Probabilistic Prediction of Tropical Cyclone Intensity from a Multiple-Linear Regression Model

Chia-Ying Lee, * Michael K. Tippett,[2] Suzana J. Camargo,[3]
Adam H. Sobel[2,3]

*International Research Institute of Climate and Society, Columbia University, Palisades, NY

[2]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY

[3]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY

*Corresponding author address: Chia-Ying Lee, International Research Institute for Climate and Socieity, Columbia University, 61 Route 9W, 202 Monell, Palisades, NY 10964.

E-mail: clee@iri.columbia.edu

ABSTRACT

The authors describe the development and verification of a statistical model relating tropical cyclone intensity to the local large-scale environment. A multiple linear regression framework is used to estimate the expected intensity of a tropical cyclone given the environmental and storm conditions. The uncertainty of the estimate is constructed from the empirical distribution of model errors. NCEP-NCAR reanalysis fields and historical hurricane data from 1981 to 1999 are used for model development, and data from 2000 to 2012 are used to evaluate model performance. Seven predictors are selected: initial storm intensity, the change of storm intensity in the past 12 hours, the storm translation speed, the difference between initial storm intensity and its corresponding potential intensity, deep-layer (850-200 hPa) vertical shear, atmospheric stability, and 200 hPa divergence. The system developed here predicts storm intensity changes in response to changes in the surrounding environment with skill comparable to existing operational tools. The probabilistic intensity predictions are shown to be reliable and skillful. Since one application of such a model is to predict changes in TC activities in response to natural or anthropogenic climate change, we examine the performance of the model using data that is readily available from global climate models, i.e., limited variables and monthly averages. We find that statistical models based on monthly data (as opposed to daily) with only a few essential predictors, e.g., the difference between storm intensity and PI, perform nearly as well at short leads as when daily predictors are used.

1. Introduction

The intensities of tropical cyclones (TCs) depend on their surrounding environments. The question of what distribution of storm intensities is consistent with a given environment is important both for short-range intensity forecasts and for long-term climate projections. As the global climate changes, environmental factors that influence TC intensity, such as large-scale circulation and atmospheric moisture, are also expected to change. The issue of how these changes could influence the climatology of TC occurrence and intensity has been the subject of a number of recent studies, either from observational (Emanuel 2005; Webster et al. 2005) or from global climate modeling (Zhao et al. 2009; Bender et al. 2010; Knutson et al. 2010; Camargo 2013; Chen and Lin 2013; Emanuel 2013) approaches. However, it remains difficult to confidently assess future changes in storm intensity. One of the primary reasons is that present climate models have horizontal resolutions that are too coarse to resolve the inner core convective structure in storms. This deficiency prevents these climate models from simulating major storms at their observed intensities and providing a reasonable distribution of storm intensity even in the current climate.

Downscaling is a strategy for connecting global scale predictions and quantities not resolved by the global model, in our case, TC intensity. Dynamical downscaling has been shown to be a very useful tool for understanding climate projections of TC activities (e.g., Emanuel 2006; Knutson et al. 2007, 2008; Bender et al. 2010, etc), but it is computationally expensive and still limited in the range of intensities that can be simulated. Statistical downscaling is much cheaper computationally. This method has been used for seasonal prediction of North Atlantic hurricane activity by Vecchi et al. (2011), who developed a basin-wide hurricane frequency forecast system based on a Poisson regression with two predictors: Atliantic main development region sea surface temperature (SST) and global tropical-mean SST from a suit of high-resolution global models. Their results from retrospective forecasts indicated that their statistical model can provide a skillful seasonal prediction.

As there is always an inherent uncertainty in TC predictions, in the absence of a per-

fect model (or perfect mathematical equations), a probabilistic forecast representing this uncertainty should be provided along with the deterministic forecast. There are two general approaches for constructing probabilistic forecasts: ensemble and statistical approaches. In the ensemble approach, a set of forecasts are conducted from either parallel multiple dynamical or statistical models (Krishnamurti et al. 1999; Puri et al. 2001; Weber 2005), or a single numerical model under various initial conditions (Zhang and Krishnamurti 1999), physical parametrization schemes, or stochastic perturbations (Chen et al. 2014). In the statistical approach, the uncertainties are based on the probability density function (PDF) of the past errors or on forecasts from a deterministic numerical or statistical model (DeMaria et al. 2009). Vecchi et al. (2011) used both ensemble and statistical approaches to estimate the uncertainties.

From the perspective of short-range TC intensity forecasting, studies from DeMaria and Kaplan (1994, 1999); DeMaria et al. (2005, 2007); DeMaria (2009) show that statistical models (e.g., Statistical Hurricane Intensity Prediction Scheme, SHIPS; Logistic Growth Equation Model, LEGM) with the environmental parameters from global models, in addition to climatological and persistent predictors, are capable of predicting storm intensity with some skill. We can think of this as essentially a form of statistical downscaling applied to individual storms in the present climate. These studies' results suggest that the statistical downscaling method may provide us with an alternative option for climate projections of TC intensity distribution: utilizing a statistical model to relate the global climate model (GCM) environment to TC intensity. In other words, we could use a system that responds correctly to the changes in the surrounding (local) environment in the current climate, and apply the current relationship between TC intensity and environment to the future when environmental conditions are projected by climate models to change.

For short-range probabilistic predictions, the National Hurricane Center (NHC) started issuing wind speed probabilistic forecasts in 2006 based on a statistical approach (DeMaria et al. 2009, 2013). They use a Monte Carlo method to generate 1000 forecast tracks based on

random samples of track errors from the past 5 years of data. A similar method is applied to account for the uncertainties in storm intensity and structure. Emanuel et al. (2006) further used synthetic TC tracks generated from an algorithm with both deterministic and random components together with an idealized axisymmetric dynamical hurricane intensity model (run along the predicted track many times) to generate a set of intensity probabilistic forecasts for a specific geographic location. This approach has now been used to study tropical cyclone risk for a number of locations and from a number of perspectives (e.g. Lin et al. 2010).

Our ultimate goal is to understand the tropical cyclone intensity distribution in the projected climate scenarios from a statistical perspective. The present study represents an initial step of our approach: the development of a probabilistic TC intensity prediction model. Following the successful example of DeMaria et al.'s work on SHIPS, we develop a SHIPS-like multiple linear regression model from global reanalysis fields derived from observations of the current climate. However, different from SHIPS, we would like to select the smallest number of predictors possible, in order to minimize the required GCM fields. Because our goal is to make projections of TC intensity under future climate change scenarios (although such projections are not yet attempted in this study), we avoid predictors which we expect will make the statistical model overly tuned to the current climate. Sea surface temperature (SST) is the most obvious example; the absolute value of SST may be a useful predictor for individual storm intensity in the present climate, but we expect the relationship of SST to TC intensity to change in future climates. We also determine whether monthly data can be used instead of daily, as this greatly reduces the data requirement for climate studies.

Detailed description of data and the development of the deterministic multiple linear regression forecast model (called 'MLR' hereafter) will be illustrated in Section 2 and 3, respectively. Examples of forecasts from the MLR, and the verification of it will be shown in Section 4. The dependence of the MLR performance (errors) on predictors will also be

discussed. In Section 5, the possibility of using monthly model output is tested through examining the sensitivity of MLR performance to global model fields. In Section 6, we will demonstrate a skillful and reliable probabilistic forecast from the MLR which is conducted with the past error distribution. We summarize our findings in Section 7.

2. Data

The best-track dataset produced by NHC for North Atlantic TCs provides historical storm information, which includes storm location, maximum wind speed, and minimum sea level pressure (Jarvinen et al. 1984; Landsea and Franklin 2013). For synoptic conditions, we use the $2.5^o \times 2.5^o$ daily and monthly National Centers for Environmental Prediction –National Center for Atmospheric Research (NCEP-NCAR) reanalysis dataset (Kalnay et al. 1996). In addition to atmospheric fields, we obtain upper ocean structure from the $\sim 1.8^o \times 1.8^o$ the NOAA NCEP Environmental Model Center (EMC) Climate Modeling Branch (CMB) global ocean data assimilation system (GODAS; Behringer and Xue 2004). We utilize the data from 1981 to 1999 for model training - both predictor selection and coefficient estimation. The data for the period 2000-2012 is used for testing model performance. We only consider storms that reach at least tropical storm strength (> 34 kt) during their lifetime. As mentioned in DeMaria and Kaplan (1994), the statistical properties of storms over land are different from those over the ocean, so we only use the data for the period when storms are not over land. In the end, we have more than 4,000 cases of 12 hour forecasts in both the training and testing datasets. The number of forecasts decreases with increasing lead time. Here we consider 12-hourly lead times from 12 to 120 hours. There are only about 2,000 cases left for both training and testing for the 120 hour lead forecast. Additionally, the official NHC and SHIPS TC forecasts from Automated Tropical Cyclone Forecasting System (ATCF; Sampson and Schrader 2000) are used for evaluating the MLR performance.

3. Multiple-linear regression model development

We begin by describing the initial pool of potential predictors for the MLR. We then choose a subset of those predictors for inclusion in the multiple linear regression model (MLR) using a forward selection (stepwise regression) procedure. The next step is to use a fitting procedure on the training data to derive the MLR, which consists of 10 linear equations for forecasts with 12-hourly lead times from 12 to 120 hours.

a. Initial pool of predictors

Four variables are chosen that represent the storm itself: initial maximum wind speed (V_0), minimum sea level pressure (MSLP), change of storm maximum wind speed in the previous 12 hours (dVdt), and the storm translation speed (trSpeed). The environmental (synoptic) conditions are represented by the vertical wind shear, moisture, environmental stability, outflow temperature, upper tropospheric divergence, sea surface temperature (SST), and upper ocean thermal structures.

The most important synoptic predictor in the SHIPS includes the maximum potential intensity (PI), which is derived empirically (DeMaria and Kaplan 1994). Here, we use the Emanuel's PI definition developed in Emanuel (1988) and modified in Emanuel (1995), and Bister and Emanuel (2002). The PI is a function of the ratio of the exchange coefficient for enthalpy to that for momentum, the ratio of outflow temperature to SST, and the vertical integral of air temperature along iso-entropy surface at the radius of maximum wind speed (Bister and Emanuel 2002). In Camargo et al. (2007) and Camargo et al. (2009), the importance of PI for predicting tropical cyclogenesis was shown using the PI computed both from daily and from monthly NCEP-NCAR reanalysis data. We use here the PI from Camargo's dataset extended to 2012. The PI is first averaged over a disk extending 500 km from the storm center and then averaged over the forecast interval (e.g., 12, 24, ..., 120 hours) along the storm track. The PI enters the MLR as the difference of the mean PI and initial

intensity V_0, denoted dPI_V_0.

We do not use SST as a predictor, nor any function of SST which includes specific threshold values based on observations of the historical relationship between SST and TC intensity. The relationship of SST to TC intensity is expected to be change with the mean climate state (e.g., Vecchi and Soden 2007; Ramsay and Sobel 2011; Johnson and Xie 2010; Emanuel and Sobel 2013). The relative SST, or difference between the local SST and an appropriately defined tropical mean, may be a better predictor than absolute SST, but even that is limited in that it cannot (by definition) capture intensity changes that are related to global mean changes(e.g., Emanuel and Sobel 2013). PI is a theoretically derived quantity whose formulation does not include any explicit tuning to the present climate, and thus in principle should be able to capture intensity changes which result from global climate changes. Emanuel's PI theory has been critiqued on theoretical grounds (e.g., Smith et al. 2008) but appears at this point to be as good as any extant theory (e.g., Bryan and Rotunno 2009). Our use of PI to the exclusion of any SST parameter thus appears to be a reasonable choice for our purpose. If advances in the theoretical understanding of TC intensity and its relationship to the environment are made in the future, our model can be easily re-derived to use any improved PI parameter that may be developed.

