Statistical Models for Tropical Cyclone Activity

James B. Elsner and Gabriele Villarini

March 10, 2011

1 Introduction

1.1 What does tropical cyclone activity mean?

Tropical cyclone activity refers to the number of cyclones in a given region over a given time period. It can also refer to the number of cyclones exceeding a threshold intensity level (e.g. 17 m s^{-1} for tropical storms and 33 m s^{-1} for hurricanes/typhoons) or the percentage of all cyclones that strike land. Or it can refer to a metric that captures several attributes at once, like the accumulated cyclone energy.

1.2 What is a statistical model?

Statistics is a way to describe and predict tropical cyclone activity. Like people all hurricanes are unique. A person's behavior might be difficult to anticipate, but the average behavior of a group of people is quite predictable. A particular hurricane may move eastward through the Caribbean Sea, but on average hurricanes travel westward with some degree of regularity. Statistical models quantify this regularity and assign a level of uncertainty to it. Indeed statistical models provide a syntax for the language of uncertainty. Preference is given to models that are simple, elegant, and explanatory.

1.3 Data

Much of what we know about tropical cyclone activity derives from data. Statistical models are built from data. Observational precision generally increases over time leading to heterogeneous data sets, but a good strategy is to include as much information as possible. One approach is to model the data directly and include terms for the probability of missing reports and changing levels of data precision. Another approach is to make adjustments to the data to reflect the potential that a cyclone was not observed. In either case, models built on a Bayesian framework are particularly flexible in this regard. The underlying principle is the accumulation of evidence. Evidence can include historical or geologic data that, by their very nature, are incomplete and fragmentary. This is particularly important in climate change studies.

1.4 Statistics and arithmetic

It might seem surprising but doing statistics is not the same as doing arithmetic. For example, a strong hurricane hitting a region might be described as a 1-in-100-year event. Doing arithmetic one might conclude that after a storm hits there is plenty of time ($100 - 1 = 99$ years) before the next big one. This is wrong. Doing statistics one can expect a storm of this magnitude or larger to occur once in every 100 years *on average*, but it could occur next year with a probability of 1 %. Moreover, storms only a bit weaker will be more frequent and perhaps a lot more so.

1.5 Statistical models and empirical methods

A statistical model is an example of an empirical method. Empirical methods include the principal component techniques taught as part of a statistical methodology course in meteorology, oceanography and geology. But a principal component analysis is not a statistical model in the sense we mean here because it lacks an estimate of uncertainty. It is not our intention to pit statistical models against alternative approaches and argue for supremacy; the best approach will invariably depend on the application and the choice should be based on the substance of the problem. Moreover the search for a universally best method is quixotic.

1.6 Use R

Foundational pillars of science include transparency and reproducibility. The powerful R language for statistical modeling makes developing, maintaining and documenting code easy. It contains a number of built-in mechanisms for organizing, graphing, and modeling data. Directions for obtaining R, accompanying packages and other sources of documentation are provided at `http://www.r-project.org/`. Anyone serious about statistical modeling should learn R. We provide some code here to help you get started.

2 Modeling tropical cyclone counts

2.1 Poisson distribution

As mentioned tropical cyclone activity typically refers to the number of storms occurring within a region over a given time period. The data set `UShur1851-2010.txt` contains a list of tropical cyclone counts by year making land fall in the United States (excluding Hawaii) at hurricane intensity. To read the data into R and save them as a data object type:

```
> landfall = read.table("UShur1851-2010.txt", header = T)
```

To examine the first six lines of the data object, type:

```
> head(landfall)

  Year US MUS G FL E
1 1851  1   1 0  1 0
2 1852  3   1 1  2 0
3 1853  0   0 0  0 0
4 1854  2   1 1  0 1
5 1855  1   1 1  0 0
6 1856  2   1 1  1 0
```

The columns include year, number of U.S. hurricanes, number of major U.S. hurricanes, number of U.S. Gulf coast hurricanes, number of Florida hurricanes, and number of East coast hurricanes in order. To make the individual columns available by column name, type:

```
> attach(landfall)
```

The total number of years in the record is obtained and saved in `n` and the average number of U.S. hurricanes is saved in `rate` using the following two lines of code.

```
> n = length(US)
> rate = mean(US)
```

By typing the names of the saved objects, the values are printed.

```
> n

[1] 160

> rate

[1] 1.69375
```

Thus the 160 years of counts indicate a mean number of U.S. hurricanes equal to 1.7.

