

## Global assessment of tropical cyclones intensity statistical-dynamical hindcasts

S. Neetu<sup>1</sup>, M. Lengaigne<sup>2,3</sup>, H. B. Menon<sup>4</sup>, J. Vialard<sup>2</sup>, M. Mangeas<sup>6</sup>, C.E. Menkes<sup>2,6</sup>, M.M. Ali<sup>5</sup>,  
I. Suresh<sup>1</sup>, J.A. Knaff<sup>7</sup>

<sup>1</sup> CSIR-National Institute of Oceanography (NIO), Goa, India

<sup>2</sup> Sorbonne Universités (UPMC, Univ Paris 06)-CNRS-IRD-MNHN,  
LOCEAN Laboratory, IPSL, Paris, France

<sup>3</sup> Indo-French Cell for Water Sciences, IISc-NIO-IITM-IRD  
Joint International Laboratory, NIO, Goa, India

<sup>4</sup> Goa University, Goa, India

<sup>5</sup> Florida State University, Tallahassee, USA.

<sup>6</sup> IRD, BP A5, Nouméa Cedex 98848, New Caledonia

<sup>7</sup> NOAA/NESDIS, Center for Satellite Research and Applications,  
Fort Collins, Colorado, USA

Submitted to Quarterly Journal of the Royal Meteorological Society

Corresponding author:

S. Neetu

Physical Oceanography Division  
National Institute Of Oceanography

Dona Paula - Goa - 403004 India

Email: [neetu@nio.org](mailto:neetu@nio.org)

### Abstract:

This paper assesses the characteristics of linear statistical models developed for Tropical Cyclones (TC) intensity prediction at global scale. To that end, multi-linear regression models are developed separately for each TC-prone basin to estimate the intensification of a TC given its initial characteristics and environmental parameters along its track. We use identical large-scale environmental parameters in all basins, derived from a 1979-2012 reanalysis product. The resulting models display comparable skill to previously-described similar hindcast schemes. Although the resulting mean absolute errors are rather similar in all basins, the models beat persistence by 20-40% in most basins, except in the north Atlantic and northern Indian Ocean, where the skill gain is weaker (10-25%). A large fraction (60 to 80%) of the skill gain arises from the TC characteristics (intensity and its rate of change) at the beginning of the forecast. Vertical shear followed by the maximum potential intensity are the environmental parameters that yield most skill globally, but with individual contributions that strongly depend on the basin. Hindcast models built from environmental predictors calculated from their seasonal climatology perform almost as well as when using real-time values. This has the potential to considerably simplify the implementation of operational forecasts in such models. Finally, these models perform poorly to predict intensity changes for Category 2 and weaker TC, while they are 2 to 4 times more skilful for the strongest TC (Category 3 and above). This suggests that these linear models do not properly capture the processes controlling the early stages of TC intensification.

### Keywords:

Tropical cyclone, Statistical model, Multiple linear regression, Intensity forecast, Atmospheric predictors.

## Introduction

Tropical Cyclones (hereafter TC) are amongst the most devastating natural phenomena and can have drastic socio-economic consequences in coastal regions around the globe (McPhaden et al. 2009; Webster et al. 2008). There are six major tropical cyclogenesis areas in the world (Fig. 1): the North Western Pacific (NWP) is the most active cyclone region (~31% of the total number of cyclones at the global scale) followed by the North Atlantic (ATL; ~22%), the North Eastern Pacific (NEP; ~17%), the Southern Indian Ocean (SIO; ~13%) and South Western Pacific (SWP; ~10%). The North Indian Ocean (NIO) is least active cyclone region (~6%) but is the basin where TC cause largest casualties (Longshore 2008).

Accurate forecasts of TC track and intensity are very important pre-requisites for a proper risk management. Tremendous progress has been made in forecasting TC tracks in the past two decades, leading to actual five-day track forecasts as accurate as three-day track forecasts a decade ago (Elsberry et al. 2013). In contrast, intensity forecasts have only modestly improved mainly for long lead-times (i.e., 3 to 5 days lead), the improvement rates being only one third of those for tracks (DeMaria et al. 2014). The intensity errors are indeed still large, reaching ~3-5kt at 12hr lead-time to 20-25kt for 120hr lead-time (Cangialosi and Franklin 2014; DeMaria et al. 2014). Along with dynamical models, the so-called “statistical-dynamical models” are routinely used for predicting TC tracks and intensity. In a similar way to the early TC intensity statistical forecast schemes (e.g. Neumann 1972; Jarvinen and Neumann 1979; Merrill 1980; Chu 1994; Aberson 1998), these statistical-dynamical models use linear statistical regression techniques to predict intensity changes from predictors derived from climatology and persistence of TC characteristics such as their current intensity and its time-derivative. They do however also include large-scale environmental parameters along the cyclone track as additional predictors. The environmental parameters along the cyclone tracks (such as vertical wind shear) are usually obtained from a dynamical model forecast, hence the “statistical-dynamical” name. Using statistical-dynamical models to predict TC intensity forecasts is competitive with and often more skilful than using dynamical models (Kucas 2010; DeMaria et al. 2007, 2014), with a considerably lower computational cost and far quicker run-time. Historically, different operational statistical-dynamical prediction schemes have been developed separately for each basin. The first operational statistical–dynamical model, the Statistical Hurricane Intensity Prediction Scheme (SHIPS model), was developed to forecast TC intensity in the Atlantic basin (DeMaria and Kaplan 1994a). It has undergone considerable changes since its introduction in 1991 (e.g., Knaff et al. 2005), including a version implemented for the Eastern Pacific (DeMaria and Kaplan 1999). A logistic growth equation model (LGEM; DeMaria et al. 2009) with the same inputs

as SHIPS is also being used for ATL and NEP since 2006. The statistical typhoon intensity prediction scheme (STIPS; Knaff et al. 2005) was then developed in 2002 for the Northwestern Pacific region. It was further adapted for the southern hemisphere in 2005 (combining the SWP and SIO regions) and referred to as the southern hemisphere statistical typhoon intensity prediction scheme (SH-STIPS model; Knaff and Sampson 2009). A model based on STIPS has also been developed for the NIO (Kotal et al. 2008).

Several large-scale environmental parameters are commonly used as predictors in these statistical-dynamical schemes, as they are known to influence the TC intensification. These parameters include the Maximum Potential Intensity (MPI) that a TC can reach, which increases with sea surface temperature (Merrill 1988; DeMaria and Kaplan 1994b; Webster et al. 2005), the mid-tropospheric relative humidity that favours TC intensification through its influence on convective buoyancy through entrainment of sub-saturated air (Emanuel et al. 2004) or the vertical wind shear that can inhibit TC intensification (Gray 1968; DeMaria 1996). Other environmental predictors in these statistical-dynamical models also include, amongst others, low-level vorticity and equivalent potential temperature as well as upper-level air temperature (Emanuel 2007; Knaff et al. 2005; DeMaria et al. 2005). Because the current statistical-dynamical models have been developed individually for each basin, the selected large-scale environmental parameters used as predictors and the database from which these variables are extracted generally differ from one basin to another. It is also important to mention that, in real-time application, these environmental predictors are affected by errors in the forecasted location (DeMaria 2010, Tien et al. 2013).

Very few studies attempted to compare the predictive skill of statistical-dynamical forecast models across different TC-prone regions. A recent study (DeMaria et al. 2014) provided a comparison of the intensity error of these models over the recent period as a function of the basin and offered an update to DeMaria et al. (2007). Their results indicate that these errors are generally larger for the NWP and southern hemisphere TC (~12kt at 24h lead-time and ~22kt at 120h lead-time) than for ATL and NEP TC (~10kt at 24h lead-time and ~15kt at 120h lead-time). This basin-wise skill dependency may however be partly related to the differences in the predictors and in datasets used in each of regional statistical schemes. The analysis performed by Lee et al. (2016) addresses this issue. They indeed developed a single global statistical model for all TC prone basins using consistent predictors derived from a single atmospheric dataset over the same period. They indeed show that a single model trained at global scale is almost as skilful as basin-wised trained models: while most basins exhibit very similar skills, a basin-wise training performs better for the NIO at long lead-time (and NEP to a lesser extend) presumably because it recognizes the short TC lifetime in this peculiar

basin. Their results also indicate a more modest basin-wise dependence than DeMaria et al. (2014). In line with De Maria et al. (2014), TC prediction for the NWP and the southern hemisphere still displays the largest errors at long lead-time (~23-25kt at 120h lead-time), while these errors are smaller in ATL (~19kt at 120h lead-time). The reasons behind this basin-wise skill dependency is however currently unknown. These differences may be related to the different TC intensity distribution in each of the basin: more intense TC in the NWP may for instance yield to larger forecast errors. This dependency could also simply be explained by differences in intrinsic TC intensity predictability in each basin.

Little is also known about the relative TC intensity statistical-dynamical hindcast skill yielded by various predictors. Several studies (DeMaria and Kaplan 1994a, 1999; Knaff et al. 2005, 2009; DeMaria et al. 2014; Lee et al. 2015) acknowledged the key contribution arising from the storm climatology and persistence predictors in most basin-wise prediction schemes, the inclusion of large-scale environmental parameters generally resulting in a modest skill improvement when compared to climatology and persistence-based models. Only a few studies provide a quantitative assessment of the respective influence of the various large-scale environmental predictors. A recent study discussing the case of the ATL basin (Lee et al. 2015) demonstrated that the initial TC intensity change is the most important predictor at short-time lead (<24h) but that MPI is the environmental parameter that yields most improvement at longer lead-times. In contrast, Sharma et al. (2013) show that the initial storm intensity is the most important parameter while the initial intensity time-derivative is not an important predictor for the NWP basin. Results for these two basins are however very difficult to compare, as the statistical model architecture, considered period and environmental datasets and predictors all differ.

Aside from the relative importance of predictors, very little is also known about the TC intensity statistical-dynamical hindcasts skill as a function of the TC characteristics, including their intensity. This assessment may however be very useful as it would allow us to better identify at which TC stage these models are less efficient and hence where a dedicated effort needs to be undertaken to improve the skill of these models.

A thorough assessment of TC intensity statistical hindcasts models is hence still lacking. The objective of this study is therefore to identify the most important predictors in TC intensity statistical models at global scale as well as to assess the sensitivity of the model skill to TC intensity. This endeavour requires a common framework for a meaningful comparison of the results between each basin. As detailed in Section 2, we hence developed statistical-dynamical TC intensity hindcast schemes for each basin using consistent predictors derived from a single atmospheric dataset over the

same period. In Section 3, we discuss the basin-wise skill differences and the respective influences of the predictors on the model skill. We will also investigate the impact of environmental predictors calculated along the forecast track versus those calculated using climatology as well as stratify model performance by initial intensities. A summary and discussion of the results is provided in Section 4.

## **2. Basin-wise statistical intensity forecast models**

### **2.1. Datasets**

Tropical cyclones location and wind intensity are derived from the International Best Track Archive for Climate Stewardship (IBTrACS; Knapp et al. 2010). IBTrACS combines the best track data from many agencies into one common format. We use the Joint Typhoon Warning Center (JTWC) data as our main database as it uses a consistent methodology for data preparation in every region (NWP, NEP, SWP, NIO and SIO basins), except in the ATL where JTWC does not operate. For this region, we use the National Hurricane Center (NHC) data. Both JTWC and NHC use 1-min average for sustained wind speed. As acknowledged in the literature (e.g. DeMaria and Kaplan 1994; Knaff et al. 2005), the statistical properties of storms over land are different from those over the ocean and specific empirical inland decay models are used for predicting the intensity change for TC after landfall. Here, we only consider TC located over the ocean (i.e. all TC track points over land are excluded from the present analysis). One caveat of these best track data is that their intensity is generally derived using satellite-based methods, such as the Dvorak technique (Dvorak 1984). The intensity archived in these datasets are estimated to the nearest 5 kt at 6-h intervals. For this reason, model formulation as well as any discussion of intensity in this paper will be discussed in knots rather than in meters per second.

We extract synoptic atmospheric conditions along the cyclone best tracks from the European Centre for Medium-range Weather Forecasts (ECMWF) Interim Re-Analysis (ERA- Interim; Dee et al. 2011) dataset. We use 34 years (1979-2012 period) of 6-hourly atmospheric ERA-I reanalysis fields at  $0.75^{\circ} \times 0.75^{\circ}$  horizontal resolution.

### **2.2. Model development**

Table 1 lists the predictors used in our TC intensity statistical hindcast schemes. These parameters are adapted from Knaff et al. (2005, 2009) and Sharma et al. (2013). Three variables account for the TC characteristics at the beginning of the forecast: the intensity (VMAX), the intensity squared (VMAX<sup>2</sup>) and the intensity change over the previous 12-hours also referred as persistence (PER). We have constructed a first set of basin-wise hindcast models that only uses those predictors (i.e. predictors 1-3 from Table 1: the initial intensity, its square, and the intensity change

12 hours prior to the start of the forecast), which we will refer to as the “baseline” models (Table 2). But we have also constructed models that also take the effects of environmental parameters along the cyclone tracks into account, referred to as Atm models (Table 2).