Strong vertical wind shear tends to have a negative effect on TC intensification (DeMaria 2006; Tang and Emanuel 2012), and is another key predictor in the SHIPS model (DeMaria et al. 2005). In this study, the deep-layer mean vertical wind shear (SHRD) is defined as the magnitude of the difference between the mean wind vectors at 200 and 850 hPa averaged over an annulus extending 200 km to 800 km around the storm center (Chen et al. 2006). Moisture is an another factor known to affect TC intensity. Kaplan and DeMaria (2003) found that rapidly intensifying TCs are often embedded in a high moisture environment. In the SHIPS model, four variables representing relative humidity at various levels (1000, 850-700, 700-500, and 500-300 hPa) have been used in the past. Until 2013, only

the 700-500 hPa mean relative humidity was used[1]. For the MLR, we initially consider two moisture variables: the mid level relative humidity (rhMid, 500-300 hPa) and the column-integrated relative humidity (rhCol; Bretherton et al. 2004). The latter accounts for the accumulated moisture from the top of the boundary layer to 200 hPa. For the environmental stability, we use the conditional instability ($d\theta_{es}/dz$), defined as the vertical gradient of saturated equivalent potential temperature (dThetaEs) averaged from boundary layer to 300 hPa. SHIPS considers convective instability (using equivalent potential temperature, θ_e) instead of conditional instability. However, convective instability is usually related to the lifting of an entire layer, not just a parcel. The upper tropospheric divergence is usually used to estimate the large-scale forcing, and we use the divergence field at 200 hPa (div200) for this purpose. While there has been some suggestion that convective available potential energy (CAPE) is not a good predictor for deep tropical convection (Zipser 2003), CAPE is included in our initial predictor pool. The last atmospheric environmental predictor is the temperature at 200 hPa (T200), representing the outflow temperature.

Similar to the environmental predictors calculated in the SHIPS, rhMid, rhCol, dThetaEs, CAPE, and T200 are first averaged over an annulus centered at the storm location at forecast time with inner radius of 200 km and outer radius of 800 km. div200 is averaged over a circle with radius of 1000 km. To account for the time variation of these variables, they are also averaged over the forecast interval and along the storm track, as described for dPI_V_0. As we are using the best-track data here, we do not apply a vortex removal scheme as used in SHIPS (DeMaria 2010) to account for differences in the instantaneous Global Forecast System (GFS) forecast fields and the NHC forecast of the storm center.

In 2004, SHIPS started incorporating predictors associated with the upper ocean (ocean heat content, ocean depth of 20 and 26 ^{o}C) from satellite altimetry data to account for the oceanic negative impact on the storm (DeMaria et al. 2005). Here, instead of upper ocean

[1] `http://rammb.cira.colostate.edu/research/tropical_cyclones/ships/docs/SHIPS_predictor_file_2013.doc`

heat content, we use another air-sea interaction quantity, the ocean temperature averaged over the top 100 m (OT100; Price 2009), which represents the lowest SST that can be induced by a storm. As daily oceanic data are not readily available, monthly data from GODAS (Behringer and Xue 2004) is used here.

Our initial pool of predictors contains 13 variables: initial maximum wind speed (V_0), initial minimum sea level pressure (MSLP), changes of storm intensity in the past 12 hours (dVdt), storm translation speed (trSpeed), difference between initial intensity and final PI: dPI_V_0, vertical shear (SHRD), mid level and column-integrated relative humidity (rhMid and rhCol), conditional stability (dThetaEs), CAPE, 200 mb divergence (div200) and temperature (T200), upper 100-m mean ocean temperature (OT100). The change of the maximum wind speed over different forecast interval (e.g. 12, 24, 48, 72, 96, and 120 hours) is the predictand. The assumption of a linear relationship between intensity change and each predictor was tested as in Tippett et al. (2011, see their Fig. 12) and found (not shown) to be adequate.

b. Forward selection procedure

In the forward selection procedure, the initial model contains no predictors other than the constant term. The reduction obtained in root-mean-square-error (RMSE) from adding a single variable is computed for each variable, and the variable that most reduces the RMSE is added. This procedure is repeated until all the variables are added. The RMSE is calculated with 10-fold cross validation; the training data are divided into 10 random subsets, 9 are used to estimate regression coefficients and the tenth is used to compute an estimate of the RMSE. Here, the random partitioning of the data is done 10 times, providing 100 estimates of the RMSE, similarly to Tippett et al. (2012). The mean RMSE as a function of the number of predictors for all lead times are shown in Fig. 1. The predictors are listed on the right of each panel in Fig. 1 in the order in which they enter the forward selection procedure. A reasonable rule for the number of predictors to include in the MLR is to stop including

predictors at the point where adding additional predictors does not significantly decrease the mean RMSE.

To estimate this stopping point robustly, we apply the forward selection procedure on the training data 10 times. While the order of the first few predictors is robust, the order of the remaining predictors is not and varies with the particular partition of the data. For example, for 12 hour forecast, the order of the 3rd - 13th predictors varies randomly each time we run the forward selection procedure, indicating little difference in the utility of the remaining predictors. Hence, we add the second stopping rule at the point (red lines in Fig. 1) when the order of selected predictors is stable.

For 12 hour forecasts (Fig. 1a), a substantial decrease in the RMSE occurs when the number of the predictors increases from 1 (dVdt) to 2 (dVdt and dPI_V_0). Additional predictors result in little reduction of the RMSE, indicating that the intensification trend and the difference between PI and the initial intensity are the most important predictors for short-lead forecasts. The failure of other environmental predictors to reduce RMSE may indicate that the predictable impact of these environmental parameters, such as shear, are implicitly included in the previous 12 hours intensity changes (dVdt). For a 24 hour lead time, dPI_V_0 is the most important predictor, followed by dVdt and SHRD. Therefore, for forecast times shorter than 1 day, three predictors, dVdt, dPI_V_0, and SHRD, are adequate, and other predictors do not substantially reduce the RMSE.

The number of predictors required in the MLR (before RMSE stops decreasing) increases with the forecasting lead time. For a forecast with 72 hours interval, we need 6 predictors before the RMSE stops decreasing sharply, and the first 7 predictors selected have a fixed order. These predictors are dPI_V_0, trSpeed, SHRD, V_0, dThetaEs, dVdt, and div200. For 96 and 120 hour forecasts, however, the sharp decrease of RMSE happens at 4, and only 5 predictors are selected in a fixed order each time we conduct the selection procedure. A possible reason could be that we need predictors other than those we have in order to further improve the 96 and 120 hours intensity forecasts. Another possible reason is that there is

a natural limitation of the linearity assumption for long-forecast intervals, i.e., we cannot continue to improve the forecast from a multiple-linear regression model with additional predictors, and a more sophisticated model is needed.

Overall we see that no more than 7 predictors (dPI_V_0,dVdt, trSpeed, V_0, SHRD, div200, dThetaEs) are needed to achieve most of the possible reduction in RMSE. The absence of any moisture parameters is noteworthy, and implies that the moisture variables do not add independent information. However, this result does not mean that moisture has no role. For instance, the impact of dry air is greatest when it reaches the core, which requires shear-dry air interaction. To capture such effect would require some other combination of shear and moisture parameters. OT100, the mean upper ocean temperature, is also not chosen. The analogous parameter in SHIPS, OHC, yields 4-8% improvement in the forecast error only when a threshold (50 KJcm^{-2}) is applied (DeMaria et al. 2005). It is possible that we need to use some kind of threshold when including OT100. Nevertheless, oceanic negative feedback also depends on the storm moving speed, trSpeed, which is selected here. Although T200 does not pass the selection test, the upper tropospheric temperature is included in the PI calculation. CAPE is eliminated as well, which is not too surprising since it has been noted that it might not be a useful predictor for tropical convection (Zipser 2003).

The signs of the coefficients (not shown) are consistent with the physical relationships between predictors and intensity change, though such a result is not guaranteed for correlated predictors. Storms are more likely to follow their historical development and therefore an intensifying storm would continue intensifying, yielding a positive coefficient of dVdt. Increasing storm speed reduces the negative oceanic feedback, and the coefficient of trSpeed is positive. There is a greater potential for a storm far from its PI value to continue intensifying, and the coefficient of dPI_V_0 is positive. Stronger shear prevents storms from strengthening, and the coefficient of SHRD is negative. The stronger the divergence and the more unstable atmosphere is, the better chance a system has to develop. Therefore the coefficients of div200 and dThetaEs are positive and negative, respectively.

4. MLR intensity forecasts and errors

a. Examples: Rita (2005), Earl (2010), and Irene (2011), Isaac (2012)

After reducing the number of predictors from 13 to 7, we test the MLR with independent data from the period 2000 to 2012. As examples, we show the intensity predictions for four storms, Hurricanes Rita (2005), Earl (2010), Irene (2011), and Isaac (2012); their tracks are shown in Fig. 2. These storms encompassed a wide range of intensities, and have tracks that are typical for North Atlantic storms. Rita was a category 5 hurricane; Earl was strong category 4; Irene was category 3; and Isaac was weak category 1. Both Rita and Earl underwent rapid intensification. Rita and Isaac passed over the Gulf of Mexico while Earl and Irene recurved and moved toward the northeast US.

The MLR creates five day forecasts at 12 hourly intervals, and we apply it every 12 hours during the lifetime of the storms (red lines in Fig. 3). For the purpose of comparison, we also plot forecasts from the SHIPS (blue line) in Fig. 3. The SHIPS predicts a faster intensification rate than the MLR does overall. Both models over-predict the intensity change (Isaac 18 to 21 Aug; Earl 25 to 28 Aug) while Isaac and Earl are in their early stage (tropical storm intensity). Aside from failing to keep a storm in its TS stage, neither the MLR nor the SHIPS is capable of capturing rapid intensification (RI). As addressed in DeMaria and Kaplan (1994), this is the limitation of a simple linear regression model. For a case where the storm strengthens gradually (Irene), both models do a fair job (Fig. 3d). In the decaying period, when the intensities drop rapidly, both models again have difficulty (after 22 Sep in Rita; after 03 Sep in Earl).

b. Dependence of forecast errors on storm characteristics and synoptic conditions

Errors in intensity predictions (from both dynamic and statistical models) are related to predictors particular to individual storms, such as initial intensity, translation speed, latitude, current intensity, PI, and wind shear (Bhatia and Nolan 2013). The relationship

between errors and predictors can differ among forecast times, and is not linear. Understanding the relationship between errors and predictors can help us to better interpret and utilize forecast results. We stratify errors by predictor value and forecast time (not shown). Overall, the error increases with lead time till around 96 hour lead time, after which the error seems to be saturated. The weaker the initial storm intensity, the larger the error can be, especially for longer lead times. Usually, strong storms are closer to PI, therefore there is less chance for them to intensify or undergo RI. Strong storms are also more resilient to shear, and are unlikely to undergo rapid weakening. Therefore, there is an upper bound on errors from both intensifying and decaying storms. We can expect that the MLR might have larger errors while predicting weak storms. We also observe that the largest errors occur when dVdt is large (strong storm intensification fast in the past 12 hours), which is believed to be related to RI processes which are unresolved in the MLR. Larger errors also occur as storms move faster. Such situation usually happens when storms recurve to the mid-latitude. Another situation with possible larger errors occurs when the storms are in an environment which are favorable for RI (large PI and large positive dPI_V_0), and when storms recurve (large negative dPI_V_0). Forecast errors are also sensitive to SHRD and atmospheric stability, and the largest errors occur in unstable and moderate environmental shear. Moderate shear and more unstable atmosphere are typical environments for storms in the tropical Atlantic (10-20^oN) and the Gulf of Mexico.

To understand the relationship between the predictors and the forecast error, in Fig. 4 we show the sparkline chart of the standard deviation (σ) of the forecast error for the individual predictors for 24 hours lead time. The larger the standard deviation's variability over the individual predictor's magnitude range, the more sensitive is the error to the predictor. Among all predictors, the forecast error seems to be most sensitive to the initial storm intensity, somewhat less to PI and dPI_V_0, and not very sensitive to the remaining predictors.

