State Space Models and the Kalman Filter

References:

RLS course notes, Chapter 7.

Brockwell and Davis, Chapter 12.

Harvey, A.C. (1989), *Forecasting. Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

A. Pole, M. West, and P.J. Harrison (1994), *Applied Bayesian Forecasting and Time Series Analysis.* Chapman-Hall, New York.

M. West and P.J. Harrison (1997, First edition 1989), *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York.

Basic equations:

$$\begin{aligned} X_t &= F_t S_t + v_t, \\ S_t &= G_t S_{t-1} + w_t, \\ v_t &\sim N[0, V_t], \\ w_t &\sim N[0, W_t], \end{aligned}$$

where

- X_t is (multivariate) observation process
- S_t is unobserved “state” process
- F_t, G_t, V_t, W_t in principle known, though may have to be estimated in practice.

Why consider such models?

- All ARMA or ARIMA models may be rewritten as state space models
- Extends to multivariate case automatically
- Many nonstationary or seasonal models...
- Natural formulation of Bayesian approach

Solution by *Kalman Filter*

Note on terminology:

A state space model is in principle any model that includes an observation process X_t and a state process S_t. The equations may be nonlinear, or non-Gaussian.

The Kalman Filter is a particular algorithm that is used to solve state space models in the linear case. This was first derived by Kalman (1960).

Some people refer to "Kalman filter models" but in my view this is imprecise terminology.

Bayesian derivation of the Kalman Filter

Follows Meinhold and Singpurwalla (1983).

We use the following fact familiar from multivariate analysis:

If

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim MVN\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right]$$

then

$$Y_1 \mid Y_2 = y_2 \sim MVN\left[\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right].$$

Let $\mathcal{X}^t$ denote all information available up to time t (the σ-algebra generated by $X_s,\ s \le t$). Suppose

$$S_{t-1} \mid \mathcal{X}^{t-1} \ \sim \ MVN[\widehat{S}_{t-1}, P_{t-1}].$$

But $S_t = G_t S_{t-1} + w_t$ so

$$S_t \mid \mathcal{X}^{t-1} \ \sim \ MVN[G_t \widehat{S}_{t-1}, R_t], \quad R_t = G_t P_{t-1} G_t^T + W_t.$$

Then

$$\begin{pmatrix} S_t \\ X_t \end{pmatrix} \mid \mathcal{X}^{t-1} \ \sim \ MVN\left[\begin{pmatrix} G_t \widehat{S}_{t-1} \\ F_t G_t \widehat{S}_{t-1} \end{pmatrix}, \begin{pmatrix} R_t & R_t F_t^T \\ F_t R_t & F_t R_t F_t^T + V_t \end{pmatrix}\right].$$

Hence

$$\begin{aligned} & S_t \mid X_t, \mathcal{X}^{t-1} \\ \sim \ & MVN[G_t \widehat{S}_{t-1} + R_t F_t^T (F_t R_t F_t^T + V_t)^{-1} (X_t - F_t G_t \widehat{S}_{t-1}), \\ & R_t - R_t F_t^T (F_t R_t F_t^T + V_t)^{-1} F_t R_t]. \end{aligned}$$

Thus we have the recursive equations

$$\begin{aligned}\widehat{S}_t &= G_t\widehat{S}_{t-1} + R_tF_t^T(F_tR_tF_t^T + V_t)^{-1}(X_t - F_tG_t\widehat{S}_{t-1}),\\ P_t &= R_t - R_tF_t^T(F_tR_tF_t^T + V_t)^{-1}F_tR_t.\end{aligned}$$

Issues:

- Initiation of $\widehat{S}_0,\ P_0$
- Prediction and smoothing: estimate S_t given $X_1, ..., X_T$. $t > T$ is *prediction problem*, $1 \le t < T$ is *smoothing problem*.
- Estimation: assume parametric forms, $F_t = F_t(\psi)$ etc. Define likelihood function using *prediction error decomposition*

$$f(X_1, ..., X_T \mid \psi) = \prod_{t=1}^{T} f(X_t \mid \mathcal{X}^{t-1}).$$

In fact, the modern derivation of exact MLE for ARMA processes is based on this approach.

Application to Financial Time Series

Let y_t be day-t return. GARCH(1,1) model:

$$y_t \;=\; \epsilon_t \sigma_t, \quad \epsilon_t \;\sim\; N[0,1], \quad \sigma_t^2 \;=\; \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

Alternatively, the simplest case of a *stochastic volatility* model Direct estimation by MLE possible.

$$y_t \;=\; \epsilon_t e^{h_t/2}, \quad h_{t+1} = \gamma_0 + \gamma_1 h_t + \eta_t, \quad \eta_t \sim N[0, H_t].$$

Solution by Monte Carlo/Bayesian methods (focus of current *Sequential Monte Carlo* program at SAMSI).

Also *generalized method of moments* (GMM) approaches popular among econometricians.

Long-Range Dependence

References:

Brockwell and Davis, Section 13.2

J. Beran (1994), *Statistics for Long-Memory Processes*. Chapman and Hall, New York.

P. Doukhan, M.S. Taqqu and G. Oppenheim (eds.) (2003), *Long-Range Dependence*. Birkhäuser, Boston.

The covariance function of any causal invertible ARMA process satisfies

$$|\gamma_k| \;\leq\; Ar^k, \quad \text{some } r \in [0,1).$$

Long-range dependence is concerned with processes that satisfy

$$\gamma_k \;\sim\; Ck^{2d-1}, \quad \text{some } C > 0, d. \tag{1}$$

Stationary and invertible if $d \in \left(-\frac{1}{2}, \frac{1}{2}\right)$.

Note that $\sum |\gamma_k| = \infty$ though $\sum \gamma_k^2 < \infty$

Also write $H = d + \frac{1}{2}$ where H is *Hurst coefficient*

$d = 0$ or $H = \frac{1}{2}$ correspond to standard short-range (including ARMA processes)

History

Hurst (1951) first proposed this model (with $H \approx 0.7$) based on a study of hundreds of years of flood volumes of the River Nile.

Mandelbrot (c. 1968) proposed a model of *Fractional Brownian Motion*: a Gaussian process with the *self-similarity* property

$$Y_{ct} \equiv c^H Y_t, \quad t \in (-\infty, \infty)$$

for any fixed $c > 0$. Differences of FBM are called *Fractional Gaussian Noise*.

Current applications include environmental (e.g. climatic) data, finance, internet traffic and many others.

Fractional Differencing

Introduced independently by Granger and Joyeux (1980) and Hosking (1981).

$$(I-B)^d Y_t \;\sim\; ARMA(p,q)$$

that has a covariance satisfying (1). Here we interpret

$$\begin{aligned}(I-B)^d Y_t &= \left\{ I - dB + \frac{d(d-1)}{2!}B^2 - \frac{d(d-1)(d-2)}{3!}B^3 + \ldots \right\} Y_t \\ &= Y_t - dY_{t-1} + \frac{d(d-1)}{2!}Y_{t-2} - \frac{d(d-1)(d-2)}{3!}Y_{t-3} + \ldots\end{aligned}$$

The case $p=q=0$ is *fractionally integrated noise*. Its spectral density and autocorrelation function are given by

$$\begin{aligned}f(\lambda) &= |1-e^{-i\lambda}|^{-2d}\frac{\sigma^2}{2\pi} = \left(2\sin\frac{\lambda}{2}\right)^{-2d}\frac{\sigma^2}{2\pi}, \\ \rho_k &= \frac{\Gamma(k+d)\Gamma(1-d)}{\Gamma(k-d+1)\Gamma(d)}\end{aligned}$$

Estimation

Assume general form of model ARIMA(p, d, q) though some of the methods are designed to work in more general cases, e.g. $f(\lambda) = |1 - e^{-i\lambda}|^{-2d} f_0(\lambda)$ with f_0 a bounded spectral density.

1. Maximum Likelihood

2. Spectral Methods

3. Wavelet Methods

Maximum Likelihood Approach (see Brockwell and Davis)

In principle can write down exact likelihood for ARMA(p,d,q) processes, treating the autoregressive parameter ϕ, the moving average parameter θ and the fractional differencing parameter d as unknown parameters $\psi = (\phi, \theta, d)$.

Classical asymptotics of MLE hold in this situation (but it's very hard to prove that!

In practice often use *Whittle approximation*

$$\log L \ \approx \ -\sum_j \left\{ \log f(\lambda_j;\psi) + \frac{I(\lambda_j)}{f(\lambda_j;\psi)} \right\}$$

where $\{\lambda_j\}$ are Fourier frequencies, I is periodogram and f is spectral density.

This has same asymptotic properties as exact MLE.

Spectral Approach

We know $f(\lambda) \sim C\lambda^{-2d}$ for λ near 0. Rewrite as

$$\log f(\lambda) \approx \log C - 2d \log \lambda. \tag{2}$$

Suggests trying to fit (2) directly.