The number of years with a particular hurricane count provides a histogram. The shape of the histogram suggests that a Poisson distribution might be a good model for hurricane counts. The Poisson distribution has the interesting property that its variance is equal to its mean. The density function of the Poisson distribution shows the probability of obtaining a count x when the mean count is λ is given by

$$p(x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}. \tag{1}$$

Thus the probability no events is $p(0) = \mathrm{e}^{-\lambda}$.

With $\lambda = 1.7$ hurricanes per year, the probability of no hurricanes in a random year is

```
> exp(-1.7)
```

```
[1] 0.1826835
```

Of course, this implies that the probability of at least one hurricane is $1 - .18$ or 82 %.

You can do this for any number of hurricanes using the `dpois` function. For example, to determine the probability of observing exactly one hurricane when the rate is 1.7 hurricanes per year, type:

```
> dpois(x = 1, lambda = rate)
```

```
[1] 0.3113602
```

Or the probability of five hurricanes expressed in percent is

```
> dpois(5, rate) * 100
```

```
[1] 2.135399
```

Note that you can leave off the argument names in the function if the argument values are placed in the default order. The argument order can be found by placing a question mark in front of the function name and leaving off the parentheses. This brings up the function's help page.

To answer the question, what is the probability of one or fewer hurricanes, we use the cumulative probability function `ppois` as follows

```
> ppois(q = 1, lambda = rate)
```

```
[1] 0.495189
```

And to answer the question, what is the probability of more than one hurricane, we add the argument `lower.tail=FALSE`.

```
> ppois(q = 1, lambda = rate, lower.tail = FALSE)
```

```
[1] 0.504811
```

You can simulate another sample of hurricane counts over this many seasons with an annual rate of 1.7 hurricanes per year. You saved the number of years and the rate, so to generate 160 random values from a Poisson distribution, type:

```
> hu = rpois(n = n, lambda = rate)
```

For comparison, we plot the time series and histogram of the actual counts above the same time series and histogram of the simulated counts in Fig. 1. The counts vary from year to year and the values for a particular year do not match values for those years from the actual sample, but the distribution is quite similar with the most active year having seven strikes and about 30 of the years having no hurricanes. This demonstrates one important use of statistical

