AMERICAN METEOROLOGICAL SOCIETY

*Monthly Weather Review*

# EARLY ONLINE RELEASE



The DOI for this manuscript is doi: 10.1175/MWR-D-16-0030.1

# Hindcasts of Integrated Kinetic Energy in Atlantic Tropical Cyclones: A Neural Network Prediction Scheme

Michael E. Kozar[1]
Risk Management Solutions
Tallahassee, FL 32304

Vasubandhu Misra
Center for Ocean-Atmospheric Prediction Studies,
Department of Earth, Ocean and Atmospheric Science,
Florida State University,
Tallahassee, FL 32306

Mark D. Powell
Risk Management Solutions
Tallahassee, FL 32304



[1] *Corresponding author address:* Michael Kozar; RMS Tallahassee; 612 Copeland St.; Tallahassee, FL 32304
Email: Michael.Kozar@rms.com

## Abstract

A new statistical-dynamical scheme is presented for predicting Integrated Kinetic Energy (IKE) in North Atlantic tropical cyclones from a series of environmental input parameters. Predicting IKE is desirable because the metric quantifies the energy across a storm's entire wind field, allowing it to respond to changes in storm structure and size. As such, IKE is especially useful for quantifying risks in large low-intensity high-impact storms such as Sandy in 2012. The prediction scheme, named the Statistical Prediction of Integrated Kinetic Energy Version 2 (SPIKE2), builds upon a previous statistical IKE scheme, by using a series of artificial neural networks instead of more basic linear regression models. By using a more complex statistical scheme, SPIKE2 is able to distinguish non-linear signals in the environment that could cause fluctuations in IKE. In an effort to evaluate SPIKE2's performance in a future operational setting, the model is calibrated using archived input parameters from Global Ensemble Forecast System (GEFS) control analyses, and is run in a hindcast mode from 1990 to 2011 using archived GEFS reforecasts. The hindcast results indicate that SPIKE2 performs significantly better than both persistence and climatological benchmarks.

# 1. Introduction

Integrated kinetic energy (IKE) is a recently developed metric that is designed to approximate the damage potential of landfalling tropical cyclones (Powell and Reinhold 2007). As its name suggests, IKE is defined as a summation of the kinetic energy within the near-surface wind field of a tropical cyclone (TC). By integrating energy across a large portion of a storm's wind field, IKE considers the overall structure of a TC. This is in stark contrast to many other existing hurricane metrics, which often quantify only a wind or pressure extreme at a single point within a TC. Intensity metrics such as maximum sustained wind speeds (VMAX) are undoubtedly useful for assessing the maximum potential damage caused by the winds in a TC (e.g. Emanuel 2005; Bell et al. 2000), but they do not paint a complete picture of storm damage potential.

In the decade following the landfall of Hurricane Wilma, no major hurricanes (VMAX > 96kts) have made landfall in the United States. This drought is thought to be a rather rare event, (Hall and Hereid 2015), depending on the metric that is used to classify major hurricanes (Hart et al. 2015). Despite this perceived quiet period of significant United States hurricane activity, there has been no shortage of damaging storms that have made landfall in the past decade. According to initial estimates from the National Hurricane Center[1], hurricanes Ike (AL092008), Irene (AL092011), and Sandy (AL182012) each caused more than $15 million in losses across the United States during the major hurricane drought despite each storm's somewhat weak landfall intensity. This disconnect between VMAX and damage often occurs because storm size and structure must also be considered to properly evaluate storm surge potential (e.g. Irish et al. 2008). Since Sandy, Irene, and Ike were such large storms, they were able to produce higher storm surge and damage totals than otherwise would be expected by storms of similar intensities. For this reason, it is likely that the IKE metric could add value to existing intensity

[1] http://www.nhc.noaa.gov/data/tcr/

metrics, by anticipating the higher damage potential of larger landfalling TCs (Powell and Reinhold 2007), especially considering that Ike, Sandy, and Irene all ranked very highly in terms of IKE relative to other storms in the historical record (Kozar and Misra 2014).

Despite the potential advantages of IKE, the concept of forecasting the energy metric in real-time is still in its infancy. Currently, operational forecasters have little to no guidance to predict IKE. Recently, Kozar and Misra (2014; hereafter KM14) explored whether or not it is feasible to fill that void with a simple statistical model in a proof-of-concept exercise. The resulting statistical model from that study was named the Statistical Prediction of Integrated Kinetic Energy (SPIKE), and it used linear regression to predict changes of IKE from a series of environmental predictors. Despite its simplicity, SPIKE was ultimately capable of outperforming a persistence forecast in a perfect-prognostic mode, indicating that statistical-dynamical forecasts of IKE might be possible in the future.

Building upon those results, the focus of this study is to further evaluate the operational potential of IKE forecasts using a more sophisticated statistical-dynamical scheme in a hindcast mode. Despite the successes of the proof-of-concept SPIKE model from KM14, linear regression is suboptimal for statistical weather prediction because the earth system is quite complex and contains several nonlinear signals. As such, the fixed linear regression coefficients in the SPIKE model will never be able to fully process the complex changing relationships between the environment and IKE variability within a TC. Therefore, a second-generation version of SPIKE is developed in this work by utilizing a more complex and nonlinear statistical framework in lieu of linear regression. More specifically, SPIKE version 2 (SPIKE2) utilizes a series of artificial neural networks (ANNs) to predict IKE tendency from a similar series of environmental input parameters. Ultimately, these networks are capable of learning and anticipating complex patterns in the environment, and as a result they are better suited to model a nonlinear system.

Furthermore, SPIKE2's evaluation will be moved from a perfect-prognostic space previously used in the initial KM14 work to a hindcast space. Obviously, in an operational setting, a statistical-dynamical forecast scheme must contend with imperfect input parameters that contain forecast errors of increasing magnitude with increasing lead-time. Therefore, by running SPIKE2 in a hindcast mode with model data from the National Oceanic and Atmospheric Administration (NOAA)'s Second Generation Global Ensemble Reforecast Archive (Hamill et al. 2013), we are able to more comprehensively measure the potential performance of the IKE prediction models.

The next section discusses the historical and reforecast data that are used to calibrate and evaluate the SPIKE2 neural network system. In the subsequent sections, the discussion shifts towards the methodology and procedures for creating, calibrating, and evaluating the neural networks used in SPIKE2. Finally, the calibration and hindcast performance of SPIKE2 is compared against various persistence and climatology benchmarks for Atlantic TCs between 1990 and 2011 in the penultimate section, preceding the concluding remarks.

## 2. Historical and Model Reforecast Data

Similar to KM14, a historical record of IKE in North Atlantic TCs is used to train and validate the SPIKE2 neural networks. This historical record covers the 1990 through 2011 hurricane seasons, and includes over 5000 six-hourly fixes from nearly 300 individual storms (Misra et al. 2013). Since gridded wind fields are not available for all of these cases, the IKE values contained in this record are all estimated from operational wind radii and intensity metrics in the extended best track dataset (Demuth et al. 2006) using a series of equations from Powell and Reinhold (2007) and Misra et al. (2013). The mean value of IKE across all of the storm fixes included in our historical archive is 35 TJs, with a standard deviation of 43 TJ. The distribution of observed IKE values takes a somewhat lognormal shape with a long tail towards higher values

126 (KM14). As such, although most storms never reach 50 TJs of IKE, Hurricane Sandy likely had
127 more than 400 TJ of IKE before it made landfall in New Jersey in 2012.
128 It should also be noted that past works have documented significant uncertainty within
129 the historical record of wind radii that fluctuates depending on the data platforms that are
130 available when analyzing each storm (e.g. Knaff et al. 2014; Landsea and Franklin, 2013).
131 Therefore our historical IKE record, which again is based on the operation wind radii, likely
132 inherits many of the same uncertainties found in the extended best track dataset.
133 Unlike KM14, SPIKE2's environmental input variables are drawn from a historical model
134 reforecast database during the same 1990 to 2011 interval. The second generation Global
135 Ensemble Reforecast Archive (Hamill et al. 2013) is selected as the source for this model data
136 because it includes model runs dating back multiple decades using a static February 2012
137 operational configuration (version 9.0.1) of National Centers for Environmental Prediction's
138 (NCEP) Global Ensemble Forecast System (GEFS).
139 These archived GEFS reforecasts include forecasts out to 16 days beyond the
140 initialization time. The first eight days of the forecast are run at T254 horizontal resolution
141 (~50km) with 42 vertical levels. The latter half of the forecast is run at a lower T190 resolution
142 (~70km), with the same 42 vertical layers. Each of the reforecast runs is initialized once daily at
143 00Z, as opposed to the six-hourly approach for generating operational GEFS forecasts. The
144 initial conditions for the reforecast dataset are produced from the Climate Forecast System
145 Reanalysis (CFSR; Saha et al. 2010) prior to February 2011 and operational Grid-Point
146 Statistical Interpolation analyses after that time.
147 The reforecast archive includes ninety-eight output fields from initial time out to F+384
148 hrs for each of the daily GEFS reforecasts[2]. The archived data is stored at 3-hourly intervals for
149 the first 72 hours of the forecast and then at 6-hourly intervals after that time. Each
150 meteorological field is bilinearly interpolated down to a somewhat coarse one-degree resolution

[2] http://www.esrl.noaa.gov/psd/forecasts/reforecast2/

global grid. In addition to the one-degree data sets, a smaller selection of twenty-eight fields is also stored in the GEFS's higher resolution native Gaussian grid (~0.5°). However, the higher resolution fields are all single-level variables, primarily near the surface. As a result of this limitation, mid- and upper-atmospheric dynamic and thermodynamic fields (winds, temperatures, humidity, etc) are only available in the one-degree grids. Therefore, to maximize consistency, we use only the one-degree data to examine the dynamical and thermodynamical processes that relate to IKE variability for our SPIKE2 neural network system.

