Section I: The Path to Better Prediction of Hurricane Economic Loss

On October 29, 2012, Hurricane Sandy made landfall and caused widespread damage along the eastern seaboard. Although it had a weak maximum wind speed of 75 mph, Hurricane Sandy led to a total loss of approximately 51.2 billion dollars. Upon watching the disturbing images of wreckage on television, I was overwhelmed with sadness and curiosity. When I heard news reporters claiming that Sandy was as extremely large in size but its wind speed was not very high, I was surprised that a seemingly weak hurricane could be so destructive. I then began to wonder the significance of hurricane size in determining the huge economic loss. After talking to my mother, an atmospheric scientist, and Dr. Lixin Zeng, an expert on hazard insurance, I learned that many empirical hurricane loss models solely rely on wind speed to determine the overall loss and ignore the role of size. I realized that I had the opportunity to discover something brand new. Inspired, I rifled through internet databases and statistical models to develop an economic loss model that uses a variety of predictors for hurricane loss.

In 2013, I constructed the multi-variate regression using Microsoft Excel from my home computer. In the summer of 2014, I was a summer intern at the California Institute of Technology and Jet Propulsion Laboratory, where I considered regional wealth and storm duration as additional factors for economic loss using MATLAB programming language. To perform the least-squares regressions, I had to study basic statistics. I learned about the correlation coefficient, explained variance, p-value, and other parameters important to regressions. After understanding these concepts, I had the foundation necessary to thoroughly analyze the data.

For weeks, I had fumbled in the dark while blindly combining wind speed and size in arbitrary function forms to fit actual losses. Unlike solving a math problem, scientific research has inherent risk because the destination cannot be foreseen. Dedicating countless hours to rifling through statistical analyses, I took the risk of chasing an answer that may not even exist. Being immersed in this unpredictability kept me intrigued as well as excited. Mathematics and science have the potential to solve the mysteries of the natural world and establish order amid seemingly random phenomena. Empowered by advanced math and science training, my naturally inquisitive mind has developed to tackle real-world problems and make original contributions to science and society.

Having an innate curiosity, I have strived to make sense of the world around me. People should pursue their intellectual curiosities and passions rather than passively shrug off innovative thoughts. When facing an obstacle, they should persevere and seek ways to overcome it or improve. If I had not actively searched for answers, I would not have made a difference in the world. Research appears to be a daunting task, but the experience is extremely rewarding. Even the smallest idea has the potential to become a scientific breakthrough. With curiosity and persistence, this potential will transform into reality.

Section II: Research on Hurricane Economic Loss

Many empirical hurricane loss models consider only the dependency of loss on maximum wind speed and neglect other factors. These models cannot explain the fact that many low intensity storms, such as Hurricane Sandy, caused substantial losses. Surprisingly, there has not been any study that quantifies the role of storm size in affecting hurricane damage. To improve the prediction of losses, I worked on using the multi-variate regression method to construct a hurricane loss model that takes into account the effects of maximum wind speed and storm size, and later added storm duration and regional gross domestic product (GDP) to these models.

I used a wide variety of data to create an optimal hurricane loss model. The US hurricane loss data and maximum wind speed data are downloaded from the ICAT Damage Estimator website (http://www.icatdamageestimator.com/). For storm size, I obtained data from the National Hurricane Center Extended Best Track (EBT) database. I used R_{34}, the radii of a storm where wind speeds at the 10-meter height above the surface are 34 knots, and R_{out}, the radius of the outer limit of a storm. In total, 73 tropical cyclones (TCs) served as the basis for my analysis.

Previous studies showed that hurricane loss approximately follows a power-law relation with maximum wind speed; therefore, I used a power-law function form for maximum wind speed, storm size and duration. For regional GDP, I found an exponential fitting captures more variance than a power-law fitting. The form of my fitting function is:

$$L = 10^{k+d\,(GDP)} {V_{max}}^{a} R^{b} \tau^{c},$$

where *k* (a scaling factor), *a*, *b*, *c*, and *d* are fitting parameters determined by the multi-variate regressions.