The dependence of the MLR error on storm location is shown in Fig. 5. In general, the error seems to show no systematic dependence on location, especially for shorter lead times.

However, there are negative biases over the Caribbean sea and western Gulf of Mexico. Over these regions, there is usually a very warm upper ocean, which is thought to be essential for RI (Lin et al. 2009; Hui and Wang 2014). Therefore such bias could be due to lacking predictors related to warm upper ocean structure in the MLR, which leads to an underestimation of storm intensity prediction. However, because OT100 (the ocean temperature averaged over the upper 100 m) does not passed the predictor selection procedure, considering upper ocean structure might only be necessary for the storms passing through the warm upper ocean area. For 96 hour forecasts, there is a positive bias (the MLR over-predicts the storm intensity) over the west Atlantic, especially along the east coast of United States.

c. Verification of the MLR

To evaluate the overall performance of the MLR, we calculate the mean absolute error (MAE) and the RMSE against the best track data from all the testing cases (Fig. 6a). Errors from the NHC official forecast (OFCL), SHIPS, and the Statistical Hurricane Intensity Forecast (SHIFOR; Knaff et al. 2003) from 2000 to 2012 are also shown. It is important to note that the specific data used by the MLR (e.g., reanalysis rather than forecast fields) gives an advantage to MLR over the operational forecasts, a point we will discuss in more detail later. Both the MAE and RMSE of the MLR increase with lead time from 12 to 96 hours forecasts. The error does not continue to grow from 96 to 120 hours in the MLR and OFCL. In theory, we expect the error to increase with forecast time, but the error can not grow indefinitely. In other words, there should be an upper bound on the forecast error, and we expect the error to level off after a certain lead time, as does the error from OFCL. However, although not significant, there is small decline of the error from 96 to 120 hours in the MLR, which could be due to the small sample size.

Conventionally, for any TC intensity prediction model to be considered skillful, the model has to provide forecasts that are better than SHIFOR. As expected, the MLR, the SHIPS, and the OFCL all have skill by this criterion at almost all lead times. An additional baseline

model is the persistence model, whose forecast intensity at all leads is simply the initial intensity (black-dashed lines in Fig. 6) here. The MLR is skillful for all forecasting times against this baseline as well (red line has smaller errors than the black dashed line in Fig. 6a, b).

Compared to the SHIPS, the MLR has smaller MAE and RMSE for the 12-hour forecasts, and the 12-hour SHIPS MAE is about 0.5 kt larger than that of the baseline model. However, this behavior is due to the fact that the baseline model (as well as the MLR) is using the storm track and initial intensity from best-track data rather the NHC intensity forecast (provided by ATCF) which is the best information available for SHIPS to use. Since all the models are verified against the best-track data, it is reasonable that our baseline model and the MLR have smaller errors than does SHIPS.

The MAE of the MLR 12-hour forecasts best-track data is 5.8 kt. Replacing the best-track data for current and past storm intensity with those from ATCF, the 12-hour MLR MAE increases to 6.5 kt (not shown), which is close to that of SHIPS. The MAE of the MLR for longer lead times (96-120 hours) increases by 1 kt (not shown) when ATCF data is used. To assess the impact of track forecast error on the performance of the MLR, we replace best-track location data with NHC forecast tracks (MLR_FstTrack, blue lines in Fig. 6c, and d). After the 12-hour forecasts, the MAE in the MLR_FstTrack forecasts is on average 2-3 kts higher than the MLR. To estimate the impact of using reanalysis data rather than GFS forecast data for environmental fields, we conduct an experiment using predictors taken from the NCEP reanalysis fields at the initial time and then keeping these values fixed (i.e. persistence) for all lead times (MLR_initNCEP, orange line in Fig. 6c, and d). MAE in the MLR_initNCEP experiment increases by about 1-2 kt in average.

Another important difference between MLR and SHIPS is the choice of predictors. SHIPS includes many more environmental parameters than the MLR does. Till 2013, SHIPS had 24 predictors (Schumacher et al. 2013). We conducted experiments to see how varying the numbers of predictors impacts the errors. Figure 6e, and f show the errors from three of these experiments. When using dPI_V_0 (dPI_V_0: blue line) as the only predictor, we

surprisingly find out that the intensity errors, both MAE and RMSE, are not too far from those obtained with the full MLR. Adding dVdt (dPI_V_0+dVdt: solid sky-blue line) as the second predictor results in a small improvement. These results suggest that with only one or two predictors, the regression model still has some degree of skill. If we continue adding predictors, we continue to see an improvement in the MLR performance, such as adding SHRD (...+SHRD: dashed-sky-blue line). Using all 13 predictors from initial pool of predictors (MLR_13: orange line), however, results in larger errors for 24 to 60 hours forecasts, an indication of over-fitting.

The difference between the performance of the MLR and SHIPS results from using best-track or NHC forecasting TC data, using global reanalysis or forecast fields, and the choice of predictors. Another difference is the data used to estimate the coefficients. SHIPS coefficients are re-estimated each year; MLR coefficients are not. The use of best-track data and reanalysis fields are the main reason why the MLR performs better than the SHIPS. It is not clear whether the use of more predictors causes better performance in the SHIPS than in the MLR for 24 to 60-hour lead times. Nevertheless, the experiments with different combination of predictors suggest that a multiple-linear regression model with even just a few of the most essential predictors has substantial skill.

5. Monthly v.s. daily reanalysis data

Given our interest in the variations of environment and TC intensity on long time-scales, an interesting question is to what extent the MLR can predict intensity changes given monthly rather than daily data. In an application where climate model output is used, for example, being able to use monthly data is a great advantage. With the same choice of predictors, we use monthly NCEP-NCAR reanalysis data to train and test the MLR (called MLR_Monthly). The forecasting errors of the MLR_Monthly are comparable to those of the MLR (Fig. 7). There is almost no discrepancy between the forecast errors of the MLR and

the MLR_Monthly for lead times less than 48 hours. For longer lead times (> 72 hours), on the other hand, the advantage of using daily data is clear.

The good performance of the MLR_Monthly suggests that there is some usable relation between the daily and monthly predictors. Comparing the predictors from monthly data to those from daily data, we find that the monthly dPI_V_0 is larger than the daily one (Fig. 8), and that this difference is a result of larger monthly PI. There is almost no relationship between monthly and daily div200, and a forward selection with monthly predictors (not shown) does not choose div200. This is because the monthly div200 is too small (and uncorrelated) compared to the daily one. There are also some positive biases for smaller values of SHRD and dThetaEs, but negative biases for larger values in monthly data. To assess which one is the primary cause of the larger error in MLR_Monthly, we replace each of the four predictors with daily data in the MLR_Monthly separately (not shown). Our results suggest that the difference in dPI_V_0 in monthly and daily data is the main cause for the differences between MLR and MLR_Monthly.

6. Probabilistic intensity prediction

a. Forecast distribution

As described in the introduction, accounting for the uncertainty of the MLR forecast increases its value and utility. The simplest and most direct estimate of forecast uncertainty is given by the empirical distribution of errors presented in the training process (using all the training data). This approach is feasible because of the size of the training dataset used here. Although we recognize that rare events may be poorly sampled and may benefit from parametric descriptions, we only consider using the empirical distribution of errors here. We condition the error distribution on forecast lead time, and show in Fig. 9a the increasing spread of the error PDF as the forecast lead time increases from 12 to 72 hours. There is no significant change in the error PDF between 96 and 120 forecast hours; the errors

appear to have saturated at that point. The errors are close to being normally distributed for all lead times. Figure 9b is the cumulative distribution function (CDF) and Fig. 9c compares the distribution of the MLR errors for 24 hour forecast to the standard normal distribution. Except for the extreme points, most of the points fit well within the theoretical normal distribution profile. All the large negative errors (< -35 kt) are from the cases with rapid intensification. For instance, the largest negative error in Fig. 9b is about -86 kt, and corresponds to a forecast of Hurricane Emily in 1987 when Emily underwent rapid intensification. Our probabilistic intensity forecasts consist of the PDF with mean value given by the MLR and spread given by the lead-time dependent error distribution (Fig. 10).

The analyses in Section 3b suggest that the forecast error and its distribution are sensitive to the predictors, and are most sensitive to V_0 (initial storm intensity). Strictly speaking, such dependence is beyond the linear regression framework, in which the error variance is assumed independent of the predictors. However, regression based-models with varying spread are common in weather forecasting (Gneiting et al. 2005; Wilks and Hamill 2007). Here, we examine the sensitivity of forecast error distribution to V_0. We illustrate the dependence of the error distribution on initial wind speed in the case of the 24 hour forecast error (Fig. 11). The standard deviation of all 24-hour forecast errors from all training data is 12 kts (dashed line in Fig. 11a, which is calculated from the blue line in Fig. 9a). However, binning 24-hour forecast errors based on the corresponding values of V_0 shows that the σ of errors increases, roughly linearly, as V_0 increases (black diamond-line in Fig. 11c). Now, if we do not account for the fact that the distribution of forecast error is a function of V_0, the joint CDF of errors between lead times and V_0 would look like Fig. 11b. If we do, the joint CDF would look like Fig. 11c. To understand how accounting for the dependence of error distribution to V_0 changes the results, we develop two probabilistic models. We name the probabilistic model with error distribution as a function of lead time only MLR_1, while the one utilizing error distribution as a function of both lead time and V_0 is named MLR_2.

In addition to MLR_1 and MLR_2, we construct another probabilistic model as a baseline

reference model from the persistence model (Fig. 6a) and its error distribution (Fig. 10b). Fig. 10b indicates that the mean of the distribution is near zero at any given lead time and that the spread increases with lead time. The spread here is larger than the spread in Fig. 10a. This is consistent with results in Fig. 6b: the σ (which can also be represented as RMSE) in persistence model is larger than that in the MLR. We call this model the probabilistic persistence model. When making a probabilistic forecast with the persistence model, we can simply shift the distribution in Fig. 10b to be centered on the initial intensity.

The MLR_1 and MLR_2 are considered skillful to the extent to which their probabilistic performance is better than that of the probabilistic persistence model. The fundamental difference is that the probabilistic persistence distribution is conditioned only on the current intensity and historical changes in storm intensity, while the MLR_1 and MLR_2 distributions are additionally conditioned on the current and forecast environment.

b. Probabilistic forecasts and verification

Probabilistic forecasts (persistence, MLR_1, and MLR_2) are made for all the cases in the testing data set. Figure 12 shows probabilistic forecasts for Hurricane Earl initialized on 0000 UTC 28 August, 2010, a period during which Earl underwent RI (36 - 60 forecast hours). For the 12 - 36 hour lead times, the observed storm intensity falls into the 25-75 percentile region in MLR_1 and MLR_2 forecast, while in the persistence model it does not. Once Earl started to undergo RI, the best-track data falls beyond the highest 50 percentile region in all models, although it is located at higher probabilities in MLR_1 and MLR_2 than in the persistence model. Qualitatively, the case in Fig. 12a-c exhibits the advantage in using MLR_1 and MLR_2 instead of the persistence model for probabilistic forecast. Knowing current and forecast environmental conditions helps probabilistic intensity prediction.