As its name suggests, the GEFS archive does not just include a single deterministic forecast. In fact, the reforecast dataset is comprised of eleven ensemble members (1 control run, and 10 perturbation runs) compared to 21 ensemble members in the operational GEFS. For the purposes of this work, only the control run in the GEFS reforecast set is considered, but future works can and should clearly expand upon these results to produce probabilistic forecasts that resolve the uncertainty in the model's initial environment.

This GEFS reforecast dataset includes some noteworthy biases with regards to resolving TCs that will be addressed here. Obviously, the GEFS reforecasts will include position and intensity errors, and the reforecasted environment is expected to be imperfect as well, with all errors increasing as lead time increases. For reference, Galarneau Jr. and Hamill (2015) analyzed track errors in the GEFS reforecast archive for TCs in the Gulf of Mexico between 1985 and 2010 and found average positions errors to be 100km with a lead time of 24 hours, 250km with a lead time of 72 hours, and 400km with a 120 hour forecast interval. Typically, these track reforecasts in the Gulf of Mexico were found to have a left and slow bias relative to the storms motion.

Furthermore, Galarneau Jr. and Hamill (2015) also indicated that the GEFS reforecasts had a significant and consistent low intensity bias. This does not come as a surprise and leads us to the potentially most concerning issue for using the GEFS reforecast database in this study. Simply put, the one-degree horizontal resolution data taken from the model will not be

sufficient to properly resolve the wind field of a TC. As a result, intensities will be underestimated, and wind fields may be too broad. In fact, the GEFS reforecasts may fail to generate a TC vortex altogether in some extreme scenarios.

However, since we are not trying to predict IKE directly from the model's wind field, but instead by relating environmental parameters to IKE variability, this low resolution data might still be sufficient, albeit less than ideal. By using the lower resolution GEFS reforecast data, we can estimate the lower bounds of skill for a real-time version of SPIKE2. Furthermore, the static model configuration provided by the GEFS reforecast dataset (Hamill et al. 2013) allows us to focus on the performance of the SPIKE2 predictive scheme independent of changes in the underlying dynamical model's configuration and performance.

## 3. Selection of Input Parameters For Statistical-Dynamical Prediction

Before the neural networks can be constructed, we must first establish which environmental and storm specific input parameters will be taken from the GEFS control reforecasts to produce predictions of IKE variability. The initial SPIKE model built in KM14 utilized a series of 14 predictors that contained a significant linear relationship with IKE variability, many of which were taken directly from the Statistical Hurricane Intensity Prediction Scheme (SHIPS)'s developmental dataset (DeMaria and Kaplan 1999). These input parameters included various environmental predictors (both dynamical and thermodynamical), storm-specific parameters (e.g. position, minimum pressure), and persistence values of IKE based on known relationships between IKE and the environment (e.g. Maclay et al. 2008; Musgrave et al. 2012). However, since these parameters were selected based on their linear relationships with IKE, it is necessary to reselect predictors to highlight the nonlinearities in the storm-environment system that hopefully can be captured by the more sophisticated neural network scheme utilized here for SPIKE2.

As was done in KM14, the goal in selecting these parameters should be to target physical processes that govern variability with a TC's structure and ultimately the IKE index. Therefore, we started with a large pool of predictors, including both predictors used in the linear model that had clear and justifiable relationships with IKE as well as control variables such as day of month. Properly tuning a nonlinear complex neural network is a bit more difficult than tuning a linear regression model as there are more weights and neurons than there are coefficients in a linear regression model. Nonetheless, as we constructed the neural network we removed predictors if network performance over the testing sample increased by subtracting the predictor. As such, each of the control parameters and a few of the other environmental predictors with weaker ties to IKE were not selected for the final version.

Ultimately, we settled on 18 input parameters for SPIKE2, each of which is related to targeted relationships between the environment and IKE, in order to maximize the neural networks potential predictive power. The specific predictors are listed in Table 1. From this point forward, each predictor will be referred to by its abbreviation in the table. This predictor list is very similar to those used in the linear SPIKE model, but does include a total of four additional predictors. As such, we acknowledge that a few of these predictors could be removed, and the performance of SPIKE2 would likely not change by an appreciable margin. However, removal of any of the predictors did not seem to improve validation performance, suggesting that the predictors were not setting the model back via overfitting. Therefore, we felt that by including some of these extra predictors, the neural network may have a better chance to resolve some of the nonlinear signals between the environment and TCs if we were careful to limit the number of neurons in the ANN, thus minimizing the chances of overfitting.

Nonetheless, we ran a series of perturbation tests and case studies to ensure that each individual variable had some physical relationship that could explain how it is affecting projections of IKE from SPIKE2. For brevity, the remaining discussion in this section is meant to highlight the physical relationships that can explain how each of the individual predictors affect

IKE variability, followed by a short explanation about how the predictors are directly calculated from the model fields.

Predictors such as D200, VORT, SHTD and SHRD are designed to represent the certain dynamical features (upper level divergence, low-level vorticity, weak easterly shear) that are favorable for TC development. These predictors were some of the more significant predictors in the linear regression SPIKE model, and their impact over SPIKE2's IKE projections remains strong. Meanwhile, SST, T150 and VMPI are meant to be tied to thermodynamical properties that govern the maximum intensity of the storm, the height of the tropopause, and how far a storm has to go before it reaches said maximum intensity (e.g. Emanuel 1988; Bister and Emanuel 1998). RHLO and RHMD capture well known relationships between moisture and TC development. MSLP, PENV, and VMAX are storm specific parameters that give some information about the TC's intensity and breadth at the validation time, wherein a more intense storm or a larger storm with all else being equal will have higher wind speeds and more IKE. LAT, LON, SDAY, and PDAY obviously give information about the storm's position and time. These can be useful for identifying climatological tendencies across the basin. Finally, predictors such as PIKE and dIKE12 give information about persistence (i.e. how much IKE the storm had previously, and was it gaining or losing IKE previously) that can be useful for predicting future trends in certain instances.

However, as alluded to in the opening section, the signals between IKE and these predictors are quite complex. Unlike traditional storm development, which has a somewhat straightforward relationship with some of these predictors (i.e. the combination of low shear and high SSTs typical translates to a stronger storm all else being equal), IKE is also tied to storm size and the many different processes that govern it. For instance, many storms tend to expand as they move poleward and interact with other baroclinic features or through extratropical transition (e.g. Evans and Hart 2008). As such, recurving TCs often gain IKE in mid-latitude

253 environments that would traditionally be considered non-favorable for development (Maclay et
254 al. 2008).

255 Considering that extratopical transition occurs in just under half of all Atlantic TCs (Hart
256 and Evans 2001), our prediction scheme must be calibrated to anticipate the correct IKE
257 tendencies from these complex signals. As a result, the nonlinear equations within the ANNs will
258 also use predictors such as LAT, SHRD, T150, RHLO, and SST to determine whether or not a
259 storm is likely to expand in size (and also in IKE) from baroclinic forcings. Encouragingly, some
260 simple case studies revealed that a hypothetical storm in the mid latitudes (high LAT), late in its
261 lifecycle (high SDAY) will actually gain IKE as expected in a more baroclinic environment with
262 lower SSTs and higher SHRD. However, if the storm is under a similar environment in the deep
263 tropics or if shear and SSTs are too prohibitive in the mid latitudes, the neural networks will
264 correctly identify that the storm is more likely to decay. Ultimately, by considering both baroclinic
265 influences and traditional developmental mechanisms from this wide-ranging predictor base
266 through a nonlinear system of equations, the ANNs should be able to improve upon the results
267 of KM14.

268 The majority of the predictors discussed above (LAT, LON, MSLP, VORT, D200, etc) are
269 calculated directly from the corresponding TC signature within 3-D atmospheric fields from the
270 GEFS's control run. However, it should be noted that the GEFS dataset by itself is insufficient to
271 calculate all eighteen of the input parameters. For instance, some of the input parameters
272 require information about the ocean surface (VMPI, SST), time and date of year (SDAY, PDAY)
273 and past values of IKE (PIKE, dIKE12). Therefore, to obtain hindcasts for each of the input
274 parameters the GEFS reforecast dataset will be supplemented with a number of other datasets.
275 Daily one-degree NOAA Optimum Interpolation SST (“OI SST”; Reynolds et al. 2007) is used to
276 estimate observed ocean surface conditions. The historical IKE record (derived from the
277 extended best track dataset) is used to produce the persistence parameters, and finally the
278 NHC best track dataset is used to get the time information for each storm fix.