First approach: Geweke–Porter-Hudak (1983) proposed an OLS regression of $\log I(\lambda_j)$ on $\log \lambda_j$, first m Fourier frequencies, some $m << \frac{T}{2}$.

Second approach: Minimize

$$\sum_{j=1}^{m} \left\{ \log(C\lambda_j^{-2d}) + \frac{I(\lambda_j)}{C\lambda_j^{-2d}} \right\}.$$

Now known as *Local Whittle Estimator* following seminal paper of P.M. Robinson (1995).

Wavelets Approach

Won't go into details but this may be the best approach overall.

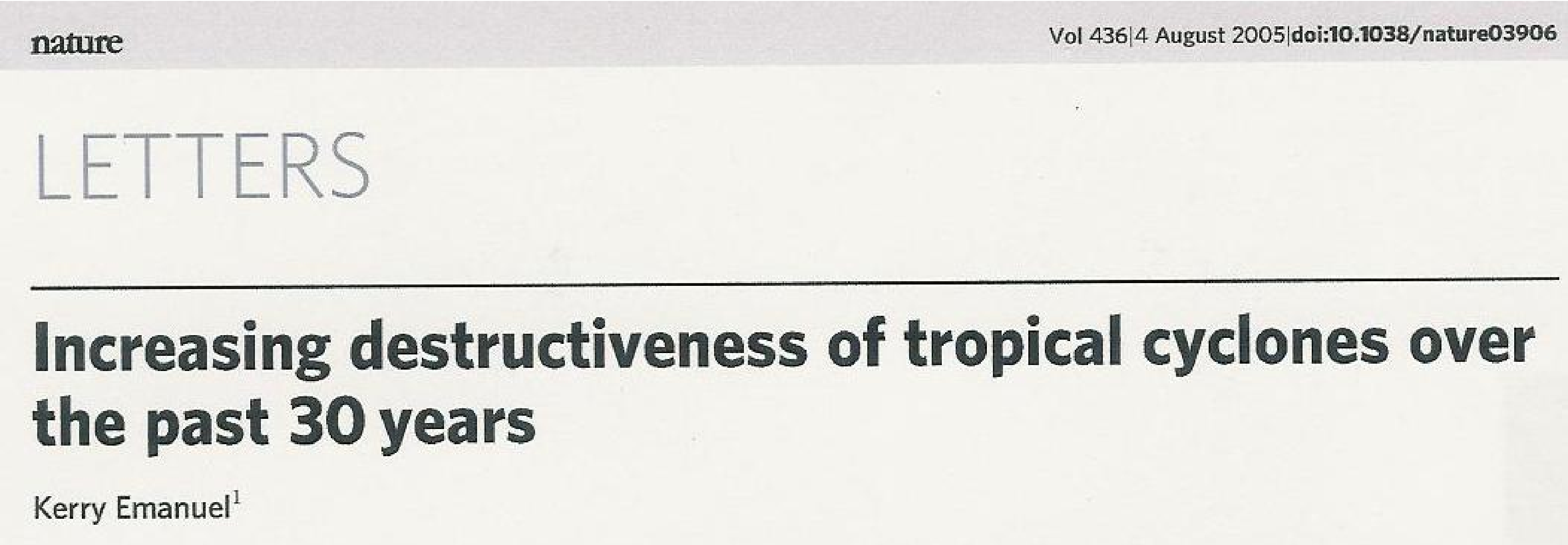
See several papers of Peter Craigmile (Ohio State), also Vladas Pipiras of UNC.

Hurricanes and Global Warming

What this is about:

A problem related to climate change that involves some very challenging applications of time series analysis!

Acknowledgements to Thomas Knutson (NOAA/GFDL) and Evangelos Evangelou (UNC).

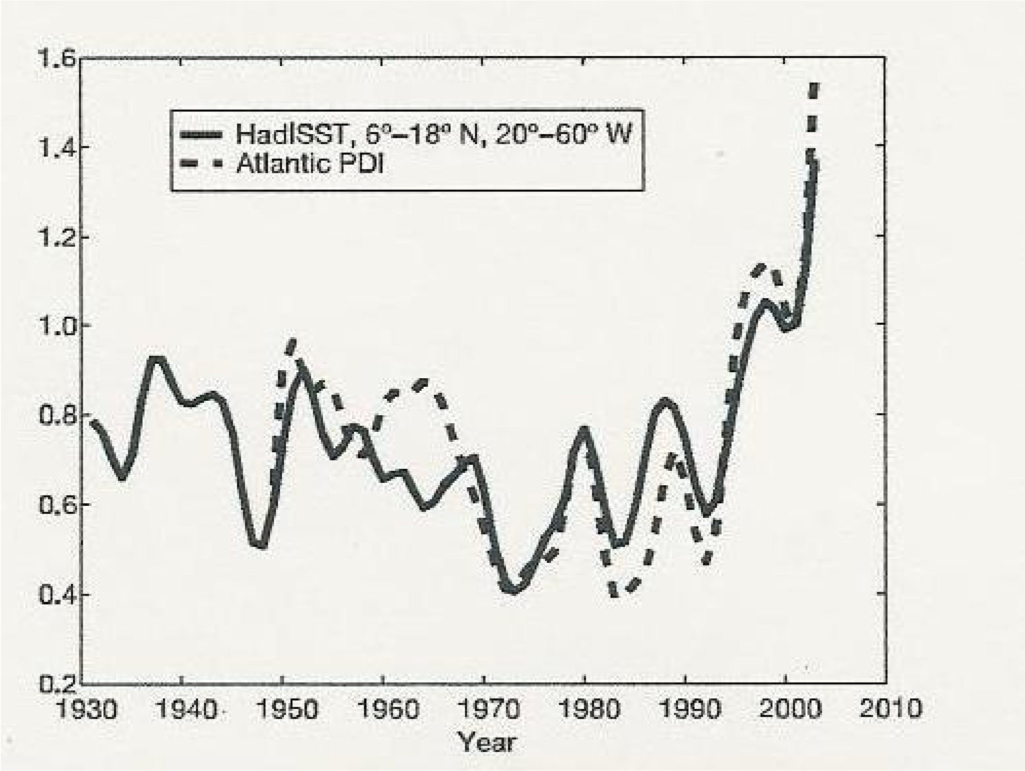


nature Vol 436|4 August 2005|doi:10.1038/nature03906

LETTERS

Increasing destructiveness of tropical cyclones over the past 30 years

Kerry Emanuel[1]



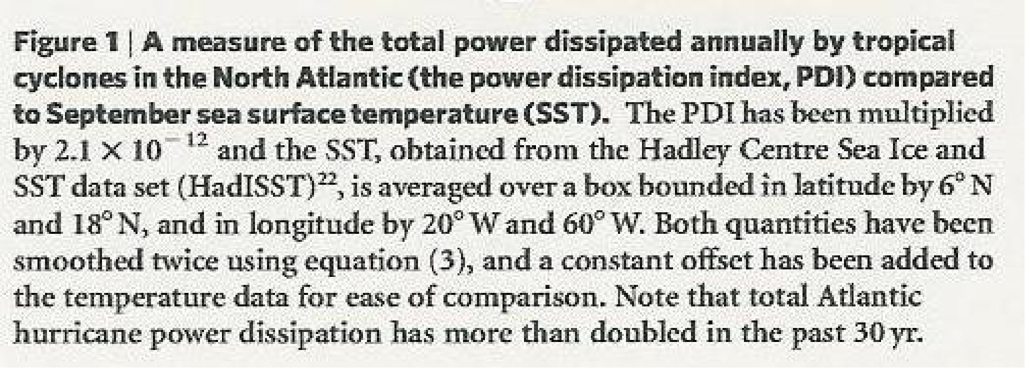


Figure 1 | A measure of the total power dissipated annually by tropical cyclones in the North Atlantic (the power dissipation index, PDI) compared to September sea surface temperature (SST). The PDI has been multiplied by 2.1×10^{-12} and the SST, obtained from the Hadley Centre Sea Ice and SST data set (HadISST)[22], is averaged over a box bounded in latitude by 6° N and 18° N, and in longitude by 20° W and 60° W. Both quantities have been smoothed twice using equation (3), and a constant offset has been added to the temperature data for ease of comparison. Note that total Atlantic hurricane power dissipation has more than doubled in the past 30 yr.

Changes in Tropical Cyclone Number, Duration, and Intensity in a Warming Environment

P. J. Webster,[1] G. J. Holland,[2] J. A. Curry,[1] H.-R. Chang[1]

We examined the number of tropical cyclones and cyclone days as well as tropical cyclone intensity over the past 35 years, in an environment of increasing sea surface temperature. A large increase was seen in the number and proportion of hurricanes reaching categories 4 and 5. The largest increase occurred in the North Pacific, Indian, and Southwest Pacific Oceans, and the smallest percentage increase occurred in the North Atlantic Ocean. These increases have taken place while the number of cyclones and cyclone days has decreased in all basins except the North Atlantic during the past decade.