We use the rank probability skill score (RPSS) to verify quantitatively whether the MLR_1 and MLR_2 have more skill than the persistence model. Rank probability score (RPS) is a squared-error score with respect to the observational probability, which is 1 if

the forecast event occurs, and 0 if the event does not occur. The cumulative forecasts and observations, denoted as P_m and O_m, are:

$$P_m = \sum_{j=1}^{m} = p_j \,, \quad m = 1, \ldots, J, \tag{1}$$

and

$$O_m = \sum_{j=1}^{m} o_j \,, \quad m = 1, \ldots, J. \tag{2}$$

where p_j is the forecast probability of the storm intensity falling in the j-th bin, the observed probability o_j is 1 is the observations falls in the j-th bin and zero otherwise, and J is the total number of bins. Here, we stratify the probabilities by the corresponding V_{max} every 5 kt from 0 to 200 kt. The RPS is the sum of the squared differences between the cumulative probabilities P_m and O_m:

$$RPS = \sum_{m=1}^{J} \left(P_m - O_m\right)^2 \,. \tag{3}$$

RPS is oriented so that smaller values indicate better forecasts. A correct forecast with no uncertainty has a RPS of 0. The RPSS compares the average RPS to that of the persistence forecast:

$$RPSS = 1 - \frac{\overline{RPS}}{\overline{RPS}_{\text{persistence}}}. \tag{4}$$

If the probabilistic model is skillful, RPSS is positive, and the larger the value is, the more skill the model has. The RPS and RPSS can be computed for each forecast. Figure 12d shows the RPSS for all of the forecasts of Earl. Compared to the persistence model, the skill of MLR_1 and MLR_2 increases with the length of forecast interval, reflecting the fact that the value of knowing the environment increases as lead time increases. As the forecast interval becomes longer, the environmental conditions become more crucial for intensity prediction. Hence, the MLR_1 and MLR_2 forecasts show greater skill compared to the persistence model with increasing lead time. For all the testing cases (Fig. 13), RPSS for the MLR_1 and MLR_2 again increases with time. The modest advantage in RPSS that the MLR_2 shows over the MLR_1 indicates that considering the sensitivity of forecast errors to initial storm intensity does improve probabilistic intensity predictions.

The performance of probabilistic forecasts of particular intensity categories (TS, category 1-2, and 3-5 hurricane) can also be verified. Figure 14 shows the reliability diagrams of MLR_1, MLR_2, the persistence model, and the NHC official probabilistic forecast (OFCL) for categorized cases from 2008 to 2012 at all lead times. Forecast probabilities at each forecast time for each intensity category are grouped into bins at 0.1 (10%) interval. The probabilistic forecasts of the occurrence of TS and hurricane intensity Saffir-Simpson categories 1-2 by the two MLR models, persistence model and OFCL all show a fair reliability as the observed frequency is close to the forecast probability (Fig. 14a, c). The MLR_1, MLR_2 and persistence model forecasts of TS occurrence show almost perfect reliability for high probabilities (> 0.6 bin) and under-forecasting for low probabilities. OFCL over-forecasts the occurrence of TS in the 0.1 - 0.6 (10% - 60%) range. MLR_1, MLR_2, and OFCL predict occurrence of hurricane intensity categories 1-2 reliably for forecasts probabilities of 0 - 0.3 (0% - 30%), while the persistence model under-forecasts it. For higher probabilities (> 0.3 bin), they all over-forecast the occurrence.

For major hurricanes (category 3-5), the reliability diagram curves are very different among these four forecast models. MLR_1 and MLR_2 are still fairly reliable. The persistence model and OFCL, however, are substantially less reliable. They both overforecast the occurrence of major hurricanes. For the persistence model, the probabilistic forecast always maintains relatively higher probabilities close to the initial storm intensity, as shown in Fig. 12a. Hence, the persistence model could easily overforecast decaying major storms. Compared to the MLR_1, MLR_2 is more reliable (the red line in Fig. 14 is closer to the one-to-one line than the pink line). Note that OFCL overestimates all categories. This implies that OFCL must underestimate tropical depression events, which could be from overforecasting the development from tropical depressions to TS or underforecasting the decaying rate of tropical storms.

7. Conclusions

This study describes the development of a probabilistic tropical cyclone (TC) intensity prediction system based on a multiple linear regression model (MLR) and the past error distribution. While such models are important for operational intensity forecasts, our interest is more directed at the capability of such model to connect TC intensity to the large-scale environment. The ability of the MLR to relate variations in the large-scale environment to TC intensity has applications beyond short-term forecasting and can be a tool towards understanding the distribution of TC intensities in a climate undergoing natural or anthropogenic changes. The MLR contains 7 predictors selected based on their physical importance to TC development and their ability to reduce forecast errors. The predictors are the initial maximum wind speed (V_0), change of storm intensity in the past 12 hours (dVdt), storm translation speed (trSpeed), the difference between PI and V_0 (dPI_V_0), 850-200 mb vertical wind shear magnitude (SHRD), conditional stability (dThetaEs), 200 mb divergence (div200). Among these predictors, dVdt is the most important predictors for forecasts with short lead times (< 24 hours). dPI_V_0 and SHRD becomes more important with increasing lead times. We try to avoid predictors whose relationship to TC intensity is likely to change in a changed global mean climate, such as sea surface temperature (SST).

We find that the magnitudes of forecast errors depends on the synoptic situation. In particular, larger forecast errors occur when the storm is weaker and when there is a higher chance of rapid intensification. The MLR does not represent RI well, similarly to other linear models. Additionally, the MLR has relatively larger errors under moderate shear, smaller upper tropospheric divergence, and a conditionally more unstable atmosphere.

A relatively simple description of the synoptic environment is able to characterize much of the predictable component of the variability of TC intensity. The errors of a single predictor (dPI_V_0) model are not too far from those of the full 7 predictors model. Likewise, a model based on monthly averaged environment performs quite well, with its primary weakness being related to the lack of daily PI information. Given these results, monthly data from

global climate models with only the most essential predictors might be sufficient for some climate applications.

In addition to the deterministic forecasts, we demonstrate the ability of the MLR to produce probabilistic predictions. We construct a forecast PDF with mean given by the MLR and uncertainty given by the empirical error distribution from the training data. Two probabilistic models are developed here. In MLR_1, the error distribution is only a function of lead time while in MLR_2, it is also a function of initial storm intensity, reflecting the dependence of forecast errors on the environment. A reference probabilistic forecast is constructed from historical storm intensity changes (independent of environment) and used to compute the rank probability score skill (RPSS). Both MLR_1 and MLR_2 have good skill, especially for a long lead forecasts when their advantage over the persistance forecast is greatest. The analysis of the reliability diagram further shows that the reliability among two MLR probabilistic models, the persistence model, and official NHC forecasts are similar to each other for TS, and hurricanes of categories 1 and 2. Nevertheless, MLR_1 and MLR_2 are much more reliable than OFCL and persistence model for major hurricanes. Furthermore, both RPSS and reliability analyses suggest that using an environment-dependent error distribution (MLR_2) improves the forecast skill.

In short, in this study we have developed a system that reliably predicts the distribution of storm intensity conditional on the surrounding local environment. We have developed and verified the system in a forecast framework using regression methods and probabilistic verification to assess resolution and reliability of the intensity distribution. This reliable probabilistic model could be a component of a system for future risk assessment.

Acknowledgments.

We thank Dr. Mark DeMaria for the useful comment. The research was supported by Office of Naval Research under the research grant of MURI (N00014-12-1-0911).

REFERENCES

Behringer, D. W. and Y. Xue, 2004: Evaluation of the global ocean data assimilation system at NCEP: The Pacific Ocean. *Eighth Symposium on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface*, AMS 84th Annual Meeting, Washington State Convention and Trade Center, Seattle, WA, 11–15.

Bender, M. A., T. R. Knutson, R. E. Tuleya, J. J. Sirutis, G. A. Vecchi, S. T. Garner, and I. M. Held, 2010: Modeled impact of anthropogenic warming on the frequency of intense Atlantic hurricanes. *Science*, **327**, 454–458.

Bhatia, K. T. and D. S. Nolan, 2013: Relating the skill of tropical cyclone intensity forecasts to the synoptic environment. *Wea. Forecasting*, **28**, 961–980.

Bister, M. and K. A. Emanuel, 2002: Low frequency variability of tropical cyclone potential intensity 1. Interannual to interdecadal variability. *J. Geophys. Res.: Atmospheres*, **107**, 4801.

Bretherton, C. S., M. E. Peters, and L. E. Back, 2004: Relationships between water vapor path and precipitation over the tropical oceans. *J. Climate*, **17**, 1517–1528.

Bryan, G. H. and R. Rotunno, 2009: Evaluation of an analytical model for the maximum intensity of tropical cyclones. *J. Atmos. Sci.*, **66 (10)**, 3042–3060.

Camargo, S. J., 2013: Global and regional aspects of tropical cyclone activity in the CMIP5 models. *J. Climate*, **26**, 9880–9902.

Camargo, S. J., K. A. Emanuel, and A. H. Sobel, 2007: Use of a genesis potential index to diagnose ENSO effects on tropical cyclone genesis. *J. Climate*, **20**, 4819–4834.

616 Camargo, S. J., M. C. Wheller, and A. H. Sobel, 2009: Diagnosis of the MJO modulation of
617 tropical cyclogenesis using an empirical index. *J. Atmos. Sci.*, **66**, 3061–3074.

618 Chen, J.-H. and S.-J. Lin, 2013: Seasonal predictions of tropical cyclones using a 25-km-
619 resolution general circulation model. *J. Climate*, **26**, 380–398.

620 Chen, S. S., F. Judt, J. Berner, C.-Y. Lee, M. Curcic, C. Snyder, and R. Rotunno, 2014:
621 Distinct characteristics of hurricane ensemble forecasts using physical parameterizations
622 vs. stochastic perturbations. AMS 94th Annual Meeting, Atlanta, GA.

623 Chen, S. S., J. A. Knaff, and F. D. Marks, 2006: Effects of vertical wind shear and storm
624 motion on tropical cyclone rainfall asymmetries deduced from TRMM. *Mon. Wea. Rev.*,
625 **134**, 3190–3208.

626 DeMaria, M., 2006: The effect of vertical shear on tropical cyclone intensity change. *J.*
627 *Atmos. Sci.*, **53**, 2076–2088.

628 DeMaria, M., 2009: A simplified dynamical system for tropical cyclone intensity prediction.
629 *Mon. Wea. Rev.*, **137**, 68–82.

630 DeMaria, M., 2010: Tropical cyclone intensity change predictability estimates using a
631 statistical-dynamical model. 29th AMS Conference on Hurricanes and Tropical Meteo-
632 rology, Tucson, AZ.

633 DeMaria, M. and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme
634 (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220.

635 DeMaria, M. and J. Kaplan, 1999: An updated Statistical Hurricane Intensity Prediction
636 Scheme (SHIPS) for the Atlantic and Eastern North Pacific basins. *Wea. Forecasting*, **14**,
637 326–337.

638 DeMaria, M., J. A. Knaff, R. Knabb, C. Lauer, C. R. Sampson, and R. T. DeMaria, 2009:

A new method for estimating tropical cyclone wind speed probabilities. *Wea. Forecasting*, **24**, 1573–1591.

DeMaria, M., J. A. Knaff, and C. Sampson, 2007: Evaluation of long-term trends in tropical cyclone intensity forecasts. *Meteorology and Atmospheric Physics*, **97 (1-4)**, 19–28.

DeMaria, M., M. Mainelli, L. K. Shay, J. A. Knaff, and J. Kaplan, 2005: Further improvements to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting*, **20**, 531–543.

DeMaria, M., et al., 2013: Improvements to the operational tropical cyclone wind speed probability model. *Wea. Forecasting*, **28**, 586–602.

Emanuel, K. and A. H. Sobel, 2013: Response of tropical sea surface temperature, precipitation, and tropical cyclone-related variables to changes in global and local forcing. *J. Adv. Model. Earth Syst.*, **5**, 1942–2466.

Emanuel, K. A., 1988: The maximum intensity of hurricanes. *J. Atmos. Sci.*, **45**, 1143–1155.

Emanuel, K. A., 1995: Sensitivity of tropical cyclones to surface exchange coefficients and a revised steady-state model incorporating eye dynamics. *J. Atmos. Sci.*, **52**, 3969–3976.