Once the input parameters are calculated from the GEFS control run for all forecast hours between initial time and T+72 hr, each parameter is normalized by its sample within the GEFS control run for all storms between 1990 and 2011. Normalizing the input parameters offers the benefit of filtering out some of the systematic biases in the GEFS, which in turn should enhance the performance of the operational IKE prediction schemes.

## 4. Setup of Artificial Neural Network for SPIKE2

With the predictors and data sources now established, this section details how the artificial neural networks are constructed, calibrated, and then run in a hindcast mode. As highlighted earlier, ANNs are chosen for SPIKE2 because of their ability to resolve and adapt to changing nonlinear signals in a certain system (e.g. Kriesel 2007). Thanks in part to their versatility, ANNs have been used in meteorology over the past several years to complete a wide array of tasks. A non-exhaustive list of tasks that ANNs have been used for includes evaluating uncertainty in hurricane wind analyses (DiNapoli et al. 2012), processing remotely sensed data (e.g. Atkinson et al. 2010), classifying circulation patterns (e.g. Cawley and Dorling 2005), predicting troposphere ozone levels (e.g. Abdul-Wahab and Al-Alawi 2002), forecasting wind speeds (Cao et al. 2012), forecasting precipitation and flooding (e.g. Hapuarachchi et al. 2011), and predicting the strength of the Indian monsoon on a seasonal scale (Shukla et al. 2011). A more detailed summary of earlier ANN applications in meteorology can be found in a review by Gardner and Dorling (1998).

### *4.1 Network Hierarchy and Algorithms*

The SPIKE2 prediction scheme will be built using a system of multiple two-layer feed-forward ANNs. Our two-layer feed-forward networks' hierarchy includes a hidden layer with twenty artificial neurons and an output layer with a single neuron that will ultimately produce the

desired results from the input parameters. Twenty neurons were chosen for the hidden layer to maximize predictive skill based on the results of an exhaustive search test, in which we found that this number of neurons corresponded to the best validation performance over the test subsample. By testing model performance with a wide varying number of neurons in this exhaustive search, we were able to find the approximate point at which ANN complexity is small enough to minimize the chance of overfitting, without compromising its ability to recognize and generalize the nonlinear signals in the TC-environment system.

In our case, the output of the neural networks will be IKE tendency for a given forecast hour, or in other words the difference between IKE at validation time and IKE at initialization time. Meanwhile, the eighteen normalized parameters discussed in Section 3 are selected as the input parameters of the neural network. As such, the goal of each ANN is to produce an estimate of IKE tendency from environmental and storm specific values within a model solution.

Ultimately, each of these ANNs within the SPIKE2 scheme are trained using a shared learning algorithm, wherein the networks are calibrated using a set of input parameters and known target (IKE tendency). The weights of the network's neurons are designed to adapt from a somewhat random initial value to a more optimal value, as the error function reaches a minimum. More specifically, the learning algorithm uses a Levenberg-Marquardt backpropagation algorithm (Marquardt 1963) to find this error minimum. This specific algorithm is designed to solve non-linear least squared problems and is typically thought to be an efficient and stable method for converging at an optimal solution in neural network learning (e.g. Hagan and Menhaj 1994).

## *4.2 Training, Validation, and Test Samples for Calibration*

To avoid overfitting and to promote generalization in the above supervised learning algorithm, the historical input and target output data series that are used to construct the ANNs

will be randomly split into three subsets. The first subset of data, named the training sample, is comprised of 70% of the input and target series. As its name suggests, the training sample is used to train the network by establishing the optimal weights within the neurons. The validation sample is a smaller subset, comprised of 15% of the historical input and target series. This subset is ultimately used to determine when the neural network can stop learning based on the network's ability to generalize effectively. As such, the learning algorithm searches for the point at which the neural network has the least amount of error over the validation subset during calibration. Finally, the third subset of input and target data is called the testing sample. This test sample is not used in the calibration of the model in any way. Instead it simply provides a more accurate measurement of out-of-sample network performance during calibration.

It should be noted that the three subset samples used in calibration are not entirely independent from one another because of storm-based serial correlation The general population of calibration data for any given forecast hour contains multiple target IKE tendency values from long-lived storms, but will not ever contain multiple sets of predictors from the same model run. Furthermore, each GEFS run that predictors are taken from is separated by at least 24hrs from the next closest analysis, as the GEFS is only initialized once daily in the NCEP reforecast data set. As KM14 showed, past IKE change did not have significant ties to future IKE tendency beyond the first 24 hours. Therefore, storm-based serial correlation between subsequent target IKE tendency values for each forecast hour should be somewhat limited. Nonetheless, these three subsets are only used in the calibration of individual neural networks. Once the weights are established with analyses as detailed in Section 5, the evaluation exercises done in Section 6, will use out of sample hindcast data to drive the neural network in an effort to best simulate how the models may perform in real time.

### *4.3 Neural Network Random Variability*

Inevitably, the methodology used to construct the neural networks introduces random variability into each individual ANN. Specifically, random variability is first introduced when the general population of input and target parameters from 1990 to 2011 is randomly split into the three separate subset samples. Additional random variability is introduced to the neural networks because the weights within the neurons are initialized somewhat randomly before arriving at their optimal weights. Ultimately, the random variability makes it all but impossible for two ANNs to be exactly identical to one another, even if they are calibrated on the exact same input and target output datasets. Each neural network weighs connections in the nonlinear system somewhat differently, and as a result, some of the networks will seem more accurate in certain situations but less accurate in other situations. Therefore, it is insufficient to base SPIKE2 off a forecast from just a single neural network.

Instead, SPIKE2 will utilize a system of one hundred individual neural networks to make its prediction of IKE tendency for each of the forecast intervals (i.e. a separate system of neural networks for each forecast interval). As shown by the schematic of SPIKE2 (Figure 1), the system of neural networks will produce 100 separate independent predictions of IKE tendency from a single set of input parameters. A deterministic forecast of IKE tendency from SPIKE2 will be taken from the median of these 100 individual predictions. Using the median from a large sample of ANNs helps to minimize the random variability present in a single neural network's forecast, thus allowing SPIKE2 to focus on the true skill of the neural networks. The overall skill of this deterministic forecast will be discussed at length in Section 6.

Not covered in this paper for brevity, but in development, are a series of probabilistic products that take each of the 100 individual ANNs within SPIKE2 into account, rather than just the median estimation. These SPIKE2 probabilistic products are used to evaluate the uncertainty within the ANN statistical scheme from a single set of input parameters, as each

ANN within SPIKE2 has a slightly different set of weights. Probability of exceedance, uncertainty ranges, and error bars, are just a small sample of some of the probabilistic utilities that are possible with SPIKE2 before even considering the idea of forcing the model with a wide array of input parameters from multiple forecast models or ensembles.

## *4.4 Comparison of Neural Network and Linear Model Performance*

Once again, the primary goal for developing SPIKE2 is to create a statistical dynamical model that is capable of predicting IKE in a hindcast mode, which would mark a significant step toward moving to real time operational forecasts of IKE. This objective differs from the goals of the linear regression version of SPIKE in KM14. That previous linear regression model was designed in a perfect prognostic mode to prove that IKE prediction is possible when given accurate environmental predictors. As a result of the different objectives and the different running environments (perfect prognostic vs hindcast), SPIKE and SPIKE2 are trained on two completely different sets of predictors, even if the storms in the calibration and evaluation sample are identical.

Therefore, to explicitly show the advantages of the new neural network methodology over the previously used linear regression model, we also created a perfect prognostic version of our neural networks using the same data source (developmental SHIPS; DeMaria and Kaplan 1999) that was used for calibration and evaluation of the SPIKE model in KM14. Unsurprisingly the added flexibility provided by the nonlinear equations allowed the neural network to outperform the linear regression model across a 1995-2011 comparison period. For instance, SPIKE2 has a mean absolute error of 12.6 TJ over its training sample at its longest 72-hour forecast interval. In contrast, the linear regression model has a comparable mean absolute error of 14 TJ at a much shorter 24-hour forecast interval. As such, we will progress onward out of the perfect prognostic space, and begin to calibrate the neural networks with predictors from numerical

406 analyses in Section 5, in an effort to prepare the neural networks for evaluation with hindcast
407 predictors in Section 6.
408