The selected large-scale environmental parameters are commonly used in the statistical-dynamical models and are derived from the parameters list proposed by Sharma et al. (2013) for the NWP region and by Knaff et al. (2009) for the southern hemisphere. They have been computed using the 6-hourly ERA-Interim dataset and averaged in time along the TC track from the initial time till the forecast hour. They include the maximum potential intensity (MPI), a theoretical upper bound for the cyclone intensity that increases with sea surface temperature (Merrill 1987; Miller 1958; DeMaria and Kaplan 1994b; Emanuel 1988; Emanuel and Nolan 2004; Holland 1997; Webster et al. 2005). We have used empirical formulae to estimate the MPI in each basin rather than the analytical formulation derived from thermodynamic principles by Emanuel (1986, 1995) because these empirical formulae are those currently used in operational statistical TC intensity forecast models. In addition, it appears that this analytical formulation results in an observed VMAX that exceeds the analytical MPI for ~10% of the TC cases (J. Knaff, personal communication). MPI is hence parameterized using an exponential function of sea surface temperature in the NWP, SWP, SIO and ATL regions, and a linear relationship in the NEP and NIO basins (DeMaria et al. 2005; Knaff et al. 2005; Knaff et al. 2009; Kotal et al. 2009). These empirical relationships and the corresponding references are listed in Table 3. The MPI is computed using 6-hourly sea surface temperature fields from ERA-Interim dataset and its maximum value set to 185kt. MPI and  $MPI^2$  are both used as predictors in our model as it is the case in most other statistical-dynamical forecast models (see Table 1). The 200-850 hPa wind shear magnitude (SHRD) and 200-850 hPa zonal wind shear magnitude (USHR) are used to account for the destructive effect of vertical shear on TC. These variables (SHRD and USHR) are averaged over an annular region within 200 to 800 km from the centre of the storm. Mid-tropospheric (i.e. average 300-500 hPa) relative humidity, temperature at 200 hPa (T200), and low-level (925 hPa) equivalent potential temperature (E925), all averaged within 200-800 km of the TC, are also used as predictors (see Table 1). Low-level relative vorticity (at 850 hPa, Z850) is also used as a predictor, but area-averaged within 1000 km from the centre of the storm following Knaff et al. (2005, 2009).

Two cross-terms, VMAX multiplied by MPI (VMXM) and VMAX multiplied by SHRD (VMXS), are also tested. Other previously-investigated cross-terms did not result in a significant TC intensity forecast improvement (DeMaria et al. 2005; Knaff et al. 2005) and are not discussed further. In line with Sharma et al. (2013) but in contrast to Knaff et al. (2005, 2009), we do not

include the storm translation speed and pressure of the initial steering motion of the wind, as it resulted in marginal improvement of our forecasts (less than 1% skill improvement). We do not use any oceanic-derived metric as predictor in the model development because the short length over which reliable oceanic datasets are available (daily ocean-reanalysis dataset are only available from 1993 onwards) would result in a considerable reduction of our training period. The primary aim of this work is indeed not to build an improved version of the existing statistical models, but rather to evaluate the importance of key predictors and regional skill variations, using a common framework. It must however be noticed that the inclusion of additional predictors does not affect the general conclusions derived from the present study.

The distributions of five of the main predictors used in the present study are shown in Figure 2, separately for each TC-prone basin. The variables that account for the TC characteristics at the beginning of the forecast (VMAX and PER; Fig. 2a, b) share a rather similar distribution in all TC-prone basins. The NWP region is however the region with the most intense TC (with an average of  $61 \text{ m.s}^{-1}$  for the highest tenth percentile in this basin; see dashed vertical lines on Fig. 2a). In contrast, the NIO displays the weakest TC. The four other basins (NEP, SWP, SIO and ATL) generally share similar TC intensity distribution (Fig. 2a). The environmental parameters distribution is more basin-dependent. For instance, the MPI distribution (Fig. 2d) exhibits its largest values in basins where the warmest surface temperature are found (NIO and NWP where their averaged MPI value exceed 150 kt; see dashed vertical lines on Fig. 2d). In contrast, MPI values are far weaker in the ATL basin, where its distribution hardly overlaps those in the NIO and NWP basin, with an averaged value lying around 60 kt. While these distributions are more similar for SHRD and RHHI environmental parameters (Fig. 2c,e), they significantly differ from one basin to another, RHHI being for instance generally lower for the ATL basin and larger for the NWP, SWP and NIO basins.

A proper interpretation of the regression coefficients arising from these statistical models and a proper assessment of the relative importance of each predictor require investigating the correlations existing between the different predictors. Table 4 displays such a cross-correlation matrix for all predictors taken at 60h (other lead-times give qualitatively similar results). VMAX and PER are only weakly correlated with other predictors, indicating that these parameters are relatively independent (except for VMAX and  $\text{VMAX}^2$  which are obviously correlated). The MPI, USHR, RHHI, E925 and MPI2 environmental predictor are generally correlated with each other, with correlations exceeding  $\pm 0.5$ . In contrast, SHRD, T200 and Z850 appear to be more independent. This implies that interpreting the values of the statistical model regression coefficients in terms of relative importance of various predictors is not straightforward, due to cross-correlations between most predictors.

In our statistical scheme, the dependent variable or the predictand is the change in intensity (DELV) between the start of the forecast and the considered lead-time (for lead times at 12 hours intervals, up to 120 hours). The predictors are the TC initial characteristics (predictors 1-3 in Table 1) and large-scale environmental parameters (predictors 4-11 in Table 1) described in the previous section. We then construct a statistical model using a multiple linear regression (MLR) technique for each TC-prone basin. A MLR model is built separately for each of the forecast lead-time, 12, 24, ..., 120 hr. Eighty percent of the TC are used for building the model (training dataset) and the remaining twenty percent are used for testing the model performance. The training and testing datasets are chosen randomly but for entire TC tracks (i.e. all the values along a TC track are either included in the training or in the testing dataset). This avoids overfitting that could result from using data from the same TC in both the training and testing datasets. For more reliable results, each regional model fit was estimated 50 times for each forecast lead-time, using 50 different randomly-selected training and testing sets. The reported values are averaged over the 50 realizations. The 90% confidence interval on these values is derived using a bootstrap technique. To do so, we compute the average of 50 randomly-selected values from the 50 realizations (re-sampling is allowed). By repeating this process 1,000 times, we obtain 1,000 estimated mean values. The 90% confidence interval on the mean value is obtained from the 5% and 95% percentiles of this distribution. This procedure is applied separately for each TC basin (NWP, NEP, SWP, NIO, SIP and ATL), using the same predictors and datasets (except for ATLTC characteristics for which NHC TC characteristics are used in place of JTWC), allowing a fair comparison of results across TC-prone basins.

We choose commonly used metrics to evaluate the model performance in order to compare our model skill with previously published results. These metrics include the mean absolute error (MAE; Lee et al. 2015; Knaff et al. 2005, 2009; Sharma et al. 2013) and the percentage of skill improvement compared to a given reference model (DeMaria et al. 2007). The MAE is defined as the mean of absolute values of the difference between predicted and observed TC intensity at each forecast hour. Consistently with DeMaria et al. (2007), the skill improvement is computed as follows,:

$$\text{Skill (\%)} = 100 * (\text{MAE}_{\text{ref}} - \text{MAE}_{\text{Model}}) / \text{MAE}_{\text{ref}} \quad (1)$$

where  $\text{MAE}_{\text{ref}}$  is the MAE derived from a reference model and  $\text{MAE}_{\text{Model}}$  is the MAE from the model being evaluated. In this paper, we will use different choices for the reference model. For example, we can use a reference model that neglects a given environmental predictor to assess its importance (Atm-VarN sensitivity experiments summarized in Table 2 and detailed later in the paper). We can also use “persistence” as a baseline model (where persistence is the simplest model of all, that simply assumes that DELV won’t change from its value at the forecast time).



In addition to the parameters used in our scheme, Knaff et al. (2005) also used two cross terms, namely VMXS and VMXM (also listed in Table 1). These terms make the assessment of the respective importance of the predictors involved in the cross-terms more complicated. To investigate whether they can be ignored, a sensitivity experiment was conducted to ascertain the importance of these terms. This model (referred to as Atm+Cross, see Table 2) is similar to our reference Atm model except that the two cross terms (VMXS and VMXM) are included in the predictor's list. The statistics shown in Figure 3 (MAE and skill improvement of Atm+Cross relative to Atm, computed as in Eqn. (1)) and in the following analyses are calculated on the testing dataset of each experiment. As it will be illustrated later in Figure 5, performing the same diagnostics separately on the testing or training datasets generally leads to similar results. In agreement with past literature, Figure 3a first illustrates that the MAE generally increase with forecast lead-time: this increase is largest for forecast lead times between 12hr to 96hr and reduces beyond 96hr. When averaging results globally, the MAE of Atm experiments ranges from 5kt at 12h to 22kt at 120h (Fig. 3a, blue curve). As expected, the Atm models are systematically more skilful than persistence (Fig. 3a, black curve). The inclusion of cross-terms leads to a modest improvement of the model performance as measured by the MAE (Fig. 3a): this improvement is indeed marginal for long lead-times and slightly larger for shorter lead times (~1kt improvement between 24h to 48h). The percentage of skill improvement of Atm+Cross as compared to Atm is further provided in Figure 3b at three lead-time (24h, 60h and 120h) for each TC-prone basin and globally. It indicates that including these cross terms results in a 5% skill improvement at 24h, 2% at 60h and 1% 120h globally. The relative influence of these terms however varies from one basin to another, being slightly larger in the southern hemisphere and in the NEP and smaller in the NIO and NWP. Despite their non-negligible skill gain, we decided not to include VMXS and VMXM as predictors in our model as these cross terms prevent a simple quantitative assessment of the relative importance of the predictors involved in these cross-terms.

### 2.3. Model Performance

The Atm models skill is first evaluated by comparing basin-wise MAE of this model with similar results reported in the literature for the NWP (Knaff et al. 2005; their Table 6) and the combined results for the SWP and SIO (Knaff et al. 2009; their Table 5) for the training dataset and for the ATL on the testing dataset (Lee et al. 2015; their Figure 5). To the author's knowledge, such error estimates are not available for NEP and NIO in the peer-reviewed literature. As displayed in Figure 4, our model skill generally compares favourably with previously published results. For the NWP region, our model is slightly less skilful than the STIPS model initially developed by Knaff et al. (2005), especially for lead-times > 48h, for which our MAE is larger by ~1-2kt (Fig. 4a). Our

model performs slightly better for the SIO and SWP regions as compared to the SH-STIPS model proposed by Knaff and Sampson (2009), with a MAE reduction up to  $\sim 2\text{-}3\text{kt}$  for long lead-time (Fig. 4b). Finally, our model performs very similarly to the recent model proposed by Lee et al. (2015) for the ATL region, our MAE being slightly weaker by  $1\text{-}2\text{kt}$  for mid-range lead-times (Fig. 4c). The modest differences between the skill of the model proposed in the present study and the skill from the models detailed in the past literature demonstrate that our model includes the most important predictors for an accurate TC prediction in each basin, giving us confidence in the robustness of the results discussed in the next section. These modest differences could arise from different sources: the choice of predictors, the source of the atmospheric data and TC database as well as the different training/testing periods. The sensitivity to the choice of the period is further discussed below.

Statistical-dynamical models presented by Knaff et al. (2005, 2009) are built on a rather short period ( $\sim 7$  years) as compared to our Atm experiment that is built from a dataset spanning 34 years. These may partly explain the skill differences seen in Figure 4ab. To investigate this further, Figure 5 compares the performance of our statistical hindcast scheme built based on a 34 years long period (27 years for training and 7 years for testing; right panels in Figure 5) to that obtained when using a 7 years period ( $\sim 5.5$  years for training and  $\sim 1.5$  years for testing; left panels in Figure 5). Using a 7 years dataset results in far weaker training errors than testing errors, especially in the southern hemisphere and ATL basin ( $\sim 5\text{kt}$  for SWP+SIO basin, and  $\sim 3\text{kt}$  for ATL at long lead-time and only  $\sim 1\text{kt}$  for WP). In contrast, models built over a longer 34-year period exhibit far weaker skill differences between the training and testing datasets (less than  $1\text{kt}$  in all the basins). These differences can be attributed to the tendency of models trained with relatively small datasets to overfit the data, which leads to accurate training skill but degraded testing skill. When using a 7-year dataset, the training performance indeed diverge more from the testing performance for basins with a lower TC density (e.g. SIO+SWP and ATL) relative to the most active TC-prone basin (NWP), for which the larger number of cyclones yields a larger training sample. As Knaff et al. (2005, 2009) provided the MAE from their training dataset of  $\sim 5$  years long, a fair comparison with our scheme would be to compare these results with our training results for the shorter training period (blue curves on the left panels of Figure 5). Doing so, although slightly less accurate, our results comes closer to Knaff et al. (2005) for the NWP basin and outperforms the results of Knaff et al. (2009) for the southern hemisphere even more than in Figure 4b.

The data length also considerably influences the stability of the skill: using a smaller dataset systematically leads to a larger spread of the model performance between the 50 models built based on random selections of the training/testing datasets (shadings of the left panels of Figure 5) relative

to bigger datasets (shadings of the left panels of Figure 5). For the southern hemisphere, the MAE on the testing dataset at 120 h forecast lead-time ranges between  $\sim 19$ kt and  $\sim 26$ kt when the model is built on a 7 years period while it ranges between  $\sim 19$ kt and  $\sim 23$ kt when the model is built for 34 years. These results illustrate the strong sensitivity of the model skill to the period used to build the model, demonstrating that building a model with larger datasets using a cross validation method will systematically result in more reliable and robust estimate of the skill. In the following sections, we will hence calculate the metrics used in the present study (MAE and skill improvement) for the entire 34 years on the testing average of the 50 randomly-selected realisations of the model. Since best track locations are used in these evaluations, the results are not influenced by track errors that occur with real-time applications of these models (DeMaria 2010, Tien et al. 2013).

