

A KNOWLEDGE DISCOVERY STRATEGY FOR RELATING SEA SURFACE TEMPERATURES TO FREQUENCIES OF TROPICAL STORMS AND GENERATING PREDICTIONS OF HURRICANES UNDER 21ST-CENTURY GLOBAL WARMING SCENARIOS

CAITLIN RACE*, MICHAEL STEINBACH*, AUROOP GANGULY**, FRED SEMAZZI***,
AND VIPIN KUMAR*

ABSTRACT. The connections among greenhouse-gas emissions scenarios, global warming, and frequencies of hurricanes or tropical cyclones are among the least understood in climate science but among the most fiercely debated in the context of adaptation decisions or mitigation policies. Here we show that a knowledge discovery strategy, which leverages observations and climate model simulations, offers the promise of developing credible projections of tropical cyclones based on sea surface temperatures (SST) in a warming environment. While this study motivates the development of new methodologies in statistics and data mining, the ability to solve challenging climate science problems with innovative combinations of traditional and state-of-the-art methods is demonstrated. Here we develop new insights, albeit in a proof-of-concept sense, on the relationship between sea surface temperatures and hurricane frequencies, and generate the most likely projections with uncertainty bounds for storm counts in the 21st-century warming environment based in turn on the Intergovernmental Panel on Climate Change Special Report on Emissions Scenarios. Our preliminary insights point to the benefits that can be achieved for climate science and impacts analysis, as well as adaptation and mitigation policies, by a solution strategy that remains tailored to the climate domain and complements physics-based climate model simulations with a combination of existing and new computational and data science approaches.

1. INTRODUCTION

The possible link between global warming and hurricane activity, while critical in terms of hazards preparedness, societal relevance and public perception, remains among the most hotly debated issues in climate science [13]. Recent studies with observations [3] and climate model simulations [1] suggest that while the overall frequency of Atlantic hurricanes may reduce under global warming scenarios, the strongest of these hurricanes may grow even more intense. However, model simulations and their interpretations do not necessarily agree [7, 1, 8], and disagreements remain about data quality issues and trends extracted from observations as well as the influence of environmental factors beyond ocean temperatures [9]. In addition, regional patterns in oceanic warming have been shown to influence hurricane activity in different ways [6].

A conceptual model for hurricane activity [4] was recently developed by downscaling simulations from the suite of models used for the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. This approach and the corresponding results suggest that while the current generation of global climate models may not be able to directly produce predictive insights on hurricanes, the simulations may nonetheless have relevant information content which can be extracted through conceptual models.

In this paper, we demonstrate the ability of data mining methods and innovative computational strategies to provide projections of hurricane activity based on warming scenarios. The prediction and uncertainty assessment strategies rely on a previous approach [4] used for regional temperature

*University of Minnesota, racex008@umn.edu, steinbac@cs.umn.edu, kumar@cs.umn.edu

**Oak Ridge National Laboratory, gangulyar@ornl.gov

***North Carolina State University, fred_samazzi@ncsu.edu.

and heat waves, where model hindcasts is used for bias correction and uncertainty quantification, model forecasts in the recent decade are used for cross-validation, and forecasts in the 21st century are used for most likely projections with uncertainty bounds. We attempt to quantify the relation between regional sea surface temperatures (SST) in the Atlantic off the coast of West Africa with tropical cyclone activity, which was hypothesized in a previous work [2].

Although each step of the methodology has room for improvement, our preliminary investigations do offer a number of interesting and novel insights. First, regional sea surface temperature patterns shows influence on hurricane activity; in particular, the average correlation of storm counts with the sea surface temperatures off West Africa is higher than what may be expected from random chance. Second, the oceanic clusters discovered in the region of interest correlate reasonably well with storm counts visually and through quantifiable metrics. Finally, projections with uncertainty bounds show that an increase in the number of storms with a warming environment, even though the uncertainties remain significant.