## 409 5. Calibration of Neural Networks Using GEFS Analyses

410 Ultimately, to establish the weights of the ANNs, we calibrate the entire system with
411 targets of IKE tendency taken from our historical record and normalized input parameters taken
412 from the control zero-hour analyses (F00) at validation time within the GEFS reforecast archive.
413 These analyses represent the best estimation for observed environmental conditions within the
414 model data's one-degree resolution. As such, the SPIKE2 system will be calibrated to accept
415 reforecast input parameters from the same coarse resolution when it is ultimately evaluated in a
416 hindcast mode. It is important to note that the F00 analyses do not include forecast errors.
417 Therefore, only the persistence IKE predictor will change with advancing forecast hour as the
418 persistence IKE value becomes further removed from the validation time.
419 Since the GEFS reforecast runs are only initialized at 00Z, the maximum sample size for
420 the F00 calibration dataset is the 1377 storm fixes that occur at 00Z between 1990 and 2011.
421 However, SPIKE2 requires persistence parameters of varying forecast lead times. Therefore,
422 the sample size of the calibration dataset will decrease with increasing forecast hour because
423 short-lived storms will not have longer-term persistence values. For comparison purposes, there
424 are 1097 fixes at a forecast hour of 12 hours and 614 fixes for the 72-hour forecast interval.
425 The performance of SPIKE2's deterministic forecast for the analysis-based calibration
426 dataset is shown in Table 2. Similar to SHIPS (DeMaria and Kaplan 1994) and SPIKE (KM14),
427 the explained variance for the targeted tendency value increases with increasing forecast hour.
428 The correlation between IKE tendency predictions from SPIKE2 with GEFS F00 data and the
429 observed historical dataset was r=0.73 at a 12-hour forecast window compared to r=0.91 at 72
430 hours. This seemingly counterintuitive result can be explained by considering that the

magnitude of IKE tendencies increases with growing forecast hour, such that random fluctuations and observational biases are less impactful at longer forecast hours. Furthermore, forecast errors are not present in any of the GEFS F00 input parameters such that the input parameters are no less accurate at 72 hours than they are at 12-hours.

In addition to predicting IKE tendency, SPIKE2 can also predict the actual value of IKE at the validation time by adding its IKE tendency prediction to the persistence IKE value from the model's initialization time. KM14 found that the IKE metric was somewhat resistant to change because it considers the energy across a storm's entire wind field. As a result, it is unsurprising that SPIKE2 performs better at predicting IKE than it does at predicting IKE tendency because it can use the decent performance of a persistence IKE forecast to its advantage, especially in short forecast windows. At a 12-hour forecast window, SPIKE2 explains 91% of the observed variance (r=.95) when using the GEFS F00 input parameters. That performance does not degrade sharply, as the explained variance remains near 80% (r=0.90) at a longer 72-hour window.

While these high correlations are promising, they are somewhat meaningless if similar performance can be achieved by simply using a persistence forecast. Encouragingly, the SPIKE2 calibration model has a lower 72-hour forecast error than does a much shorter 24-hour persistence forecast. To provide another metric for comparison, we have evaluated the mean squared error (MSE) from SPIKE2 over its calibration dataset against a persistence forecast at each corresponding forecast hour in Figure 2. Overall, SPIKE2 has lower MSE than does persistence by a fair margin (45% at a 12 hour forecast window, climbing up to 82% by 72 hours). The improvements over persistence are statistically significant at a p=0.025 level for all forecast intervals based on a two-sample bootstrapping test.

Also shown in Figure 2 are the reproduced results from the original linear version of SPIKE detailed in KM14. These results are also calculated over the model's calibration interval, 1990-2011, but as noted earlier, the two models used predictors from entirely different datasets

making this comparison uneven. Nonetheless, the calibration statistics indicate that the linear SPIKE model simply cannot measure up to the neural networks in SPIKE2, although both models offer substantial improvement over persistence. Like the results of Section 4.4, this evidence continues to support our hypothesis that the neural networks will be superior to simple linear regression because it can account for the nonlinearities in the TC-environment system.

Although these initial calibration results appear to be encouraging, it should once again be noted that the hindcast version of SPIKE2 discussed in the following section will use imperfectly reforecasted input parameters from the GEFS control runs. As such, it would be unfair to expect SPIKE2's hindcasts in the following section to achieve these high performance benchmarks. Instead, the performance metrics shown in Table 2 can be viewed as the maximum potential skill that can be obtained by SPIKE2. The intent of these performance benchmarks is to determine how the model will degrade when forecast errors are introduced to the model input fields. Nonetheless, the exercise proved useful by identifying a set of weights within the artificial neurons that can be used to produce hindcasts of IKE from the GEFS reforecasts.

## 6. Performance of SPIKE2 Hindcasts Using GEFS Reforecasts

In this section, we will adapt the SPIKE2 ANN system to run in a hindcast mode with the GEFS reforecast control run from 1990 to 2011. As just discussed, the network will retain the same neuron weights that were calibrated in the previous exercise with GEFS control analyses. However, unlike the calibration exercises the neural networks will be given imperfect input parameters from the GEFS reforecast control run at various lead times out to 72 hours. This will enable us to determine how forecast errors affect SPIKE2's ability to predict IKE. We can understand from this analysis of predictive skill whether or not SPIKE2 might offer skillful operational support in a real time environment.

Much like the last section, we will evaluate the deterministic forecast from SPIKE2 using the target IKE tendency and IKE values as the historical baseline. Statistics such as correlations and mean absolute errors will be used to detect the magnitude of performance deterioration relative to the maximum potential performance levels obtained in the calibration exercise. As was done in the earlier calibration exercises and in KM14, SPIKE2's deterministic skill will be evaluated relative to simple persistence forecasts. However, in addition, a new more challenging benchmark will also be introduced by way of a simple statistical model that considers climatology and other non-forecast parameters

Such a benchmark model would follow in the footsteps of the Statistical Hurricane Intensity Forecast model (SHIFOR), which uses seven known parameters at initialization time to set the baseline performance for operational intensity forecasts (Jarvinen and Neumann 1979; Knaff et al. 2003). The exact parameters of SHIFOR include: Julian day, initial storm intensity, previous 12-hour intensity change, initial latitude, initial longitude, initial zonal component of storm motion, and initial meridional component of storm motion. These SHIFOR climatology and persistence predictors are somewhat relevant to IKE tendency as well. Therefore, an IKE statistical persistence model named the "Benchmark of Integrated Kinetic Energy (BIKE)" is created to predict IKE tendency in a simple linear regression model using the same seven input parameters, with two exceptions. First, the 12-hr intensity change parameter will be switched out for a 12-hr IKE change parameter. Second, the initial or persistence value of IKE will be added as an eighth predictor. This BIKE regression model is trained using all 00Z storm fixes from 1990-2011, such that its calibration fit will be compared to the GEFS-SPIKE2 hindcasts at lead times of 24hrs, 48hrs, and 72hrs for the same 1990-2011 interval.

Case studies act as a good first step to evaluate the SPIKE2 hindcasts relative to their assortment of benchmarks in an effort to see how IKE forecasts might perform during significant landfalling events. To that end, Figure 3 contains a plot of SPIKE2 hindcasts shown against historical values of IKE just prior to landfall for Hurricanes Floyd (AL081999), Katrina

(AL122005), Ike (AL112008), and Irene (AL122011). Each of these four storms gained considerable IKE as they approached land, and as a result, a persistence forecast would have greatly underestimated the storm's destructive potential at landfall. BIKE proves to be a more challenging benchmark for SPIKE2 in these four case studies, as it arguably outperforms SPIKE2 for Hurricane Floyd. Nonetheless, the SPIKE2 hindcasts outperform BIKE in most other cases. The SPIKE2 hindcasts for Irene were particularly impressive as the green curve representing the hindcast remains very close to the black line representing the observations throughout the 72 hour forecast period. One final item of note, in nearly each case, the SPIKE2 hindcast using reforecasted predictors performs worse than the SPIKE2 calibration model using predictors from analyses. This result is expected, as it suggests that the performance of the ANNs will degrade with the introduction of forecast errors in the series of input predictors.

Moving to a more general perspective, mean error and correlation statistics are shown on a line plot in Figure 4 for all of the storm fixes within the 1990-2011 evaluation sample. The SPIKE2 hindcasts are capable of explaining more than 80% of the variance in the historical IKE record with a day of lead time, and mean absolute errors are approximately 12 TJs in the same 24-hour forecast window. As lead time increases, hindcast performance expectedly decays, but the model is still capable of explaining 70% of the historical IKE variance at 48hrs and 62% at 72-hours, with errors of 16.6 and 20.7 TJs at those times respectively.

The performance of the hindcast easily exceeds the performance benchmark set forth by a persistence forecast. For instance, a 72-hour SPIKE2 hindcast has comparable error on average to a half-as-long 36-hour persistence forecast. Mean squared error statistics paint a similar picture (Figure 5), as the SPIKE2 model offers a 60% reduction of MSE at 24-hours relative to persistence. This reduction in MSE relative to persistence holds steady as forecast hour increases, fluctuating between 50% and 70% between 24hrs and 72hrs of lead time. The lack of a trend with advancing forecast hour in this MSE reduction metric (outside of the first 12-24hrs, where persistence forecasts excel) is likely attributed to a balance between rapidly

increasing persistence error (lowering the benchmark), and increasing forecast errors in the input data holding back the SPIKE2 scheme (decreased hindcast performance). The significance of these improvements is once again tested with a two-sample bootstrapping exercise. Results, indicate that the SPIKE2 hindcasts are significantly better than persistence at the p=0.05 level for all forecast windows greater than 12hrs, and at the p=0.01 level for all forecast windows greater than or equal to 48hrs.

