HURRICANE FORECASTING: A NOVEL MULTIMODAL MACHINE LEARNING APPROACH

Léonard Boussioux*, **Cynthia Zeng***
Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA
leobix@mit.edu

Théo Guénais
School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

Dimitris Bertsimas
Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA

ABSTRACT

This paper describes a novel machine learning (ML) framework for tropical cyclone intensity and track forecasting, combining multiple ML techniques and utilizing diverse data sources. Our multimodal framework, called Hurricast, efficiently combines spatial-temporal data with statistical data by extracting features with deep-learning encoder-decoder architectures and predicting with gradient-boosted trees. We evaluate our models in the North Atlantic and Eastern Pacific basins on 2016-2019 for 24-hour lead time track and intensity forecasts and show they achieve comparable mean absolute error to current operational forecast models while computing in seconds. Furthermore, the inclusion of Hurricast into an operational forecast consensus model could improve over the National Hurricane Center's official forecast, thus highlighting the complementary properties with existing approaches. In summary, our work demonstrates that utilizing machine learning techniques to combine different data sources can lead to new opportunities in tropical cyclone forecasting.

1 INTRODUCTION

A tropical cyclone (TC) is a low-pressure system originating from tropical or subtropical waters and develops by drawing energy from the sea. It is characterized by a warm core, organized deep convection, and a closed surface wind circulation about a well-defined center. Every year, tropical cyclones cause hundreds of deaths and billions of dollars of damage to households and businesses (Grinsted et al., 2019). Therefore, producing an accurate prediction for TC track and intensity with sufficient lead time is critical to undertake life-saving measures.

The forecasting task encompasses the track, intensity, size, structure of TCs, and associated storm surges, rainfall, and tornadoes. Most forecasting models focus on producing track (trajectory) forecasts and intensity forecasts, i.e., intensity measures such as the maximum sustained wind speed in a particular time interval. Current operational TC forecasts can be classified into dynamical models, statistical models, and statistical-dynamical models (Cangialosi, 2020). Dynamical models, also known as numerical models, utilize powerful supercomputers to simulate atmospheric fields' evolution using dynamical and thermodynamical equations (Biswas et al., 2018; ECWMF, 2019). Statistical models approximate historical relationships between storm behavior and storm-specific features and, in general, do not explicitly consider the physical process (Aberson, 1998; Knaff et al., 2003). Statistical-dynamical models use statistical techniques but further include atmospheric variables provided by dynamical models (DeMaria et al., 2005). Lastly, consensus models typically

*Equal contribution

combine individual operational forecasts with a simple or weighted average (Cangialosi, 2020; Cangialosi et al., 2020).

In addition, recent developments in Deep Learning (DL) enable Machine Learning (ML) models to employ multiple data processing techniques to process and combine information from a wide range of sources and create sophisticated architectures to model spatial-temporal relationships. Several studies have demonstrated the use of Recurrent Neural Networks (RNNs) to predict TC trajectory based on historical data (Moradi Kordmahalleh et al., 2016; Gao et al., 2018; Alemany et al., 2019). Convolutional Neural Networks (CNNs) have also been applied to process reanalysis data and satellite data for track forecasting (Mudigonda et al., 2017; Lian et al., 2020; Giffard-Roisin et al., 2020) and storm intensification forecasting (Chen et al., 2019; Su et al., 2020).

This paper introduces a machine learning framework, called Hurricast (HUML), for both intensity and track forecasting by combining several data sources using deep learning architectures and gradient-boosted trees.

Our contributions are three-fold:

1. We present novel multimodal[1] machine learning techniques for TC intensity and track predictions by combining distinct forecasting methodologies to utilize multiple individual data sources. Our Hurricast framework employs XGBoost models to predict using statistical features based on historical data and spatial-temporal features extracted with deep learning encoder-decoder architectures from atmospheric reanalysis maps.
2. Evaluating in the North Atlantic (NA) and Eastern Pacific (EP) basins, we demonstrate that our machine learning models produce comparable results to currently operational models for 24-hour lead time for both intensity and track forecasting tasks.
3. Based on our testing, adding one machine learning model as an input to a consensus model can improve the performance, suggesting the potential for incorporating machine learning approaches for hurricane forecasting.