### **3. Results**

#### **3.1. Basin-wise performance**

As discussed in the introduction, DeMaria et al. (2014; their Figure 4) and Lee et al. (2016; their Figure 1) provided a recent basin-wise evaluation of the statistical-dynamical models performances. DeMaria et al. (2014) showed that the NWP and the southern hemisphere basins generally exhibit larger MAE ( $\sim 22$ - $24$ kt at 120h lead-time) than the ATL basin ( $\sim 15$ kt at 120h lead-time). These differences may however be partly due to the different training/testing periods, predictors and atmospheric datasets used to derive the large-scale environmental parameters. The common framework allows addressing the basin-wise skill differences more thoroughly. As compared to DeMaria et al. (2014), Figure 6a indicates that the MAE derived from our common modelling framework only modestly differs from a basin to another: they systematically range between 8 and 11kt at 24h and between 20 and 24kt at 120h. The MAE values are however slightly larger for the NEP region at short time-lead and for the NWP region at long time-lead. In contrast, they are weaker for the NIO and SIO regions at short time-lead and for SWP at long time-lead. While results from DeMaria et al. (2014) indicate that errors are generally far larger for southern hemisphere TC than for NEP TC, it is not the case in our analysis, implying that the differences discussed in DeMaria et al. (2014) are hence likely attributable to the differences in the modelling framework (the LGEM model used for ATL and NEP regions includes the same inputs as SHIPS but utilizes a more sophisticated prediction equation; see DeMaria et al. 2009) rather than to the inherent predictability of TC in these basins. In contrast, the larger MAE in the NWP region as compared to ATL basin, also noticeable in DeMaria et al. (2014), cannot be explained by differences in the model architecture and are hence likely to arise from the specific TC characteristics in this basin (e.g., the fact that this basin hosts generally stronger TC) or to the inherent predictability in this region. The

basin-wise differences of the errors derived from our models discussed above are generally consistent with the analysis derived by the common modelling framework also developed by Lee et al. (2016; their Figure 1).

Comparing the basin-wise models performances based on the percentage of improvement compared to simple persistence (Fig. 6c) provides a different picture from the comparison based on the MAE (Fig. 6a). Based on this metric, four basins (NWP, SWP, SIO and NEP) exhibit similarly high percentages of improvement relative to persistence, ranging from ~20 to 40% at lead-times between 48 and 120 hours. In contrast, the skill improvement compared to persistence is considerably weaker in the ATL and to a lesser extent in the NIO basin, where it ranges ranging between 15 and 25%. This finding qualitatively agrees with the analysis performed by Lee et al. (2016; their figure 1c). While the ATL model exhibits one of the weakest MAE, it also exhibits one of the weakest relative improvement compared to persistence (along with NIO). This apparent discrepancy can be explained by the performance of the persistence model in each basin. Figure 6b indeed shows that MAE from persistence is actually weaker for both the ATL and NIO basins, illustrating that TC in these basins undergo smaller intensity changes than in other basins. The NWP model exhibits one of the strongest MAE but displays one of the largest improvements relative to persistence (along with SWP). The persistence yields large MAE in the NWP (Figure 6b), likely due to intense TC in this basin that can hence undergo strong intensity changes (see Figure 2a). Similarly, the persistence model yields a weak MAE in the NIO, most likely because this basin on average exhibits the weakest TC (see Fig. 2a). This interpretation does not however hold for the ATL basin where TC intensity distribution is similar to that in the NEP and SIO basins (see Fig. 2a) but its persistence skill is larger than in these basins. It implies that, for similar TC intensity, the TC intensity is more steady than in other basins.

We further investigate differences in the basin-wise models by comparing the coefficients of the multiple linear regression at 60h lead for each basin for some key predictors in Figure 7. This figure first reveals that these regression coefficients are generally significantly different from zero (see error bars in Figure 7) and exhibit consistent signs and order of magnitude across basins, suggesting qualitatively similar influences of each predictor. For instance, PER regression coefficients are systematically positive, illustrating that an intensifying TC is likely to further intensify in the near future. Similarly, regression coefficients for RHHI are systematically positive, demonstrating the favourable impact of large mid-tropospheric humidity on TC intensification. In contrast, regression coefficients for SHRD are systematically negative, highlighting that a strong vertical shear is detrimental to TC intensification in all basins. Results related to MPI are more

difficult to interpret as they involve two predictors (MPI and  $MPI^2$ ) that are strongly correlated (see Table 4). Regression coefficients are generally negative for MPI, and positive for  $MPI^2$ , except for SIO where these two signs reversed. This change of sign could result from the strong correlation existing between these two parameters (MPI and  $MPI^2$ ; see Table 4) that could result in coefficient estimates that may drastically change from one basin to another. Because VMAX and  $VMAX^2$  are also strongly correlated, their regression coefficients are also difficult to interpret. Despite this overall consistency of the regression coefficients amongst the basins, the exact values of these coefficients vary significantly from one basin to another. While SHRD and PER regression values are rather close to each other in most basins, it is not the case for instance for RHHI whose regression coefficients can vary by an order of magnitude from one basin to another. As shown in Figure 2d, the rather different distribution of this parameter amongst the basins could explain the basin-wise differences in the regression coefficients, i.e. the influence of a given environmental parameter may differ depending the range of this parameter variation in each basin.

Figure 7 clearly indicates that the linear regression coefficients can vary from a basin to another, suggesting that it is better to construct TC intensity statistical hindcast schemes regionally than globally. To quantify this, we constructed a hindcast scheme that was trained collectively on all the basins (Glob model; see Table 2). The performance of the regional models relative to that of a global model is shown in Figure 8. As expected from the basin-wise disparity for some of the regression coefficients, regional models generally outperform the global model. However, when comparing these skills at global scale, the improvement arising from a regional training compared to a global one is rather modest, reaching respectively 2%, 3% and 1% at 24h, 60h and 108h (Fig. 8), suggesting that it may be possible to build an efficient statistical model for TC intensity forecast at global scale. These relative improvements however strongly vary from one basin to another: some regions such as SIO and SWP exhibit very similar skills when trained regionally or globally while the improvement gained from a regional training is larger for the NIO basin, especially at mid to long time-lead (up to 5%). A similar finding has already been reported in Lee et al. (2016) and has been attributed to the ability of regionally-trained model to recognize the short TC lifetime in this peculiar basin. As expected, the regression coefficients for each predictor derived from the global model generally falls within the range of those derived for the regional ones (Fig. 7).

### **3.2 Relative importance of predictors**

Following this basin-wise performance assessment, our next objective is to compare the respective contribution of each predictor to the overall skill of these models. We first assess the contribution of the TC characteristics at the beginning of the hindcast, since they strongly contribute

to the model skill (e.g. DeMaria and Kaplan 1994a, 1999; Knaff et al. 2005, 2009; Lee et al. 2015). In Figure 9, we compare the skill of the “Atm” models with all environmental predictors to that of “baseline” models (Table 2) built by only accounting for the TC initial characteristics (predictors 1, 2 and 3 from Table 1). This figure reveals that TC characteristics (i.e., the baseline models) explain a large fraction of the MAE decrease when compared to persistence (Fig. 9a): the average MAE at 60 h averaged over all basins decreases by 6 kt (25 kt to 19 kt) from persistence to baseline model and by only further 2.5kt (to 16.5 kt) for the Atm model. When averaged over all basins, the relative improvements compared to persistence range from 19% for 24hr lead to 27% for 108-h lead time for the baseline model, while they reach 24% for 24-h lead time to 36% for 108-h lead time for the Atm model (Fig. 9b). The inclusion of the large-scale environmental parameters in the predictors hence results in an additional skill improvement relative to the baseline model ranging from ~5% at 24hr to ~9% at 108hr. That is, accounting for environmental parameters accounts for 20 to 40% of the Atm model performance. The improvement brought by the inclusion of environmental parameters is qualitatively similar from one basin to another (Fig. 9b). As a summary, Figure 9 generally illustrates that the TC initial characteristics account for a large fraction of the TC hindcasts skill of the statistical-dynamical linear models in all TC basins, in agreement to previously published literature.

To further quantify the relative contribution of each large-scale environmental parameter listed in Table 1 (predictors 4 to 11), a series of experiments excluding each of these environmental parameters is carried out (Atm-VarN experiments in Table 2). The contribution of each large-scale environmental parameter is assessed by calculating the percentage of skill improvement of the Atm model relative to Atm-VarN, using Eq. (1). For VMAX and MPI parameters, both the term and its square are removed from the predictors list when carrying out the Atm-MPI and Atm-VMAX experiments. Figure 10 displays this result for a particular lead-time (60 h). When averaged over all basins, SHRD is the most important environmental parameter, yielding to a ~4% skill improvement. The most important parameters are then MPI (~2.5%) followed by Z850, USHR, RHHI and T200 (~1%). Finally, E925 contributes very weakly to the skill improvement at global scale (~0.25%). When looking at these contributions basin-wise, the skill improvement is also generally dominated by a subset of environmental parameters. As it is the case at global scale, SHRD is generally the most important environmental parameter in all TC-prone basins, except in the NEP and NWP where the improvement brought by MPI is larger. Its contribution is particularly large in the Indian Ocean basin (NIO and SIO) where it yields to ~8% of skill improvement versus 2 to 4% improvement in other basins. While the MPI contribution is the second largest when assessed at global scale, its contribution varies considerably from one basin to another: it is relatively large in the NWP, NEP

and NIO (~2 to 4% improvement) but far weaker in SWP, SIO (less than 1%) and ATL (less than 2%). Finally, some parameters yield negligible improvements in most basins but are skilful for some specific basin: it is for instance the case for RHHI which does not contribute much in most basins, while this parameter is amongst the two most important (with SHRD) for the SWP basin. This probably translates some differences in the relative importance of processes that control the TC intensity between various basins as well as the basin-wide climatology and variability of these quantities. The basin-wise differences above are qualitatively similar at all forecast lead times, except in the NIO and ATL basins where the relative importance of each parameters considerably vary as a function of the lead-time (not shown). While this sensitivity is most likely related to the reduced dataset used to build the model for the NIO basin, the reasons behind this sensitivity are unclear for the ATL and would require further investigation for this specific basin.

### **3.3 Real-time vs climatological environmental parameters**

The environmental parameters used as predictors in the present and previously published TC intensity statistical-dynamical hindcast models are systematically calculated based on data at the lead-time of the forecast (and at previous lead-times, since some variables are averaged from the forecast time to the lead time). These daily data hence include seasonal variations but also perturbations at other timescales, ranging from synoptic features to large-scale interannual anomalies associated with climate modes such as the El Niño Southern Oscillation. In this section, we evaluate the relative contribution of the mean climatological seasonal cycle and non-seasonal variability to the performance of the model. To reach that goal, we perform additional regional experiments in which the climatological environmental parameters are used in place of the actual environmental parameters at the time of the hindcast (Atm\_Clim experiment; see Table 2). In the case of Atm\_Clim experiments, the parameters have been computed along the TC track from a 6-hourly ERA-Interim climatological dataset. As for the Atm model, these parameters are then averaged in time along the TC track from the initial time till the forecast hour. Comparing Atm\_Clim with Atm experiments hence allows assessing the added value of considering the environmental parameters actual values instead of considering those derived from a climatological environment for each basin. Figure 11a illustrates the model using a climatological seasonal cycle for environmental predictors (Atm\_Clim) yields a very similar skill compared that using actual values (Atm): the globally-averaged MAE is nearly identical for lead-times greater than 84 h, where errors tend to saturate, and only slightly reduced for shorter lead-times. Figure 9b allows a basin-wise quantification of that feature. At global scale, using real-time predictors does not improve the model skill at long lead-time but modestly improves it at short to mid-range lead-time (1 to 3%). The added value of using daily parameters in

place of climatological parameters varies from one basin to another: it does not improve the skill for NWP and NEP and marginally for the ATL basin (~2%). In contrast, the use of daily parameters leads to a larger improvement for the southern hemisphere TC, especially in the SIO basin where this improvement reaches ~8% for medium to long lead-time. The good performance of the *Atm\_Clim* hindcast scheme suggests that TC intensity changes are far more constrained by the spatial patterns and seasonal variations of environmental predictors rather than by their non-seasonal component related to synoptic signals and large-scale climate modes, especially for the northern hemisphere.

### **3.4 Skill as a function of TC intensity**

We finally evaluate the model skill as a function of the TC intensity. For this purpose, the predictands are divided into two subsets based on the TC intensity. The TC up to category 2 (<96 kt) are referred to as “moderate” (~60% of the dataset) category 3-5 ( $\geq 96$  kt) are considered as “strong” (~40% of the dataset). Rather surprisingly, Figure 12 indicates that the model skill strongly depends on the TC intensity, with considerably weaker skill for moderate relative to strong TC. The global skill improvement relative to persistence indeed reaches ~35% to 45% for the strongest TC (Fig. 12a) but only ~10% to 20% for moderate TC (Fig. 12b). That is, the TC hindcast model is about three to four times more skilful for strong than for moderate TC. This result holds for all basins but is particularly striking for ATL where the model skill is marginal for moderate TC. We further checked that results related to the basin-wise performance of the model (Fig. 6), the importance of TC initial characteristics (Fig. 9), the relative environmental parameters (Fig. 10) and the good performance of *Atm\_Clim* compared to *Atm* (Fig. 11) obtained for the entire TC database hold true for the strongest TC (not shown).

## **5. Summary and discussion**

The prediction of tropical cyclone intensity is still a challenging problem. Along with dynamical models, TC intensity prediction relies on statistical-dynamical linear models that use linear statistical regression techniques to predict intensity changes from predictors derived from initial TC characteristics and from large-scale environmental parameters along the TC track. Given the tremendous consequences of TC worldwide, there is a need to better assess the performance of these regional models. However, it is difficult to compare these models across TC basins, as these models have generally been developed independently, using different predictors, atmospheric datasets and training periods. In this paper, we develop statistical-dynamical linear hindcast models for each TC-prone basin using the same set of predictors, atmospheric dataset and extended training/validation period (1979-2012). TC intensity statistical-dynamical linear hindcast models are



hence built consistently for all TC-prone basins. A similar consistent methodology has recently been developed in Lee et al. (2016) but as a first step to develop an autoregressive model able to simulate the climatological distribution of global TC intensity, that can ultimately be used for projecting future changes in TC in response to climate change. The goal of the present study is different as we use this common framework to assess the relative importance of TC initial characteristics vs. environmental parameters in the models skill, identify the most skilful environmental predictors in each basins and investigate the sensitivity of the skill to the TC strength. The models in this study display similar skills to those previously published, giving confidence in the conclusions drawn from this analysis. We have shown that the long period that we consider in the present study ensures a better robustness of the results. Re-conducting analyses in this paper over 1990-2012, during which intensity records (e.g. Kossin et al. 2013) and tracks (Velden et al. 2006; Knapp and Kossin 2007) are more reliable, does not change the overall conclusions of the present study (not shown).