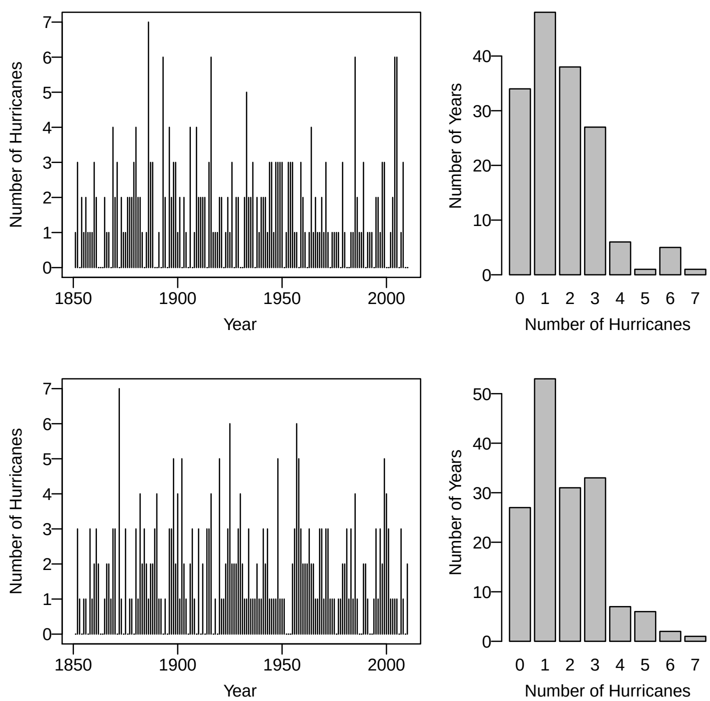

Figure 1: Time series and distribution of hurricane counts. The counts are the annual number of U.S. land falling hurricanes. The top two panels show the actual counts and the bottom two show a single realization from a random Poisson model with a rate equal to the actual long-term land fall rate.

models; to simulate random samples with the same statistical characteristics as the observations.

This can be useful at the next level of analysis. An example is the use of hurricane counts over time as input to a deterministic model of coastal sediment transport. Summary characteristics of a 100 years of tropical cyclones at the location of interest may be of little value, but running a coastal sediment transport model with a large number of 100-year cyclone samples may give a realistic assessment of the uncertainty in hurricane-induced transport caused by natural variation in activity.

As mentioned, tropical cyclone activity sometimes refers to a metric that captures several characteristics at once. Different metrics have been suggested with the two most common being Accumulated Cyclone Energy (ACE) and Power Dissipation Index (PDI). Empirical distributions of the PDI and ACE suggest asymmetric probability density functions (specifically, a positive skewness) so a useful model for the PDI might be a gamma or a Weibull distribution.

The components of these metrics including cyclone frequency, intensity, and duration are controlled by different environmental factors and each has a different statistical distribution, so it would be possible to combine separate (not necessarily independent) models for each component. A statistical model of aggregated hurricane wind losses uses this approach where the number of loss events is modeled as a Poisson distribution with the amount of loss given an event modeled as an independent log-normal distribution.

2.2 Regression models

We saw in the previous section that the parameter of interest in a Poisson model for hurricane counts is the rate. Given the rate, we can generate a probability distribution corresponding to the possibility of having any number of hurricanes. But what if the rate of hurricanes depends on climate variables or is trending? Note how the question is worded. Focus is on the rate of hurricanes. Given the rate, the counts follow a Poisson distribution.

If interest is on the relationship between climate variables and hurricane frequency then a regression model that specifies that the logarithm of the annual rate is used. The logarithm of the rate is linearly related to the predictors, expressed statistically as:

$$\log(\lambda) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon. \tag{2}$$

Here there are p predictors indicated by the x_i's and $p + 1$ coefficients (β_i's). The vector ε is a set of independent and identically distributed residuals. The assumption of independence can be checked by examining whether there is temporal correlation or other patterns in the residual values.

Note the above Poisson regression model uses the logarithm of the rate as the response variable. This is different than a regression on the logarithm of counts. Poisson regression is nonlinear in the regression function, but linear in regression structure and the model coefficients are determined by the method of maximum likelihoods rather than the method of least squares used in ordinary regression.

As a side note, the word "predictor" is the generic term for an explanatory variable in a statistical model. A further distinction is sometimes made between covariates, which are continuous-valued predictors and factors, which can take on only a few values that may or may not be ordered.

The Poisson regression model in equation 2, in which the logarithm of the rate of occurrence is a linear function of predictors, is a generalized linear model (GLM). It can be extended by considering smooth functions for terms on the right-hand side. This is known as a generalized additive model (GAM) and the functions can be parametric or non-parametric. This allows a covariate to have a nonlinear relationship with the rate function.

2.3 Choosing predictors

Statistical modeling usually refers to choosing a set of variables that can be used to predict the response. A particular set defines a candidate model. A problem arises when there are a large number of candidate models. How do you choose? There are some strategies.