2 DATA

In this study, we employed three kinds of data dated since 1980: historical storm data, reanalysis maps, and operational forecast data. We use all storms from the seven TC basins since 1980 that reach 34 kt maximum intensity at some time, i.e., are classified at least as a tropical storm, and where more than 60 h of data are available after they reached the speed of 34 kt for the first time.

First, we collected historical storm data through the post-season storm analysis dataset IBTrACS (Knapp et al., 2010) maintained by the National Oceanic and Atmospheric Administration. Second, we obtained reanalysis data from the ERA-5 data set, which contains hourly high spatial resolution reanalysis maps (ERA5, 2017). We extracted nine reanalysis maps for each TC time step, corresponding to three different features, geopotential z, u and v components of the winds, at three atmospheric altitudes (see Figure 1).

Finally, to use as benchmark, we obtained operational forecast data from the Automated Tropical Cyclone Forecasting (ATCF) data set, maintained by the US National Hurricane Center (NHC) (Sampson & Schrader, 2000; National Hurricane Center, 2021). We selected the strongest operational forecasts with a sufficient number of cases concurrently available. More details on all data sources are in Appendix A.

3 METHODOLOGY

Our Hurricast framework makes predictions based on time-series data with different formats: three-dimensional vision-based reanalysis maps and one-dimensional historical storm data consisting of numerical and categorical features.

[1]Multimodality in machine learning refers to the simultaneous use of different data formats, including, for example, tabular data, images, time series, free text, audio.

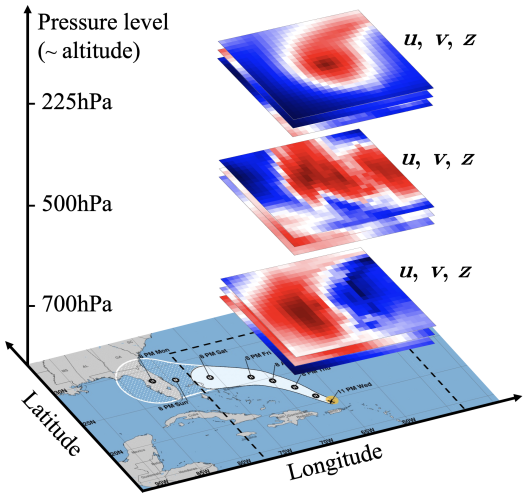

Figure 1: Representation of the nine reanalysis maps extracted for each time step, corresponding to geopotential z, u and v components of the winds, repeatedly extracted at three atmospheric altitudes, 225, 500, and 700hPa. Each map is of size $25° \times 25°$, centered on the TC center location, and each pixel corresponds to the average field value at the given latitude and longitude degree.

Overall, we adopt a three-step approach to combine the multiple data sources. We first extract a one-dimensional feature representation (embedding) from each reanalysis maps sequence, using encoder-decoder architectures. Second, we concatenate this one-dimensional embedding with the statistical data to form a one-dimensional vector. Third, we make our predictions using gradient-boosted tree XGBoost models (Chen & Guestrin, 2016) trained on the selected features.

At a given time step (forecasting case), we perform two 24-hour lead time forecasting tasks: intensity prediction, i.e., predicting the maximum sustained wind speed at a 24-hour lead time; and displacement prediction, i.e., the latitude and longitude storm displacement in degrees between given time and forward 24-hour time. Figure 2 illustrates the three-step pipeline.

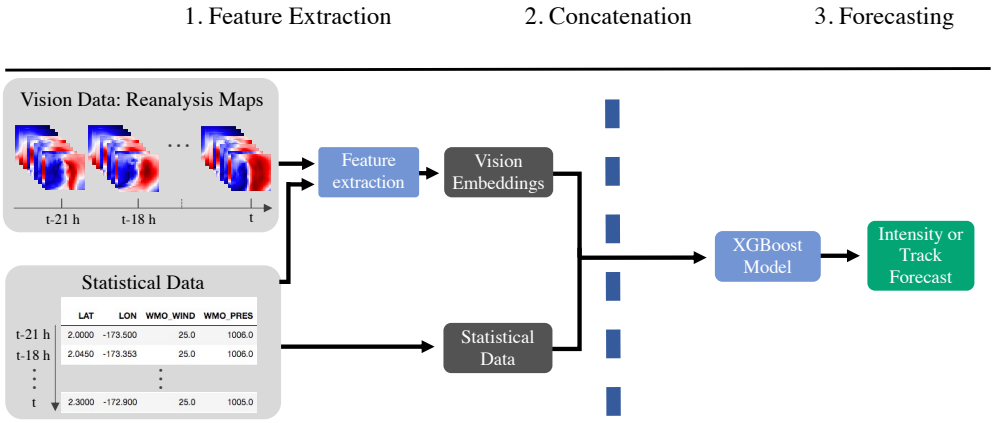

Figure 2: Representation of our multimodal machine learning framework using the two data sources.