Our results first reveal that the mean absolute errors derived from these regional models are rather similar amongst the different TC-prone basins. However, the relative skill improvement brought by using this multi-linear statistical framework as compared to simple persistence is larger for northwest and northeast Pacific, the southwest Pacific and the southern Indian Ocean than for the Atlantic and Northern Indian Ocean basins. This relative skill difference can confidently be attributed to weaker errors for persistence in the Atlantic and Northern Indian Ocean basins, i.e. TC intensity is more steady in those basins. These results qualitatively agree with the analysis performed by Lee et al. (2016). Some differences can however be noticed. Our model outperforms theirs in the southern hemisphere for all lead times and in the NIO for long lead times (Figure 6a), while their configuration behaves better for long lead times in the Atlantic (their Figure 1b). These differences may arise due to differences in the selection of predictors and training/testing period. We further show that that the regression coefficients for each predictor of these models generally exhibit consistent signs and order of magnitude across basins, suggesting qualitatively similar influences of each predictor. Comparing results from models trained for each TC-prone regions to those from a model trained over the entire globe demonstrates that the improvement arising from a regional training compared to a global one is rather modest, reaching respectively 2%, 3% and 1% at 24h, 60h and 108h, suggesting that it may be possible to build an efficient statistical model for TC intensity forecast at global scale, in line with Lee et al. (2016).

Several studies already mentioned that, amongst all predictors, the storm characteristics at the beginning of the forecast (intensity and its time-derivative) are dominant contributors to the skill. Our results confirm those findings and quantifies them more precisely: TC initial characteristics

contributes to ~ 60 to 80% to the model skill improvement relative to persistence in all TC-prone basins, with large-scale environmental predictors accounting for the rest (~20 to 40 %). At global scale, we find that the most important environmental predictor is the vertical shear, followed by the maximum potential intensity. Other environmental parameters contribute less to the skill improvement. Vertical wind shear plays a key role in most TC-prone basins, but the relative contribution of environmental parameter varies considerably from a basin to another. Vertical wind shear has a very large impact in the Indian Ocean, contributing to ~8% of the skill increase relative to persistence. Our result for the northwest Pacific broadly agrees with the STIPS model result that also showed that the vertical shear and the maximum potential intensity are the most important predictors (Knaff et al. 2005). Lee et al. (2015) do not use the 850hPa vorticity as a predictor in their model, which is the second most important predictor in our model for the Atlantic. However, in the absence of this parameter in their model, vertical shear and a form of maximum potential intensity are the two most important atmospheric predictors, which are first and third most important parameters in our analysis for this basin. A more quantitative comparison with earlier studies is however not possible due to various differences arising from the choice of predictors, data source and period.

One of the key results of the present study could lead to a considerable simplification of the implementation of current statistical forecast of TC intensity. Our results indeed reveal that a model built from climatological environmental parameters along the TC track yields nearly the same predictive skill as a model built from real-time values. The improvement of using real-time environmental parameters is marginal for the northern hemisphere TC (less than 2%), but larger for the southern hemisphere TC (up to 9% in the SIO). This result echoes Lee et al. (2015) who demonstrated that their statistical hindcast model has a comparable skill when using environmental predictors derived from monthly rather than daily environmental data for the case of the Atlantic. In agreement with our results, this suggests that subseasonal variations of environmental parameters only marginally contribute the skill improvement of the model. Our results further demonstrate that the same conclusion holds for interannual anomalies, such as those associated with large-scale climate modes such as the El Niño Southern Oscillation. This result is potentially interesting for statistical operational forecasting of TC intensity. For implementation purposes, it is indeed far simpler to use atmospheric predictors calculated from their seasonal climatology than to retrieve the real-time forecasted values of those predictors along the forecasted TC track. This strategy may however not be adapted to the southern hemisphere, where our results suggest the largest skill gains when accounting for real-time environmental parameter variations.

Finally, our analysis also demonstrates that these forecasting schemes are skilful to predict intensity changes for strongest TC (from Cat 3 to 5), while it is hardly the case for moderate TC (Cat-2 and below). The skill improvement relative to persistence is indeed three to four times weaker for moderate than for the strong TC. It is plausible that the strength of the control of a given environmental parameter is varying as a function of the TC intensity, a non-linear behaviour that our linear scheme is not able to capture. It will be interesting to investigate whether non-linear statistical prediction schemes are able to improve the hindcasts for moderate cyclones. The use of non-linear schemes such as Artificial Neural Networks (Fine, 1999) or Support Vector Machines (Cortes and Vapnik, 1995) may be beneficial to further improve the performance of these statistical TC intensity models.

In real-time applications of these methods it is important to note that both uncertainties in TC location and intensity and the use of a forecast track will result in additional sources of errors. Nonetheless, results of this study will still be valid.

**Acknowledgments:** NS and IS thank Director, CSIR-NIO for his encouragement. NS, ML and JV thank IFCPAR (Indo French Centre for Promotion of Advanced Research), New Delhi for funding of the proposal 4907-1. This work was done while ML was a visiting scientist at the CSIR-National Institute of Oceanography (NIO) in Goa, under Institut de Recherche pour le Développement (IRD) funding. JV also acknowledges IRD support for regular visits to NIO. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official National Oceanic and Atmospheric Administration or U.S. Government position, policy, or decision. This is NIO contribution number xxxx.

### **References:**

- Aberson SD. 1998. Five-day tropical cyclone track forecasting in the North Atlantic Basin. *Wea. Forecasting*. 13: 1005-15.
- Cangialosi JP, Franklin JL. 2014. `2013 National Hurricane Center Forecast Verification Report.` NOAA/NWS/NCEP/National Hurricane Center. [Available online at [www.nhc.noaa.gov/verification/pdfs/Verification\\_2013.pdf](http://www.nhc.noaa.gov/verification/pdfs/Verification_2013.pdf)].
- Chu J-H. 1994. `A regression model for the western North Pacific tropical cyclone intensity forecasts.` NRL Memo. Rep. 7541-94-7215, Naval Research Laboratory, 33 pp. [Available from Naval Research Laboratory, 7 Grace Hopper Avenue, Monterey, CA 93943-5502]
- Cortes C, Vapnik V. 1995. Support-vector networks. *Machine Learning*. 20 (3): 273-297. DOI: 10.1007/BF00994018.
- Dee DP, Uppala SM, Simmons AJ et al. 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. Roy. Meteorol. Soc.* 137, 656: 553-597.

- DeMaria M, Mainelli M, Shay LK, Knaff JA, Kaplan J. 2005. Further improvement to the Statistical Hurricane Intensity Prediction Scheme (SHIPS). *Wea. Forecasting*. 20: 531–543.
- DeMaria M, and Kaplan J. 1999. An updated statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic and eastern north Pacific basins. *Wea. Forecasting*. 14: 326-337.
- DeMaria M, Kaplan J. 1994a. A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*. 9: 209–220.
- DeMaria M, Kaplan J. 1994b. Sea surface temperature and the maximum intensity of Atlantic tropical cyclones. *J. Climate*. 7: 1324–1334.
- DeMaria M, Knaff JA, Sampson CR. 2007. Evaluation of long-term trend in tropical cyclone intensity forecasts. *Meteor. Atmos. Phys.* 97: 19-28.
- DeMaria M, Sampson CR, Knaff JA, Musgrave KD. 2014. Is tropical cyclone intensity guidance improving? *Bull. Amer. Meteor. Soc.* 95: 387-398.
- DeMaria M. 1996. The effect of vertical shear on tropical cyclone intensity change. *J. Atmos. Sci.* 53: 2076–2087.
- DeMaria M. 2009. A simplified dynamical system for tropical cyclone intensity prediction. *Mon. Wea. Rev.* 137: 68-82.
- DeMaria M. 2010. 'Tropical cyclone intensity change predictability estimates using a statistical-dynamical model.' 29th AMS conference on hurricanes and tropical meteorology, Tucson, AZ. 5 pp.
- Dvorak VF. 1984. 'Tropical cyclone intensity analysis using satellite data.' NOAA Tech. Rep. 11, 45 pp.
- Elsberry RL, Chen LS, Davidson J, Rogers R, Wang Y, Wu L. 2013. Advances in
- Emanuel K. 2007: Environmental factors affecting tropical cyclone power dissipation. *J. Climate*. 20: 5497-5509.
- Emanuel KA, DesAutels C, Holloway C, Korty R. 2004. Environmental control of tropical cyclone intensity. *J. Atmos. Sci.* 61: 843–858.
- Emanuel KA, Nolan DS. 2004. 'Tropical cyclone activity and global climate,' *26th Conf. on Hurricanes and Tropical Meteorology*, Miami, FL, Amer. Meteor. Soc. 240–241.
- Emanuel KA. 1988. The maximum intensity of hurricanes. *J. Atmos. Sci.* 45: 1143–1155.
- Evans AD, Falvey RJ. 2013. Annual tropical cyclone report, 2012. Joint Typhoon Warning Center, Pearl Harbor, Hawaii, pp. 1–118. [Available online at [www.usno.navy.mil/NOOC/nmfc-ph/RSS/jtwc/atcr/2012atcr.pdf](http://www.usno.navy.mil/NOOC/nmfc-ph/RSS/jtwc/atcr/2012atcr.pdf)].
- Fine TL. 1999. *Feedforward neural network methodology*. Springer-Verlag, New York; 356p.

- Frank WA, Ritchie EA. 1999. Effects of environmental flow upon tropical cyclone structure. *Mon. Wea. Rev.* 127: 2044-2061.
- Frank WA, Ritchie EA. 2001. Effects of environmental flow on the intensity and structure of numerically simulated hurricanes. *Mon. Wea. Rev.* 129: 2249–2269.
- Gray WM. 1968. A global view of the origin of tropical disturbances and storms. *Mon. Wea. Rev.* 96: 669-700.
- Holland GJ. 1997. The maximum potential intensity of tropical cyclones. *J. Atmos. Sci.* 54: 2519 – 2541.
- Holland GJ. 1980. An analytic model of the wind and pressure profiles in hurricanes. *Mon. Wea. Rev.* 108: 1212–1218
- Jarvinen BR, Neumann C J. 1979. `Statistical forecasts of tropical cyclone intensity for the North Atlantic basin.` NOAA Tech. Memo. NWS NHC –10, 22 pp. [Available from NTIS, Technology Administration, U.S. Dept. of Commerce, Springfield, VA 22161]
- Kaplan J, DeMaria M. 2003. Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. *Wea. Forecasting.* 18: 1093–1108.
- Knaff JA, DeMaria M, Sampson CR Gross JM. 2003. Statistical, five-day tropical cyclone intensity forecasts derived from climatology and persistence. *Wea. Forecasting,* 18, 80-92.
- Knaff JA, Sampson CR, DeMaria M. 2005. An operational Statistical Typhoon Intensity Prediction Scheme for the western North Pacific. *Wea. Forecasting.* 20: 688–699.
- Knaff JA, Sampson CR. 2009. Southern Hemisphere tropical cyclone intensity forecast methods used at the Joint Typhoon Warning Center. Part II: Statistical–dynamical forecasts. *Aust. Meteor. Oceanogr. J.* 58: 9–18.
- Knapp KR, Kossin JP. 2007. New global tropical cyclone data set from ISCCP B1 geostationary satellite observations. *J. Appl. Remote Sens.* 1. 013505, DOI: 10.1117/1.2712816.
- Knapp KR, Kruk MC, Levinson DH, Diamond HD, Neumann CJ. 2010. The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying Tropical Cyclone Data, *Bull. Amer. Meteor. Soc.* 10. DOI: 10.1175/2009BAMS2755.1.
- Kossin JP, Knapp KR, Vimont DJ, Murnane RJ, Harper BA. 2007. A globally consistent reanalysis of hurricane variability and trends. *Geophys. Res. Lett.* 34: L04815. DOI: 10.1029/2006GL028836.
- Kossin JP, Olander TL, Knapp KR. 2013. Trend analysis with a new global record of tropical cyclone intensity. *J. Climate.* 26: 9960–9976. DOI: 10.1175/JCLI-D-13-00262.1.
- Kotal SD, Roy Bhowmik SK, Kundu PK, Das AK. 2008. A Statistical Cyclone Intensity Prediction (SCIP) Model for Bay of Bengal. *J. Earth Sys. Sci.* 117: 157-168.
- Kucas ME. 2010. `Challenges of forecasting tropical cyclone intensity change at the Joint Typhoon Warning Center.` Preprints, *29th Conf. on Hurricanes and Tropical Meteorology,*

Tucson, AZ, *Amer. Meteor. Soc.* 9C.6. [Available online at <https://ams.confex.com/ams/pdfpapers/168732.pdf>.]