16 SEPTEMBER 2005 VOL 309 SCIENCE www.sciencemag.org

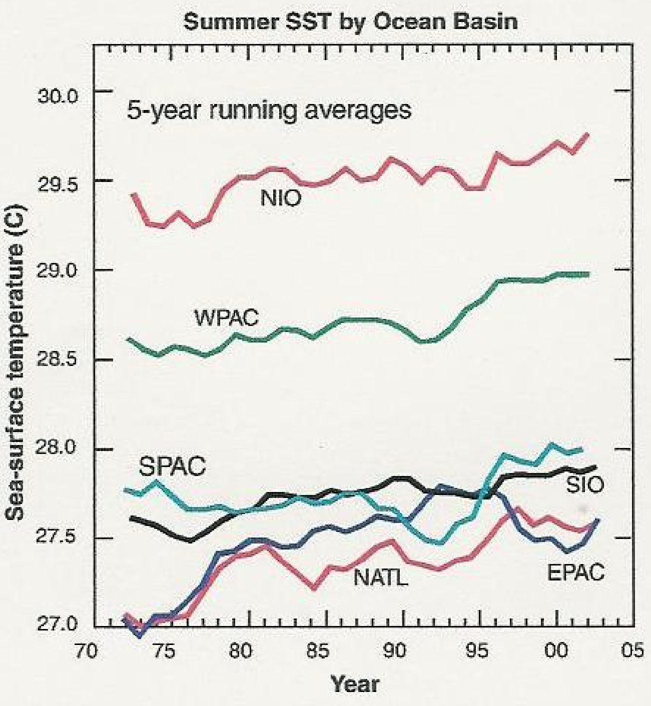


Fig. 1. Running 5-year mean of SST during the respective hurricane seasons for the principal ocean basins in which hurricanes occur: the North Atlantic Ocean (NATL: 90° to 20°E, 5° to 25°N, June-October), the Western Pacific Ocean (WPAC: 120° to 180°E, 5° to 20°N, May-December), the East Pacific Ocean (EPAC: 90° to 120°W, 5° to 20°N, June-October), the Southwest Pacific Ocean (SPAC: 155° to 180°E, 5° to 20°S, December-April), the North Indian Ocean (NIO: 55° to 90°E, 5° to 20°N, April-May and September-November), and the South Indian Ocean (SIO: 50° to 115°E, 5° to 20°S, November-April).

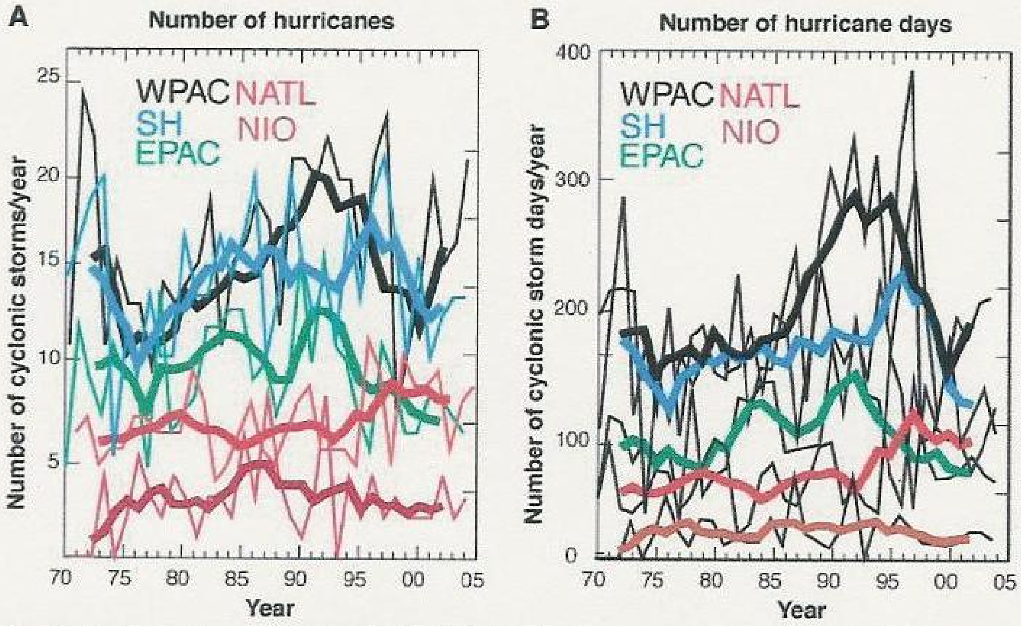


Fig. 3. Regional time series for 1970–2004 for the NATL, WPAC, EPAC, NIO, and Southern Hemisphere (SIO plus SPAC) for (A) total number of hurricanes and (B) total number of hurricane days. Thin lines indicate the year-by-year statistics. Heavy lines show the 5-year running averages.

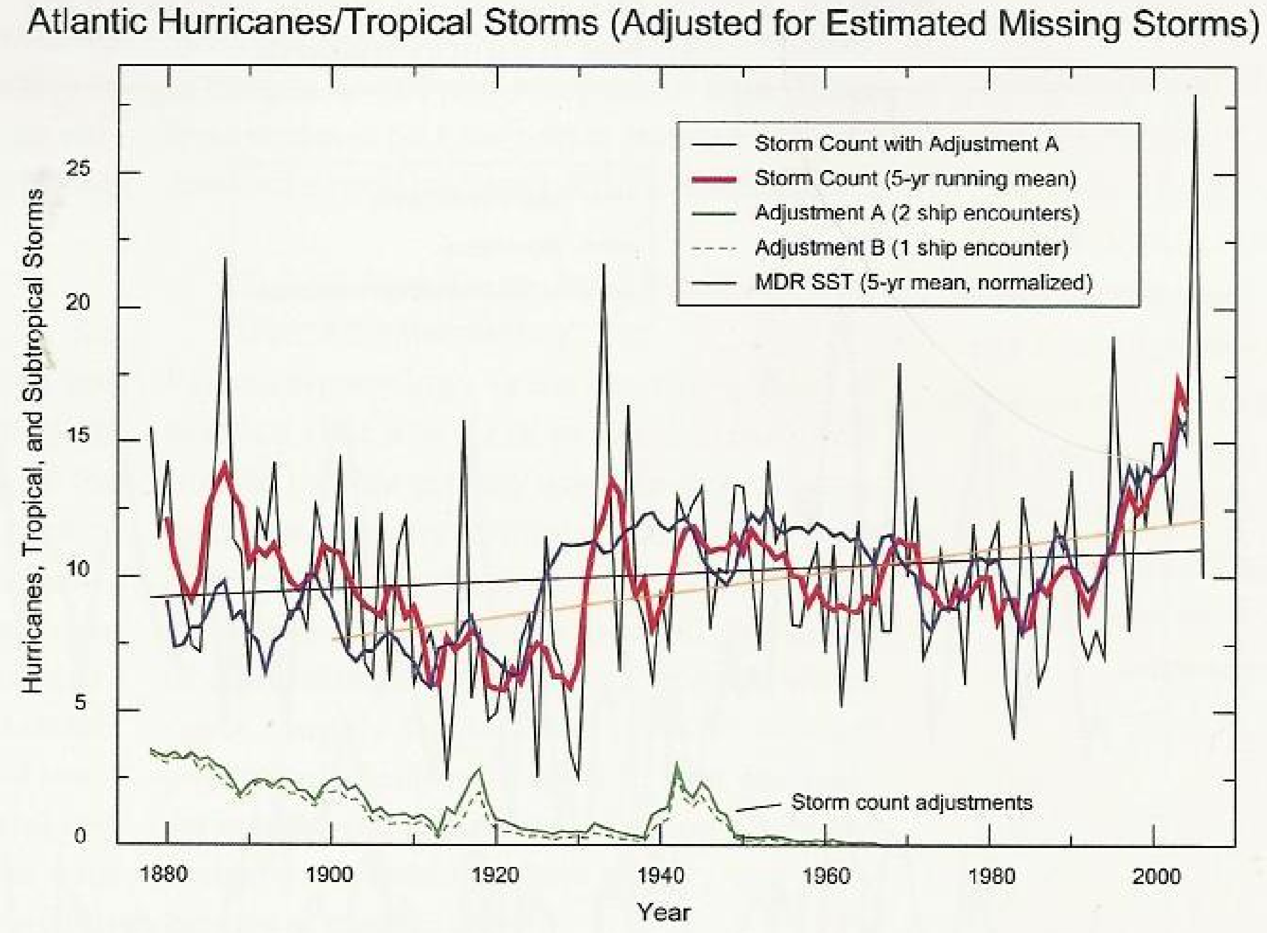


Figure 2.16 Atlantic hurricanes and tropical storms for 1878-2006, using the adjustment method for missing storms described in the text. Black curve is the adjusted annual storm count, red is the 5-year running mean, and solid blue curve is a normalized (same mean and variance) 5-year running mean sea surface temperature index for the Main Development Region of the tropical Atlantic (HadISST, 80-20°W, 10-20°N; Aug.-Oct.). Green curve shows the adjustment that has been added for missing storms to obtain the black curve, assuming two simulated ship-storm "encounters" are required for a modern-day storm to be "detected" by historical ship traffic for a given year. Straight lines are least squares trend lines for the adjusted storm counts. (Adapted from Vecchi and Knutson, 2007).