3.1 Feature Extraction

We experimented with encoder-decoder architectures trained with a supervised learning fashion to perform feature extraction. The encoder component consists of a Convolutional Neural Network. We compare two decoder variations. The first one relies on Gated Recurrent Units (GRUs) (Chung et al., 2014), a well-suited recurrent neural network to model temporal dynamic behavior in sequen-

tial data. The second one uses Transformers (Vaswani et al., 2017), a state-of-the-art architecture for sequential data. While the GRUs model the temporal aspect through a recurrence mechanism, the Transformers utilize attention mechanisms and positional encoding (Bahdanau et al., 2015; Vaswani et al., 2017) to model long-range dependencies.

First, we trained the encoder-decoder architectures end-to-end using the Adam optimizer (Kingma & Ba, 2014). We used a mean squared error loss with either an intensity or track objective and added an $L2$ regularization on the network's weights. We then froze the encoder-decoder's weights after training was completed.

To perform feature extraction from a given input sequence of reanalysis maps and statistical data, we passed them through the whole frozen encoder-decoder, except the last fully-connected layer (see Figure 3). The second fully connected layer after the GRU or the pooling layer after the Transformer output a vector of relatively small size, e.g., 1024 or 128 features, to compress information and provide predictive features. This vector constitutes our one-dimensional reanalysis maps embedding that we extract from the initial 45,000 ($8 \times 9 \times 25 \times 25$) features forming the spatial-temporal input. Figure 3 illustrates the encoder-decoder architectures (see Appendix B for more details).

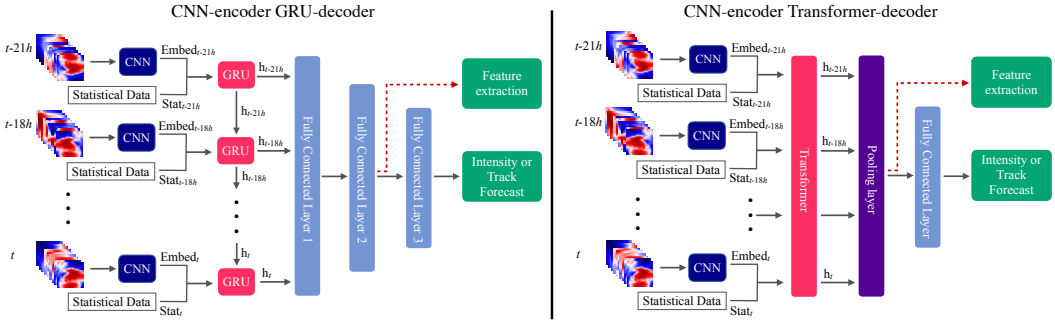

Figure 3: Schematic of the encoder-decoder networks for an 8-time step TC sequence. The CNN produces a one-dimensional representation of the reanalysis maps at each time step. Then, we concatenate these embeddings with the corresponding statistical features to create a sequence of inputs fed sequentially to the GRU or as a whole to the Transformer. In the GRU-decoder, the GRU sequentially outputs hidden states passed to the next time step. Finally, we concatenate all the successive hidden states and pass them through three fully connected layers to predict intensity or track with a 24-hour lead time. We extract our spatial-temporal embeddings as the output of the second fully connected layer. In the Transformer-decoder, the Transformer outputs a new 8-timestep sequence that we average feature-wise and feed into one fully connected layer to predict intensity or track. We extract our spatial-temporal embeddings as the output of the pooling layer.