Unsurprisingly, the BIKE model is indeed a tougher benchmark than just a simple persistence forecasts as noted by both the correlation and mean error metrics. For instance, BIKE has a 12% lower mean absolute error at 72-hours than does persistence. Nonetheless, the SPIKE2 hindcasts still clearly outperform BIKE. For instance, BIKE's mean absolute errors are more than 30% higher than the SPIKE2 hindcast errors at all three of the shown forecast windows.

On the other hand, the performance of the hindcasts falls short of the higher performance levels found during the calibration exercises. Again, this result was expected because statistical-dynamical prediction schemes are only as accurate as the input data going into the statistical model. In this case, the GEFS reforecasts include forecast errors that were not present in the analyses, which results in this degradation of performance. Furthermore, a lesser decrease in performance should also be expected just by running the ANNs on a dataset that they were not calibrated with (i.e. the GEFS F00 analyses).

Nonetheless, the drop in performance from the calibration tests to the hindcast tests is not a hindrance. Mean errors only increased by less than 15% and correlations only decreased by less than 7% inside the shorter 12 and 24-hour windows. Growing inaccuracies in the GEFS input variables, led to a more dramatic decrease in performance at larger longer forecast windows relative to the maximum potential performance level in the calibration exercises. However, once again, these hindcasts are still convincingly skillful relative to a persistence forecast. In fact, the hindcast performance metrics (green curve) are much closer to the

potential performance metrics in the calibration runs (blue curve) than they are to the persistence performance benchmarks (red curves).

## 7. Conclusions and Outlook for Future Operational Development

Despite the promise of the hindcast results presented above, there is still some work left to be done to adapt this model for operational use. For example, the neural networks would likely need be recalibrated to receive operationally predicted input parameters from a desired model in real-time unless the targeted model is similar to the GEFS reforecast data used here (such as the operational GEFS). If recalibration is needed, a sufficiently long historical database will once again be needed to normalize the predictors and set the neuron weights. This limitation is one of the primary reasons for using the control run in the available GEFS reforecast database. Despite its coarse one-degree resolution, the GEFS archive contained a long record of data from a static version of the same model. Unfortunately, few operational models have long archives of forecasts or hindcasts that are readily available. Therefore, adapting SPIKE2 to be used with a higher resolution operational model or model ensembles is dependent upon securing an archive for the desired model. As such, adapting SPIKE2 to the rest of the GEFS ensemble members at a similar resolution or to archived model data stored in the Observing System Research and Predictability Experiment (THORPEX)'s Interactive Grand Global Ensemble (TIGGE) archive would be easier to accomplish than would be adapting SPIKE2 to work with predictors from the Hurricane Weather Research and Forecasting Model (HWRF).

In addition to calibration, future work must focus on determining whether or not SPIKE2 forecasts can be made in a timely manner. SPIKE2’s products almost certainly cannot be issued instantaneously at initialization time. Although the neural networks themselves can be run fairly quickly, an operational version of SPIKE2 still requires dynamically forecasted input parameters,

and unfortunately, the output from most modern dynamical models is not available until a few hours after their initialization time. Therefore, statistical-dynamical models dependent upon forecast model data, such as Model Output Statistics (MOS) and in the future SPIKE2, cannot come out until the dynamical models' run time concludes. As a result, a 72-hour SPIKE2 forecast using GFS or GEFS data would be already a few hours into its forecast period by the time it was issued, thus shortening it to a 66-70 hour forecast depending on the actual issuance time.

A large delay between issuance and initialization would be detrimental to the usefulness of SPIKE2 because most operational forecasters are required to issue their forecasts at regular intervals. To alleviate this concern some operational statistical-dynamical models, are run in a so called "early cycle" mode, wherein each product uses environmental predictors from the previous dynamical model run, which is typically initialized six hours earlier (i.e. the 00Z forecast uses dynamical predictors from an 18Z model). Adapting this early cycle approach to SPIKE2 will ensure that its IKE forecasts are in advance of each forecast advisory. Consequently, SPIKE2's dynamical predictors in an early cycle mode would be several hours old before SPIKE2 is even issued. As such, the need to forecast the input parameters an additional few hours into the future would likely result in a slight degradation to model skill. Considering that SPIKE2 hindcasts outperform much shorter persistence forecasts, this may not have a substantial effect on performance. Nonetheless, it will be necessary to compare the pros and cons of running SPIKE2 in this early cycle mode against attempting to develop an interpolator to issue SPIKE2 forecasts in a "late cycle" mode as real time development of SPIKE2 continues.

Nonetheless, the results presented in the earlier sections serve as a proof of concept, suggesting that SPIKE2 could be a viable product in an operational setting once these hurdles are cleared. In calibration exercises, the deterministic scheme is capable of explaining the majority of variance in the historical IKE archive, and offers a significant improvement over persistence. Importantly, the addition of imperfectly predicted input parameters from a coarse

GEFS control run archive did not cause SPIKE2's performance to drop off severely. Instead, SPIKE2 hindcasts from 1990 to 2011 still exhibit significant skill over any known IKE persistence or climatology metrics, despite the inclusion of forecast errors from a rather coarse resolution GEFS dataset. Not to mention, the briefly discussed SPIKE2 probabilistic products add value to the deterministic IKE forecasts by offering a quantitative estimate of uncertainty in the statistical neural network scheme.

With the inclusion of input parameters from a higher resolution dataset that is capable of better resolving some of the storm specific predictors, it may be possible to improve SPIKE2's skill even further. Nonetheless, if even the level of performance by SPIKE2 with the GEFS reforecast data can be maintained when adapting SPIKE2 for operations, it would surpass the ability of any known guidance specifically targeted for deterministic IKE prediction.

## Acknowledgements

Thanks to Robert Hart, Phillip Sura, Allan Clarke, Ming Ye, and Mark Bourassa for their helpful comments and feedback. This work was graciously supported by grants from NOAA (NA12OAR4310078, NA10OAR4310215, NA10OAR4320143). Finally, we greatly appreciate the constructive comments and suggestions given to us by two anonymous reviewers during the submission process.

# References


Abdul-Wahab, S.A., and S.M Al-Alawi, 2002: Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. *Environmental Modelling & Software,* **12**, 219-228.

Atkinson, P.M., and A.R.L. Tatnall, 1997: Introduction Neural networks in remote sensing. *International J. of Remote Sensing*, **18**, 699-709.

Bell, G. D., and Coauthors, 2000: Climate assessment for 1999. *Bull. Amer. Meteor. Soc.,***81**, 1328–1378.

Bister, M. and K.A. Emanuel, 1998: Dissipative heating and hurricane intensity. *Meteor. Atm. Phys.,* **52**, 233-240.

Cao Q., B.T. Ewing, M.A. Thompson, 2012: Forecasting wind speed with recurrent neural networks. *European J. of Operational Res.*, **221**, 148-154.

Cawley G.C. and S.R. Dorling, 1996: Reproducing a subjective classification scheme for atmospheric circulation patterns over the United Kingdom using a neural network. *Proceedings International Conference on Neural Networks,* 281-286.

Demuth, J.L., M. DeMaria., and J. A. Knaff, 2006: Improvement of advanced microwave sounder unit tropical cyclone intensity and size estimation algorithms. *J. Appl. Meteorol. Climatol.,* **45**, 1573-1581.

661 DiNapoli S.M, M.A. Bourassa, and M.D. Powell, 2012: Uncertainty and Intercalibration Analysis
662 of H*Wind. *J. Atmos. and Ocean. Tech.,* **29**, 822-833
663
664 Emanuel K., 1988: The Maximum Intensity of Hurricanes. *J. Atmos. Sci.*, **45**, 1143-1155.
665
666 Evans, C., and R.E. Hart, 2008: Analysis of the Wind Field Evolution Associated with the
667 Extratropical Transition of Bonnie (1998). *Mon. Wea. Rev.*, **136**, 2047-2065.
668
669 ——, 2005: Increasing destructiveness of tropical cyclones over the past 30 years. *Nature*, **436**,
670 686–688.
671
672 Galarneau, T., and T. Hamill, 2015: Diagnosis of Track Forecast Errors for Tropical Cyclone
673 Rita (2005) Using GEFS Reforecasts. Wea. Forecasting, in press.
674
675 Gardner M.W., and S.R. Dorling, 1998: Artificial neural networks (the multilayer perceptron)-a
676 review of applications in the atmospheric sciences. *Atmos. Environment,* **32**, 2627-2636.
677
678 Hagan, M.T. and Menhaj, M.B., 1994: Training feedforward networks with the Marquardt
679 algorithm. *IEEE Transactions on Neural Networks*, **5**, 989-993.
680
681 Hall T., and K. Hereid, 2015: The frequency and duration of U.S. hurricane droughts. *Geo. Res.*
682 *Let,* **42**, 3482-3485.
683
684 Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M.l Fiorino, T. J. Galarneau, Y. Zhu,
685 and W. Lapenta, 2013: NOAA's Second-Generation Global Medium-Range Ensemble
686 Reforecast Dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565.