- Landsea CW, Harper BA, Hoarau K, Knaff JA. 2006. Can we detect trends in extreme tropical cyclones? *Science*. 313: 452–454.
- Lee, C-Y, Tippett MK, Camargo SJ, Sobel AH. 2015. Probabilistic Multiple Linear Regression Modeling for Tropical Cyclone Intensity. *Mon. Wea. Rev.* 143: 933–954. DOI: <http://dx.doi.org/10.1175/MWR-D-14-00171.1>
- Longshore D. 2008. *Encyclopedia of Hurricanes, Typhoons, and Cyclones*. Checkmark, New York; 468 pp..
- Mainelli M, DeMaria M, Shay LK, Goni G. 2008. Application of oceanic heat content estimation to operational forecasting of recent Atlantic category 5 hurricanes. *Weather and Forecasting*. 23: 3–16.
- McPhaden MJ, Foltz GR, Lee T, Murty VSN, Ravichandran M, Vecchi GA, Vialard J, Wiggert JD, Yu L, 2009. Ocean–atmosphere interactions during cyclone Nargis. *EOS Trans. AGU* 90: 53–54. DOI: 10.1029/2009EO070001
- Merrill RT. 1980. `A statistical tropical cyclone motion forecasting system for the Gulf of Mexico.` NOAA Tech. Memo. NWS NHC 14, 21 pp. [Available from NTIS, Technology Administration, U.S. Dept. of Commerce, Springfield, VA 22161]
- Merrill RT. 1987. An experiment in statistical prediction of tropical cyclone intensity change. NOAA Tech. Memo. NWS NHC 34, 34 pp. [Available from the National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22151.]
- Merrill RT. 1988. Environmental influence of hurricane intensification. *J. Atmos. Sci.* 45: 1678–1687.
- Miller BI. 1958. On the maximum intensity of hurricanes. *J. Meteor.* 15: 184–195.
- Neumann CJ. 1972. `An alternate to the HURAN (hurricane analog) tropical cyclone forecasting system.` NOAA Tech Memo. NWS SR 62, 23 pp. [Available from NTIS, Technology Administration, U.S. Dept. of Commerce, Springfield, VA 22161]
- Sharma N, Ali MM, Knaff JA, Chand P. 2013. A soft-computing cyclone intensity prediction scheme for the Western North Pacific Ocean. *Atmosph. Sci. Lett.* 14: 187–192. DOI: 10.1002/asl2.438
- Tang B, Emanuel K. 2012. Sensitivity of tropical cyclone intensity to ventilation in an axisymmetric model. *J. Atmos. Sci.* 69: 2394–2413.
- Tien DD, Ngo-Duc T, Mai HT, Kieu C. 2013. A study of the connection between tropical cyclone track and intensity errors in the WRF model. *Meteorol. Atmos. Phys.* 122: 55–64. DOI: 10.1007/s00703-013-0278-0.

Velden C, Herndon D, Kossin J, Hawkins J, DeMaria M. 2006. 'Consensus estimates of tropical cyclone intensity using integrated multispectral (IR and MW) satellite observations.' Preprints, *26th AMS Conf. on Hurr. and Trop. Meteor.*, Monterey, CA, Amer. Meteor. Soc.

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

Webster PJ. 2008. Myanmar's deadly daffodil. *Nat. Geosci.* 488-490.

Whitney LD, Hobgood JS. 1997. The relationship between sea surface temperatures and maximum intensities of tropical cyclones in the eastern North Pacific Ocean. *J. Climate*. 10: 2921–2930.

Zeng Z, Wang Y, Wu CC. 2007. Environmental dynamical control of tropical cyclone intensity – an observational study. *Mon. Wea. Rev.* 135:38 –59

### Figure Captions:

**Figure 1:** Tropical Cyclones (TC) climatological density (per 4°x4° bin) global map. The six red frames indicate the TC-prone regions for which individual statistical TC intensity prediction models are build (NWP: North Western Pacific, NEP: North Eastern Pacific, SWP: Southwestern Pacific, ATL: Atlantic, SIO: Southern Indian Ocean, NIO: Northern Indian Ocean). The numbers in parenthesis indicate the total number of TC considered for each region over the 1979-2012 period.

**Figure 2:(a)** Normalized distribution of TC intensity for each basin. Vertical dashed lines indicate the mean of the upper tenth percentile of the distribution for each basin. Normalized distribution of **(b)** SHRD, **(c)** MPI and **(d)** RHHI at 12h lead-time for each basin. Vertical dashed lines indicate the mean of the distribution for each basin.

**Figure 3:(a)** MAE as a function of lead-time, averaged over all basins for persistence (black line), Atm model (blue line) and Atm+Cross model (red line). **(b)** Percentage of Atm+Cross improvement relative to Atm at 24h, 60h and 108h lead-times, for each basin and globally. Atm+Cross model is similar to Atm model, except that the two cross terms (VMXS and VMXM) are included in the predictors list. Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.

**Figure 4:(Left panels)** MAE as a function of lead-time for persistence (black line), Atm model (blue line) and previously published results (red line) in **(a)** the NWP for the training dataset (compared to Knaff et al. 2005; their Table 6), **(c)** in the southern hemisphere (SWP+SIO) for the training dataset (compared to Knaff et al. 2009; their Table 5) and **(e)** the ATL for the testing dataset (compared to Lee et al. 2015; their Figure 5).

**Figure 5:(Left panels)** MAE as a function of lead-time of the training dataset (blue line) and testing dataset (red line) for the Atm model built from a 7 year period in the **(a)** NWP, **(b)** SWP+SIO and **(c)** ATL. MAEs from already published results are also shown as dashed lines. **(Right panels)** Same as left panels but for a model build over a 34 years period. Thick lines indicate MAE resulting from averaging results from 1000 simulations with different randomly

selected training/testing datasets while shading indicate the 5% lower and 95% higher bound of the ensemble distribution.

**Figure 6:** MAE as a function of lead-time for **(a)** Atm model and **(b)** persistence for each basin (i.e. NWP, NEP, SWP, ATL, SIO, NIO). **(c)** Percentage of skill improvement of these models relative to persistence as a function of lead-time for each basin. See section 2.2 for a definition of persistence metric and the skill metrics shown on panel c. Error bars on panels a and c give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.

**Figure 7:(a)** Regression coefficients for the key predictors used in the Atm model at 60h for each basin. Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2. These regression coefficients have been multiplied by 100, 1.5, -0.5, 100, -1 and 2 for VMAX2, PER, MPI, MPI2, SHRD and RHHI respectively for a better readability.

**Figure 8:**Percentage of model skill improvement relative to persistence for basin-wise trained Atm models (light colors; see Table 2) and the globally-trained Glob model (plain colors; see Table 2) at 24h, 60h and 108h, as a function of the basin. Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.

**Figure 9:(a)**MAE as a function of lead-time, averaged over all basins, for persistence (black curve), Atm model (blue curve) and baseline model (red curve). **(b)** Percentage of skill improvement of baseline (plain colors) and Atm models (light colors) relative to persistence at 24h, 60h and 108h lead for each basin and globally. Only the TC initial characteristics (predictors 1 to 3 in Table 1) are used as predictors in the baseline model (cf table 2). Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.

**Figure 10:**Percentage of skill improvement of Atm relative to Atm-VarN (see table 2) at 60h for each basin and globally. This is a measure of the respective contributions of the SHRD, MPI, PHHI, T200, USHR, E925 and Z850 predictors to the overall skill. Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.

**Figure 11:(a)**MAE as a function of lead-time, averaged over all basins, for persistence (black curve), Atm model (blue curve) and Atm\_Clim model (red curve). **(b)** Percentage of skill improvement of Atm relative to Atm\_Clim at 24h, 60h and 108h lead-times for each basin and globally. In Atm\_Clim, the environmental parameters are calculated from their climatology, rather than on their actual value at the forecast time (cf. Table 2). Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.

**Figure 12:**Percentage of Atm model skill improvement relative to persistence for **(a)** strong and **(b)** moderate TC at 24h, 60h and 108h, as a function of the basin. Here, TC up to category 2 (< 96 kt) are considered as “moderate” and those under category 3-5 ( $\geq 96$  kt) are considered as “strong”. Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.



S. No.	Predictor	Description
1	VMAX	Initial intensity
2	VMAX <sup>2</sup>	Initial intensity squared
3	PER	Intensity change during previous 12 hour
4	MPI#	Maximum potential intensity
5	MPI <sup>2</sup> #	Maximum potential intensity squared
6	SHRD*#	200 to 850 hpa Wind shear magnitude
7	USHR*#	200 to 850 hpa Zonal wind shear magnitude
8	RHHI*#	500 to 300 hpa Average relative humidity
9	T200*#	200 hpa Temperature
10	E925*#	925 hpa Equivalent potential temperature
11	Z850**#	850 hpa Vorticity
12	<i>VMXS</i>	<i>VMAX*SHRD</i>
13	<i>VMXM</i>	<i>VMAX*MPI</i>

**Table 1:** List of the predictors used in the present study. The predicted variable is DELV, i.e. the intensity change since the forecast start, at 12, 24, ... , 120 hours into the forecast. The variables marked with a \* are estimated from an area-average within 200 to 800 km of the cyclone track. The variables marked with a \*\* are estimated from an area-average within 1000 km of the cyclone track. The variables marked with a # are time-averaged from the initial to the forecast time. The variables in black (No 1 to 11) are used in the Atm reference model presented throughout the paper. The variables in *italics* (No 12 and 13; i.e. the cross-terms) are commonly used in statistical-dynamical forecasts but are discarded from the final list of predictors in the present study to allow a proper assessment of the relative importance of each of the predictors (see text for details).

Name	Predictors used
Atm	Regional models built with predictors 1-11 in each basin
Atm+Cross	Regional models built with predictors 1-13 (Table 1) in each basin
Baseline	Regional models built with predictors 1-3 in each basin
Atm-VarN	As Atm, without predictor N; For Vmax and MPI, both the term and its square are removed from the predictors list
Atm_Clim	As Atm model but atmospheric predictors (4-11) are calculated from climatological fields
Glob	Single global model built with predictors 1-11

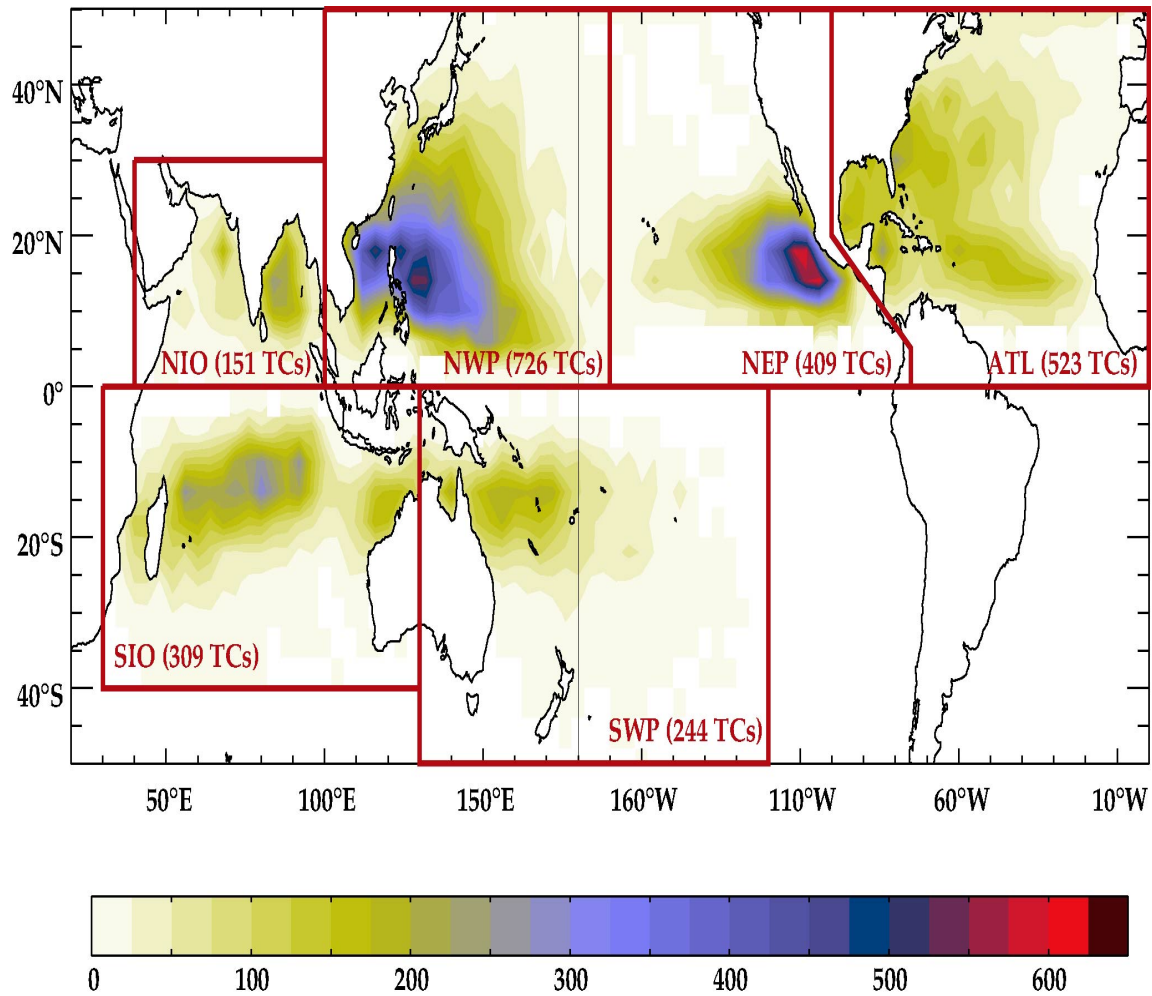
**Table 2:** List of different sensitivity experiments performed and related predictors used. The numbering of the predictors used refers to the parameters listed in Table 1