These insights demonstrate the potential to inform adaptation decisions through preparedness efforts for hurricane-induced disasters as well as mitigation policies through projections of hurricane activity based on warming projections which in turn rely on greenhouse-gas emissions scenarios. However, further developments of the methodologies and thorough data analysis and mining of multi-sensor observations, reanalysis data sets and multi-model simulations are required to confirm or reject these preliminary insights.

2. DATA AND METHODS

The observed data for this research are monthly averages of reanalysis SST data, obtained from the National Center for Atmospheric Research (NCAR) at <http://dss.ucar.edu/datasets/ds090.2> [5]. This data is available on a Gaussian grid ($1.875^\circ \times 1.904^\circ$) in Kelvin for every month from 1948-2007. The tropical cyclone count data are a vector of cyclone counts occurring in the area of interest each August for the 25 years from 1982 to 2007. These counts are shown in Figure 1.

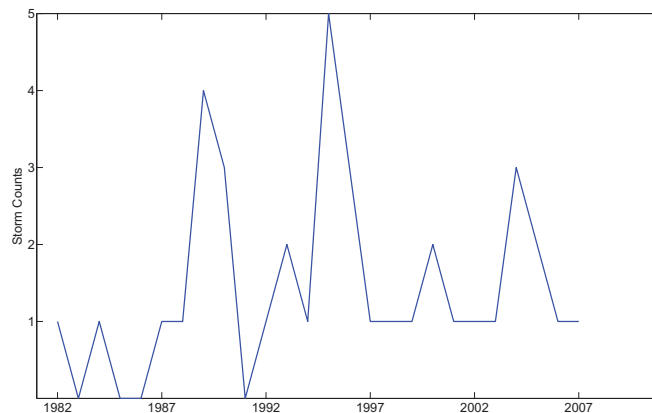


FIGURE 1. Tropical storm counts off the Western Coast of Africa for 1982-2007.

The simulation data used were outputs of the Community Climate System Model 3.0 (CCSM) from <http://www.earthsystemgrid.org/>, all having surface temperature components measured in Kelvin on a 1.406° grid, averaged by month. The specific outputs are the Climate of the 20th Model (20c3m) available from 1870-1999, and B1, A1FI, A2, and A1B, all available from 2000-2099. While 20c3m is hindcast data, B1, A1FI, A2, and A1B are all simulations for the future based on the Intergovernmental Panel for Climate Change (IPCC) Special Report on Emissions Scenarios (SRES)

[10]. In the cases (B1, A2, A1B, and 20c3m) where there are multiple initial condition ensembles, we used the average of all available ensembles to account for uncertainty [10]. The B1 scenario, sometimes called the best case for climate change, focuses on clean, efficient technologies with global problem-solving techniques. On the other hand, the A1FI scenario, or the worst case, assumes rapid economic growth and a convergence among global regions that use fossil-fuel intensive energy sources. The A2 and A1B cases are more moderate, however. A1B is similar to A1FI except it includes a balance of energy sources. A2 is a scenario where the economic focus is on local, rather than global, markets, with slower economic and technological growth than the other scenarios.

All analyses were done in MATLAB and R. The details of the various analysis steps are described in more detail in each section.

3. RELATIONSHIP OF SST TO TROPICAL STORM COUNTS

A visual indication of the strength of the relationship between SST for the month of August and tropical cyclone counts is provided by a Figure 2. To generate this figure we computed the correlation of the storm counts with August SST for each of the 12,134 locations in the reanalysis SST data set.¹ The resulting correlations were then mapped to a longitude–latitude grid using the locations that accompany each SST time series. Note that the correlation varies relatively smoothly as the location changes, as we would expect from physical considerations. Some locations, including those off the Western coast of Africa, have relatively high correlation—up to 0.7 for some points—to the storm counts. (Note that we have drawn a box around a region of SST in that area and will perform further analysis on it shortly.) Although other regions of the globe also show moderately high levels of correlation, this may be because of teleconnections or just due to spurious correlation.