To help decide whether to include a variable in a model, it is important to have a statistic that accounts for a better model fit (reduced bias) and the loss of a degree of freedom (increased variance). The degrees of freedom are the sample size minus the number of model parameters. Adding a variable to the statistical model reduces the bias but increases the variance. This is known as the bias-variance trade-off. A commonly used statistic that preforms this trade-off calculation is called the Akaike Information Criterion (AIC) given by

$$\mathrm{AIC} = 2(p+1) + n[\log(\mathrm{SSE}/n)], \tag{3}$$

where p is the number of predictors and SSE is the residual sum of squares. We can compare the AIC values when each predictor is added or removed from a given model. For example, if after adding a predictor, the AIC value for the model increases then the trade-off is in favor of the extra degree of freedom and against retaining the predictor.

Stepwise regression is a procedure (not a model) for automating model selection. It is useful for finding the best model from a large set of candidate models under the condition that the models depend linearly on their unknown parameters (models are nested).

Another issue to consider is the correlation among predictors (collinearity). Correlation between predictors ($r > .6$) can result in an unstable model because the standard errors on the predictor coefficients are not estimated with enough precision. This means that it might be hard to tell whether the response increases or decreases for a given change in the predictor even if there exists strong correlation between the response and the predictor variable. An informative prior understanding of the partial correlation may help argue in favor of retaining two highly-correlated predictors, but in the usual case of a vague

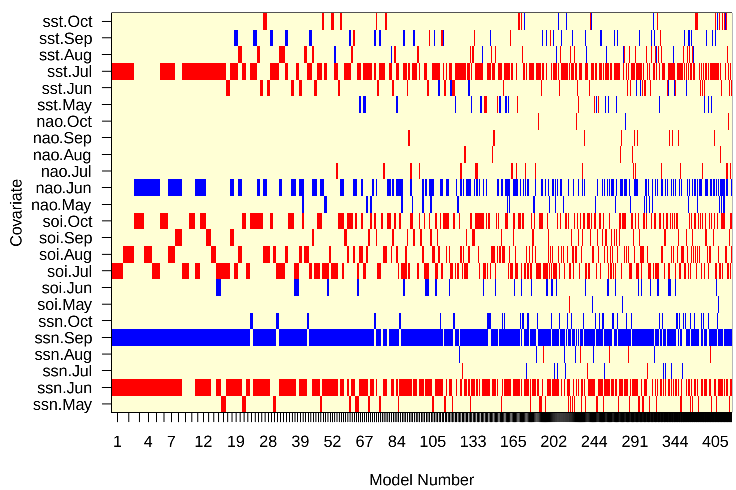

Figure 2: Model covariates versus model number. The covariates are individual monthly (May through October) averages of North Atlantic sea-surface temperature (sst), the North Atlantic oscillation (nao), the Southern Oscillation Index (soi), and sun spot numbers (ssn). If a covariate with a positive (negative) relationship with U.S. hurricanes is included in the model it is indicated by a red (blue) bar. The bar height is constant and the bar width is proportional to the probability of the model given the data. Model probabilities decrease with increasing model number.

prior it is better to eliminate the predictor whose relationship with the response variable is harder to explain physically or that has the smaller correlation with the response variable.

2.4 Using a consensus model

A consensus model is an attractive alternative to choosing a single best model. Using Bayes' rule, which states that the posterior probability is proportional to a prior distribution times the likelihood function, a probability is computed for a set of candidate models given the data. As an example, Fig. 2 shows 24 predictors (vertical axis) of U.S. hurricane frequency versus model number.

A predictor is colored red or blue if it is included in the model. Predictor inclusion is based on selection criterion similar to the AIC. The models are ordered from left to right by decreasing probability. The predictor candidates are listed on the vertical axis. The bar color corresponds to the parameter sign (red for positive and blue for negative). The signs indicate the chance of a U.S.

hurricane increases with July SST, July SOI, and June SSN and decreases with September SSN. The width of the bar is proportional to the model's probability so the bars become narrower with increasing model number.