Emanuel, K. A., 2005: Increasing destructiveness of tropical cyclones over the past 30 years. *Natural*, **436**, 686–688.

Emanuel, K. A., 2006: Climate and tropical cyclone activity: A new model downscaling approach. *J. Climate*, **19**, 4797–4802.

Emanuel, K. A., 2013: Downscaling CMIP5 climate models shows increased tropical cyclone activity over the 21st century. *PNAS.*, published online.

Emanuel, K. A., S. Ravela, E. Vivant, and C. Risi, 2006: A statistical determinatic approach to hurricane risk assessment. *Bull. Amer. Meteor. Soc.*, 299–313.

662 Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic
663 forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon.*
664 *Wea. Rev.*, **133**, 1098–1118.

665 Hui, W. and Y. Wang, 2014: A numerical study of Typhoon Megi (2010). Part I: Rapid
666 Intensification. *Mon. Wea. Rev.*, **142**, 29–48.

667 Jarvinen, B. R., C. J. Neumann, and M. A. S. Davis, 1984: A tropical cyclone data tape for
668 the North Atlantic basin, 1886-1983: Contents, limitations, and uses. NOAA Technical
669 Memorandum NWS NHC 22, NOAA, Coral Gables, Florida, 21pp.

670 Johnson, N. C. and S.-P. Xie, 2010: Changes in the sea surface temperature threshold for
671 tropical convection. *Natural Geosci.*, **3**, 842–845.

672 Kalnay, E., et al., 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor.*
673 *Soc.*, **77**, 437–471.

674 Kaplan, J. and M. DeMaria, 2003: Large-scale characteristics of rapidly intensifying tropical
675 cyclones in the North Atlantic basin. *Wea. Forecasting*, **18**, 1093–1108.

676 Knaff, J. A., M. DeMaria, C. R. Sampson, and J. M. Gross, 2003: Statistical, 5-day tropical
677 cyclone intensity forecasts derived from climatology and persistence. *Wea. Forecasting*,
678 **18**, 80–92.

679 Knutson, T. R., J. J. Sirutis, S. T. Garner, I. M. Held, and R. E. Tuleya, 2007: Simulation of
680 the recent multidecadal increase of Atlantic hurricane activity using a 18-km-grid regional
681 model. *Bull. Amer. Meteor. Soc.*, **88**, 1549–1565.

682 Knutson, T. R., J. J. Sirutis, S. T. Garner, G. A. Vecchi, and I. M. Held, 2008: Simulated
683 reduction in Atlantic hurricane frequency under twenty-first-century warming condition.
684 *Natural Geosci.*, **1**, 359–364.

685 Knutson, T. R., et al., 2010: Tropical cyclones and climate change. *Natural Geosci.*, **3**,
686 157–163.

687 Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E.
688 Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate
689 forecast from multimodel superensemble. *Science*, **285**, 1548–1550.

690 Landsea, C. W. and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and
691 presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592.

692 Lin, I.-I., C.-H. Chen, I.-F. Pun, W. T. Liu, and C.-C. Wu, 2009: Warm ocean anomaly, air
693 sea fluxes, and the rapid intensification of Tropical Cyclone Nargis (2008). *Geophys. Res.*
694 *Lett.*, **36**, L03 817.

695 Lin, N., K. A. Emanuel, J. A. Smith, and E. Vanmarcke, 2010: Risk assessment of hurricane
696 storm surge for New York City. *J. Geophys. Res.: Atmospheres*, **115**, D18 121.

697 Price, J. F., 2009: Metrics of hurricane-ocean interaction: vertically-integrated or vertically-
698 averaged ocean temperature? *Ocean Science*, **5 (3)**, 351–368.

699 Puri, K., J. Barkmeijer, and T. Palmer, 2001: Ensemble prediction of tropical cyclones using
700 targeted diabatic singular vectors. *Q. J. R. Meteor. Soc.*, **127**, 709–731.

701 Ramsay, H. A. and A. H. Sobel, 2011: Effects of relative and absolute sea surface temperature
702 on tropical cyclone potential intensity using a single-column model. *J. Climate*, **24**.

703 Sampson, C. R. and A. J. Schrader, 2000: The automated tropical cyclone forecasting system
704 (version 3.2). *Bull. Amer. Meteor. Soc.*, **81**, 1231–1240.

705 Schumacher, A., M. DeMaria, and J. Knaff, 2013: Summary of the new statistical-dynamical
706 intensity foreast models for the Indian Ocean and Southern Hemisphere and resulting
707 performance. Tech. rep., JTWC.

Smith, R. K., M. T. Montgomery, and S. Vogl, 2008: A critique of Emanuel's hurricane model and potential intensity theory. *Quart. J. Roy. Meteor. Soc.*, **134 (632)**, 551–561, doi:10.1002/qj.241.

Tang, B. and K. A. Emanuel, 2012: A ventilation index for tropical cyclones. *Bull. Amer. Meteor. Soc.*, **93**, 1901–1912.

Tippett, M., S. J. Camargo, and A. H. Sobel, 2011: A possion regression indec for tropical cyclone genesis and the role of large-scale vorticity in genesis. *J. Climate*, **21**, 2335–2357.

Tippett, M. K., A. H. Sobel, and S. J. Camargo, 2012: Association of monthly U.S. tornado occurrence with large-scale atmospheric parameters. *Geophys. Res. Lett.*, **39**, L02 801.

Vecchi, G. A. and B. J. Soden, 2007: Effect of remote sea surface temperature change on tropical cyclone potential intensity. *Natural*, **450**, 1066–1070.

Vecchi, G. A., M. Zhao, G. Villarini, R. Anthony, A. Kumar, I. M. Held, and R. Gudgel, 2011: Statisitcal-dynamical predictions of seasonal north atlantic hurricane activity. *Mon. Wea. Rev.*, **139**, 1070–1082.

Weber, H. C., 2005: Probabilistic prediction of tropical cyclones. Part II: Intensity. *Mon. Wea. Rev.*, **133**, 1853–1864.

Webster, P. J., G. Holland, J. A. Curry, and H.-R. Chang, 2005: Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science*, **309**, 1844–1846.

Wilks, D. S. and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390.

Zhang, Z. and T. N. Krishnamurti, 1999: A perturbation method for hurricane ensemble predictions. *Mon. Wea. Rev.*, **127**, 447–469.

730 Zhao, M., I. M. Held, S.-J. Lin, and G. A. Vecchi, 2009: Simulations of global hurricane cli-
731 matology, interannual variability, and response to global warming using a 50-km resolution
732 GCM. *J. Climate*, **22**, 6653–6678.

733 Zipser, E., 2003: *Some views on "Hot Towers" after 50 years of tropical field programs and*
734 *two years of TRMM data.* 29, Amer. Meteor. Soc., 49-57 pp.

List of Tables

TABLE 1. Descriptions of experiments.

experiment	description
MLR	a multiple linear regression model with all 7 selected predictors, NHC best-track data, daily NCEP-NCAR reanalysis fields
MLR_FstTrack	same as MLR, but with environmental predictors calculated along the NHC official forecast track
MLR_InitNCEP	same as MLR, but with environmental predictors calculated from NCEP-NCAR reanalysis fields at the initial time
dPI_V_0	a linear regression model with only dPI_V_0 as a predictor, NHC best-track data, daily NCEP-NCAR reanalysis fields
dPI_V_0+dVdt	same as dPI_V_0, but with one more predictor: dVdt
... + SHRD	same as dPI_V_0+dVdt, but with one more predictor: SHRD
MLR_Monthly	same as MLR, but with environmental predictors calculated from monthly NCEP-NCAR reanalysis fields

List of Figures

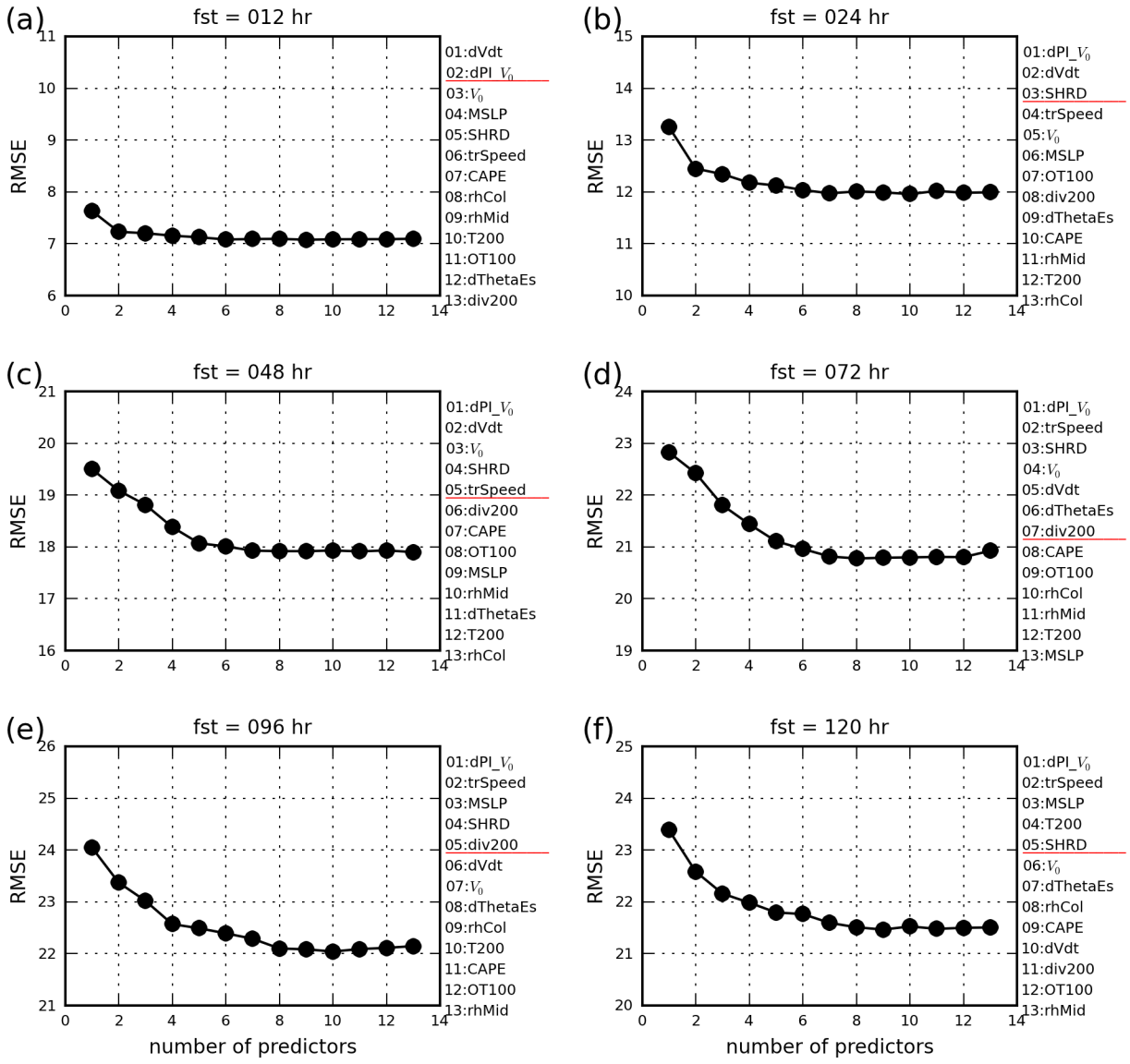

FIG. 1. Root-mean-square-error (RMSE) as a function of number of predictors used in MLR for forecast lead time: (a) 12, (b) 24, (c) 48, (d) 72, (e) 96, and (f) 120 hours. The predictors are listed on the right of each panel in the order that they are selected by the forward selection procedure. The order of the predictor selection is stable above the red line.

