational data (hazard or losses)
- Dynamical models (weather and climate GCMs)
- Calibration techniques to estimate risk

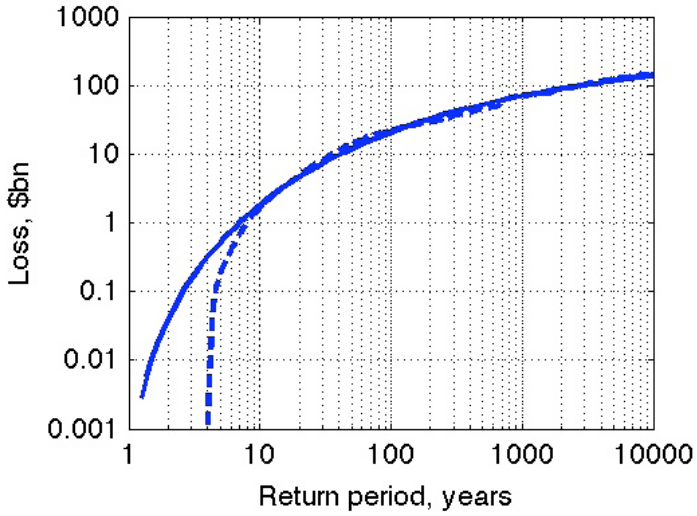


Catastrophe risk

Willis Research Network | THE SCIENCE OF MANAGING EXTREMES

- Member of the Willis Research Network working on parsimonious models for catastrophe risk assessment
- Assessment of economic losses arising from disasters
- Demonstration of ability of atmospheric models to generate synthetic hurricanes
- Potential for global windstorm risk assessment





Loss estimates from calibrated HIGEM model (solid) versus commercial CAT model (dashed).