Predictions are made using each model and then averaged. The average is weighted by the model's probability. The assumption is that each model explains the data to some degree or another. By using many models the uncertainty associated with model selection is automatically included in the prediction variance.

2.5 Cross validation

If the statistical model will be used to make forecasts, a cross validation is needed. Cross validation is a procedure to assess how well a prediction algorithm (whether or not it involves model selection) will do in forecasting the unknown future. In the case of independent hurricane seasons, the cross validation typically involves withholding a season's worth of data, developing the algorithm on the remaining data, then using the algorithm to predict data from the withheld season. The result is an estimate of out-of-sample error that will more precisely capture actual forecast errors when the model is used to predict the future.

3 Modeling tropical cyclone intensity

3.1 Parametric models

How strong a hurricane gets depends on a number of conditions including but not limited to the amount of ocean heat, proximity to land, and wind shear. One indicator of the accumulated effect of these conditions over the life of a hurricane is its lifetime strongest wind speed. A more practical concern might be the highest tropical cyclone wind speed over a limited area (a city) over a given year.

A useful statistical model for wind speeds from tropical storms affecting a given area is the Weibull distribution. Expressed in terms of a cumulative distribution where $\Pr(W > v)$ is the probability of observing a wind speed from a tropical cyclone exceeding some threshold wind speed value v, the survival function of the Weibull distribution is

$$\Pr(W > v) = \exp\left[-\left(\frac{v}{b}\right)^a\right] \tag{4}$$

where a is the shape parameter and b is the scale parameter.

For example, for the set of lifetime stongest wind speeds for all North Atlantic tropical cyclones (over 17 m s^{-1}) over the period 1981–2006, a is estimated (using the method of maximum likelihoods) to be 2.64 and b is estimated to be 42.5 m s^{-1}. Thus given the data and the statistical model, the probability that a random tropical cyclone will have wind speeds stronger than 50 m s^{-1} is given by

```
> exp(-(50/42.5)^2.64)

[1] 0.2152839
```

or 21.5 %. A limitation of the Weibull model for these wind speeds is the data have a lower bound. The bound is used to define the event as a tropical cyclone.

For the strongest wind speeds a generalized Pareto distribution (GPD) is a better alternative. For example, the exceedances $W - u$ (peaks over threshold) are modeled as samples from a GPD, so that for a hurricane with wind speed W,

$$\Pr(W > v|W > u) = \left(1 + \frac{\xi}{\sigma}[v - u]\right)^{-1/\xi} \tag{5}$$

where $\sigma > 0$ and $\sigma + \xi(v - u) \geq 0$.

In the Weibull and GPD cases, a Poisson model for the frequency of storms above the threshold wind speed is needed to specify a return period or annual probability for a given wind speed or stronger. The model parameters including the shape and scale for the Weibull and ξ and σ for the generalized Pareto distribution and the frequency rate λ can be adjusted or regressed to study how climate variables influence wind speed return periods. The GPD model is particularly useful in the context of examining sediment records of storminess since the marine signature is interpreted as a threshold event indicating a prehistoric storm of this magnitude or greater.

3.2 Quantile regression

Quantile regression is an alternative approach for examining climate influences on wind speed exceedances. Quantile regression extends ordinary regression to quantiles of the response variable. Quantiles are points taken at regular intervals from the cumulative distribution function of a random variable. Quantiles mark a set of ordered data into equal-sized data subsets. To illustrate, suppose in a sample of wind speeds 25 % of them are less than 20 m s^{-1}, 50 % of them are less than 35 m/s and 75 % are less than 45 m s^{-1}. Thus we state there is an equal number of wind speeds between 20 and 35 m s^{-1} as there is between 35 and 45 m s^{-1}.

Ordinary regression is a model for the conditional mean, where the conditioning is on the value of the explanatory variable. Likewise, quantile regression is a model for the conditional quantiles. It can be considered a semi-parametric technique because it relies on quantiles (non-parametric), but uses parameters to assess the relationship between the quantile and the covariates. In general, parametric models are more informative. A parametric model involves more stringent assumptions, but it is a good strategy to start with stronger assumptions and back off toward weaker assumptions when necessary. However, parametric models are more sensitive to outlying data points and the parameters may not be as easy to understanding physically.