4 Experiments and Results

We separated the data set chronologically into training (80% of the data, ranging from 1980 to 2011), validation (10% of the data, ranging from 2012 to 2015), and testing (10% of the data, ranging from 2016 to 2019). The training and validation sets consist of TCs from all basins, whereas we restrict the testing set to only NA and EP basins where operational forecasts are available to compare performances. We computed the mean absolute error (MAE) on the predicted 1-minute maximum sustained wind speed in 24 hours to evaluate our intensity forecasts' performance. To evaluate track forecasts' performance, we computed the mean geographical distance error in kilometers between the actual position and the predicted position in 24 hours, using the Haversine formula[2].

Standalone ML models produce a comparable performance to standalone operational models. For both intensity and track forecasting tasks, the best Hurricast model HUML-(stat/viz, xgb/cnn/transfo) achieve better performance than operational statistical models, and comparable

[2]The Haversine metric calculates the great-circle distance between two points — in our case, the shortest distance between these two points over the Earth's surface.

results to operational dynamical models. Results for 24-hour lead time intensity and track forecasting are displayed in Table 1 and Table 2 respectively. Details on the operational models and HUML models can be found in Appendix in Tables 4 and 5. These results highlight that machine learning approaches can emerge as a new methodology to currently existing forecasting methodologies in the field. In addition, we believe there is potential for improvement if given more available data sources.

Machine learning models bring additional insights to consensus models. Consensus models often perform better than individual models by averaging errors and biases. In particular, the official NHC forecast (OFCL) is a consensus model. Therefore, to experiment if ML models can complement operational models to improve the official forecast, we built two types of consensus models: (i) OP-consensus, a simple average of the standalone operational models included in our benchmark; (ii) HUML/OP-consensus, a simple average of HUML-(stat/viz, xgb/cnn/transfo) and the other standalone operational models included in our benchmark. As shown in Tables 1 and 2, consensus models have a lower MAE than standalone models. In addition, the HUML/OP-average consensus demonstrates the potential to improve the NHC Official Forecast (OFCL) by including an ML model.

Table 1: Mean absolute error (MAE) and standard deviation of the error (Error sd) of standalone Hurricast models, standalone operational forecasts, operational consensus, and consensus experiments on the same test set between 2016 and 2019, for 24-hour lead time intensity forecasting task. Bold values highlight the best performance in each category.

Model Type	Model Name	Eastern Pacific Basin Comparison on 877 cases		North Atlantic Basin Comparison on 899 cases	
		MAE (kt)	Error sd (kt)	MAE (kt)	Error sd (kt)
Hurricast (HUML) Methods	HUML-(stat, xgb)	10.6	10.5	10.7	9.3
	HUML-(stat/viz, xgb/cnn/gru)	**10.3**	10.0	10.8	9.2
	HUML-(stat/viz, xgb/cnn/transfo)	**10.3**	**9.8**	**10.4**	**8.8**
Standalone Operational Forecasts	GFSO	15.7	14.7	14.2	14.1
	Decay-SHIPS	11.7	**10.4**	10.2	9.3
	HWRF	**10.6**	11.0	**9.7**	**9.0**
Operational Consensus	FSSE	**9.7**	**9.5**	**8.5**	**7.8**
	OFCL	10.0	10.1	**8.5**	8.1
Consensus Experiments	OP-average consensus	9.6	9.7	8.5	7.9
	HUML/OP-average consensus	**9.2**	**9.0**	**8.3**	**7.6**

Table 2: Mean absolute error (MAE) and standard deviation of the error (Error sd) of standalone Hurricast models, standalone operational forecasts, operational consensus, and consensus experiments on the same test set between 2016 and 2019, for 24-hour lead time track forecasting task. Bold values highlight the best performance in each category.

Model Type	Model Name	Eastern Pacific Basin Comparison on 837 cases		North Atlantic Basin Comparison on 899 cases	
		MAE (km)	Error sd (km)	MAE (km)	Error sd (km)
Hurricast (HUML) Methods	HUML-(stat, xgb)	81	47	144	108
	HUML-(stat/viz, xgb/cnn/gru)	**72**	**43**	111	79
	HUML-(stat/viz, xgb/cnn/transfo)	**72**	**43**	**109**	**71**
Standalone Operational Forecasts	CLP5	121	67	201	149
	HWRF	67	42	75	**49**
	GFSO	65	45	**71**	54
	AEMN	**60**	**37**	73	55
Operational Consensus	FSSE	56	47	**69**	**53**
	OFCL	**54**	**33**	71	56
Consensus Experiments	OP-average consensus	55	37	64	48
	HUML/OP-average consensus	**50**	**32**	**61**	**42**