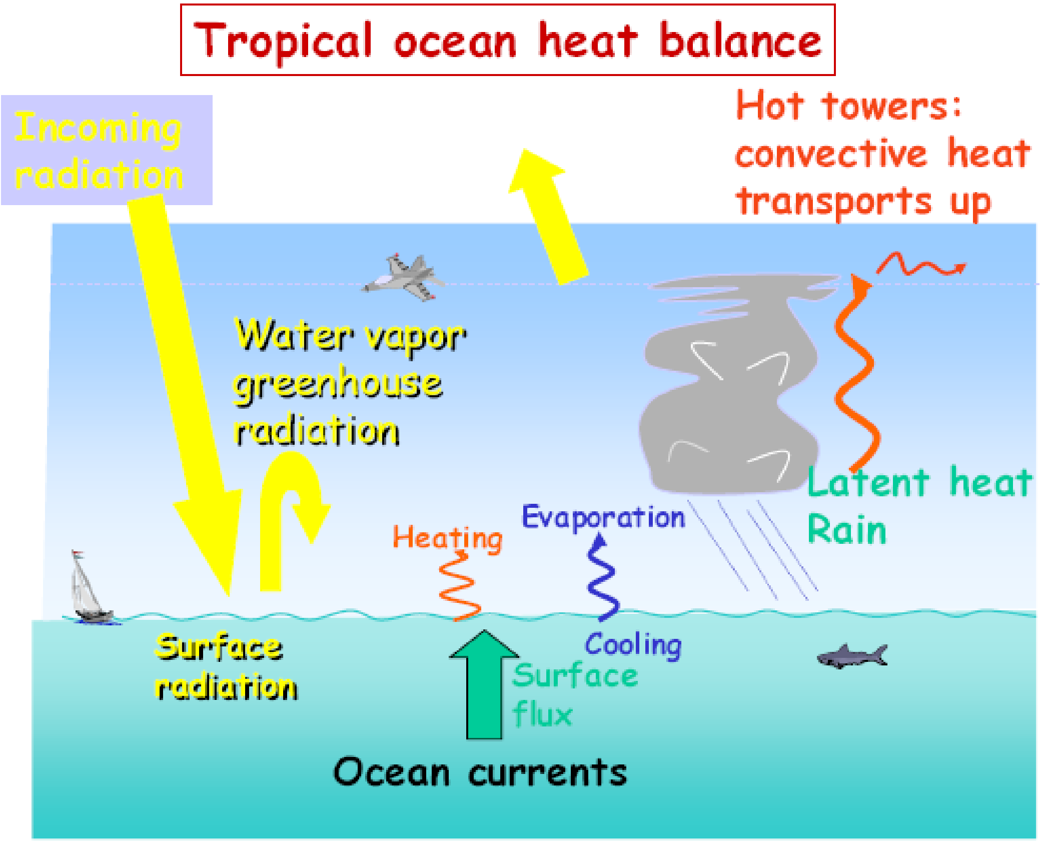
Some Basic Facts About Hurricanes

(From a presentation by Kevin Trenberth)

- SSTs $> 26^{o}$C (80^{o}F)
- High water vapor
- Weak wind shear
- Weak static stability
- Pre-existing disturbance

Also:

Large variability from year to year
El Ninõ means more activity in Pacific, less in Atlantic
Large interdecadal variability in Atlantic (AMO)



Source: Presentation by Kevin Trenberth

- Hurricanes play a key role in climate, but are not in models and are not parameterized
- Competition between thunderstorms and convection, but these are not resolved and are treated as sub-grid-scale phenomena
- Climate models have premature onset of convection
- Result: Existing models are likely to underpredict hurricanes

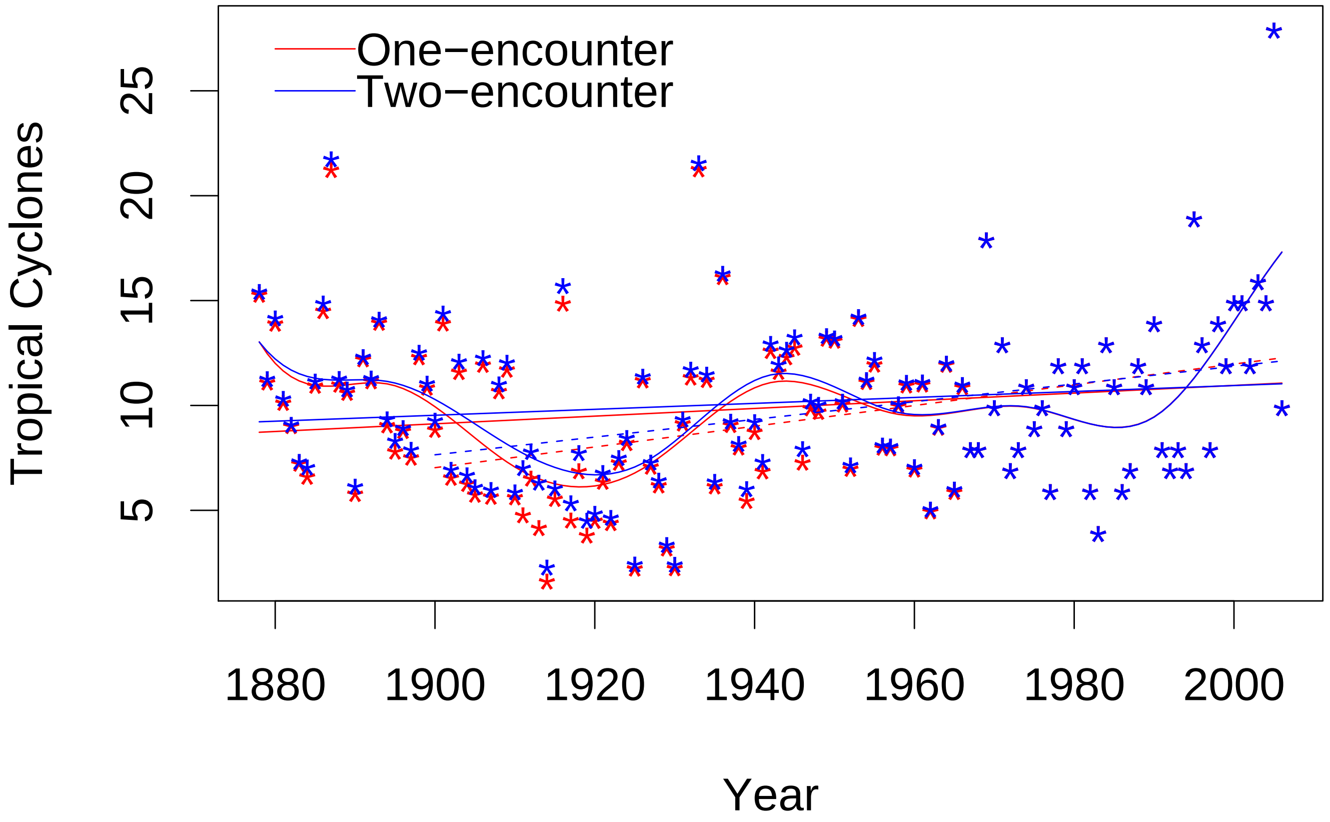
This analysis compares two reconstructions of the TC series (Vecchi and Knutson)

- One-encounter: Assumes a single encounter between a modern storm and a historical ship track is sufficient to count the storm as one that would have been observed historically