The fitting coefficients and their p-values, explained variances (R^2) and root-mean-squares (RMSs) of fitting residuals were compared for each regression in order to examine which regression models provide the most accurate and statistically significant fittings. The explained variance quantifies how well a regression model captures the variance of the loss data. The *p*-value measures the probability of obtaining the same results through a random fitting, or the statistical significance of the results. An optimal regression model would have a high explained-variance but a low *p*-value, i.e., the model provides a close estimate of the actuals that is unlikely duplicated through a fitting of randomly generated data. The RMS of the fitting residuals is a metric of the accuracy of the model when estimating the actual loss. A low RMS means the model's estimated values are close to the actual values while a high RMS means the model's estimated values are far off from the actual values. Therefore, a low RMS is preferred.

By comparing the explained variances of various regression models, I found that adding storm size has the most significant improvement to the traditional models. Using wind speed and size together as predictors captures about 70% of the variance of hurricane losses, which is significantly larger than the 39% explained variance of the model that uses wind speed only. I conducted bi-variate regressions for subsets of the 73 TCs based on V_{max} thresholds since stronger storms may have different dependencies compared to those of weaker storms. As the maximum wind speed threshold increased, the fitting coefficients tended to increase as well.

Table 1. Regression results using V_{max} and R_{34} as predictors for loss, following the function form $L=10^k V_{max}{}^a R_{34}{}^b$, compared to the regression results using V_{max} and R_{34} alone. The subsamples are determined by the V_{max} values. R^2 is the explained variance of loss by a regression model.

Threshold V_{max}	Sample Size	R^2	a	b	k	R^2 (V_{max} only)	a (V_{max} only, b=0)	R^2 (R_{34} only)	b (R_{34} only, a=0)
>=35	73	0.45	4.19	1.25	-1.83	0.39	5.27	0.26	2.36
>=60	64	0.58	6.78	1.43	-7.31	0.52	7.77	0.23	2.57
>=65	60	0.55	6.92	1.44	-7.62	0.48	7.69	0.18	2.32
>=70	53	0.62	6.29	1.82	-7.11	0.49	7.60	0.31	2.75
>=75	43	0.69	4.98	2.66	-6.22	0.40	7.11	0.51	3.36
>=80	38	0.75	6.53	2.61	-9.30	0.57	9.01	0.51	3.92
>=85	30	0.75	6.82	2.48	-9.64	0.50	8.07	0.41	3.10
>=90	27	0.74	7.80	2.59	-11.90	0.44	8.42	0.37	2.85
>=100	24	0.64	8.82	3.13	-15.17	0.30	6.73	0.16	2.09
>=110	15	0.75	11.97	4.44	-24.62	0.23	6.54	0.16	2.17
>=115	13	0.80	12.11	4.34	-24.72	0.25	6.92	0.20	2.31

During the summer of 2014, I further tested the sensitivity of the bi-variate fittings to using R_{out} instead of R_{34}. I compared the explained variances and fitting coefficients between the regressions using R_{34} and R_{out}, as shown in Table 2.

Table 2: Regression results using Best Track V_{max} and R_{34} versus using V_{max} and R_{out} (73 storms).

		R_{34}				R_{out}			
Threshold V_{max}	Sample Size	*a*	*b*	*k*	R^2	*a*	*B*	*k*	R^2
≥35	73	4.19	1.25	-1.83	0.45	4.77	1.85	-4.68	0.45
≥60	64	6.78	1.43	-7.31	0.58	7.51	2.27	-11.08	0.61
≥65	60	6.92	1.44	-7.62	0.55	7.65	2.37	-11.59	0.58
≥70	53	6.29	1.82	-7.11	0.62	7.46	2.61	-11.77	0.64
≥75	43	4.98	2.66	-6.22	0.69	6.89	2.90	-11.27	0.63
≥80	38	6.53	2.61	-9.30	0.75	8.48	2.26	-13.06	0.68
≥85	30	6.82	2.48	-9.64	0.75	7.99	2.36	-12.30	0.69
≥90	27	7.80	2.59	-11.90	0.74	8.85	2.63	-14.70	0.71
≥100	24	8.82	3.13	-15.17	0.64	9.47	2.73	-16.23	0.57
≥110	15	11.97	4.44	-24.62	0.75	14.24	3.89	-29.01	0.67
≥115	13	12.11	4.34	-24.72	0.80	14.17	3.68	-28.39	0.68