5 CONCLUSION

This study develops a novel multimodal machine learning framework for tropical cyclone intensity and track forecasting utilizing historical storm data and meteorological reanalysis data. We present

a three-step pipeline to combine multiple machine learning approaches, consisting of (1) deep feature extraction, (2) concatenation of all processed features, (3) prediction. We demonstrate that a successful combination of deep learning techniques and gradient-boosted trees can achieve strong predictions for both track and intensity forecasts, producing comparable results to current operational forecast models, especially in the intensity task.

We show that multimodal encoder-decoder architectures can successfully serve as a spatial-temporal feature extractor for downstream prediction tasks. In particular, this is also the first successful application of a Transformer-decoder architecture in tropical cyclone forecasting. Furthermore, consensus models that include machine learning models could benefit the NHC's official forecast for both intensity and track, thus demonstrating the potential value of developing machine learning approaches as a new branch methodology for tropical cyclone forecasting. Moreover, once trained, our models run in seconds, showing practical interest for real-time forecast, the bottleneck lying only in the data acquisition. In conclusion, our work demonstrates that machine learning can be a valuable approach to providing alternatives in tropical cyclone forecasting.

REFERENCES

Sim D. Aberson. Five-day tropical cyclone track forecasts in the north atlantic basin. *Weather and Forecasting*, 13(4):1005 – 1015, 1998.

Sheila Alemany, Jonathan Beltran, Adrian Perez, and Sam Ganzfried. Predicting hurricane trajectories using a recurrent neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 468–475, 2019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

Gu-Feng Bian, Gao-Zhen Nie, and Xin Qiu. How well is outer tropical cyclone size represented in the era5 reanalysis dataset? *Atmospheric Research*, 249:105339, 2021. ISSN 0169-8095.

Mrinal K Biswas, Sergio Abarca, Bernardet Ligia, Ginis Isaac, Grell Evelyn, Iacono Michael, Kalina Evan, Liu Bin, Avichal Mehra Kathryn Newman Jason Sippe Vijay Tallapragada Biju Thomas Weiguo Wang Henry Winterbottom Qingfu, Liu Timothy Marchok, and Zhan Zhang. Hurricane weather research and forecasting (hwrf) model: 2018 scientific documentation. *Developmental Testbed Center*, 2018.

John P. Cangialosi. National hurricane center forecast verification report. *National Hurricane Center*, 2020. URL `https://www.nhc.noaa.gov/verification/pdfs/Verification_2020.pdf`.

John P. Cangialosi, Eric Blake, Mark DeMaria, Andrew Penny, Andrew Latto, Edward Rappaport, and Vijay Tallapragada. Recent Progress in Tropical Cyclone Intensity Forecasting at the National Hurricane Center. *Weather and Forecasting*, pp. 1–30, 07 2020. ISSN 0882-8156.

Rui Chen, Xiang Wang, Weimin Zhang, Xiaoyu Zhu, Aiping Li, and Chao Yang. A hybrid cnn-lstm model for typhoon formation forecasting. *GeoInformatica*, 2019.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794. ACM, 2016. ISBN 978-1-4503-4232-2.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.

Mark DeMaria and John Kaplan. A statistical hurricane intensity prediction scheme (ships) for the atlantic basin. *Weather and Forecasting*, 9(2):209–220, 1994.

Mark DeMaria, Michelle Mainelli, Lynn K. Shay, John A. Knaff, and John Kaplan. Further improvements to the statistical hurricane intensity prediction scheme (ships). *Weather and Forecasting*, 20(4):531 – 543, 2005.

ECWMF. *PART III: Dynamics and Numerical Procedures*. Number 3 in IFS Documentation. ECMWF, 2019. URL https://www.ecmwf.int/node/19307.

ERA5. Era5 reanalysis, 2017. URL https://doi.org/10.5065/D6X34W69. Accessed 2020-09-20.

Song Gao, Peng Zhao, Bin Pan, Yaru Li, Min Zhou, Jiangling Xu, Shan Zhong, and Zhenwei Shi. A nowcasting model for the prediction of typhoon tracks based on a long short term memory neural network. *Acta Oceanologica Sinica*, 37:8–12, 2018.