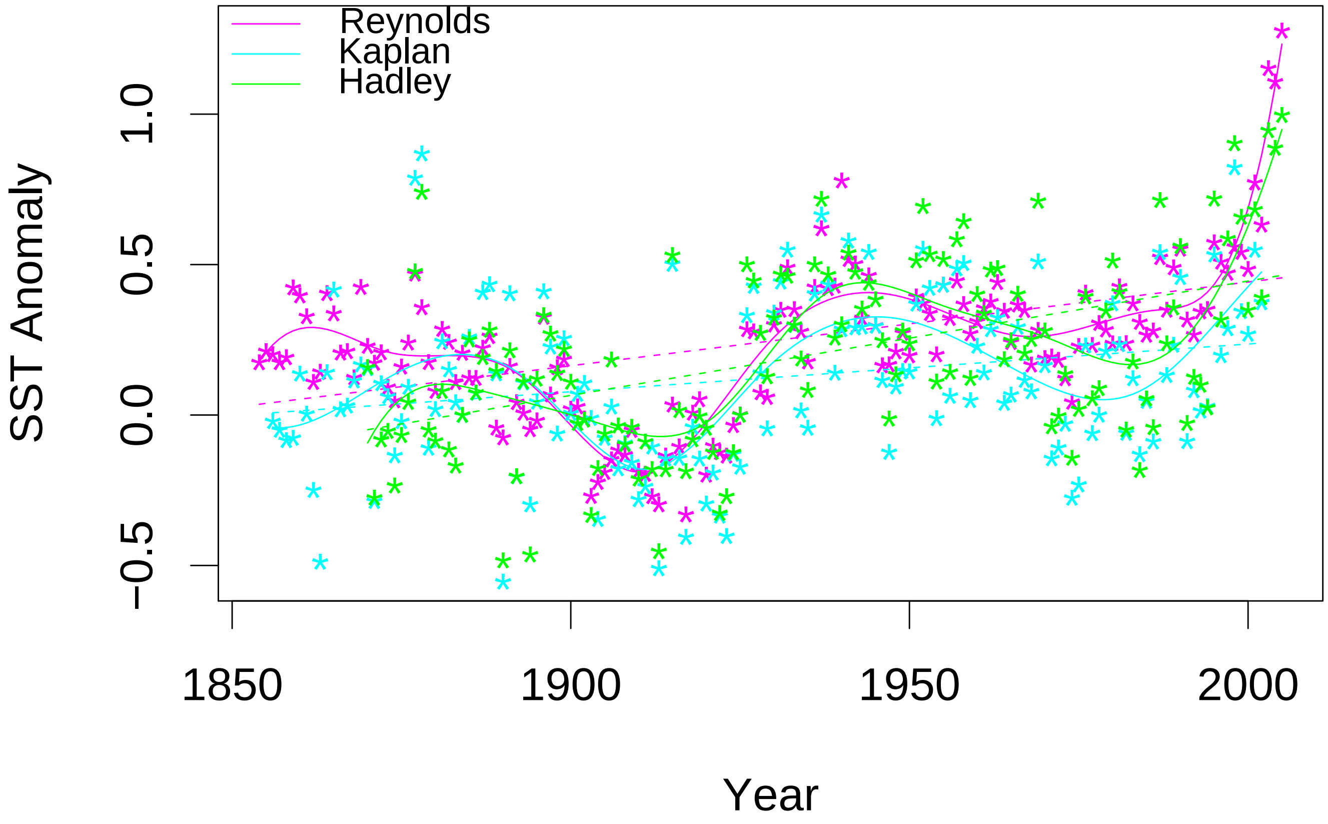
- Two-encounter: Similar, but requires two ship × storm encounters

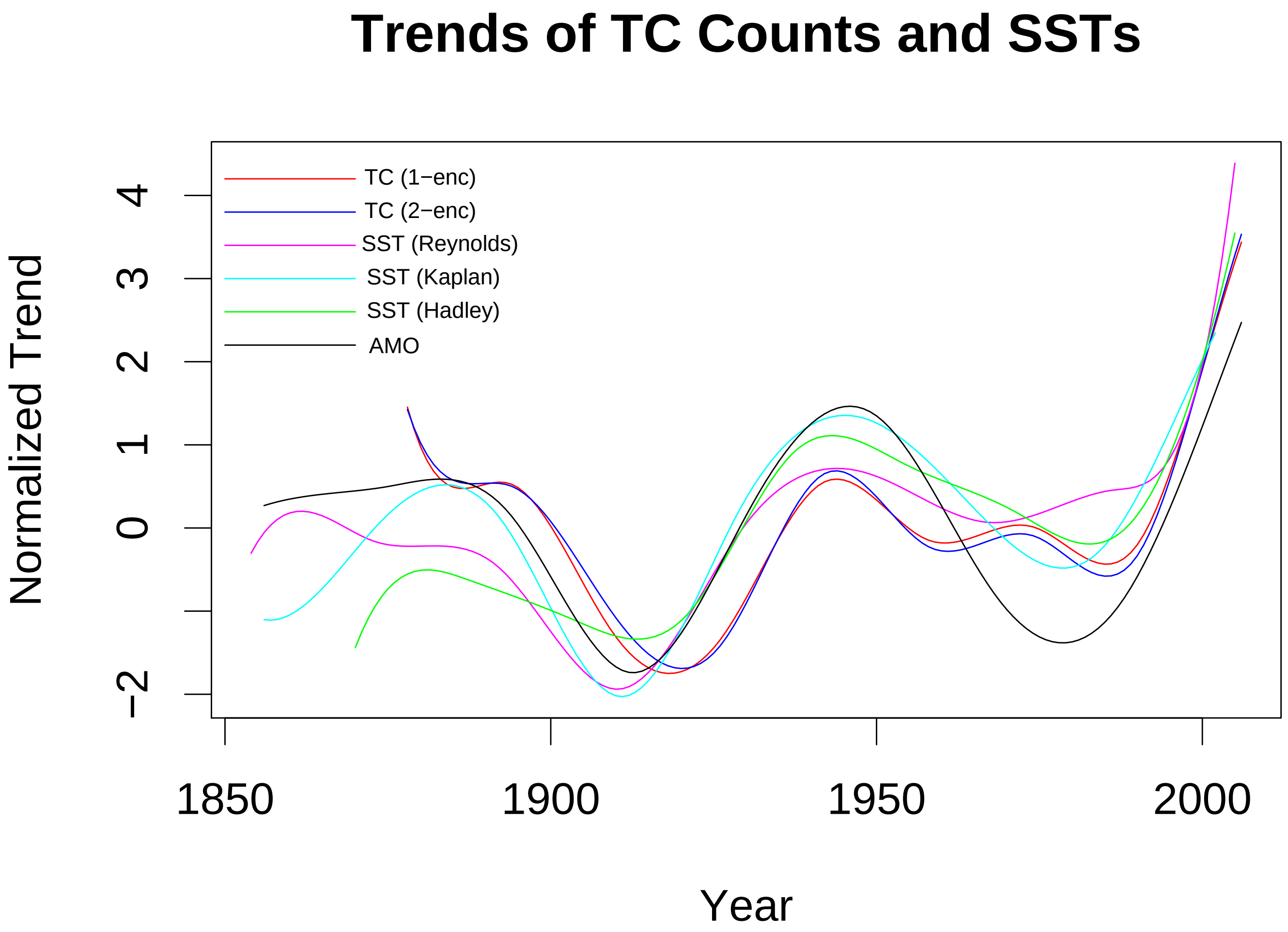
The "two-encounter" model applies a stronger correction to the historical record

Tropical Cyclones by Year
(Data by Vecchi and Knutson)