For relatively weaker storms with V_{max} < 75 mph (the 3rd to 6th rows in Table 3), using R_{out} (marked in orange) explains slightly higher variance than using R_{34} (marked in yellow). For Category 1 and stronger hurricanes with $V_{max} \geq$ 75 mph, using R_{34} (marked in light green) explains noticeably higher variance than using R_{out} (marked in dark green). A reason for this may be that the full range of spiraling winds account for storm losses when the maximum wind speeds are relatively weak. When maximum wind speeds are very high, the inner part of the storms, i.e., the area covered by R_{34}, are more destructive than the outer part of the storms beyond R_{34}. Thus, hurricane losses are more related to the radii of stronger winds. The power-law dependencies, *a* and *b*, are generally similar between the cases using R_{34} or R_{out}, with somewhat larger coefficients when R_{out} is used. I also found that the exponent *a* is always much larger than *b*, implying that wind speed has a greater impact on loss than size.

From the Best Track data, I calculated storm duration (τ) from landfall to decay when a tropical cyclone (TC) is no longer being tracked in the database. Surprisingly, adding storm duration had very little impact on the regression results in terms of the explained variances. The *p*-values were much greater than 0.05, suggesting that the results were not statistically significant. Therefore, this factor can be omitted in the loss models. However, this conclusion needs further verification with varying definitions of storm duration in the future.

Regional GDP data were acquired from the Geographically-based Economic (G-Econ) Database created by the team led by Professor William Nordhaus at Yale University. The averages of individual GDP values over the grid cells covered by landfalling hurricanes are calculated to represent the regional economic wealth affected by the storms. The regression coefficient *d* for GDP is positive, consistent with the expectation that hurricane loss increases with the wealth of the area being hit. Although the values of *d* appear to be small, about 0.01-0.02, it is non-negligible because hurricane damage follows the increase of GDP exponentially and the regional differences in GDPs can be quite large, such as between different countries. Although regional GDP is not a dominant factor in determining the loss of a tropical cyclone, explicitly including the dependency of loss on regional GDP separates the human impact from storm physical factors and thus enables us to compare the economic damage of tropical cyclones that occur in different ocean basins.

Besides the size data from the EBT database, Professor Kerry Emanuel and his former student Dr. Daniel Chavas of the Massachusetts Institute of Technology provided me another set of R_{out} data based on the NASA Quick Scatterometer (QuikSCAT) satellite measurements. When using QuikSCAT R_{out} in the regression models, the explained variance is approximately the same as that from using the Best Track R_{out}, but the fitting coefficients *a* and *b* are somewhat different. This suggests that accurate measurements of storm physical parameters are important to TC damage estimates.

Table 3. Regression coefficients and explained variances from the bi-variate loss models with Best Track R_{out} and QuikSCAT R_{out} (33 storms) following the function form $L = 10^k V_{max}{}^a R_{out}{}^b$.

	a	*b*	*k*	R^2
Best Track R_{out}	5.71	2.16	-7.21	0.55
QuikSCAT R_{out}	5.60	1.80	-6.70	0.55

Figure 1 summarizes the explained variances for different loss models with various combinations of maximum wind speed, size, duration and regional GDP. The first 9 models are created by the regressions on the 73 TCs. The last 9 models are created by the regressions on the 43 hurricane cases. Within each of the two groups, the first model uses V_{max} as a predictor only (the traditional model); the second model uses V_{max} and R_{34}; the third model uses V_{max}, R_{34} and τ; the fourth model uses V_{max}, R_{34} and GDP; the fifth model uses V_{max}, R_{34}, τ and GDP. Then Best Track R_{out} is used instead of R_{34} in the following 4 models. This figure clearly shows the role of each factor:

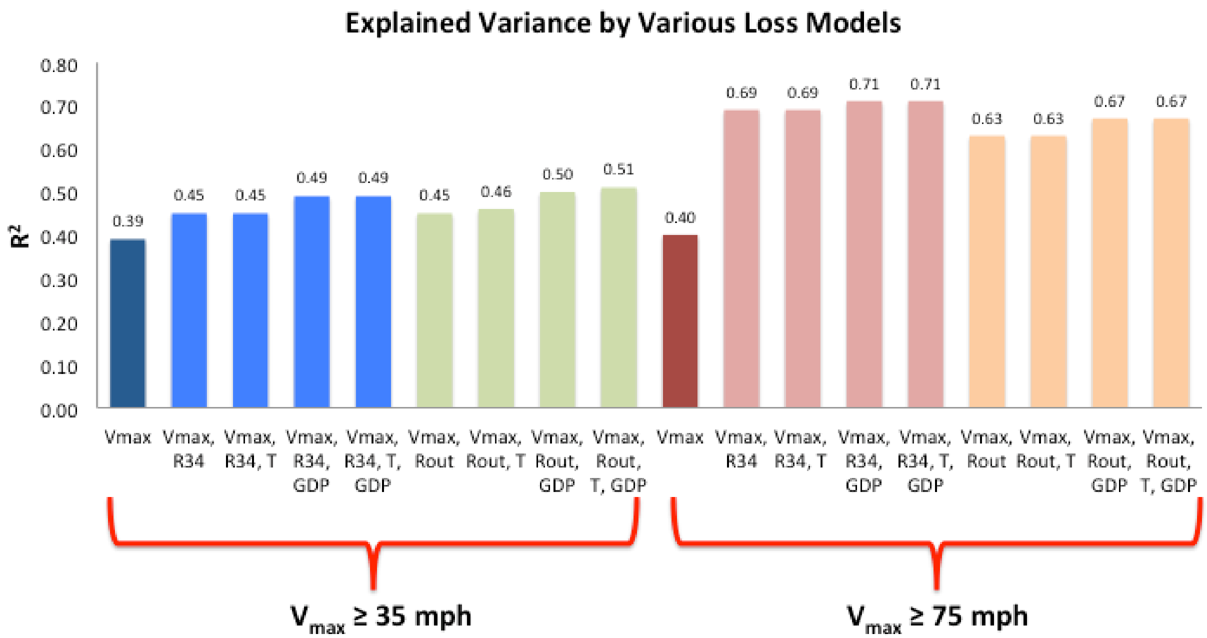
(i) for all storms, a noticeable improvement (R^2 increased by 6%) by adding storm size in the loss model and a similar degree of improvement (R^2 increased by 4%) by further considering regional GDP;

(ii) for hurricanes, a drastic improvement by adding storm size in the loss model with R^2 increased by about 30% (about 70% relatively) and a small improvement in explained variance through further adding regional GDP (by 2%-5%, up to 10% relatively);

(iii) the impact of storm duration is minimal;

(iv) using R_{out} or R_{34} produces very similar results.

Figure 9. Explained variance, R^2, by the loss models with different combinations of maximum wind speed, storm size, storm duration and regional GDP as predictors. See text for details.



The "best fit" model that explains the most variance and all regression coefficients are statistically significant at 95% is the one that uses V_{max}, R_{34} and GDP as predictors for hurricane losses. The corresponding equation is $L=10^{-5.58+0.01\,(GDP)}V_{max}^{4.70}R_{34}^{2.54}$. Using this model, I compared the relative contributions of maximum wind speed, storm size, and regional GDP to the economic damages of Hurricane Sandy (2012) and Typhoon Haiyan (2013). Haiyan's maximum wind speed is about 2 times of Sandy's while Sandy's size is 3 times of Haiyan's. Compounded by the regional GDP difference between US and Philippines, Sandy produced a greater monetary loss than Haiyan did, highlighting the importance of storm size and regional wealth in determining economic damage of a tropical cyclone.

People can use this model to predict the economic impact of a particular storm, thus provide timely financial assistance and relief to the storm-affected areas. Scientists can also use this model to predict the impact of climate change on hurricane loss and to help policy-making, such as whether to implement new building codes or limit coastal population and development to reduce the economic damage related to hurricanes. The development of improved loss models may provide better quantitative justifications for policies related to mitigation and adaption of global warming.