Basin	MPI Equation	Reference
NWP	$A+Be^{C(T-T_0)}$ , A=38.21 kt, B=170.72 kt C=0.1909 °C <sup>-1</sup> , T <sub>0</sub> =30.0 °C	Knaff et al., <i>Wea. Forecasting</i> , 2005
NEP	A+BT, A=-79.17262 m s <sup>-1</sup> , B=5.361814 m s <sup>-1</sup> °C <sup>-1</sup>	DeMaria et al., <i>Wea. Forecasting</i> , 2005
SWP, SIO	$A+Be^{C(T-T_0)}$ , A=-42.1 kt, B=220.58 kt C=0.0792 °C <sup>-1</sup> , T <sub>0</sub> =30.0 °C	Knaff et al., <i>Aust. Met. Oceanogr. J.</i> , 2009
NIO	A+BT, A=-889.64 m s <sup>-1</sup> , B=35.714 m s <sup>-1</sup> °C <sup>-1</sup>	Kotal, et al., <i>Meteorol. Appl.</i> , 2009
ATL	$A+Be^{C(T-T_0)}$ , A=28.2 m s <sup>-1</sup> , B=55.8 m s <sup>-1</sup> C=0.1813 °C <sup>-1</sup> , T <sub>0</sub> =30.0 °C	DeMaria et al., <i>Wea. Forecasting</i> , 2005

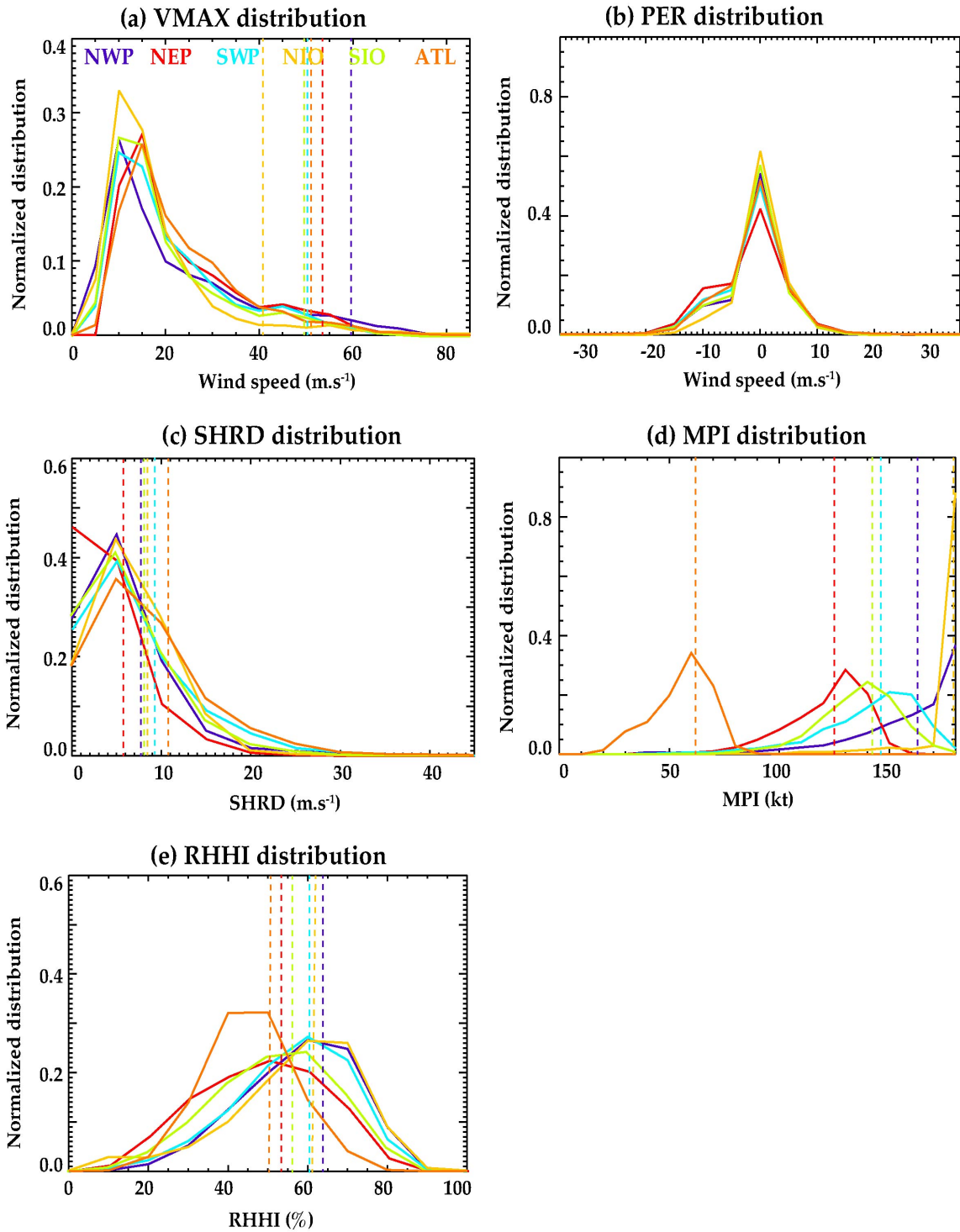
**Table 3:** Empirical formulation of the Maximum Potential Intensity (MPI) for each TC-prone basin and related references.

	VMAX	PER	MPI	SHRD	USHR	RHHI	T200	E925	Z850	VMAX <sup>2</sup>	MPI <sup>2</sup>
VMAX	<b>1.00</b>	0.14	-0.20	0.03	0.21	-0.23	0.23	-0.09	0.08	<b>0.96</b>	-0.23
PER		<b>1.00</b>	0.12	-0.19	-0.11	0.19	0.00	0.18	0.02	0.13	0.13
MPI			<b>1.00</b>	-0.32	<b>-0.53</b>	<b>0.50</b>	0.44	<b>0.72</b>	0.09	-0.12	<b>0.99</b>
SHRD				<b>1.00</b>	0.45	-0.18	-0.05	-0.42	0.02	0.00	-0.29
USHR					<b>1.00</b>	-0.48	-0.17	<b>-0.50</b>	-0.09	0.17	<b>-0.52</b>
RHHI						<b>1.00</b>	0.21	<b>0.52</b>	0.14	-0.17	<b>0.52</b>
T200							<b>1.00</b>	0.46	0.15	0.24	0.42
E925								<b>1.00</b>	0.04	-0.04	<b>0.70</b>
Z850									<b>1.00</b>	0.10	0.16
VMAX <sup>2</sup>										<b>1.00</b>	-0.15
MPI <sup>2</sup>											<b>1.00</b>

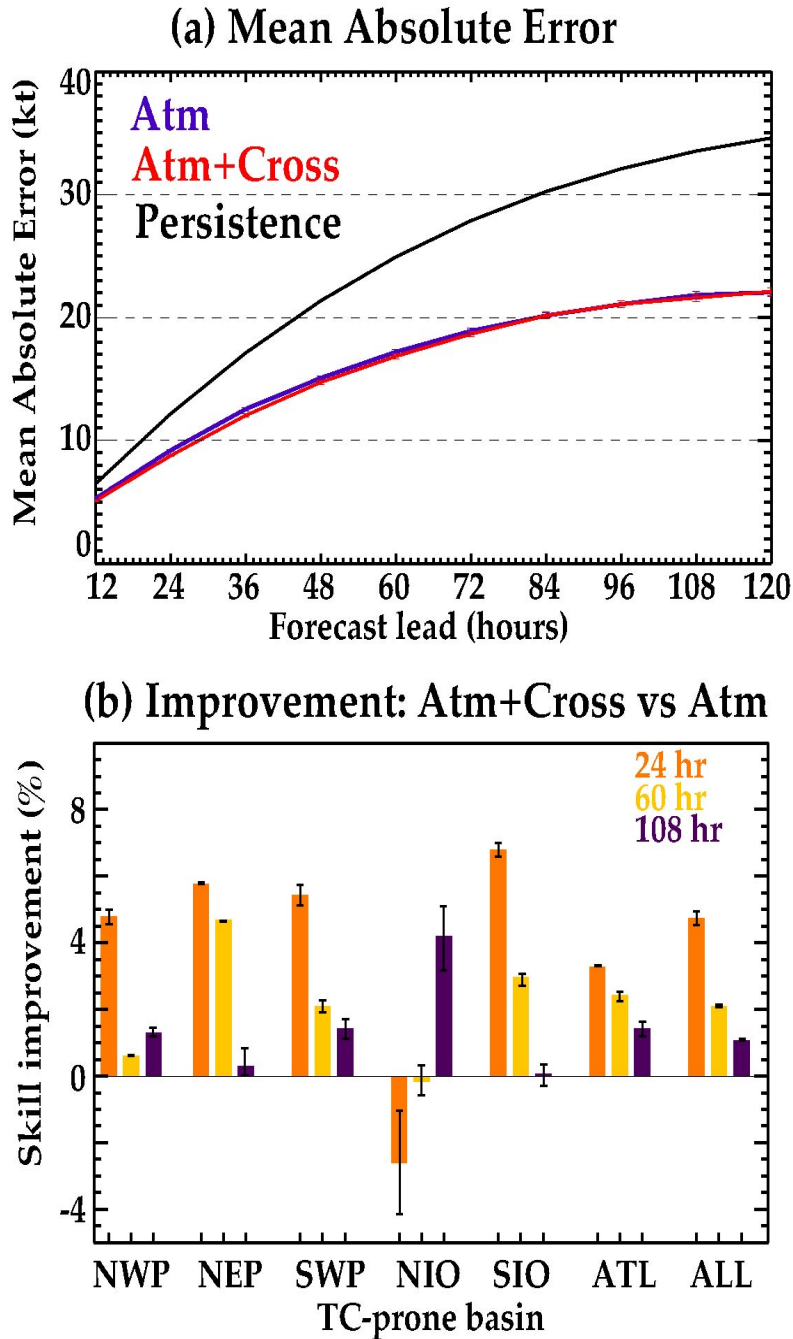
**Table 4:** Correlation table between predictors at 60h at global scale. Performing these correlations per basin provides similar qualitative results. Bold numbers shows the correlations significant at 99% statistical significance level from t- test.



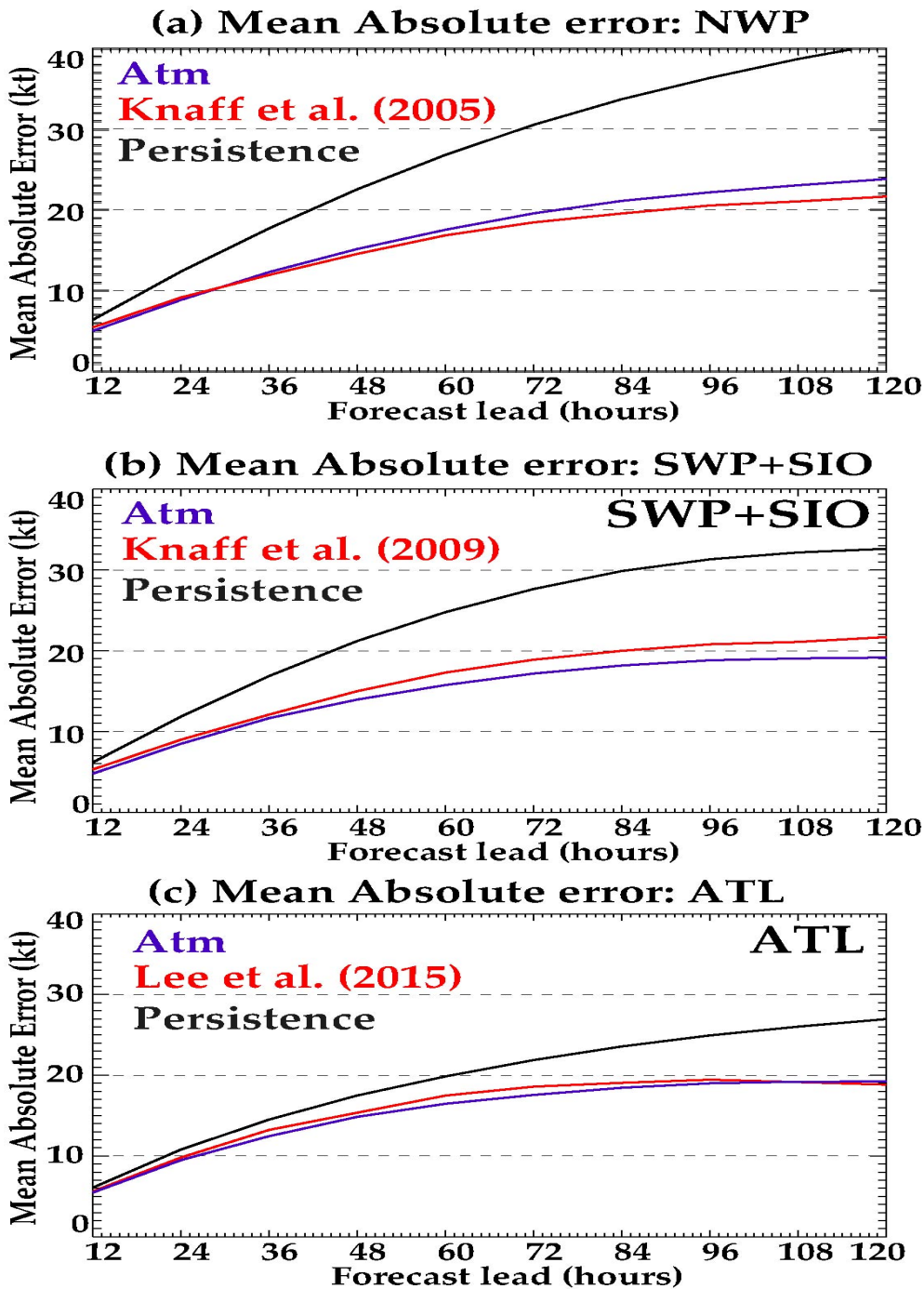
**Figure 1:** Tropical Cyclones (TC) climatological density (per  $4^{\circ} \times 4^{\circ}$  bin) global map. The six red frames indicate the TC-prone regions for which individual statistical TC intensity prediction models are built (NWP: North Western Pacific, NEP: North Eastern Pacific, SWP: Southwestern Pacific, ATL: Atlantic, SIO: Southern Indian Ocean, NIO: Northern Indian Ocean). The numbers in parenthesis indicate the total number of TC considered for each region over the 1979-2012 period.