687 Hapuarachchi, H.A.P., Q.J. Wang, and T.C. Pagano, 2011: A review of advances in flash flood
688 forecasting. *Hydrol. Process.,* **25**, 2771-2784.
689
690 Hart R.E., and J.L. Evans, 2001: A Climatology of the Extratropical Transition of Atlantic
691 Tropical Cyclones. *J. Climate*, **14**, 546-564.
692
693 ——, D.R. Chavas, and M.P. Guishard, 2015: The arbitrary definition of the current Atlantic
694 major hurricane landfall drought. *Bull. Amer. Soc.*, in press.
695
696 Irish, J.L., D T. Resio, and J.J. Ratcliff, 2008: The influence of Storm Size on Hurricane Surge.
697 *J. Phys. Oceanogr.,* **38**, 2003-2013.
698
699 Jarvinen, B. R., and C. J. Neumann, 1979: Statistical forecasts of tropical cyclone intensity for
700 the North Atlantic basin. NOAA Tech. Memo. NWS NHC-10, 22 pp.
701
702 Knaff J.A., M. DeMaria, C.R. Sampson, and J.M. Gross, 2003: Statistical, 5-Day Tropical
703 Cyclone Intensity Forecasts Derived from Climatology and Persistence. *Wea. Forecasting*,
704 **18**, 80–92.
705
706 ——, S.P. Longmore, D.A. Molenar, 2014: An Objective Satellite-Based Tropical Cyclone Size
707 Climatology, *J. Cli.,* **27**, 455-476.
708
709 Kozar, M.E. and V. Misra, 2014: Statistical Prediction of Integrated Kinetic Energy in Atlantic
710 Tropical Cyclones, *Mon. Wea. Rev.,* e-view.
711

712 Kriesel D., 2007: A Brief Introduction to Neural Networks. [available online:
713 http://www.dkriesel.com/en/science/neural_networks]
714
715 Landsea, C.W., and J.L. Franklin, 2013: Atlantic Hurricane Database Uncertainty and
716 Presentation of a New Database Format. *Mon. Wea. Rev.*, **141**, 3576-3592.
717
718 Maclay K.S., M. DeMaria, and T.H. Vonder Haar, 2008: Tropical Cyclone Inner-Core Kinetic
719 Energy Evolution. *Monthly Weather Review* **136**:12, 4882-4898.\
720
721 Marquardt D., 1963: An algorithm for least squares estimation of non-linear parameters. *J. Soc.*
722 *Ind. Appl. Math,***11**, pp.431 -441
723
724 Musgrave K.D., R. K.Taft, J. L.Vigh, B. D.McNoldy, *and* W. H.Schubert, 2012: Time evolution of
725 the intensity and size of tropical cyclones, J. *Adv. Model. Earth Syst*., **4**, M08001.
726
727 Powell, M.D., and T.A. Reinhold, 2007: Tropical Cyclone Destructive Potential by Integrated
728 Kinetic Energy. *Bull. Amer. Meteor. Soc.*, **88**, 513–526.
729
730 Reynolds, R. W., T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, and M. G. Schlax, 2007: Daily
731 high-resolution blended analyses for sea surface temperature. J. Climate, 20, 5473-5496.
732
733 Saha S., and coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer.*
734 *Meteor. Soc.*, **91**, 1015-1057.
735
736 Shukla R.P., K.C. Tripathi, A.C. Pandley, and I.M.L. Das, 2011: Prediction of Indian summer
737 monsoon rainfall using Niño indices: A neural network approach. *Atmos. Res.*, **102**, 99-109.

## Figure Captions

**Figure 1:** Schematic of the SPIKE2 neural network system. A single set of input parameters is passed into each of the one-hundred independent artificial neural networks (ANN1, ANN2, … ANN100) that make up SPIKE2. Each network produces its own separate prediction of IKE tendency based on the same input parameters. The median of these predictions is used as SPIKE2's deterministic prediction, but each individual prediction can be used for probabilistic forecasting.

**Figure 2:** Evaluation of SPIKE2 skill in a calibration mode with GEFS analyses relative to a persistence forecast. Calibration skill is measured as a percent reduction of mean squared error (MSE) for the model's deterministic predictions from 1990 to 2011, with respect similar MSE calculations for a persistence forecast at various forecast hours. A reduction of MSE is plotted as a positive percentage, indicating that the model has outperformed persistence. SPIKE2 has significantly lower MSE than persistence at the p=0.05 level for all forecast hours. For reference, also shown is the reproduced results of the linear regression version of SPIKE as detailed in KM14.

**Figure 3:** 72-hour runs of SPIKE2 plotted against benchmarks and the observed IKE values (black lines) for notable hurricanes immediately prior to their landfalls. SPIKE2 hindcasts utilize GEFS reforecasted predictors from a run initialized at the time specified in each legend. SPIKE2 calibration runs utilize analyzed predictors from the GEFS archive valid at each forecast time. The benchmark model is initialized at the same time as the SPIKE2 hindcast run for direct comparison purposes. Not explicitly shown is a persistence forecast which would be

represented by a horizontal line stretching from the first observed IKE value on the left through the entire 72hr period.

**Figure 4:** Performance statistics for SPIKE2. Correlation (panel A) and mean absolute error values, in units of TJ, (panel B) are shown between the historical record and SPIKE2 in various modes or a persistence forecast. The correlation value that is shown in these plots is for IKE, not IKE tendency. Calibration statistics are identical to those in Table 2, and are used as a maximum potential reference point to determine the degradation of skill when forecast error is introduced to the model in the hindcast runs via the input parameters from the GEFS reforecast. Also shown for reference is the performance of a persistence forecast and the statistical climatological and persistence model, BIKE.

**Figure 5:** Evaluation of SPIKE2 skill relative to a persistence forecast. Performance is once again measured as a percent reduction of mean squared error (MSE) for the model's deterministic predictions from 1990 to 2011, with respect similar MSE calculations for a persistence forecast at various forecast hours. The calibration skill is reproduced from Figure 2 and is shown alongside the skill of the SPIKE2 hindcasts with reforecasted input parameters relative to persistence. The hindcast model is significantly better than persistence at the p=0.05 level for all forecast hours greater than or equal to 24-hours in length.

# Figures and Tables

**Table 1:** Variables used in the SPIKE2 neural networks. These input parameters are obtained from GEFS reforecasts and analyses, NOAA OI SSTs, and the historical record. Many of these predictors are inspired by the parameters contained in the SHIPS developmental dataset.

| Variable | Definition | Units |
|---|---|---|
| PIKE | persistence of IKE | TJ |
| dIKE12 | previous 12hr change of IKE | TJ |
| VMAX | maximum sustained wind speed | kts |
| VMPI | Difference between maximum potential intensity and VMAX | kts |
| LAT | latitude of storm's center | °N |
| LON | longitude of storm's center | -°W |
| MSLP | minimum sea level pressure | hPa |
| PENV | average surface pressure *(averaged from r=0-800km)* | hPa |
| VORT | 850 hPa vorticity *(r=0-1000km)* | $10^{-7}$ $s^{-1}$ |
| D200 | 200 hPa divergence *(r=0-1000km)* | $10^{-7}$ $s^{-1}$ |
| SHRD | 850-200 hPa shear magnitude *(r=0-800km)* | kts |
| SHTD | 850-200 hPa shear direction *(r=0-800km)* | ° |
| RHLO | 850-700 hPa relative humidity *(r=0-800km)* | % |
| RHMD | 700-500 hPa relative humidity *(r=0-800km)* | % |
| T150 | 150 hPa temperatures *(r=0-800km)* | °C |
| SST | sea surface temperatures | °C |
| SDAY | time after tropical storm genesis | days |
| PDAY | time from peak of season (Sept. 10) | days |

**Table 2:** Performance of SPIKE2's deterministic forecast when evaluated with the calibration input set from the GEFS F00 analyses. Sample size indicates the amount of storm fixes that were included at each forecast hour. $R_{tendency}$ measures the correlation between the observed IKE tendency value and the predicted IKE tendency value from SPIKE2's output. $R_{IKE}$ measures the correlation between the observed IKE value at validation time and the predicted IKE value calculated by adding SPIKE's tendency prediction to the existing persistence value from initialization time. Mean error is simply the mean absolute difference between the predictions from SPIKE2 and the observed IKE values.

| Forecast Hr | Sample Size | $R_{tendency}$ | $R_{IKE}$ | Mean Error |
|---|---|---|---|---|
| 12 | 1097 | 0.73 | 0.95 | 7.8 TJ |
| 24 | 974 | 0.83 | 0.92 | 10.7 TJ |
| 36 | 859 | 0.83 | 0.90 | 12.5 TJ |
| 48 | 773 | 0.86 | 0.89 | 13.4 TJ |
| 60 | 679 | 0.89 | 0.91 | 13.2 TJ |
| 72 | 614 | 0.91 | 0.90 | 14.1 TJ |