Sophie Giffard-Roisin, Mo Yang, Guillaume Charpiat, Christina Kumler Bonfanti, Balázs Kégl, and Claire Monteleoni. Tropical cyclone track forecasting using fused deep learning from aligned reanalysis data. *Frontiers in Big Data*, 3:1, 2020. ISSN 2624-909X.

Aslak Grinsted, Peter Ditlevsen, and Jens Hesselbjerg Christensen. Normalized us hurricane damage estimates using area of total destruction, 1900-2018. *Proceedings of the National Academy of Sciences*, 116(48):23942–23946, 2019. ISSN 0027-8424.

Kevin Hodges, Alison Cobb, and Pier Luigi Vidale. How well are tropical cyclones represented in reanalysis datasets? *Journal of Climate*, 30(14):5243 – 5264, 2017.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

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

K. R. Knapp, M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann. The international best track archive for climate stewardship (ibtracs): Unifying tropical cyclone best track data, 2010.

Jie Lian, Pingping Dong, Yuping Zhang, Jianguo Pan, and Kehao Liu. A novel data-driven tropical cyclone track prediction model based on cnn and gru with multi-dimensional feature selection. *IEEE Access*, 2020.

Mina Moradi Kordmahalleh, Mohammad Gorji Sefidmazgi, and Abdollah Homaifar. A sparse recurrent neural network for trajectory prediction of atlantic hurricanes. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, pp. 957–964, 2016.

M. Mudigonda, Sookyung Kim, A. Mahesh, S. Kahou, K. Kashinath, D. Williams, Vincent Michalski, T. O'Brien, and M. Prabhat. Segmenting and tracking extreme climate events using neural networks. 2017.

National Hurricane Center. Automated tropical cyclone forecasting system (atcf), 2021. URL https://ftp.nhc.noaa.gov/atcf/. Accessed: 2021-04-06.

C. Sampson and Ann J. Schrader. The automated tropical cyclone forecasting system (version 3.2). *Bulletin of the American Meteorological Society*, 81:1231–1240, 2000.

Benjamin A. Schenkel and Robert E. Hart. An examination of tropical cyclone position, intensity, and intensity life cycle within atmospheric reanalysis datasets. *Journal of Climate*, 25(10):3453 – 3475, 2012.

Udai Shimada, Hiromi Owada, Munehiko Yamaguchi, Takeshi Iriguchi, Masahiro Sawada, Kazumasa Aonashi, Mark DeMaria, and Kate D. Musgrave. Further Improvements to the Statistical Hurricane Intensity Prediction Scheme Using Tropical Cyclone Rainfall and Structural Features. *Weather and Forecasting*, 33(6):1587–1603, 11 2018. ISSN 0882-8156.

Hui Su, Longtao Wu, Jonathan H. Jiang, Raksha Pai, Alex Liu, Albert J. Zhai, Peyman Tavallali, and Mark DeMaria. Applying satellite observations of tropical cyclone internal structures to rapid intensification forecast with machine learning. *Geophysical Research Letters*, 47(17), 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. 2017.

A DATA

A.1 HISTORICAL STORM DATA SET

We obtained historical storm data from the National Oceanic and Atmospheric Administration through the post-season storm analysis dataset IBTrACS (Knapp et al., 2010). Among the available features, we have selected time, latitude, longitude, and minimum pressure at the center of the TC, distance-to-land, translation speed of the TC, direction of the TC, TC type (disturbance, tropical, extra-tropical, etc.), basin (North-Atlantic, Eastern Pacific, Western Pacific, etc), and maximum sustained wind speed from the WMO agency (or from the regional agency when not available). Our overall feature choice is consistent with previous statistical forecasting approaches (DeMaria & Kaplan, 1994; DeMaria et al., 2005; Giffard-Roisin et al., 2020). In this paper, we refer to this data as *statistical data* (see Table 3).

We processed statistical data through several steps before inputting it into machine learning models. First, we treated the categorical features using the one-hot encoding technique: for a specific categorical feature, we converted each possible category as an additional binary feature, with 1 indicating the sample belongs to this category and 0 otherwise. We encoded the basin and the nature of the TC as one-hot features. Second, we encoded cyclical features using cosine and sine transformations to avoid singularities at endpoints. Features processed using this smoothing technique include date, latitude, longitude, and storm direction.