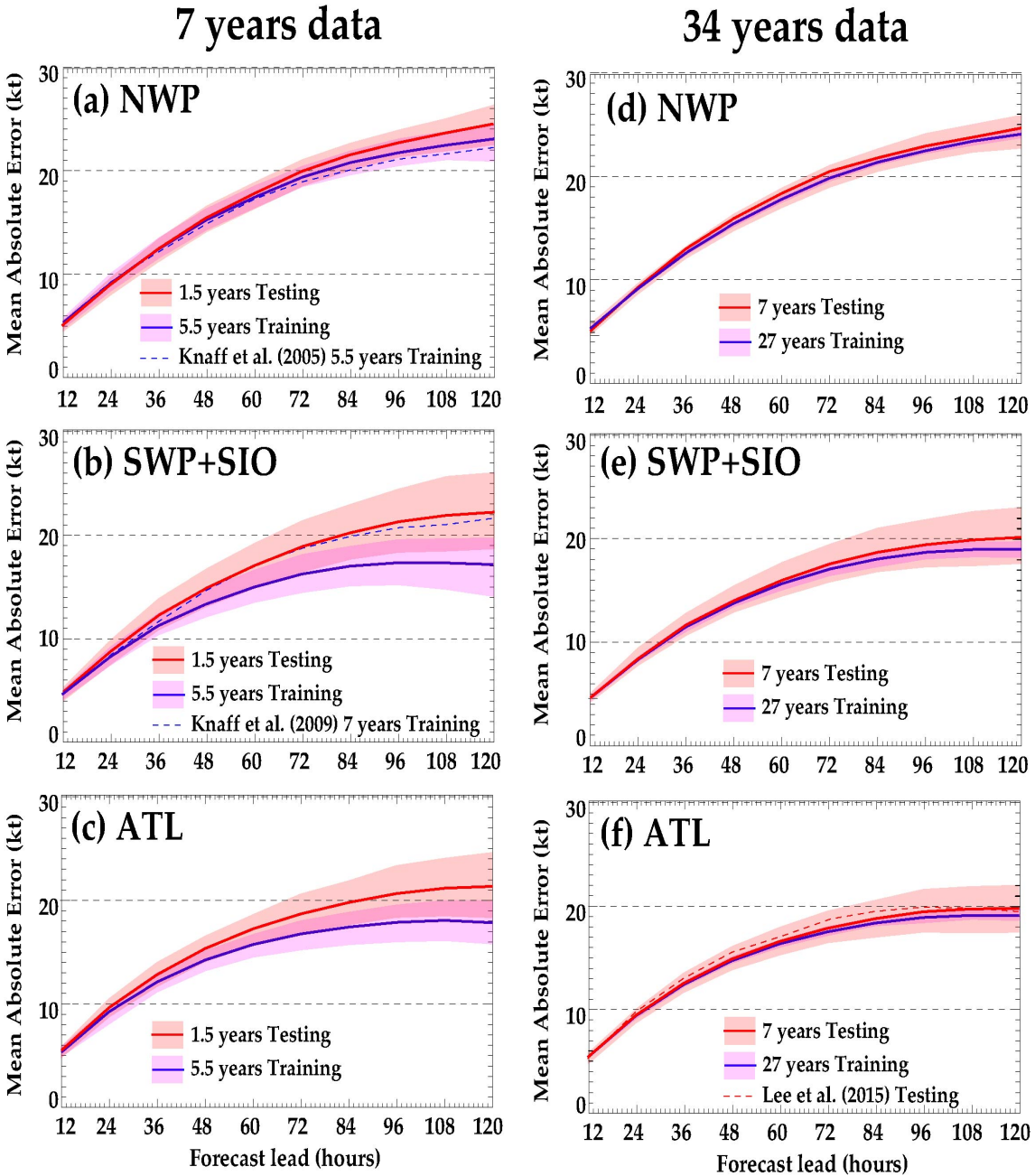
**Figure 2:**(a) Normalized distribution of TC intensity for each basin. Vertical dashed lines indicate the mean of the upper tenth percentile of the distribution for each basin. Normalized distribution of (b) SHRD, (c) MPI and (d) RHHI at 12h lead-time for each basin. Vertical dashed lines indicate the mean of the distribution for each basin.



**Figure 3:**(a) MAE as a function of lead-time, averaged over all basins for persistence (black line), Atm model (blue line) and Atm+Cross model (red line). (b) Percentage of Atm+Cross improvement relative to Atm at 24h, 60h and 108h lead-times, for each basin and globally. Atm+Cross model is similar to Atm model, except that the two cross terms (VMXS and VMXM) are included in the predictors list. Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.

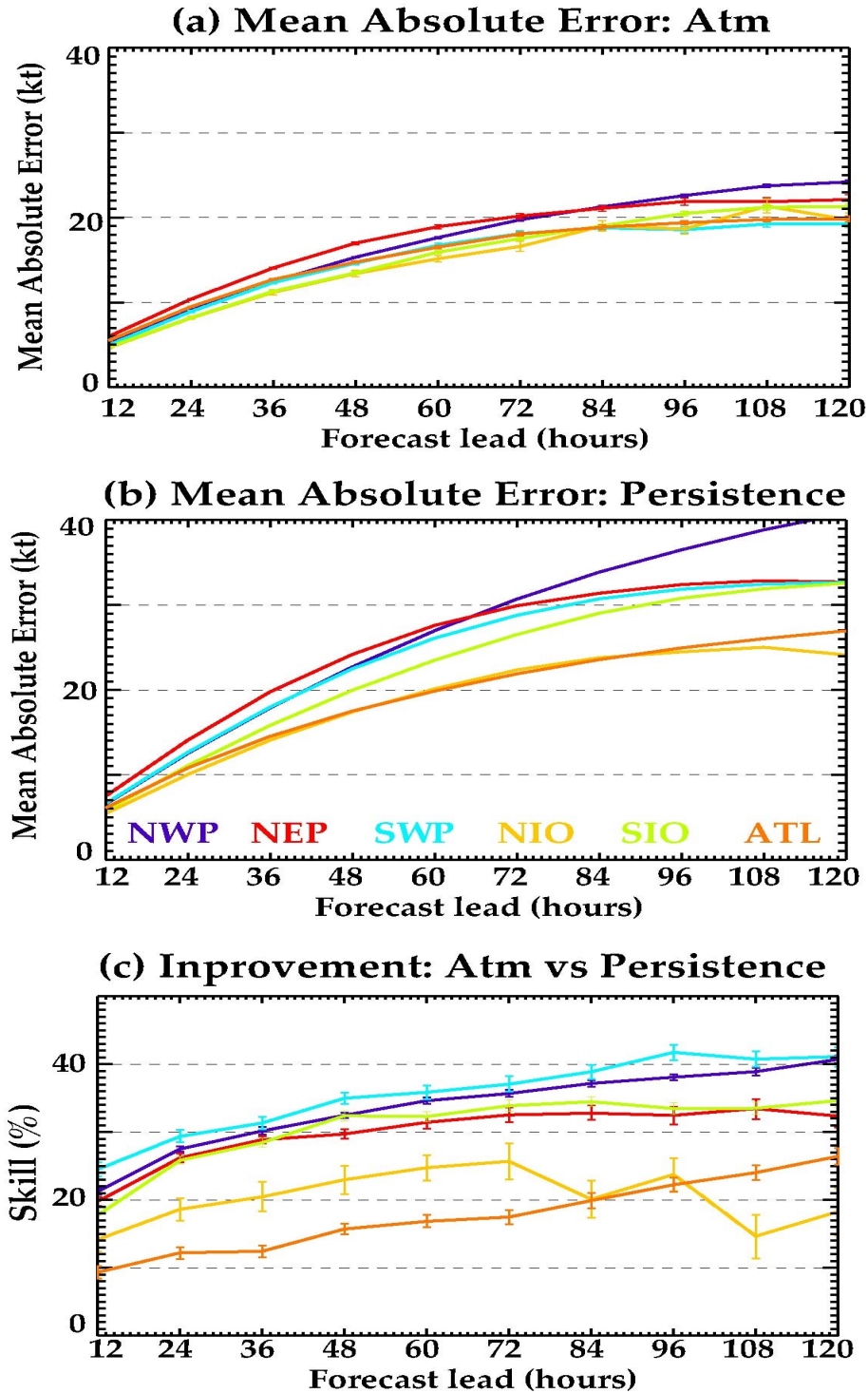


**Figure 4:** (Left panels) MAE as a function of lead-time for persistence (black line), Atm model (blue line) and previously published results (red line) in (a) the NWP for the training dataset (compared to Knaff et al. 2005; their Table 6), (c) in the southern hemisphere (SWP+SIO) for the training dataset (compared to Knaff et al. 2009; their Table 5) and (e) the ATL for the testing dataset (compared to Lee et al. 2015; their Figure 5).



**Figure 5:** (Left panels) MAE as a function of lead-time of the training dataset (blue line) and testing dataset (red line) for the Atm model built from a 7 year period in the (a) NWP, (b) SWP+SIO and (c) ATL. MAEs from already published results are also shown as dashed lines. (Right panels) Same as left panels but for a model build over a 34 years period. Thick lines indicate MAE resulting from averaging results from 1000 simulations with different randomly selected training/testing datasets while shading indicate the 5% lower and 95% higher bound of the ensemble distribution.

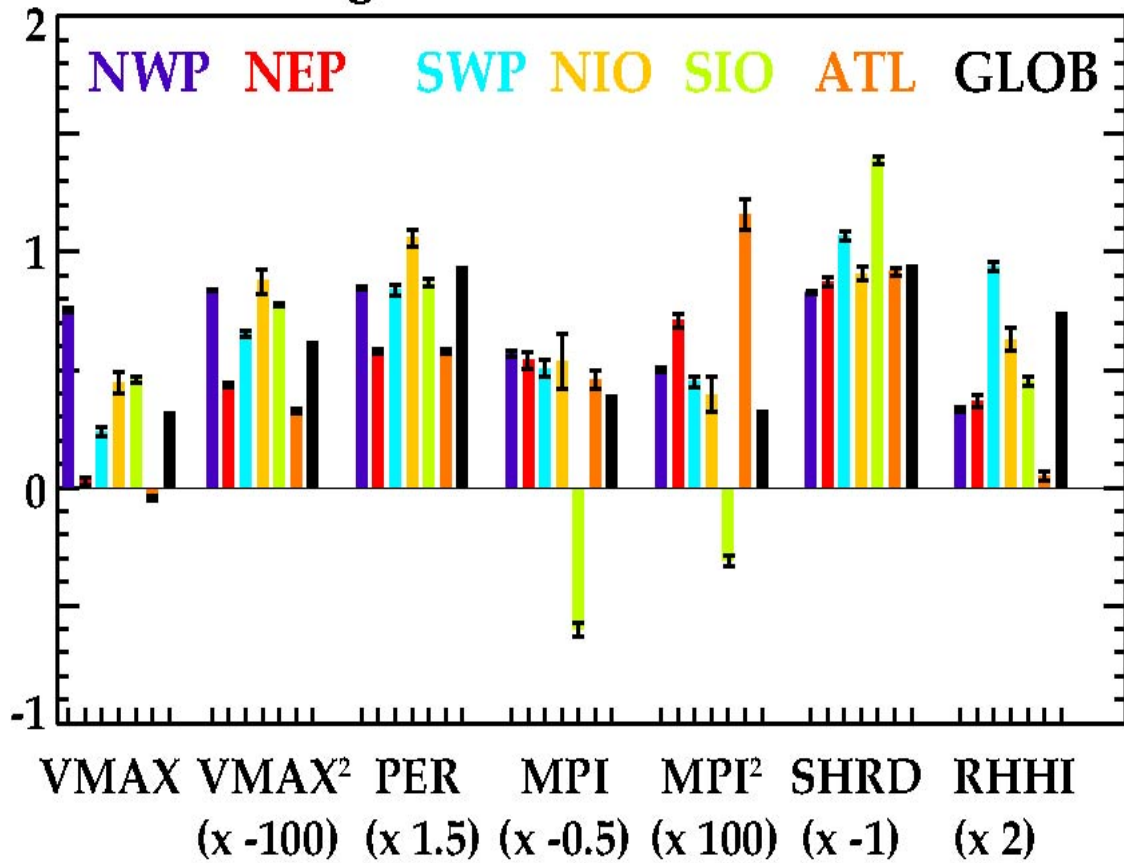




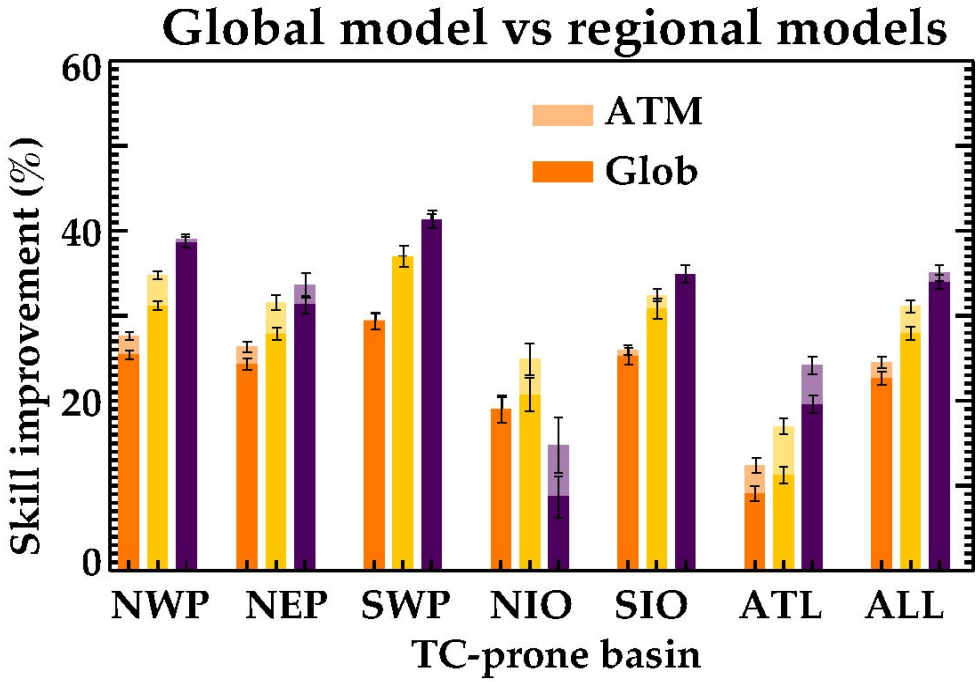
**Figure 6:** MAE as a function of lead-time for **(a)** Atm model and **(b)** persistence for each basin (i.e. NWP, NEP, SWP, ATL, SIO, NIO). **(c)** Percentage of skill improvement of these models relative to persistence as a function of lead-time for each basin. See section 2.2 for a definition of persistence metric and the skill metrics shown on panel c. Error bars on panels a and c give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.