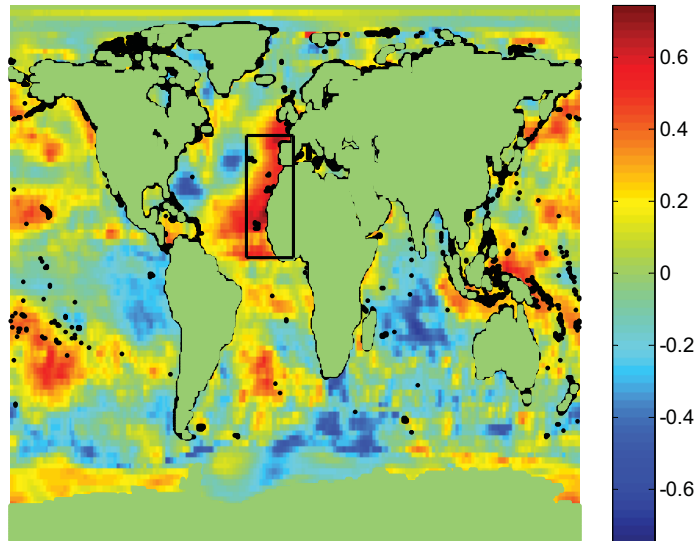


FIGURE 2. Correlation of August SST versus August storm count for 1982-2007. Best viewed in color.

Indeed, even though it is physically plausible that SST off the West African coast may be related to storm counts, it is important to check that it is not spurious. A full evaluation would require a more extensive study, but for this initial study, we compared the correlation of the SST in the boxed region to randomized storm counts. More specifically, to test the statistical significance of the correlations of the SST in the boxed region, a randomization test was performed by comparing

¹August was chosen since it gave the highest correlations to storm counts of any month. July and September also showed high correlations, whereas winter months did not.

the mean correlation of the actual SST data and tropical cyclone counts to a distribution of mean correlations of SST and randomized counts. Because the actual mean correlation of the region, which is 0.28, is greater than 99.41% of all the correlations from the randomized storm counts, we take this as reasonably strong evidence that there is a non-spurious connection. See figure 3.

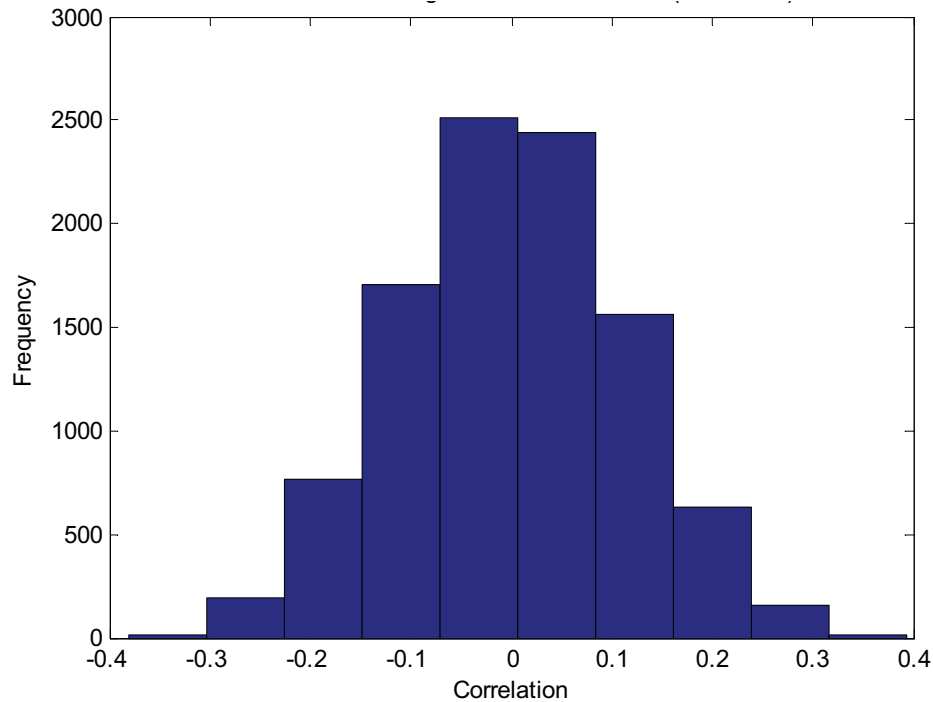


FIGURE 3. Histogram of mean correlations of August SST off the West African Coast to randomized August hurricane counts. $n=10,000$.