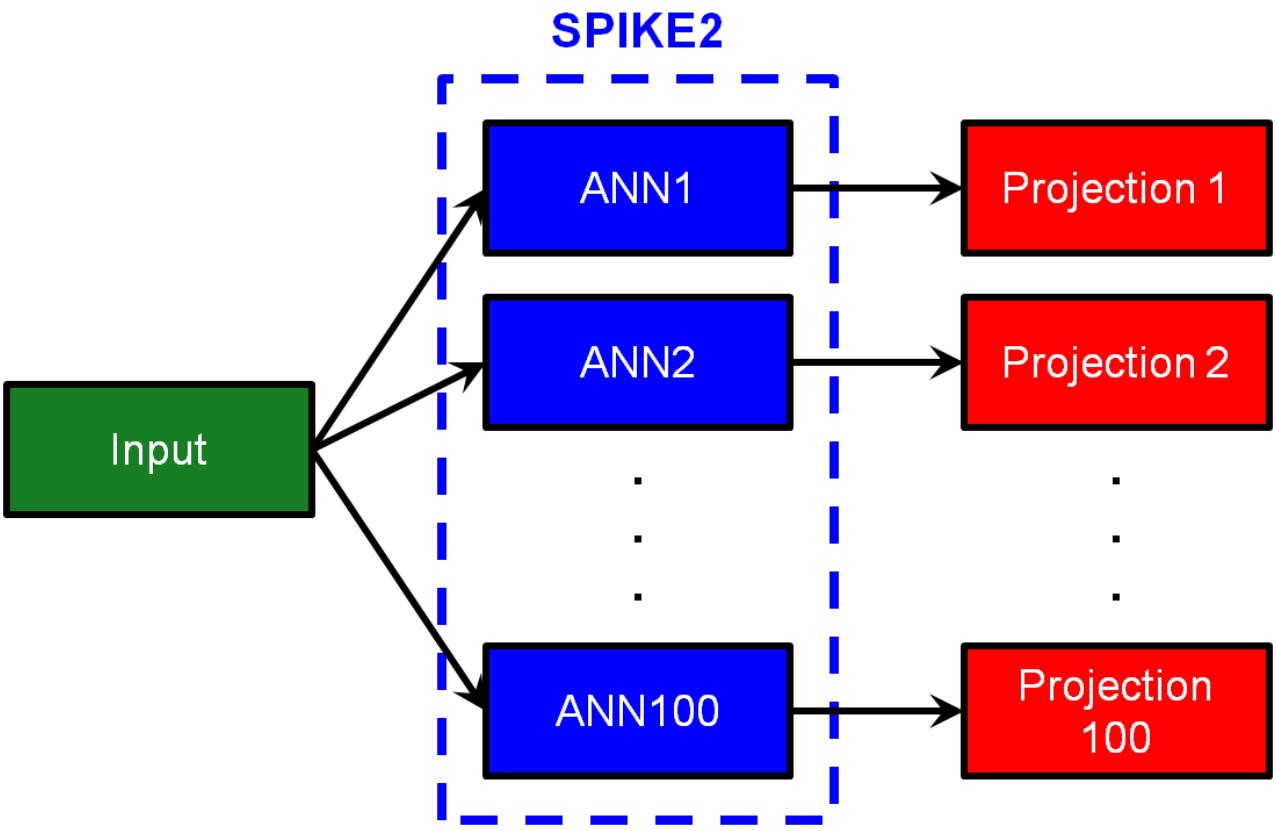



**Figure 1:** Schematic of the SPIKE2 neural network system. A single set of input parameters is passed into each of the one-hundred independent artificial neural networks (ANN1, ANN2, … ANN100) that make up SPIKE2. Each network produce its own separate prediction of IKE tendency based on the same input parameters. The median of these predictions is used as a SPIKE2's best deterministic prediction, but each individual prediction can be used for probabilistic forecasting.

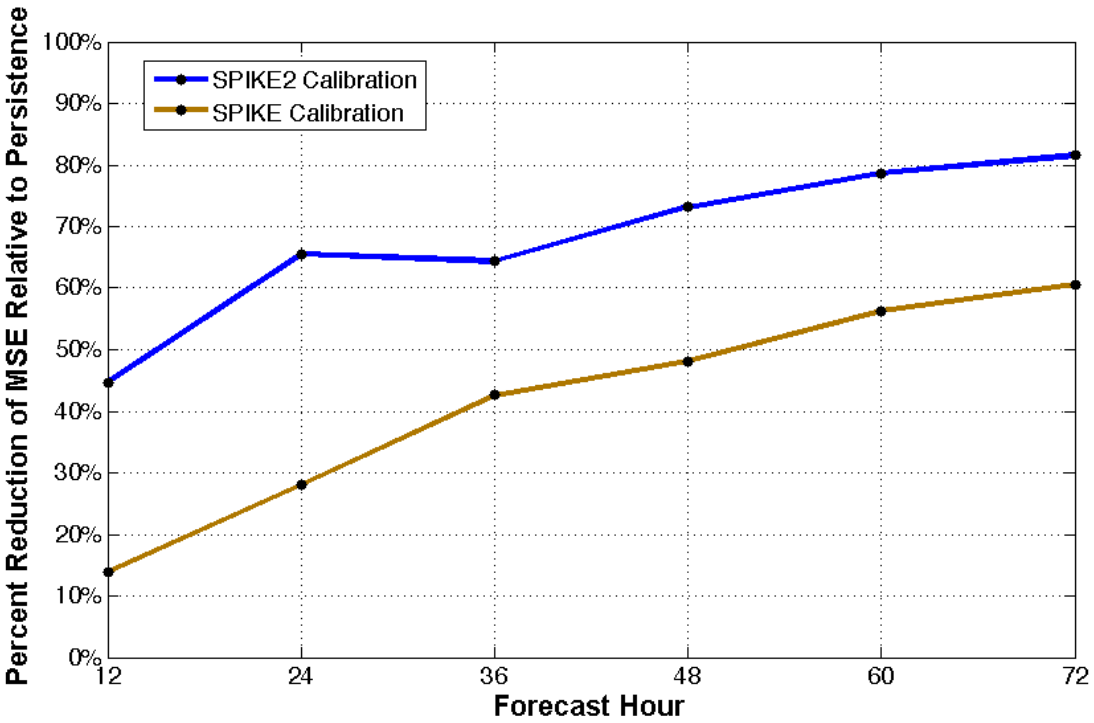



**Figure 2:** Evaluation of SPIKE2 skill in a calibration mode with GEFS analyses relative to a persistence forecast. Calibration skill is measured as a percent reduction of mean squared error (MSE) for the model's deterministic predictions from 1990 to 2011, with respect similar MSE calculations for a persistence forecast at various forecast hours. A reduction of MSE is plotted as a positive percentage, indicating that the model has outperformed persistence. SPIKE2 has significantly lower MSE than persistence at the p=0.05 level for all forecast hours. For reference, also shown is the reproduced results of the linear regression version of SPIKE as detailed in KM14.