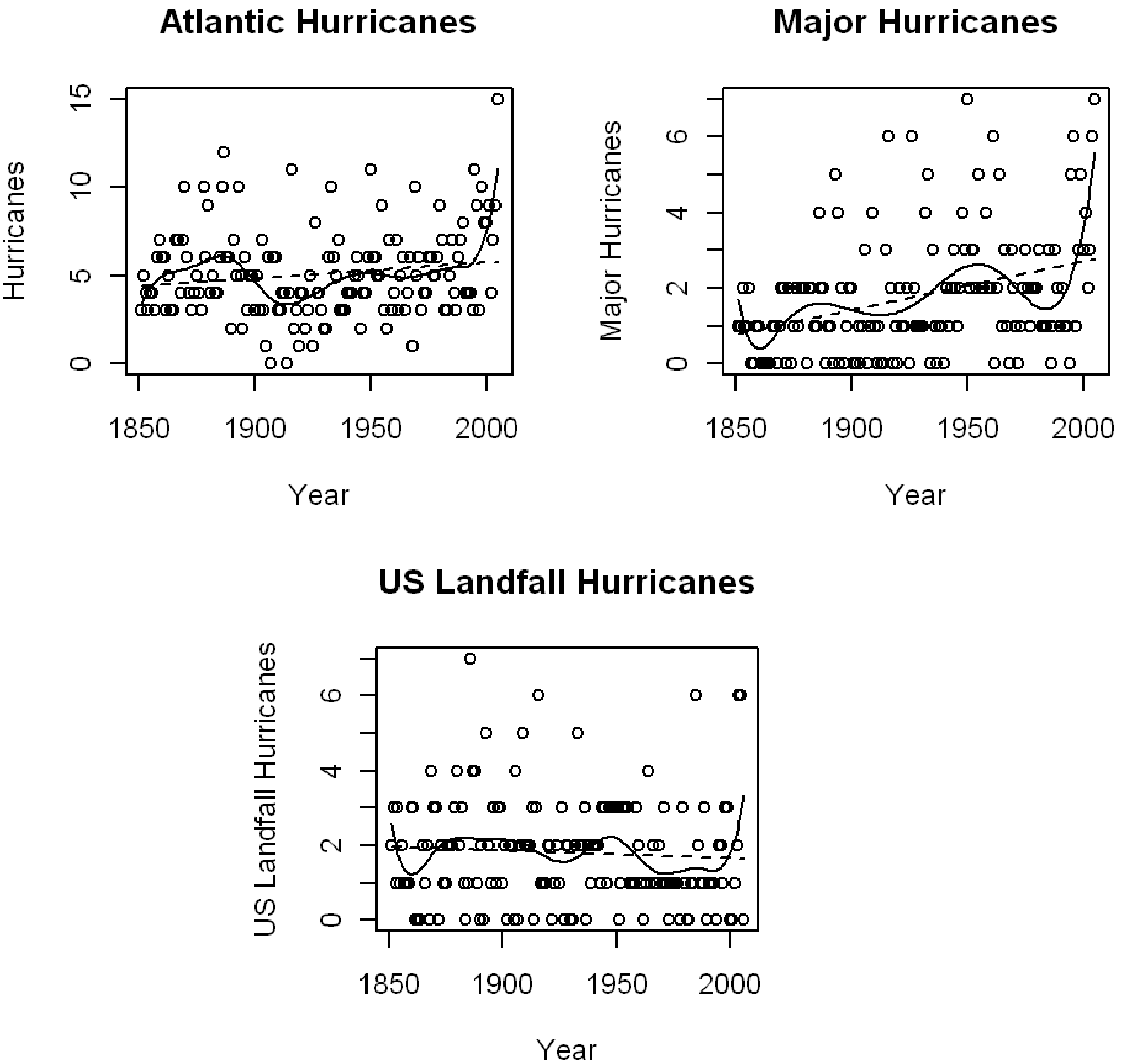
Three Reconstructions of SST




Trends of TC Counts and SSTs
TC (1−enc)
TC (2−enc)
SST (Reynolds)
SST (Kaplan)
SST (Hadley)
AMO
Normalized Trend
−2
0
1
2
3
4
1850
1900
1950
2000
Year

Trends in hurricane counts are similar to those in TCs but the data are much sparser

US landfalling hurricanes show no trend or a slight decrease but this can be explained as a sampling effect

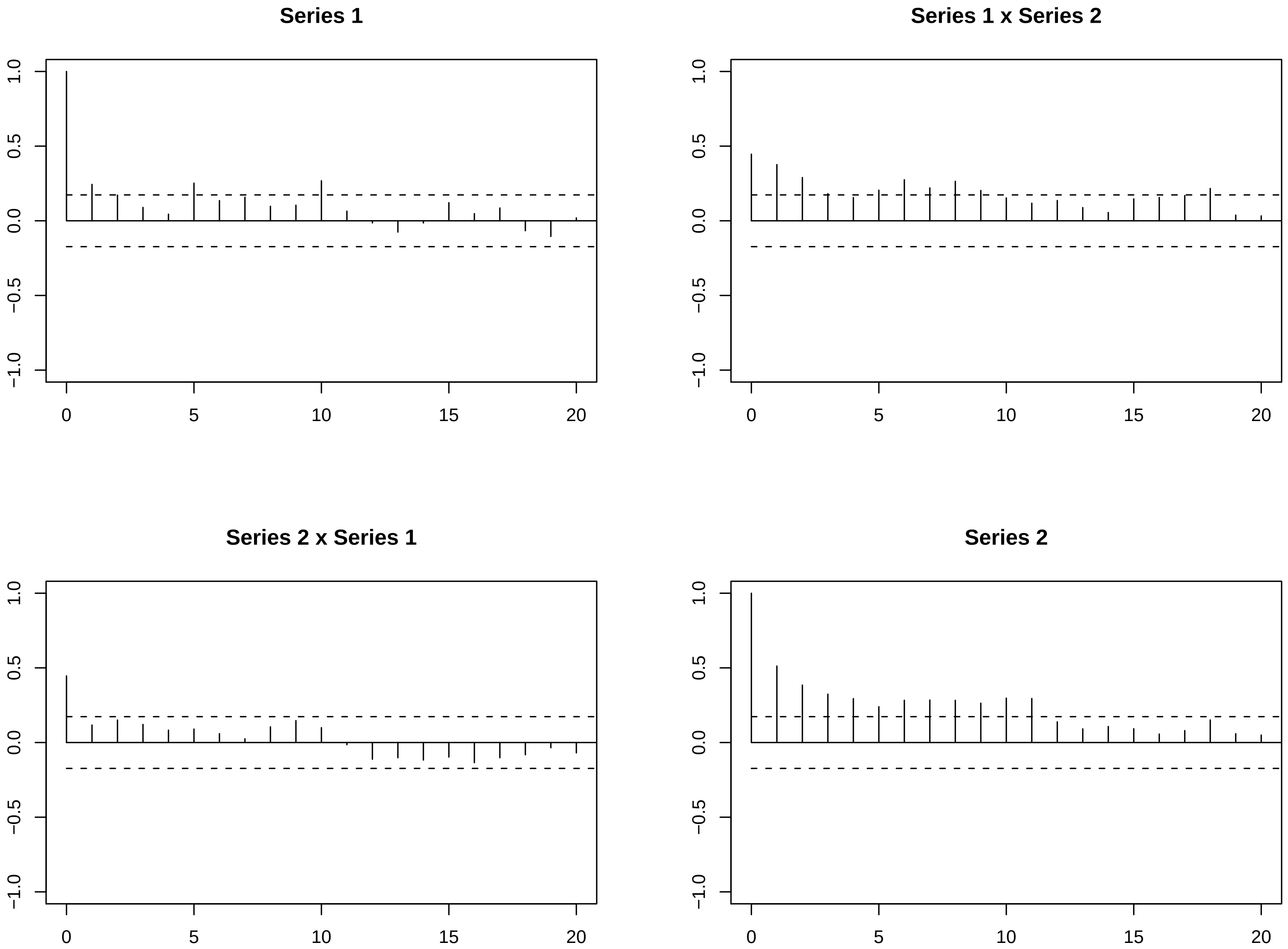


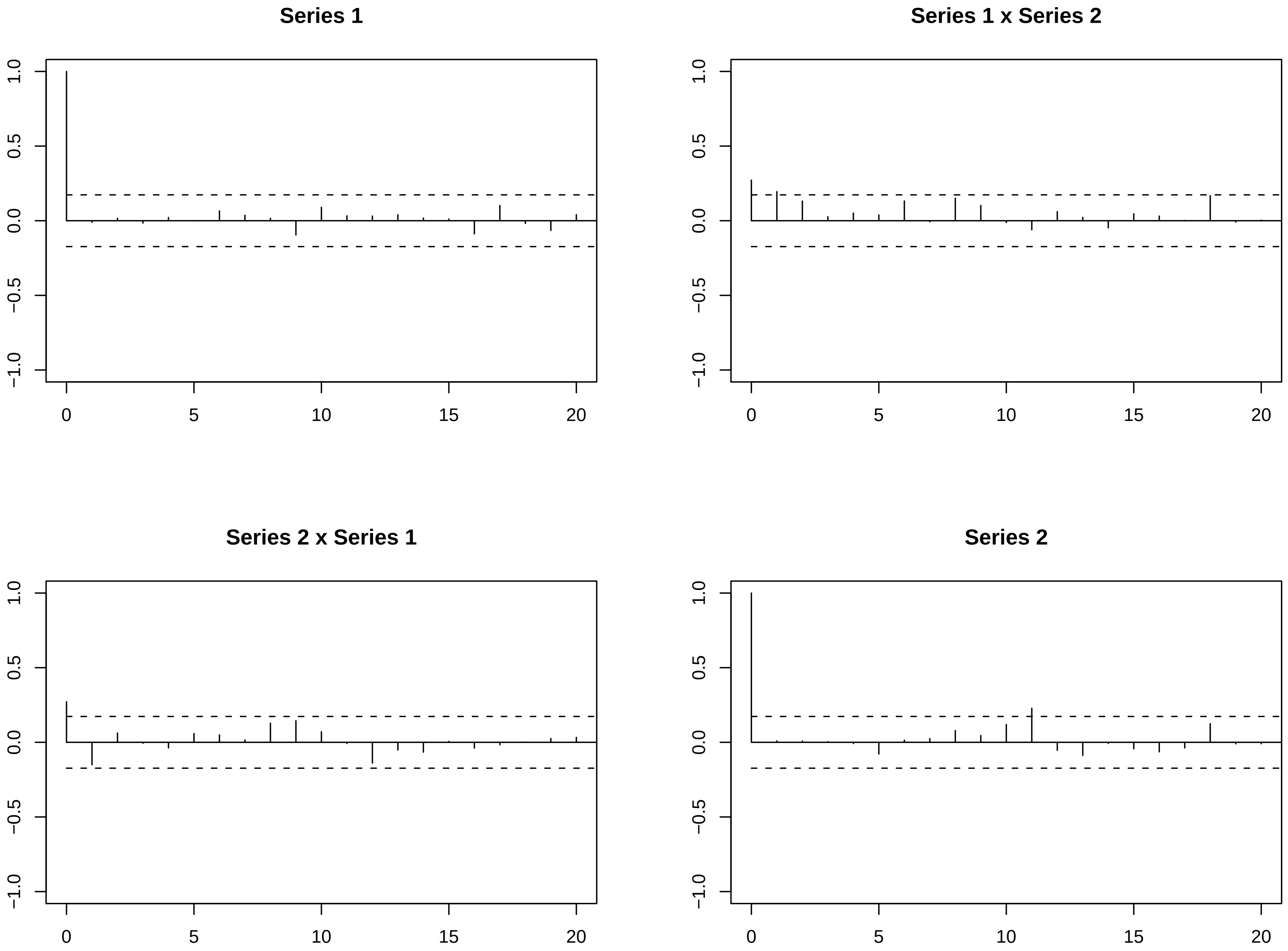
Linear and Nonlinear Trends in Three Series