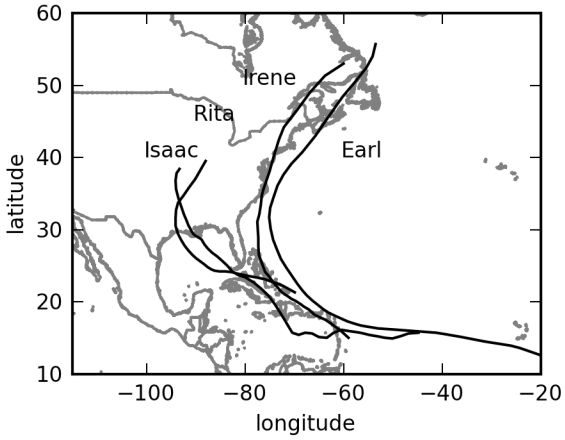

FIG. 2. Tracks for hurricanes Rita (2005), Earl (2010), Irene (2011) and Isaac (2012). The intensity predictions for each storm are shown in Fig. 3.

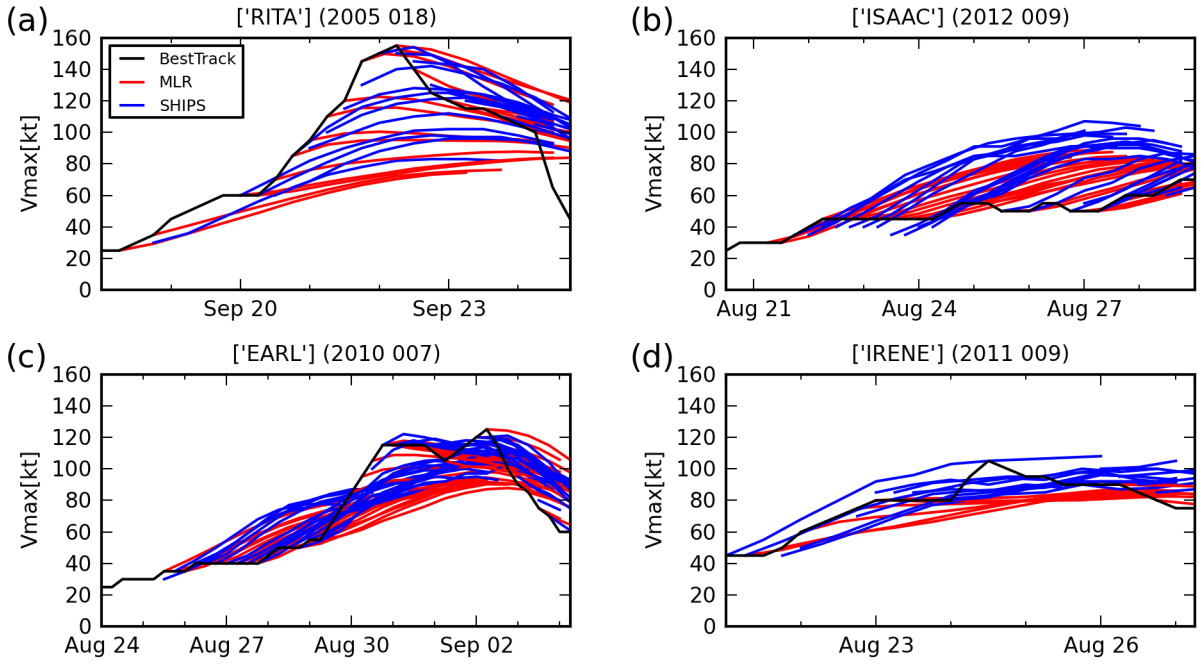

FIG. 3. Five-days intensity forecasts made every 12 hours by MLR (red) and SHIPS (blue) of (a) Rita, (b) Isaac, (c) Earl, and (d) Irene up to landfall. The black line shows the best-track intensity of each storm.

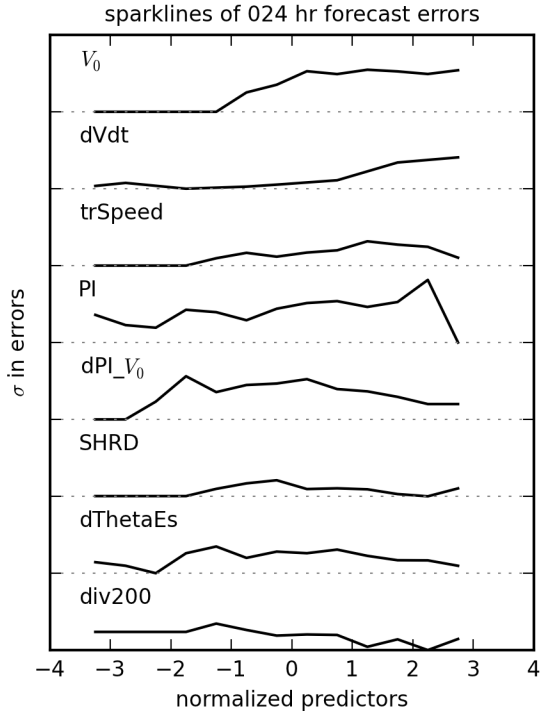

FIG. 4. A sparkline chart of the standard deviation (σ) of forecast errors for 24 hours forecasts. The σ is calculated with data that have been binned based on the magnitude of each individual predictor, and therefore the σ of errors is a function of the predictor here. The unit of the x-axis is one standard deviation of the individual predictor.

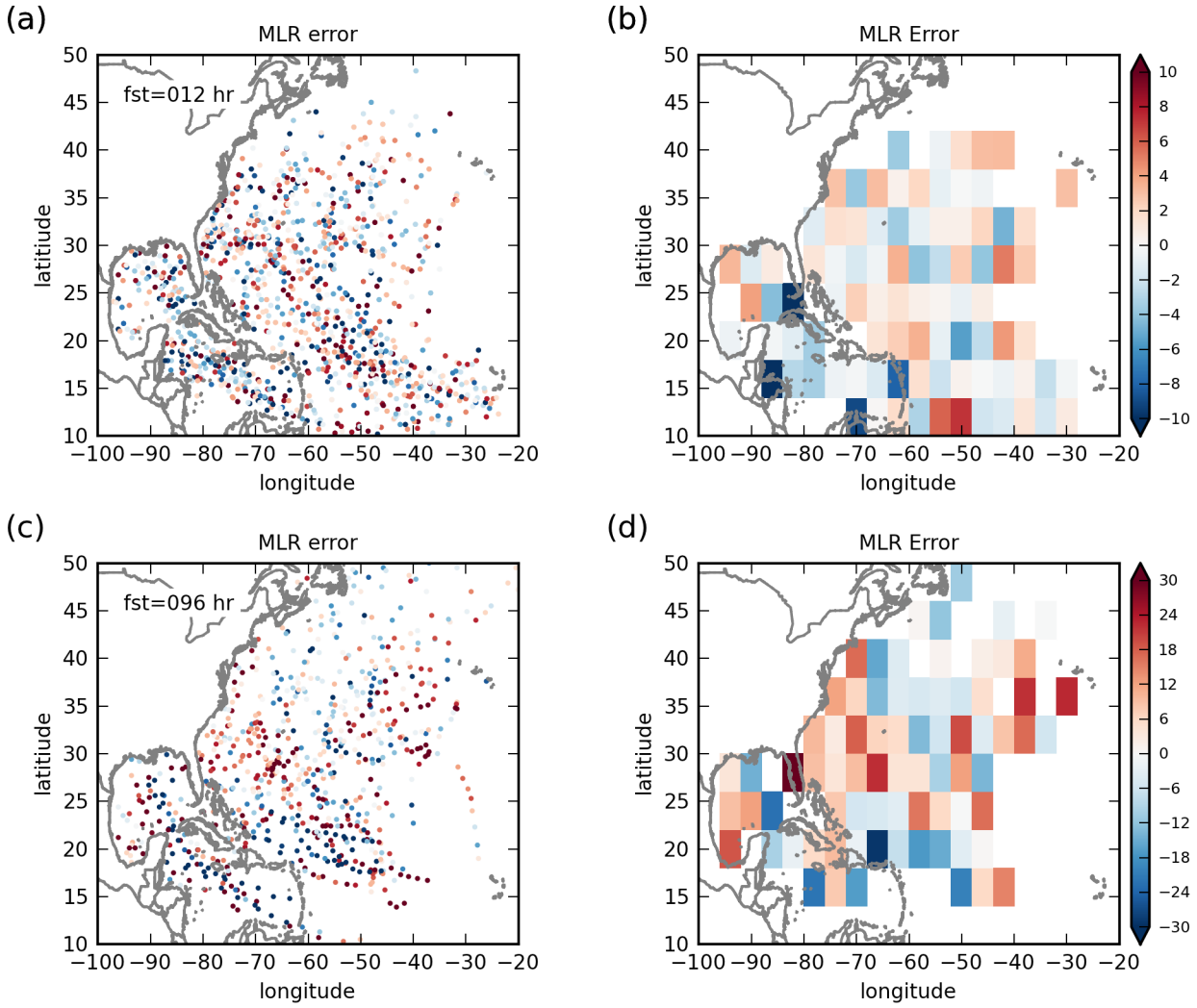

FIG. 5. (a) Horizontal map of storm locations of all 12 hours forecasts, where the color indicates the intensity prediction errors from the MLR. (b) Horizontal map of the MLR errors from (a). Each tile represents a $4^o \times 4^o$ box; the color indicates the mean error averaged over each tile. We show only the boxes with more then 5 data points. (c) and (d) are similar to (a) and (b), but for the 96 hours intensity prediction errors.