Figure 3 shows the results of a quantile regression model using North Atlantic lifetime maximum intensity as the response variable and Atlantic SST

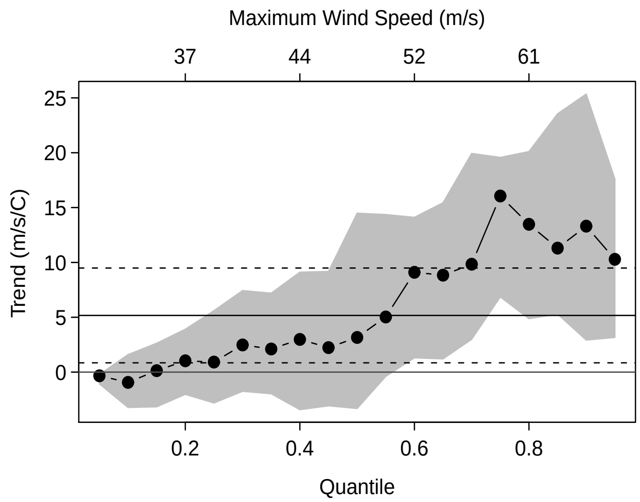

Figure 3: Quantile trends with respect to Atlantic SST controlling for ENSO. For an increase in SST there are increases in hurricane intensity with increases generally larger (above 10 m s^{-1} per °C) at higher hurricane intensities. Statistically significant trends are noted for quantile values above the median hurricane intensity.

and Pacific SOI as the explanatory variables. The trend values are plotted for percentile values of wind speed between 5 % and 95 % in intervals of 5 %. Trend values for Atlantic SST range from near zero for weaker hurricanes (lowest quantiles) to between 10 and 15 m s^{-1} per °C for stronger hurricanes. The trends are statistically significant for hurricanes above the 60th percentile (on average, above 52 m/s) as can be seen by the 90 % confidence band above the zero trend line. The trend peaks at 16 m s^{-1} per °C at the 75th percentile with a 90 % confidence interval of between 7 and 20 m s^{-1} per °C. The mean regression line indicates a trend of about 5 m s^{-1} per °C which is statistically significant above the zero trend line as indicated by the dashed lines (90 % confidence intervals).

3.3 Changes in frequency and intensity

There tends to be some confusion about the relationship between changes in the intensity and changes in the frequency of tropical cyclones. Today on average approximately 20 % of the strongest cyclones globally exceed 49 m s^{-1}. Suppose that with a 1 °C rise in ocean temperature, 20 % of the strongest cyclones exceed 51 m s^{-1} in the future. This means that the 80th percentile intensity increases by only 2 m s^{-1}. However, in terms of frequency, today 17 cyclones per year exceed 49 m s^{-1} and 13 exceed 51 m s^{-1}. After a 1 °C warming 51 m s^{-1} is the new 80th percentile thus, without a change in the overall number of cyclones, 13 becomes 17. So although the theory relates warmer oceans to stronger hurricanes, a future increase in the percentile intensity might mean more strong hurricanes.

4 Modeling tropical cyclone activity spatially

Hurricane climate studies are hampered by the relatively short archive of storms. The sample size can be increased by modeling hurricane activity spatially. Spatial statistical models are widely used in epidemiology, but have yet to be employed in hurricane climatology. The key components are a method to spatially aggregate storm data and models that make explicit use of the resulting spatial autocorrelation.

A scaffolding is need to colocate track-relative attributes with climate data. A tessellation of the tropical cyclone domain using hexagonal grids is one example. Hexagons are more efficient than squares at covering storm tracks meaning for equal area grids, it takes fewer hexagons than squares to cover the track.