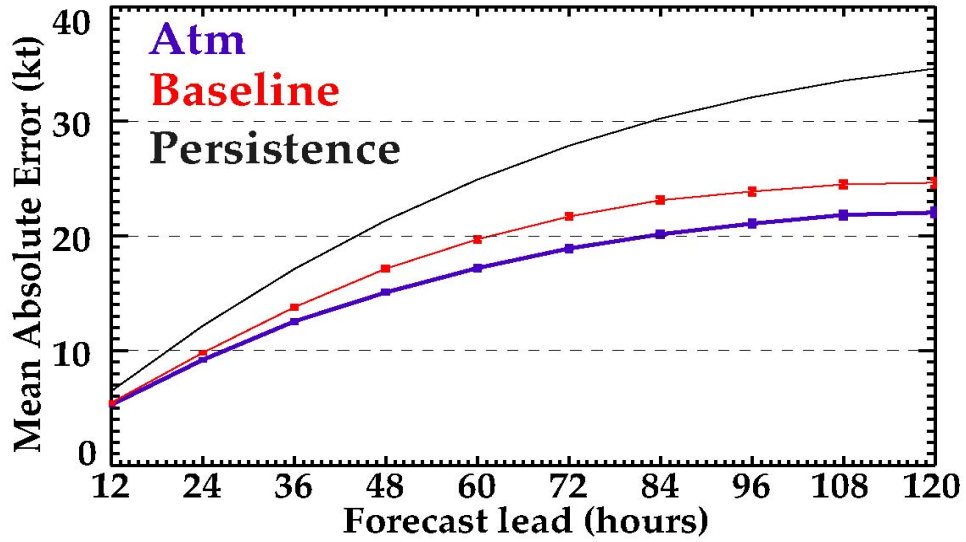
## Regression coefficients



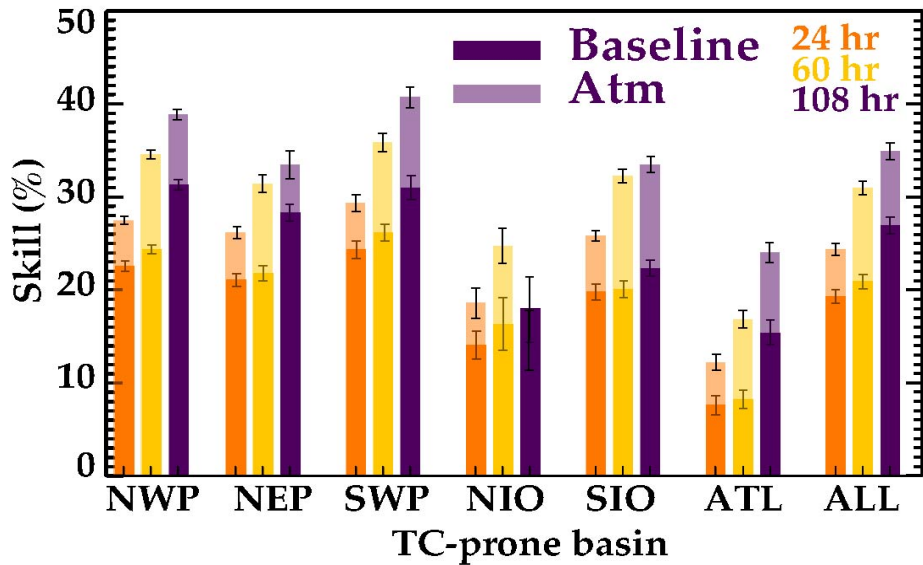
**Figure 7:(a)** Regression coefficients for the key predictors used in the Atm model at 60h for each basin. Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2. These regression coefficients have been multiplied by 100, 1.5, -0.5, 100, -1 and 2 for VMAX2, PER, MPI, MPI2, SHRD and RHHI respectively for a better readability.



**Figure 8:** Percentage of model skill improvement relative to persistence for basin-wise trained Atm models (light colors; see Table 2) and the globally-trained Glob model (plain colors; see Table 2) at 24h, 60h and 108h, as a function of the basin. Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.

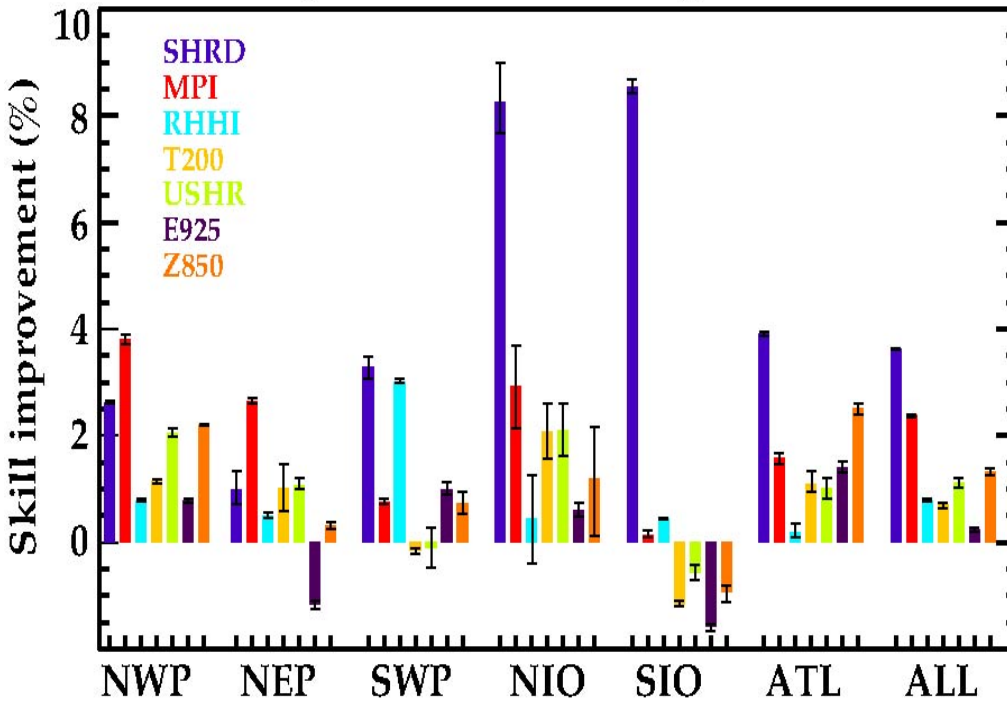


(b) Improvement relative to persistence

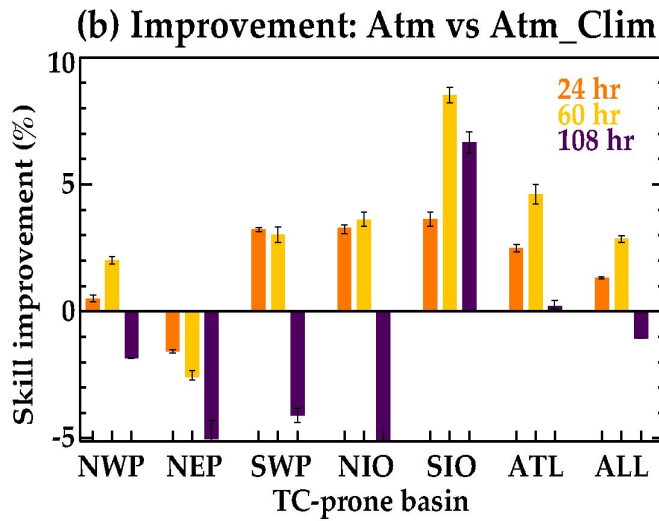
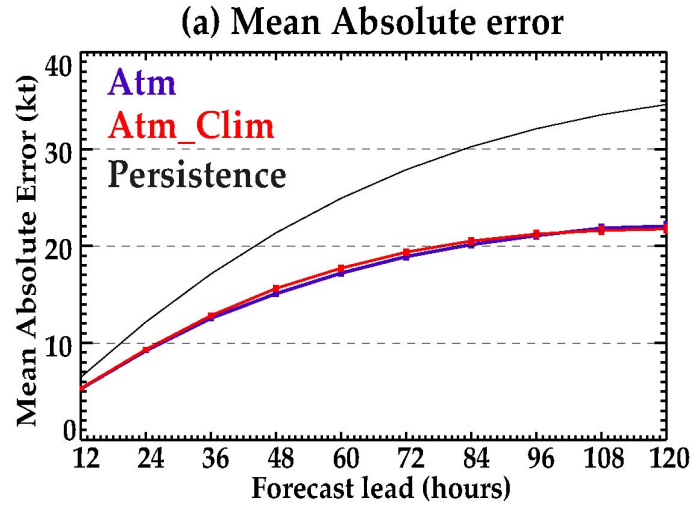


**Figure 9:**(a)MAE as a function of lead-time, averaged over all basins, for persistence (black curve), Atm model (blue curve) and baseline model (red curve). **(b)** Percentage of skill improvement of baseline (plain colors) and Atm models (light colors) relative to persistence at 24h, 60h and 108h lead for each basin and globally. Only the TC initial characteristics (predictors 1 to 3 in Table 1) are used as predictors in the baseline model (cf table 2). Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.

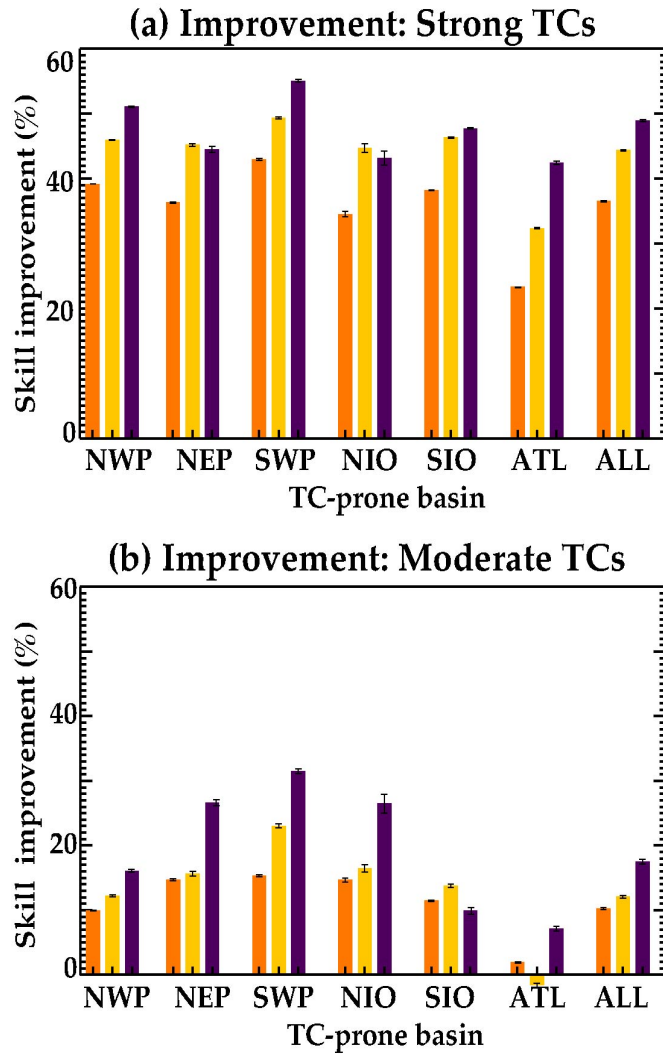
## Relative importance of atmospheric variables



**Figure 10:** Percentage of skill improvement of Atm relative to Atm-VarN (see table 2) at 60h for each basin and globally. This is a measure of the respective contributions of the SHRD, MPI, PHHI, T200, USHR, E925 and Z850 predictors to the overall skill. Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.



**Figure 11:**(a)MAE as a function of lead-time, averaged over all basins, for persistence (black curve), Atm model (blue curve) and Atm\_Clim model (red curve). (b) Percentage of skill improvement of Atm relative to Atm\_Clim at 24h, 60h and 108h lead-times for each basin and globally. In Atm\_Clim, the environmental parameters are calculated from their climatology, rather than on their actual value at the forecast time (cf. Table 2). Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.



**Figure 12:**Percentage of Atm model skill improvement relative to persistence for **(a)** strong and **(b)** moderate TC at 24h, 60h and 108h, as a function of the basin. Here, TC up to category 2 (< 96 kt) are considered as “moderate” and those under category 3-5 ( $\geq 96$  kt) are considered as “strong”. Error bars give the 95% confidence interval estimated from a bootstrap technique detailed in section 2.2.