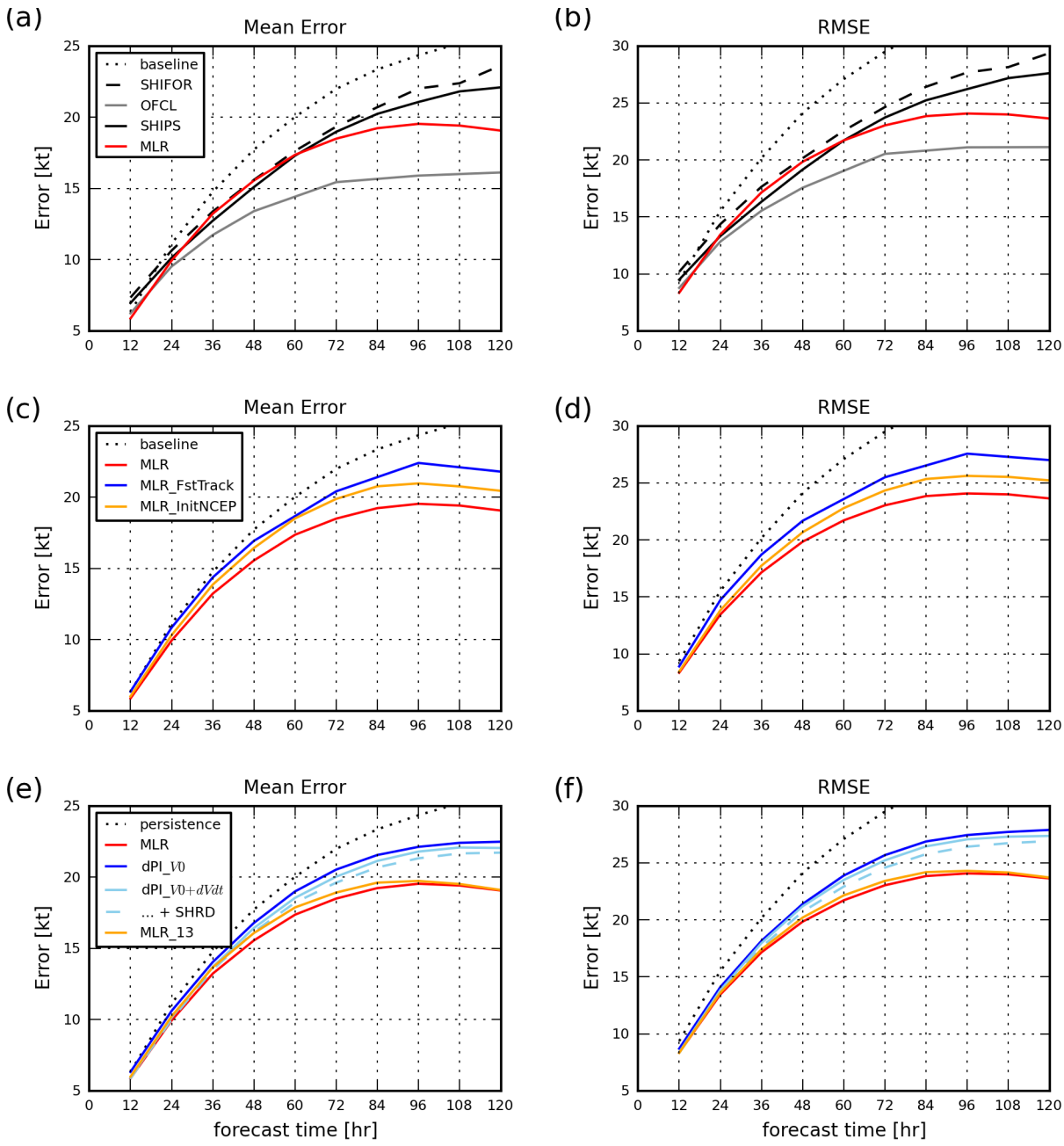

FIG. 6. (a) Mean absolute error (MAE) of intensity prediction from the persistence (baseline, black-dotted line), SHIFOR (black-dashed line), NHC (OFCL, gray-solid line), SHIPS (black-solid line), and MLR forecasts. (c) and (e) are similar to (a), but for different sets of experiments, which are described in Table 1. (b), (d) and (f) are the same as (a), (c), and (e), but for RMSE.

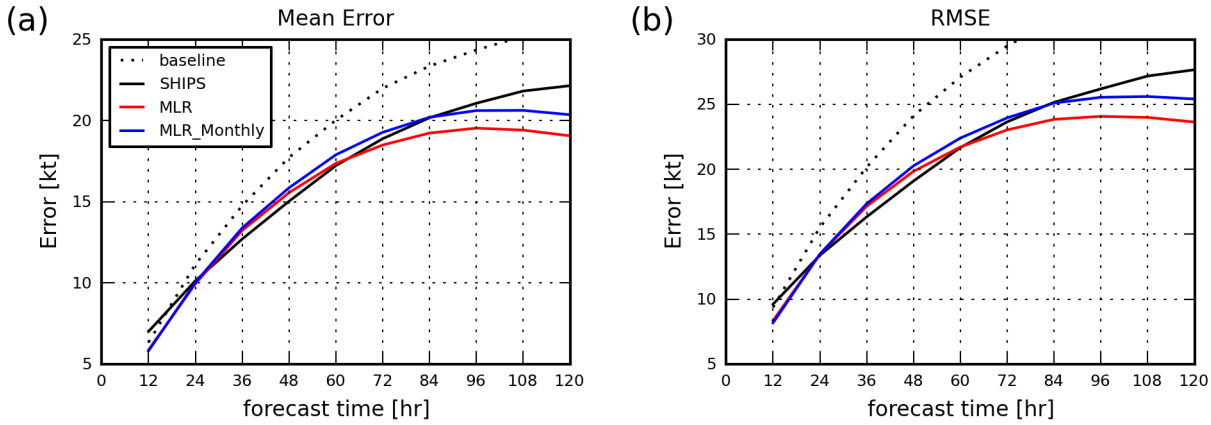

FIG. 7. (a) MAE of intensity prediction from the persistence (baseline, black-dotted line), SHIPS (black-solid line), MLR (MLR with daily data, red), and MLR_Monthly (MLR with monthly data, blue)

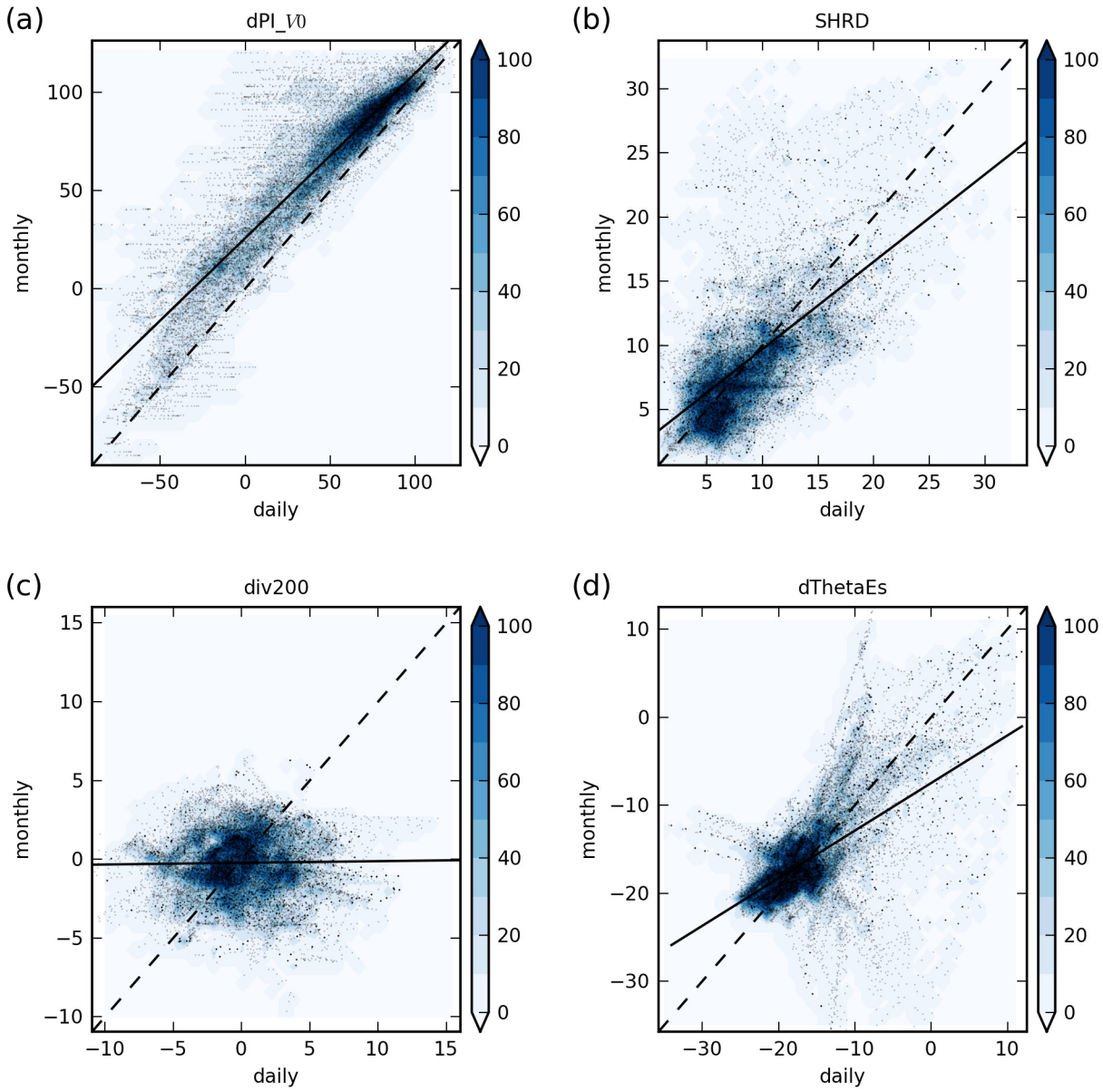

FIG. 8. Scatter plots overlaid on two-dimensional histograms of daily (x-axis) and monthly (y-axis) environmental predictors used in MLR: (a) the difference between PI and initial maximum wind speed: dPI_V_0, (b) the vertical wind shear: SHRD, (c) the 200 hPa divergence: div200, and (d) the conditional stability: dThetaEs. The black-solid lines indicate the best-fit lines while the black-dashed lines are the one-to-one lines.

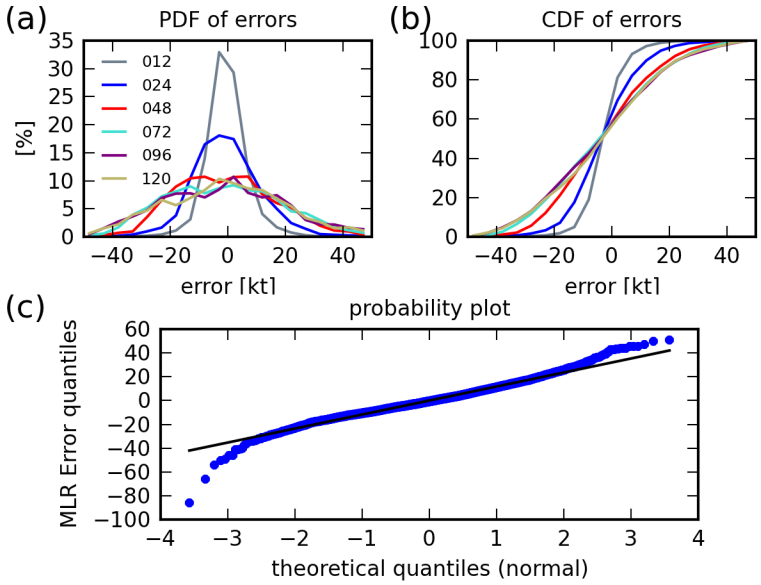

FIG. 9. (a) Probability density function (PDF) of the MLR forecast errors from all the training data. (b) Similar to (a) but for cumulative density function (CDF). In (a) and (b), colors indicate various forecast times as listed on the legend. (c) Quantile-Quantile plot showing the fit of the MLR errors for 24 hours forecast (from training data) to the normal distribution. The quantiles of the MLR errosr are shown on y-axis while the quantiles of the normal distribution are on the x-axis. The unit of x-axis is one standard deviation of MLR errors.

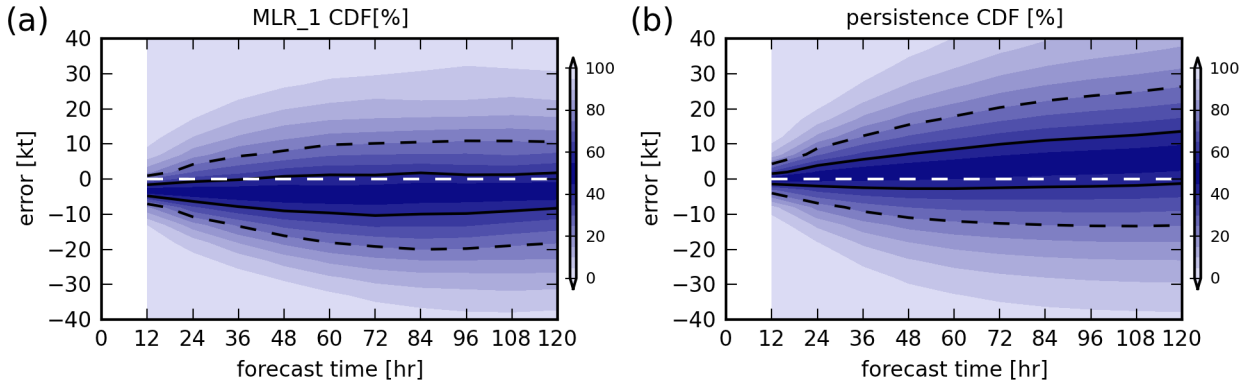

FIG. 10. Joint CDF (%) of the magnitude of errors and lead time from (a) MLR_1, and (b) the probabilistic persistence model based on training data. In both (b) and (a), the white-dashed lines indicate where the mean error is, and the black solid and dashed lines show where the highest 10% and 50% probabilities are.