Values obtained by aggregating observations within contiguous grids will, in general, be self (auto) correlated. Neighboring grid values will tend to be more similar than grid values farther away. Like temporal autocorrelation, spatial autocorrelation is about this proximity, but it is more complex because of the extra dimension.

Spatial statistical models incorporate spatial autocorrelation. This makes the model parameters stable and statistical tests more reliable. For instance, confidence intervals on a regression slope from a spatial regression model will have the proper coverage probabilities and the prediction errors will be smaller compared with the non-spatial alternative.

Spatial dependency can enter the regression model directly by adding a lagged variable to the model or by adding a spatially correlated error term. Spatial dependency can also enter the model by allowing the relationship between the response and the predictors to vary across the tiling. This is called geographically-weighted regression and the model parameter values vary with grid location.

This is accomplished by weighting the grid values near the grid of interest more than the grid values farther away. For example, a regression of storm intensity on SST is performed at a particular hexagon grid using the paired intensity and SST values at each grid across the domain, with the paired values inversely weighted by their distance to the particular grid. Geographically-weighted regression is thus equivalent to local linear regression in the space of predictor variables.

Interest may be storm intensity as a function of SST, but the highest storm intensity within the grid will also depend on the number of observations. In general a grid with a larger number of storm hours will have a higher intensity. Thus the spatial model includes SST and storm hours as covariates in which case the SST coefficient is the marginal effect on intensity after accounting for storm hours.

Figure 4 shows the results. The grids are colored according to the value of the SST coefficient. Hexagons with positive coefficients indicating a direct relationship between storm strength and ocean warmth in m s^{-1} per °C are displayed in yellow to red colors and those with negative coefficients are shown with blue hues. Grids with the largest positive coefficients (greater than 4 m s^{-1} per °C) are found over the Caribbean Sea extending into the southeast Gulf of Mexico and east of the Lesser Antilles. Positive coefficients extend over much of the Gulf of Mexico and northeastward up the eastern seaboard. A region of negative coefficients (greater than -4 m s^{-1} per °C) is noted over the central North Atlantic.

To assess model adequacy the residuals from the model are examined. A residual is the difference between the intensity observed in the grid and intensity predicted by the geographically-weighted regression model. Other spatial regression models can be employed using this framework. For instance, if interest is on cyclone counts or the presence/absence of cyclones then a Poisson or logistic model is used.

5 Future work

Recent upward trends in hurricane activity have spurred a healthy debate on the possible connection between hurricane activity and climate change. Much of the contention derives from results using bi-variate analysis with basin-wide data. Results from multivariate statistical models will clarify issues in this debate. Models capable of handling data that are not all equally precise will also help. In particular, the spatial modeling framework described here is quite useful for combining numerical model output with observations. Output from a global circulation model can be respecified from latitude/longitude coordinates to equal-area hexagons providing a basis for comparing storm climatologies from different models. Careful application of statistical models will lead to a better understanding of the physical mechanisms underlying the relationship between hurricanes and climate.

6 Selected references

- Elsner, J. B., and C. P. Schmertmann, 1994: Assessing forecast skill through cross validation, *Weather and Forecasting*, **9**, 619–624.

- Elsner, J. B., and B. H. Bossak, 2001: Bayesian analysis of U.S. hurricane climate, *Journal of Climate*, **14**, 4341–4350.

- Elsner, J. B., and T. H. Jagger, 2004: A hierarchical Bayesian approach to seasonal hurricane modeling, *Journal of Climate*, **17**, 2813–2827.

- Elsner, J. B., and T. H. Jagger, 2006: Prediction models of U.S. hurricane activity, *Journal of Climate*, **19**, 2935–2952.