- One analysis is a simple bivariate time series analysis of TCs and SSTs

- Apply square root transformation to TCs to normalize variances

- After prewhitening both series, lag-0 cross-correlation is 0.27 with a p-value of about .002


Series 1
Series 1 x Series 2
Series 2 x Series 1
Series 2
1.0
0.5
0.0
−0.5
−1.0
0
5
10
15
20


Series 1
Series 1 x Series 2
Series 2 x Series 1
Series 2
1.0
0.5
0.0
−0.5
−1.0
0
5
10
15
20

Why Linear Trends?

The gold standard for this kind of analysis is *detection and attribution analysis* (D&A), which is a statistical technique devised by climatologists that is used to apportion an observed climate signal as a combination of different forcing factors.

Outline of D&A Method

- Use a climate model to generate “signals” under different forcing factors. The main forcing factors used in practice are (a) greenhouse gases, (b) atmospheric particles (aerosols), (c) solar fluctuations, (d) volcanoes.

- Perform a multiple regression analysis to decompose the observed climate signal as a linear combination of the model signals. The details of this analysis are quite complicated, because the signals have dimension ≈ 5000.

- If the coefficient due to greenhouse gases is statistically significant, we say the greenhouse gas signal is “detected”. If this condition is satisfied, the regression coefficients are then interpreted as an “attribution” of the observed signal over different forcing factors.

This method has been applied many times to temperature series, and more recently to rainfall. However there are difficulties applying it to hurricane data, because hurricanes are not well reproduced by climate models.

- Conflicting evidence from climate models, e.g. the GFDL model suggests a levelling off of hurricane activity as SST increases (Knutson) but the NCAR model suggests the opposite conclusions (Trenberth)

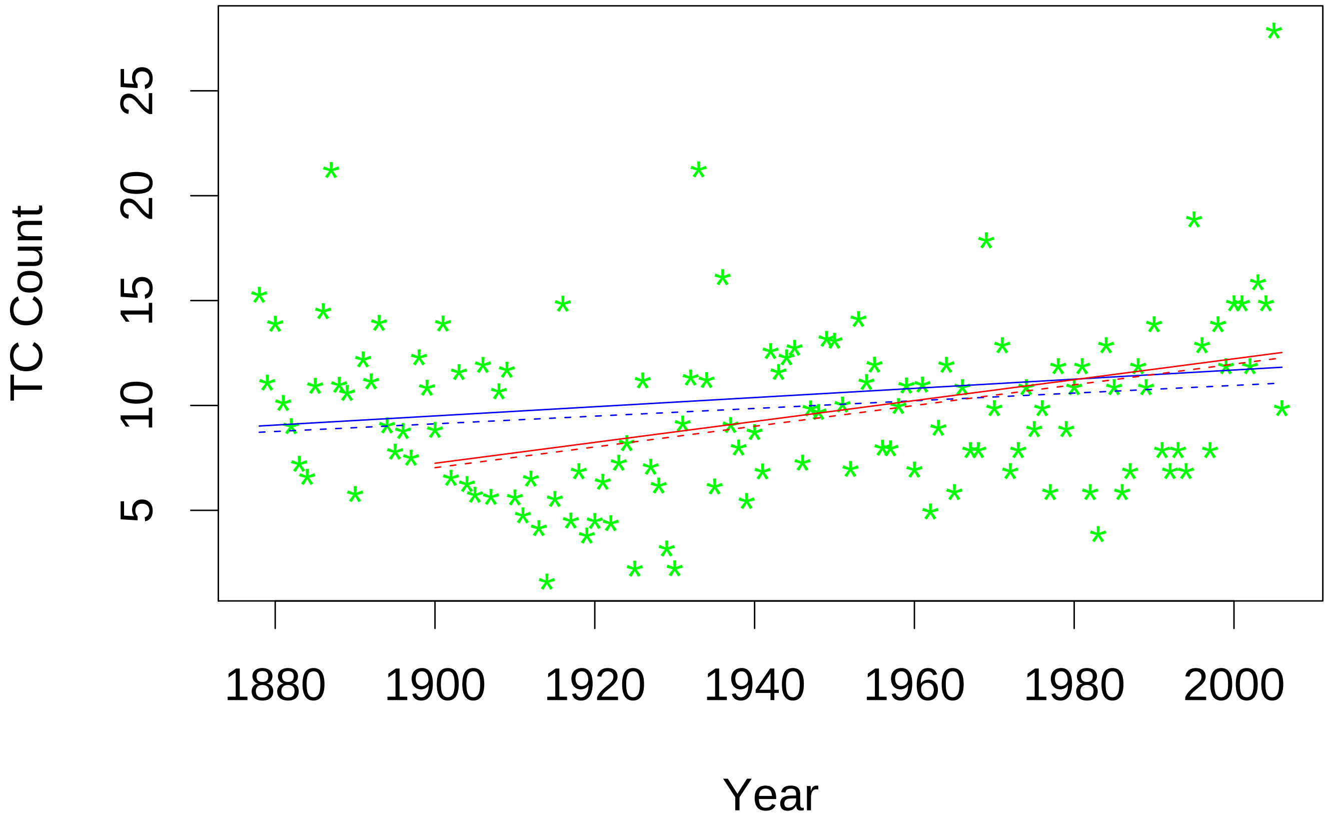
- In the absence of an agreed definition of the “greenhouse gas signal”, we use a linear trend as an approximation.

- The idea is to decompose observed SST and hurricane/storm counts as a combination of linear trend and long-term fluctuations (AMO)

Trends Fitted to Tropical Cyclones

Period	ARMA	Trend	SE	Trend/SE	p-value
1878–2006	(0,0)	0.018	0.0094	1.94	0.055
1878–2006	(9,2)	0.022	0.022	0.97	0.33
1900–2006	(0,0)	0.049	0.012	4.11	8×10^{-5}
1900–2006	(9,4)	0.050	0.020	2.54	0.011

Tropical Cyclones with Trends
OLS and ARMA Regression



Bivariate Time Series Model

x_t: SST average in year t

y_t: square root TC count in year t

Model:

$$\begin{aligned}
\begin{pmatrix} x_t \\ y_t \end{pmatrix} &= \begin{pmatrix} \beta_{0,0} \\ \beta_{1,0} \end{pmatrix} + \begin{pmatrix} \beta_{0,1} \\ \beta_{1,1} \end{pmatrix} t + \begin{pmatrix} w_t \\ z_t \end{pmatrix}, \\
\begin{pmatrix} w_t \\ z_t \end{pmatrix} &= \sum_{j=1}^{p} \begin{pmatrix} a_{0,0,j} & a_{0,1,j} \\ a_{1,0,j} & a_{1,1,j} \end{pmatrix} \begin{pmatrix} w_{t-j} \\ z_{t-j} \end{pmatrix} + \begin{pmatrix} \epsilon_{t,0} \\ \epsilon_{t,1} \end{pmatrix}, \\
\begin{pmatrix} \epsilon_{t,0} \\ \epsilon_{t,1} \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{0,0} & \sigma_{0,1} \\ \sigma_{1,0} & \sigma_{1,1} \end{pmatrix} \right].
\end{aligned}$$

Calculate $\beta_{1,1}$ and its standard error.