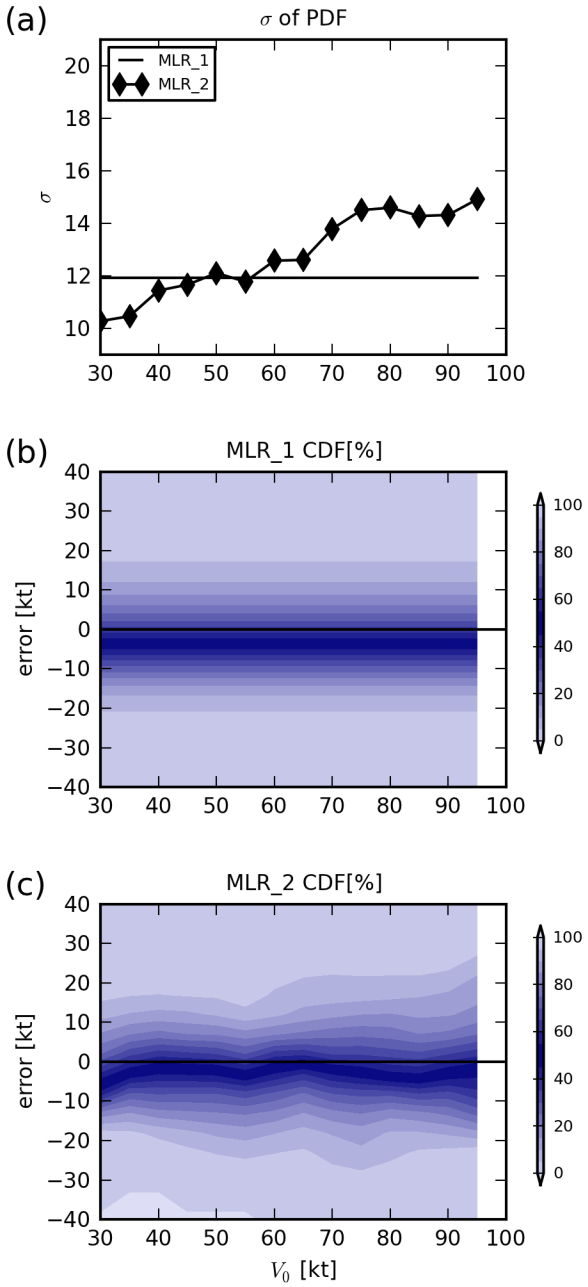

FIG. 11. (a) Standard deviation (σ) of error distribution for 24 hours intensity predictions from the training data. Black-dashed line is calculated with all the errors (MLR_1), while black-diamond line is calculated with errors that have binned based on the initial storm intensity (V_0) (MLR_2). (b) Joint CDF (%) from MLR_1. Note that the CDF is not a function of V_0. (c) Similar to (b), but now the CDF is is calculated from MLR_2, and is also a function of V_0.

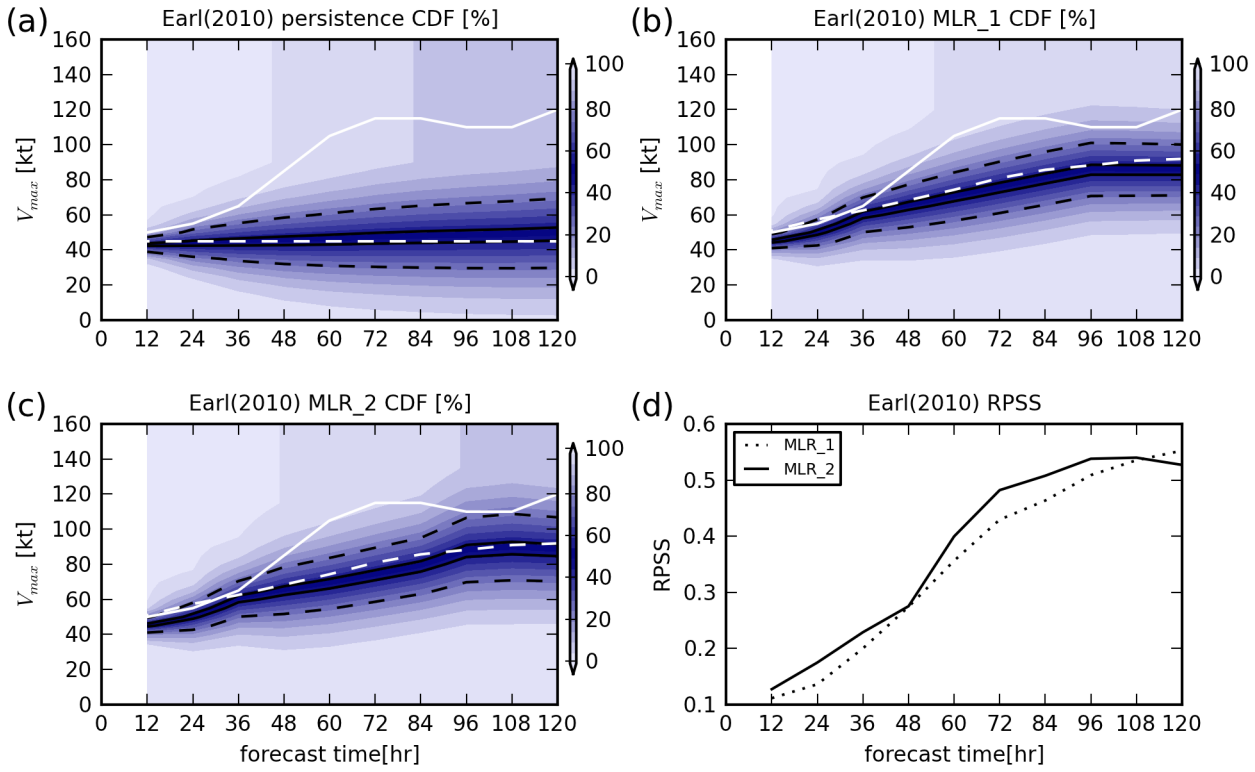

FIG. 12. (a) Intensity probabilistic prediction in CDF for Hurricane Earl (2010) based on the probabilistic persistence model (Fig. 10b). (b) Same as (a), but the CDF is based on MLR_1 (Fig. 10a and Fig. 11b). (c) Similar to (b), but the CDF is based on MLR_2 (Fig. 11c). The initial time for (a-c) is 0000 UTC 28 August, 2010. The white-solid lines indicate the observed maximum wind speed, while the white-dashed lines are the MLR and the persistence model predicted intensity, which is also where the mean error is in Fig. 10. Black solid and dashed lines show where the highest 10% and 50% probabilities are. (d) Ranked probability skill score (RPSS) of the two MLR probabilistic predictions relative to those from the probabilistic persistence model for all forecasts in Earl's life cycle.

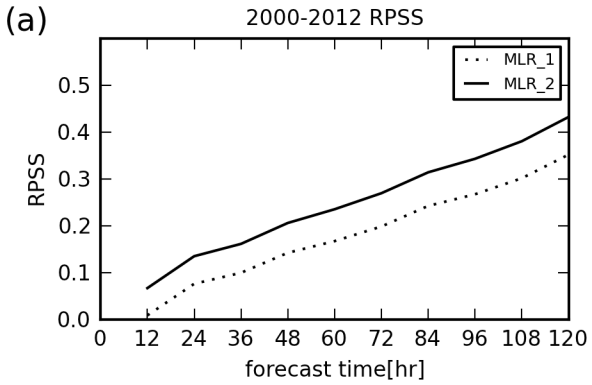

FIG. 13. Rank probability skill score (RPSS) of the two MLR probabilistic predictions relative to those from probabilistic persistence model for all testing cases.

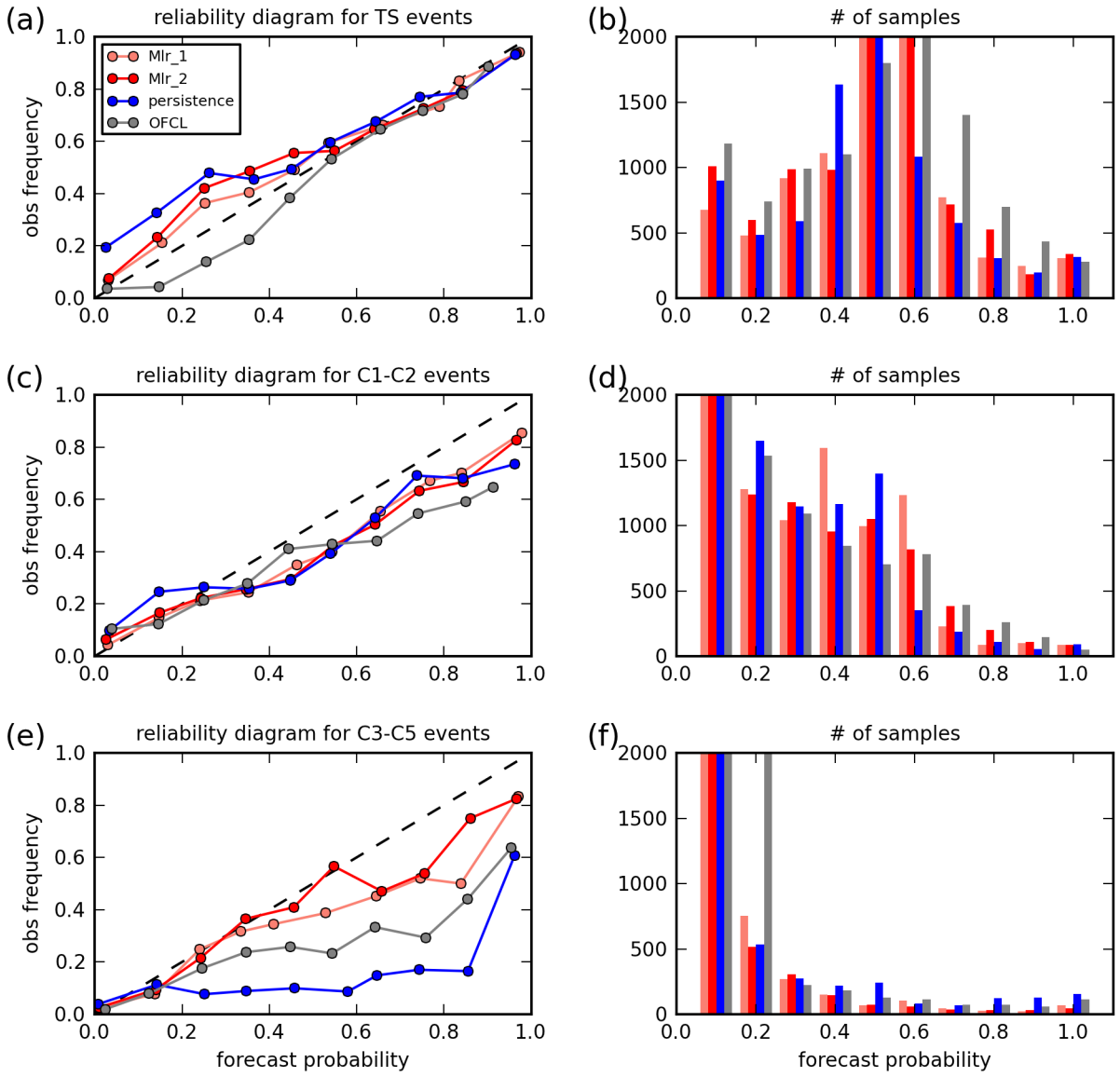

FIG. 14. Reliability diagrams(a, c, e) showing observed frequency as a function of forecast probabilities for (a) tropical storm (TS), (b) category 1-2 hurricane, and (c) category 3-5 hurricane, respectively. Colors indicates forecasts from MLR_1(pink), MLR_2 (red), persistence (blue) and NHC (OFCL, gray) from 2008 - 2012. (b), (d), and (f) show the sample size used in (a), (c), (e) for each forecast.