4. RELATIONSHIP OF KNOWN CLIMATE INDICES TO TROPICAL STORM COUNTS

A climate index is a time series that summarizes the behavior of the oceans and/or atmosphere in some region of the world. Climate researchers have used climate indices to investigate the connections between different parts of the climate system, e.g., the impact of El Nino on droughts in Australia. It is well known that many climate indices have strong connections to the SST in various regions of the globe. For example, climate indices such as the Southern Oscillation Index (SOI) have a strong connection to SST temperatures off the western coast of South America. Indeed, some indices, e.g., NINO12, are defined in terms of SST. Thus, it is worth investigating whether they may also be a relationship between known climate indices and the tropical cyclone counts.

We took a collection of well-known climate indices and computed the correlation between their August values and the storm counts for the 1982–2007 period. However, correlating the tropical cyclone counts to known climate indices produces insignificant results at $p=0.05$ (Table 4). It should be noted, however, that this lack of correlation does not necessarily mean a lack of connection since we were examining a subset of tropical storms and only for a particular month. Also, we didn't check for non-linear or time-lagged relationships, although we plan to do this in further work. Others (e.g., see <http://www.cpc.noaa.gov/products/outlooks/hurricane.shtml>) have used climate indices such as El Nino for predicting hurricanes, but again, the problem they consider is more general in both time and space than ours.

TABLE 1. R^2 values of correlating known climate indices with storm counts.

Index Name	R^2	p value
SOI	0.004	0.76
NAO	0.010	0.62
AO	0.0002	0.95
PDO	0.017	0.53
PNA	0.024	0.45
QBO	0.009	0.64
WP	0.027	0.42
NINO12	0.073	0.18
NINO3	0.042	0.31
NINO4	0.009	0.64
NINO34	0.015	0.55

5. CLUSTERING

Results discussed in Section 3 indicate a non-random correlation of SST off the Western coast of Africa and tropical storm counts. However, to quantify this relationship, it would be useful to derive a single time series summarizing SST behavior. However, summarizing the region specified above with a single time series, which is like creating a climate index from this region, does not yield the ideal results since the correlation of individual SST time series within this region to storm counts varies widely. Also, the choice of this region was somewhat arbitrary, leaving such an approach open to question. Thus, we decided to investigate whether we could find a climate index that was (1) defined in a non-arbitrary manner, but (2) in this general region, and (3) had a good correlation to storm counts. We provide a brief summary of this approach and then present the results.

In the past, Earth scientists have used observation and, more recently, eigenvalue analysis techniques, such as principal components analysis (PCA) and singular value decomposition (SVD), to discover climate indices [12]. These techniques are only useful for finding a few of the strongest signals and impose a condition that all discovered signals must be orthogonal to each other. We have developed an alternative methodology [11] for the discovery of climate indices that overcomes these limitations and is based on clusters that represent geographic regions with relatively homogeneous behavior. These clusters are found using the shared nearest neighbor (SNN) clustering approach we had previously developed [11]. The centroids of these clusters are time series that summarize the behavior of these geographical areas.

Figure 4 shows the clusters produced by shared nearest neighbor (SNN) clustering of SST for the period 1982-2007 [11]. Many pairs of clusters in this clustering are highly correlated with the known climate indices. In particular, some of these clusters are very highly correlated (correlation > 0.9) with well-known climate indices (some of the El Nino indices) and are located in approximately the same location as where these indices are defined. See [11] for more details. The SST clusters that are less well correlated with known indices may represent new Earth science phenomena or weaker versions or variations of known phenomena. Indeed, some of these cluster centroids provide better ‘coverage,’ i.e., higher correlation to land temperature, for some areas of the land [11]. Only one cluster produced by this clustering was in the same area as the region discussed above and had significant correlation with the storm counts. See the red-circled cluster in Figure 4.