Table 3: List of features included in our statistical data.

Feature	Range	Unit	Type	Processing	Description
Latitude	[-90.0, 90.0]	deg north	numerical	spline interpolation by IBTrACS, standardize	Latitude of the center of the hurricane.
Longitude	[-180.0, 180.0]	deg east	numerical	spline interpolation by IBTrACS, standardize	Longitude of the center of the hurricane.
WMO Wind	[10, 165]	knots	numerical	linear interpolation, conversion to 1-min, standardize	Maximum sustained wind speed from the WMO agency for the current location.
WMO Pressure	[880, 1022]	mb	numerical	linear interpolation, standardize	Wind pressure from the WMO agency for the current location.
Distance to Land	[0, 4821]	km	numerical	standardize	Distance to land from the current position. The IBTrACS land mask includes islands larger than 1400 km^2.
Storm Speed	[0, 69]	knots	numerical	standardize	Translation speed of the system as calculated from the positions in latitude and longitude.
Storm Direction	[0, 360]	deg	numerical	cosine & sine encoding	Translation direction of the system, as calculated from the positions, pointing in degrees east of north.
Storm Displacement Latitude	[-2.68, 3.13]	deg	numerical	standardize	Engineered feature, indicating latitude change since the last time step (3 hours ago).
Storm Displacement Longitude	[-3.83, 4.28]	deg	numerical	standardize	Engineered feature, indicating longitude change since the last time step (3 hours ago).
Basin	[NA, EP, WP, NI, SI, SP, SA]	N/A	categorical	one-hot encoding	Basins include: NA - North Atlantic, EP - Eastern North Pacific, WP - Western North Pacific, NI - North Indian, SI - South Indian, SP - Southern Pacific, SA - South Atlantic
Storm Type	[DS, TS, ET, SS, MX]	N/A	categorical	one-hot encoding	Storm types include: DS - Disturbance, TS - Tropical, ET - Extratropical, SS - Subtropical, NR - Not reported, MX - Mixture (contradicting nature reports from different agencies)

A.2 REANALYSIS MAPS

Reanalysis maps are used extensively for atmospheric monitoring, climate research, and climate predictions. They are assimilated using observational data and provide a comprehensive record of how weather and climate evolve, based on dynamical aspects of the Earth systems, such as the air pressure, humidity, and wind speed.

In our work, we used the extensive ERA5 reanalysis data set (ERA5, 2017) developed by the European Centre for Medium-Range Weather Forecasts (ECWMF). ERA5 provides hourly estimates of a large number of atmospheric, land, and oceanic climate variables. The data cover the Earth on a 30 km grid and resolve the atmosphere using 137 levels from the surface up to a height of 80 km.

We extracted $(25^\circ \times 25^\circ)$ maps centered at the storm locations across time, given by the IBTrACS dataset described previously, of resolution $1^\circ \times 1^\circ$, i.e., each cell corresponds to one degree of latitude and longitude, offering a sufficient frame size to capture the entire storm. We obtained nine reanalysis maps for each TC time step, corresponding to three different features, geopotential z, u and v components of the winds, at three atmospheric altitudes, 225, 500, and 700 hPa (see Figure 1). We chose the three features to incorporate physical information which would influence the TC evolution, and this choice is motivated by previous literature in applying ML techniques to process reanalysis maps (Shimada et al., 2018; Chen et al., 2019; Giffard-Roisin et al., 2020).

As a remark, we acknowledge two main limitations from using reanalysis maps for TC forecasting. First, since they are reanalysis products, they are not available in real-time and thus significantly hinder operational use. Second, they have deficiencies in representing tropical cyclones (Schenkel & Hart, 2012; Hodges et al., 2017; Bian et al., 2021); for instance, with large TC sizes particularly being underestimated (Bian et al., 2021).