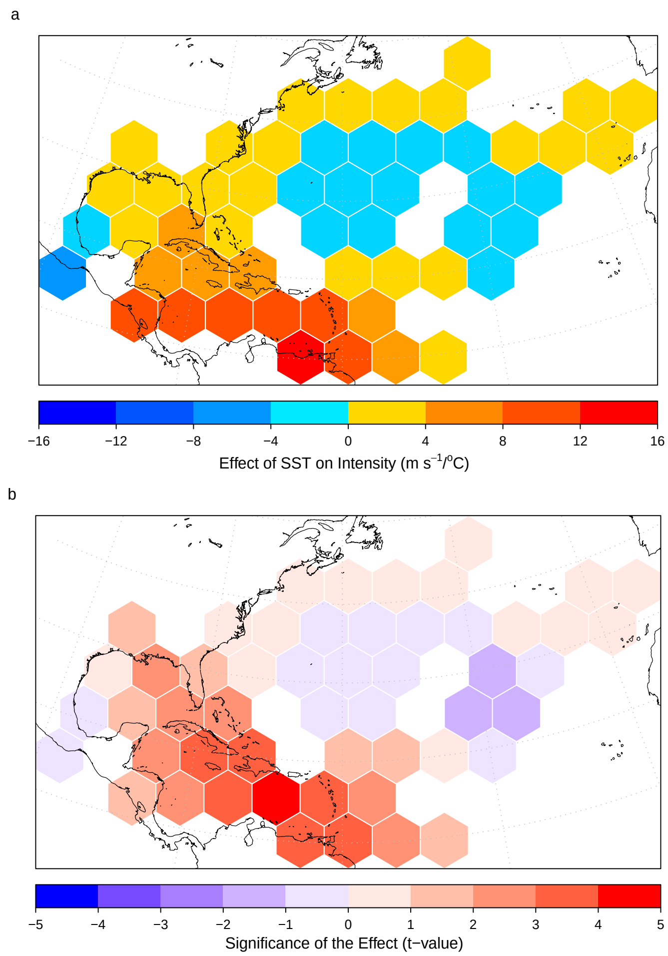

Figure 4: Effect of SST on storm intensity (a) Trends in grid maximum intensity as a function of SST holding storm count constant. A geographically-weighted regression is used to borrow information from neighboring grids. The bandwidth for the smoothing is 256.5 km. (b) The ratio of the trend to the standard error of the trend as a measure of statistical significance. Values greater than an absolute value of 2 are considered significant above a trend of zero and are shaded using an inverted color transparency level.

- Elsner, J. B., J. P. Kossin, and T. H. Jagger, 2008: The increasing intensity of the strongest tropical cyclones, *Nature*, **455**, 92–95.

- Elsner, J. B., T. H. Jagger, and K.-b. Liu, 2008: Comparison of hurricane return levels using historical and geological records, *Journal of Applied Meteorology and Climatology*, **47**, 368–374.

- Jagger, T. H., J. B. Elsner, and X.-F. Niu, 2001: A dynamic probability model of hurricane winds in coastal counties of the United States, *Journal of Applied Meteorology*, **40**, 853–863.

- Jagger, T. H., X.-F. Niu, and J. B. Elsner, 2002: A space-time model for seasonal hurricane prediction, *International Journal of Climatology*, **22**, 451–465.

- Jagger, T. H., and J. B. Elsner, 2006: Climatology models for extreme hurricane winds near the United States, *Journal of Climate*, **19**, 3220–3226.

- Jagger, T. H., and J. B. Elsner, 2008: Modeling tropical cyclone intensity with quantile regression, *International Journal of Climatology*, **28**, doi:10.1002/joc.1804.

- Jagger, T. H., and J. B. Elsner, 2010: A consensus model for seasonal hurricane prediction, Journal of Climate, 23, 6090–6099.

- Malmstadt, J. C., J. B. Elsner, and T. H. Jagger, 2010: Risk of strong hurricane winds to Florida cities, *Journal of Applied Meteorology and Climatology*, **49**, 2121–2132.

- Villarini, G., G. A. Vecchi, and J. A. Smith, 2010: Modeling of the dependence of tropical storm counts in the North Atlantic Basin on climate indices, *Monthly Weather Review*, **138**, 2681–2705.