Testing the predictive value of this cluster centroid against the vector of storm count values yields an R^2 value of 0.3106 ($p < 0.01$). To visually display the correlation, we plotted the normalized cluster centroid and the storm counts for the years 1982-2007. See Figure 5. Qualitatively, the two time series seem to track each other fairly well.

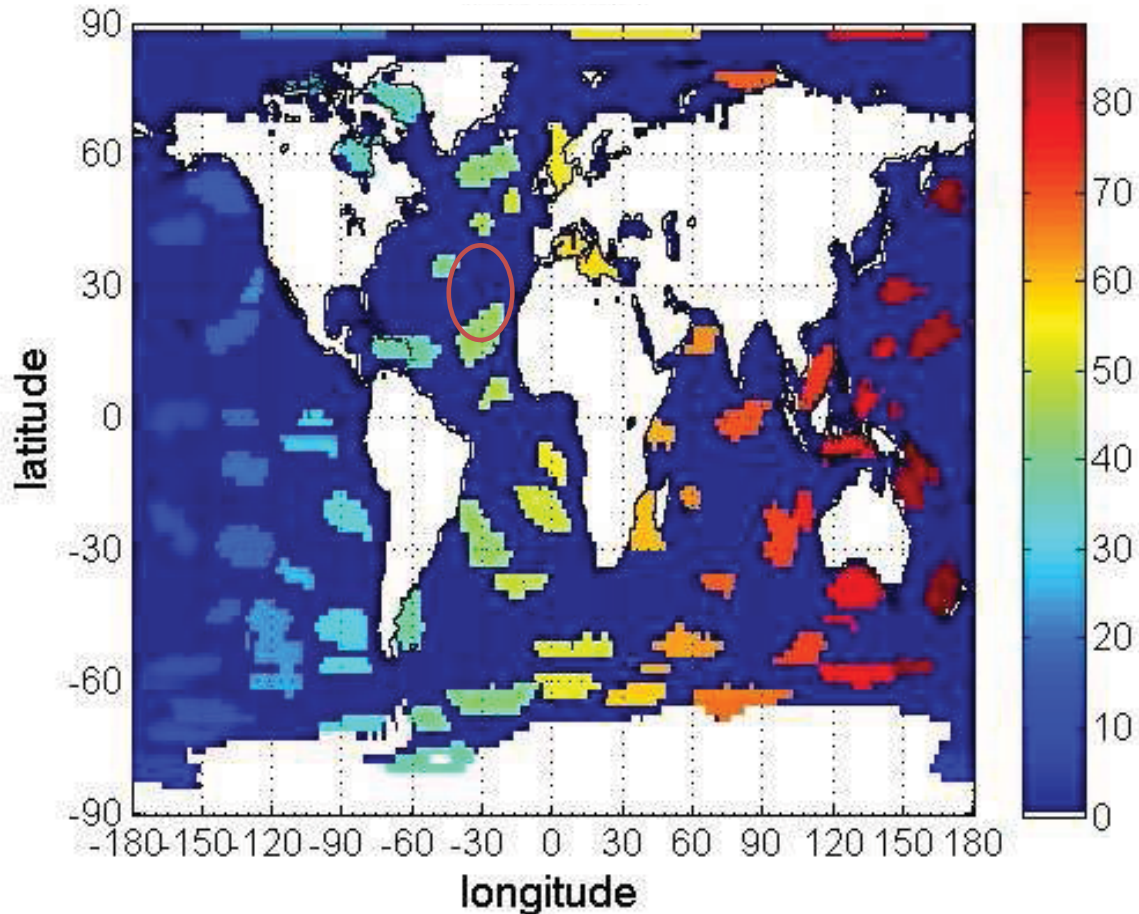


FIGURE 4. SNN clusters of SST using reanalysis data from 1982-2007. The different colors are used to distinguish different clusters.