Selection of AR order p

p	AIC
0	−219.7
1	−240.8
2	−236.8
3	−230.2
4	−222.5
5	−228.5
6	−223.8
7	−219.8
8	−217.2
9	−211.9
10	−218.4

Conclusion: $p = 1$ seems best

Trend Coefficient $\beta_{1,1}$ and p-value

Start	One-encounter		Two-encounter	
Year	Trend	p-value	Trend	p-value
1878	.20	.25	.14	.40
1890	.51	.006	.43	.02
1900	.71	.0005	.61	.002
1910	.99	.00003	.88	.0002
1920	.86	.003	.78	.007
1930	.43	.17	.35	.27
1940	.45	.20	.35	.32
1950	.78	.08	.74	.10
1960	1.58	.006	1.55	.007
1970	2.22	.008	2.22	.008
1980	4.21	.002	4.21	.002

Still get different results corresponding to the start time

Questions About This Approach

There are clear signs of “unit root” behavior. I tried fitting multivariate ARMA models (in SAS - PROC VARMAX) but these generally don’t converge, because of unit root difficulties.

For the hurricane series on its own, ARMA models fit much better than AR, but still difficulties with AIC and similar measures (indicative of non-stationarity?)

I also tried fractionally differenced processes but this doesn’t seem to resolve the issues either.

Current Approach — Conditional ARMA

Recall
x_t: SST in year t
y_t: square root TC in year t

Also let T_t denote some modeled trend (initially linear, but later we consider alternatives)

Model:

$$\begin{aligned} x_t &= \alpha_0 + \alpha_1 T_t + u_t \text{ } (u_t \text{ ARMA}) \\ \widehat{u}_t &= x_t - \widehat{\alpha}_0 - \widehat{\alpha}_1 T_t, \text{ (residuals)} \\ y_t &= \beta_0 + \beta_1 T_t + \beta_2 \widehat{u}_t + \beta_3 \widehat{u}_{t-1} + \beta_4 \widehat{u}_{t-2} + v_t \text{ } (v_t \text{ ARMA}) \end{aligned}$$

First attempt: T_t linear

Use TC1 dataset (“1-encounter”), Hadley SST
ARMA(1,1) for u_t
ARMA(7,2) for v_t

Focus on coefficient of trend in y_t (β_1 parameter), various start years ending in 2005

Start Year	$\hat{\beta}_1$	SE	t ratio	p value
1880	1.18	0.93	1.28	0.20
1890	1.66	0.87	1.92	0.05
1900	2.19	0.84	2.62	0.01
1910	3.14	0.93	3.36	0.00
1920	2.68	0.99	2.71	0.01
1930	1.71	1.08	1.59	0.11
1940	2.87	1.97	1.46	0.14
1950	3.38	2.23	1.52	0.13
1960	6.41	2.51	2.55	0.01
1970	8.28	3.00	2.76	0.01

Conclusions from this analysis

The results are not a big advance on fitting linear trends without involving SST (cp. Vecchi-Knutson 2008)

We find significant linear trends from starting times near the bottom of the AMO (e.g. 1900, 1960), but not from others such as 1880, 1940.

No clear-cut conclusion of an anthropogenic trend.

Second attempt: T_t based on a GCM

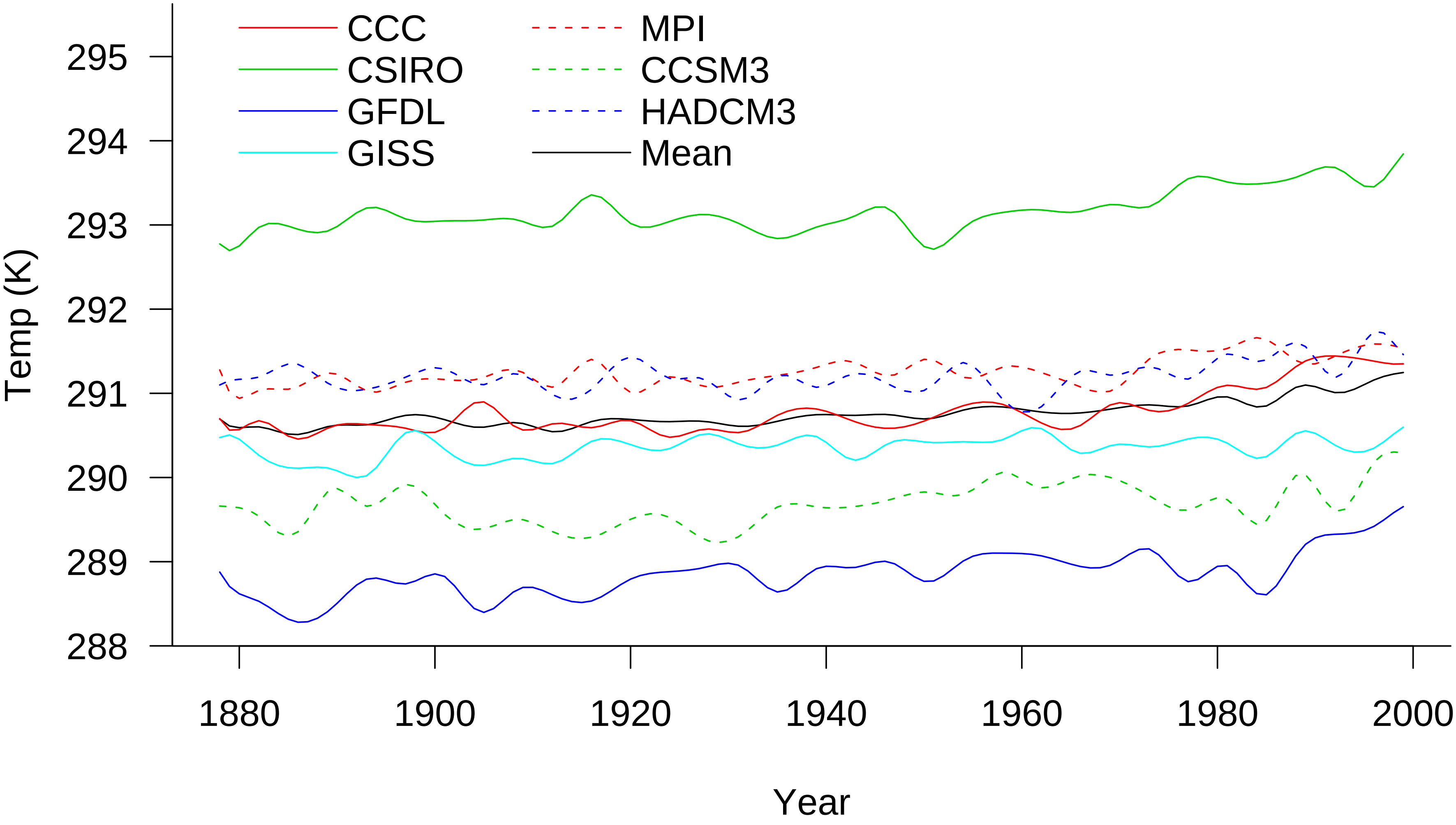
We use 20th Century runs from seven climate models (N.B. one run from each, though in several cases there are multiple model runs available)

Compute mean temperature over 0–60 $^{\circ}$N and 7.5–75 $^{\circ}$W as proxy for North Atlantic SST

Apply spline smoother to remove local variation

Also plot mean of all 7 models

SST 20th Century Model Projections



We fit same model using smoothed mean of seven GCMs as T_t

This time use AR(1) for u_t, AR(5) for v_t.

Start Year	$\widehat{\beta}_1$	SE	t ratio	p value
1880	1.39	0.50	2.78	0.005
1890	1.70	0.47	3.66	0.0003
1900	1.82	0.47	3.85	0.0001
1910	1.96	0.50	3.90	0.0001
1920	1.81	0.54	3.34	0.0008
1930	1.37	0.50	2.74	0.006
1940	1.98	0.65	3.05	0.002
1950	2.08	0.64	3.26	0.001
1960	2.43	0.64	3.81	0.0001
1970	2.56	0.52	4.95	10^{-6}

Conclusion: a statistically significant trend ($p < 0.01$) from all starting times

Other Statistical Issues

- This analysis assumes TCs and SSTs are normal after taking square route transformations of SSTs — ignores the fact that TCs are counts

- Possible alternative model based on Poisson counts

- Existing literature (Elsner, Tsonis, Jagger) has used Bayesian MCMC methods to create hierarchical model

- We are looking at an alternative approach extending Davis's and Dunsmuir's GLARMA modeling approach (Evangelou, in progress)

Conclusions

- Simple bivariate time series analysis shows strong cross-correlation between TCs and SSTs, but this could be a by-product of AMO — not directly indicative of anthropogenic trend

- Simple linear trend analysis shows conflicting conclusions — results highly sensitive to start time, analysis ignores interdependence between TCs and SSTs

- Preferred analysis takes both time series and trend effects into consideration

- Correlating observed time series with trends from climate models seems most effective approach. However, there are still arguments about the data.