A.3 OPERATIONAL FORECAST MODELS

We obtained operational forecast data from the ATCF data set, maintained by the National Hurricane Center (NHC) (Sampson & Schrader, 2000; National Hurricane Center, 2021). The ATCF data contains historical forecasts by operational models used by the NHC for its official forecasting for tropical cyclones and subtropical cyclones in the North Atlantic and Eastern Pacific basins. To compare the performance of our models with a benchmark, we selected the strongest operational forecasts with a sufficient number of cases concurrently available: including Decay-SHIPS, GFSO, HWRF, FSSE, and OFCL for the intensity forecast; CLP5, HWRF, GFSO, AEMN, FSSE, and OFCL for the track forecast (see detailed list in Table 4).

Table 4: Summary of all operational forecast models included in our benchmark.

Model ID	Model name or type	Model type	Forecast
CLP5	CLIPER5 Climatology and Persistence	Statistical (baseline)	Track
Decay-SHIPS	Decay Statistical Hurricane Intensity Prediction Scheme	Statistical-dynamical	Intensity
GFSO	Global Forecast System model	Multi-layer global dynamical	Track, Intensity
HWRF	Hurricane Weather Research and Forecasting model	Multi-layer regional dynamical	Track, Intensity
AEMN	GFS Ensemble Mean Forecast	Ensemble	Track
FSSE	Florida State Super Ensemble	Corrected consensus	Track, Intensity
OFCL	Official NHC Forecast	Consensus	Track, Intensity

B ENCODER-DECODER ARCHITECTURES DETAILS

The CNN-encoder At each time step, the corresponding nine reanalysis maps are fed into the CNN-encoder, which produces one-dimensional embeddings. The CNN-encoder consists of three convolutional layers, with ReLU activation and MaxPool layers in between, then followed by two fully connected layers.

Next, we concatenate the reanalysis maps embeddings with processed statistical data corresponding to the same time step. Note that at this point data is still sequentially structured as 8 time steps to be passed on to the GRU-decoder or the Transformer-decoder.

The GRU-decoder Our GRU-decoder consists of two unidirectional layers. The data sequence embedded by the encoder is fed sequentially in chronological order into the GRU-decoder. For each time step, the GRU-decoder outputs a hidden state representing a "memory" of the previous time steps. Finally, a track or intensity prediction is made based upon these hidden states concatenated all together and given as input to fully-connected layers (see Figure 3).

The Transformer-decoder Conversely to the GRU-decoder, the sequence is fed as a whole into the Transformer-decoder. The time-sequential aspect is lost since attention mechanisms allow each hidden representation to attend holistically to the other hidden representations. Therefore, we add a *positional encoding* token at each timestep-input, following standard practices (Vaswani et al., 2017). This token represents the relative position of a time-step within the sequence and re-introduces some information about the inherent sequential aspect of the data and experimentally improves performance.

Then, we use two Transformer layers that transform the 8 time steps (of size 142) into an 8-timestep sequence with similar dimensions. To obtain a unique representation of the sequence, we average the output sequence feature-wise into a one-dimensional vector, following standard practices. Finally, a track or intensity prediction is made based upon this averaged vector input into one fully-connected layer.

B.1 Summary of Hurricast Models

This section lists all the Hurricast models reported in this paper and Table 5 summarizes the methodologies employed.

Table 5: Summary of the various versions of the Hurricast framework for which we report results. Models differ in architecture and data used and are named based on these two characteristics.

N°	Name	Data Used	ML Methods
1	HUML-(stat, xgb)	Statistical	XGBoost
2	HUML-(stat/viz, xgb/cnn/gru)	Statistical, Vision embeddings	XGBoost, Feature extraction with CNN, GRU
3	HUML-(stat/viz, xgb/cnn/transfo)	Statistical, Vision embeddings	XGBoost, Feature extraction with CNN, Transformers
4	HUML/OP-average	Operational forecasts, HUML-(stat/viz, xgb/cnn/transfo)	Simple average

Models 1-3 are variations of the three-step framework described in Figure 2, with the variation of input data source or processing technique. Model 1, HUML-(stat, xgb), has the simplest form, utilizing only statistical data. Models 2-3 utilize statistical and vision data and are referred to as multimodal models. They utilize vision features extracted with the encoder-decoder, with GRU and Transformer decoders, respectively. Model 4 is a simple average consensus of a few operational forecasts models used by the NHC and our Model 3, HUML-(stat/viz, xgb/cnn/transfo). We use Model 4 to explore whether the Hurricast framework can benefit current operational forecasts by comparing its inclusion as a member model.