6. USING A MODEL RELATING SST AND STORM COUNTS TO PREDICT FUTURE STORM COUNTS

To get a forecasting equation, we regressed the mean of the selected SNN cluster centroid from reanalysis SST data to a vector of storm counts from 1982-2007. In order to do predictions, cluster centroids from hindcast and simulated data were normalized to have the same mean value as the original SST data. Those data sets were then used to build the predictive model. Validation on this model was then done by comparing the actual storm counts vs. the predicted storm counts for hindcast and simulated data during the years where all sets of data were available. The forecasted storm counts for the hindcast data are within the range of the actual storm counts, so we proceeded to use our model for the 21st century. Figure 6 shows these predictions.

The A1FI scenario is a "worst case" scenario. As expected, the number of storms increases significantly over the course of the century, as storms are positively correlated with SST in the model we have created, and in the worst case, global sea surface temperature increases significantly with time. The B1 scenario is regarded as a "best case" scenario, and the number of predicted storms stays relatively constant. For the less extreme cases of A2 and A1B, the results show an increasing storm count over the century, but at a slower rate than that of the A1FI scenario. Note that A1B seems to level off at the end of the century, while A2 appears to increase exponentially with predicted counts approaching that of the A1FI scenario.

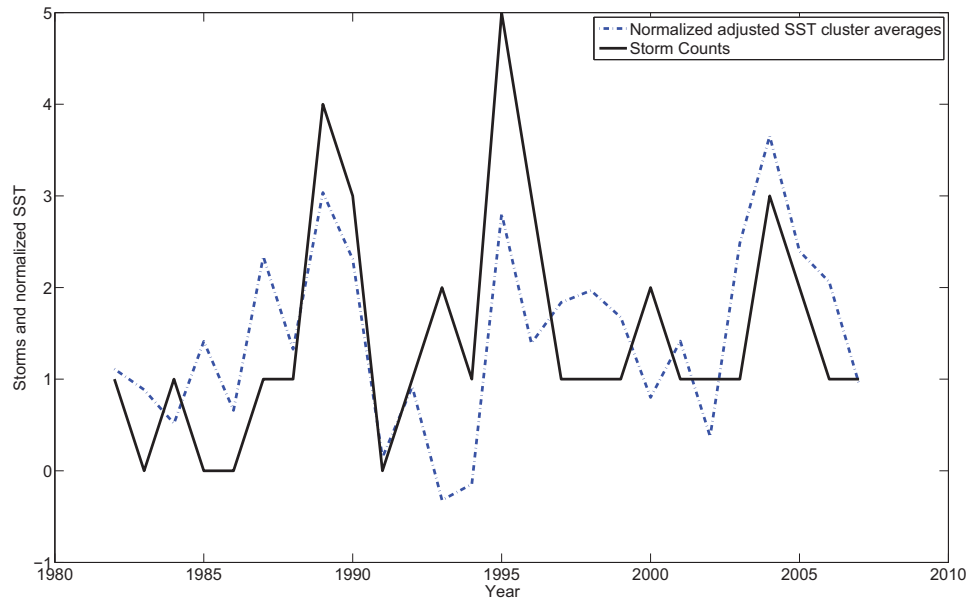


FIGURE 5. Storm counts and mean-adjusted, standard-normalized cluster centroid averages: August, 1982-2007

We would emphasize the limited scope of this exploratory analysis and the need for further investigation to take into account additional types of data, e.g., Saharan dust and wind data, additional modeling approaches, e.g., nonlinear regression, and the previous work that was mentioned earlier in this paper.