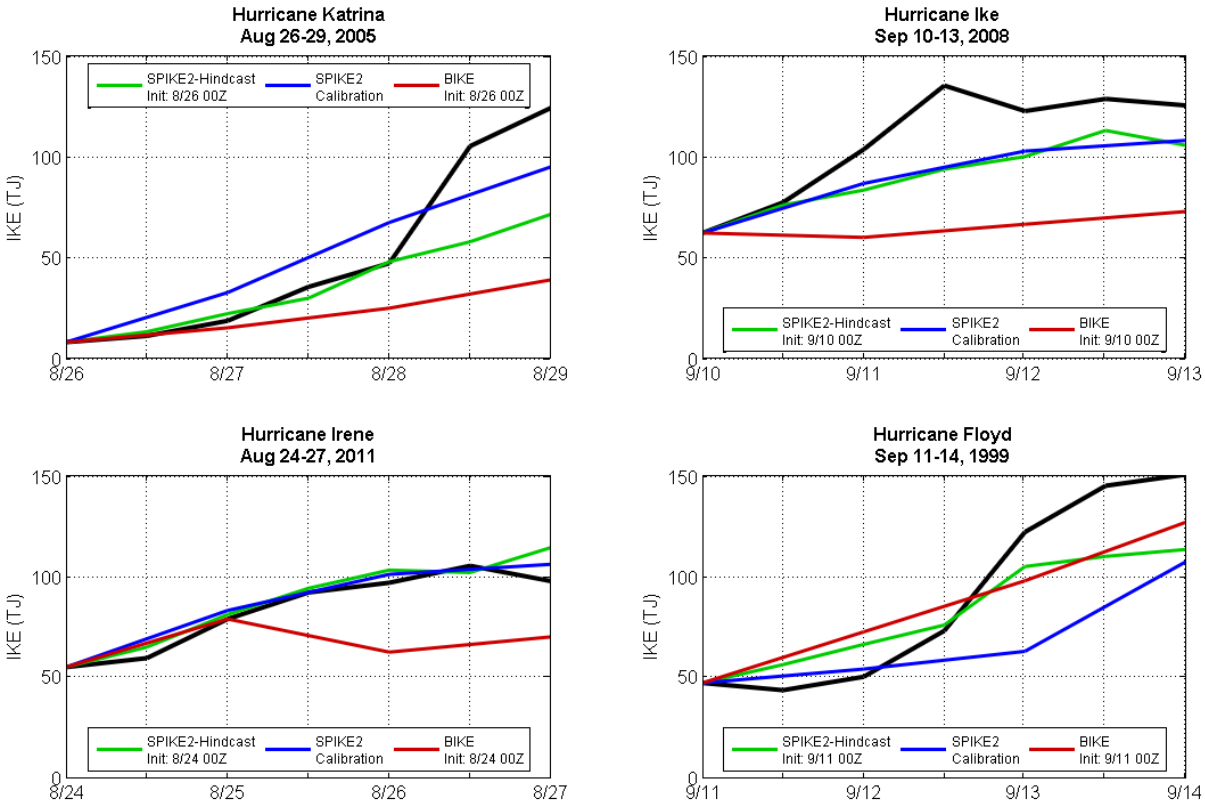
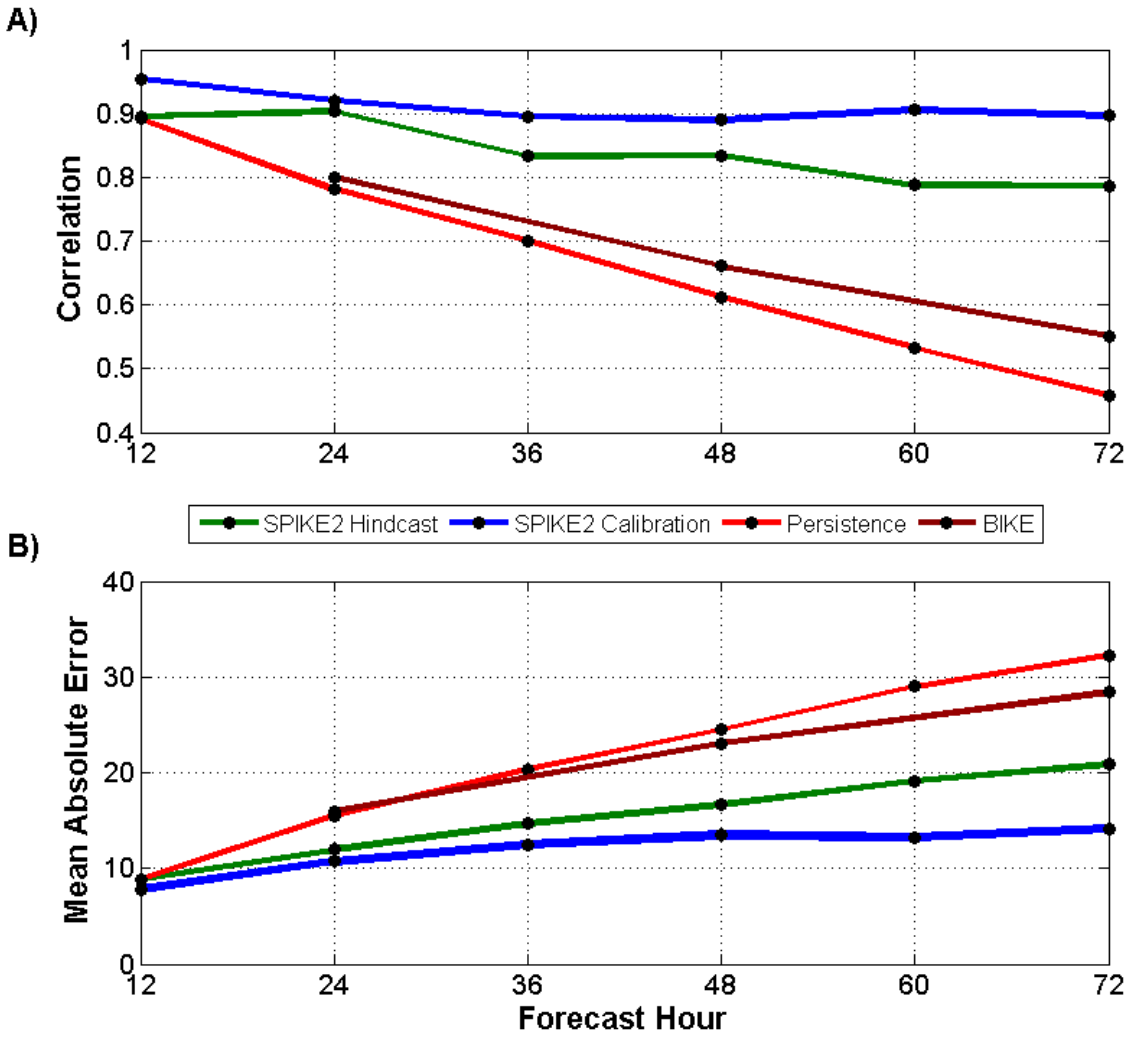



**Figure 3:** 72-hour runs of SPIKE2 plotted against benchmarks and the observed IKE values (black lines) for notable hurricanes immediately prior to their landfalls. SPIKE2 hindcasts utilize GEFS reforecasted predictors from a run initialized at the time specified in each legend. SPIKE2 calibration runs utilize analyzed predictors from the GEFS archive valid at each forecast time. The benchmark model is initialized at the same time as the SPIKE2 hindcast run for direct comparison purposes. Not explicitly shown is a persistence forecast which would be represented by a horizontal line stretching from the first observed IKE value on the left through the entire 72hr period.




861

862 **Figure 4:** Performance statistics for SPIKE2. Correlation (panel A) and mean absolute error
863 values, in units of TJ, (panel B) are shown between the historical record and SPIKE2 in various
864 modes or a persistence forecast. The correlation value that is shown in these plots is for IKE,
865 not IKE tendency. Calibration statistics are identical to those in Table 2, and are used as a
866 maximum potential reference point to determine the degradation of skill when forecast error is
867 introduced to the model in the hindcast runs via the input parameters from the GEFS reforecast.
868 Also shown for reference is the performance of a persistence forecast and the statistical
869 climatological and persistence model, BIKE.

870

871

872

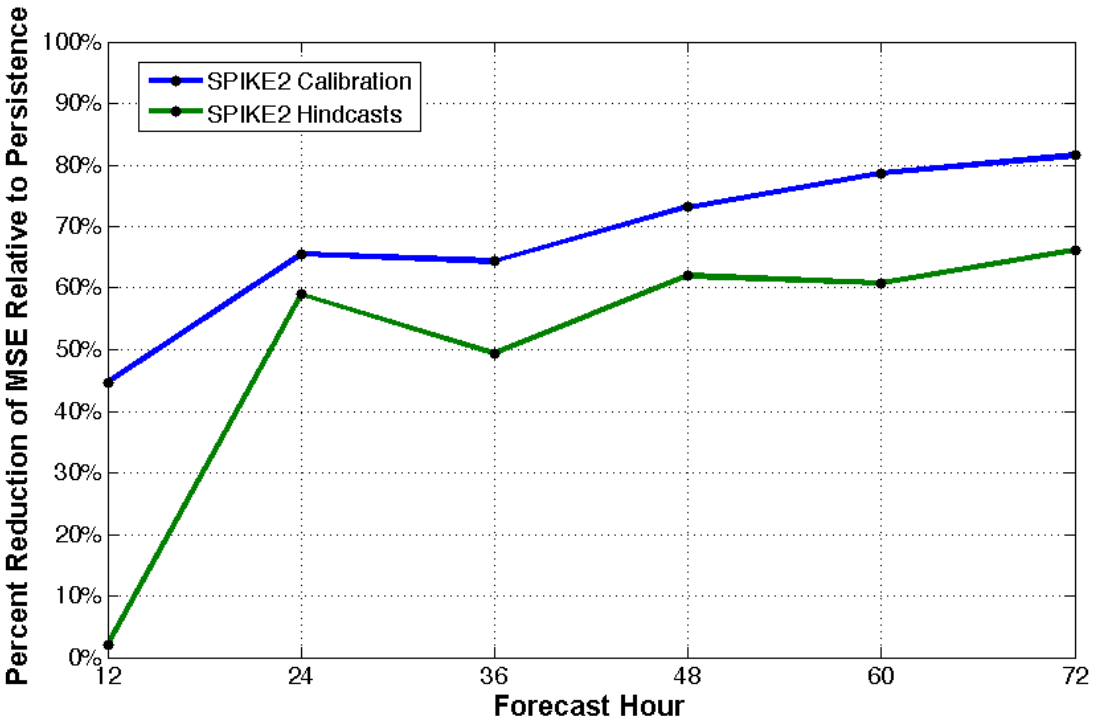



**Figure 5:** Evaluation of SPIKE2 skill relative to a persistence forecast. Performance is once again measured as a percent reduction of mean squared error (MSE) for the model's deterministic predictions from 1990 to 2011, with respect similar MSE calculations for a persistence forecast at various forecast hours. The calibration skill is reproduced from Figure 2 and is shown alongside the skill of the SPIKE2 hindcasts with reforecasted input parameters relative to persistence. The hindcast model is significantly better than persistence at the p=0.05 level for all forecast hours greater than or equal to 24-hours in length.