7. CONCLUSION AND FUTURE WORK

The primary contribution of this paper is a proof-of-concept demonstration that shows that attempts to address one of the most pressing gaps in climate change science and among the most hotly debated issues in the context of shaping public perception and informing policy: How does global warming impact hurricane frequency and can we generate credible projections of storm counts under warming scenarios? The debates in the scientific community clearly indicate that the climate modeling community has not been able to arrive at a clear consensus, while the discussions in the public sphere point to the perceived and real importance of this issue for perceptions, preparedness and emissions policy. The fact that we were able to develop predictions with uncertainty bounds by combining physics-based climate model simulations with data-guided insights from observations and simulations represents an important step forward, which could not have been achieved based on either physics-based or data-guided models on their own.

The development of hurricane projections relies on three hypotheses: first, there is information content in sea surface temperatures relevant for hurricane counts which can be extracted from observed data; second, climate model simulations of sea surface temperatures retain information about hurricane counts, and third, the data-guided strategy developed for extracting information content from observations can be generalized to model simulations. The first hypothesis is tested by investigating observed storm counts with reanalysis datasets, which in this case are assumed to represent surrogate observations, and used to develop a data-guided model which relates sea surface temperature clusters to storm counts. The second and third hypotheses are validated by examining the

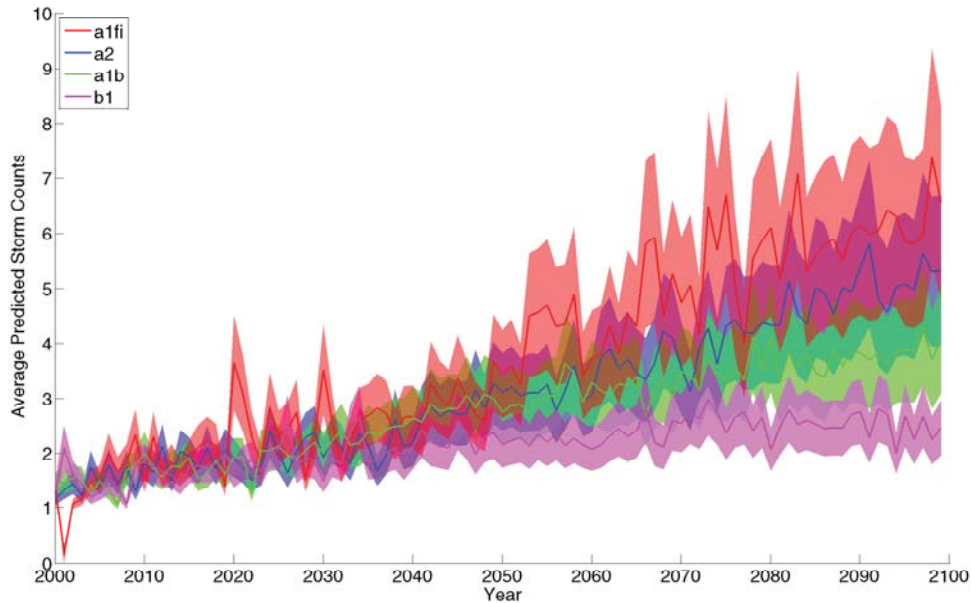


FIGURE 6. Predicted storm counts and accompanying 66% confidence intervals for some IPCC scenarios. Best viewed in color.

degree to which the results based on reanalysis datasets agree with climate model simulations generated in hindcast mode till 2000 and in forecast mode in the current decade. These hypotheses and the results derived based on these provide the necessary information for producing the projections of hurricane frequencies in the 21st-century, which are conditioned both on the underlying physics encapsulated within the climate models as well as the predictive insights extracted from observed and model-simulated data. Our approach suggests that relating the methodological development to the science challenges and the underlying hypotheses or critical science gaps may be as important (if not more important) to the development of an integrated and strategic computational solution than producing relatively incremental innovations in any specific methodologies. However, this conclusion is expected to be case-specific and may need to be tested further in climate and other multidisciplinary settings.

Future research is urgently motivated in two directions: climate science and computational science. From a climate science perspective, there is a need to further validate the hypotheses and enhance the predictions along with uncertainty quantifications by using multiple climate model simulations as well as observations or their surrogates. Thus, the entire suite of IPCC AR4 global climate models, as well as multi-sensor observations or reanalysis datasets, need to be used for longer historical periods to further validate the hypotheses and develop credible projections. From a knowledge discovery perspective, we believe we have barely explored the tip of the iceberg. The possibilities presented on this paper can be further expanded by developing or utilizing new and state-of-the-art tools, for example in the context of network analysis, clustering and regression, as well as by refining and fine-tuning the integrated knowledge discovery strategy. Additional variables, such as wind, also need to be considered. These issues will be explored in depth in our future work.

8. ACKNOWLEDGEMENTS

This work was supported by NSF grants III-0713227 and IIS-0905581, and by NOAA grant NA06OAR4810187, which funds the Interdisciplinary Scientific Environmental Technology (ISET)

Cooperative Research and Education Center, with which Semazzi, Kumar and Steinbach are affiliated. Computing resources were provided by the Minnesota Supercomputer Institute.

Co-author Ganguly was funded by the Laboratory Directed Research & Development (LDRD) Program of the Oak Ridge National Laboratory (ORNL), which in turn is managed by UT-Battelle, LLC, for the US Department of Energy under Contract DE-AC05-00OR22725. The United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

REFERENCES

- [1] M. A. Bender, T. R. Knutson, R. E. Tuleya, J. J. Sirutis, G. A. Vecchi, S. T. Garner, and I. M. Held. Modeled impact of anthropogenic warming on the frequency of intense atlantic hurricanes. *Science*, 327(5964):454–458, 2010.
- [2] M. Diaz and F. Semazzi. The role of west african coastal upwelling in the genesis of tropical cyclones: A new mechanism. Newsletter of the Climate Variability and Predictability Programme, December 2008.
- [3] J. Elsner, J. Kossin, and T. Jagger. The increasing intensity of the strongest tropical cyclones. *Nature*, 455(7209):92–95, 2008.
- [4] A. Ganguly, K. Steinhäuser, D. Erickson, M. Branstetter, E. Parish, N. Singh, J. Drake, and L. Buja. Higher trends but larger uncertainty and geographic variability in 21st century temperature and heat waves. *Proceedings of the National Academy of Sciences*, 106(37):15555, 2009.
- [5] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, et al. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–472, 1996.
- [6] H.-M. Kim, P. J. Webster, and J. A. Curry. Impact of shifting patterns of pacific ocean warming on north atlantic tropical cyclones. *Science*, 325(5936):77–80, 2009.
- [7] T. R. Knutson, J. L. McBride, J. Chan, K. Emanuel, G. Holland, C. Landsea, I. Held, J. P. Kossin, A. K. Srivastava, and M. Sugi. Tropical cyclones and climate change. *Nature Geoscience*, 3:157–163, 2010.
- [8] T. R. Knutson, J. J. Sirutis, S. T. Garner, G. A. Vecchi, and I. M. Held. Simulated reduction in Atlantic hurricane frequency under twenty-first-century warming conditions. *Nature Geoscience*, 1:359–364, 2008.
- [9] J. M. Shepherd and T. Knutson. The current debate on the linkage between global warming and hurricanes. *Geography Compass*, 1(1):1–24, 2007.
- [10] S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, M. Tignor, and e. Miller, HL. *Ippc, 2007: Climate change 2007: The physical science basis. contribution of working group i to the fourth assessment report of the intergovernmental panel on climate change.* Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
- [11] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of climate indices using clustering. In *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD=2003*, pages 446–55, August 2003.
- [12] H. Von Storch and F. Zwiers. *Statistical analysis in climate research.* Cambridge Univ Pr, 2002.
- [13] P. J. Webster, G. J. Holland, J. A. Curry, and H.-R. Chang. Changes in Tropical Cyclone Number, Duration, and Intensity in a Warming Environment. *Science*, 309(5742):1844–1846, 2005.