A STOCHASTIC DISCREPANCY DATA-PHYSICS FUSION FRAMEWORK FOR TROPICAL CYCLONE MODEL AND APPLICATIONS TO TYPHOON WIND HAZARD ASSESSMENT

X. ZHONG[1] AND J. ZHANG[2]

[1]Department of Civil and Environmental Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China
xzhongan@connect.ust.hk

[2]Department of Civil and Environmental Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China
cejize@ust.hk

Key words: Data-Physics Fusion Framework, Tropical Cyclone, Hazard Assessment.

Abstract. *This study proposes a stochastic discrepancy framework to enhance physics-based tropical cyclone (TC) models and risk assessments by systematically integrating observational data. Current deterministic TC models exhibit structural biases from oversimplified physics and fail to quantify epistemic uncertainties inherent in their formulations. We reframe existing models as prior knowledge and derive parameter-efficient stochastic governing equations to characterize their discrepancies against TC observations. Validated through hindcasts of historical TCs in the Western North Pacific, the framework improves track and intensity simulations while quantifying uncertainty. For wind hazard assessment, our estimated wind speed generally agrees with the code recommendations with well calibrated confidence intervals to include results from alternative models. Overall, the proposed framework provides a principled approach to enhance the physics-based TC models, paving the future way for more informed TC simulation under changing climates.*

1 INTRODUCTION

Tropical cyclones (TCs) are highly destructive natural phenomena that may form strong winds, huge waves, extreme rainfall, and storm surges, posing significant threats to coastal communities. It is of the utmost importance to reliably assess the potential TC-induced risks. Due to the shortage of reliable observations, TC risk assessment usually requires generating synthetic TC scenarios to enlarge the dataset [1]. Full-track models utilize basin-wise historical storms to simulate synthetic TCs from genesis to lysis [2], consisting of three components: a genesis model for storm formation, a track model for storm propagation, and an intensity model for storm intensity evolution along the track. Our focus is on track and intensity components, both essential for full-track TC dynamic propagation.

For TC track and intensity models, an ideal trait is the dependence on storm environment variables to accommodate changing climates. However, existing approaches, e.g., regression-based empirical track methods [2], heavily rely on historical TC characteristics with limited use

of environmental variables such as sea surface temperature and mix-layer depth. Their simulations can match historical observations, but the lack of TC-environment interactions makes them less robust, interpretable, and unsuitable under changing climates.

As possible remedies, physics-based TC models have been recently investigated for their ability to generate synthetic TCs driven by local climate conditions. This is pioneered by the statistical-deterministic model [3]. Storm tracks will be propagated by a beta-and-advection model from local winds, while their intensities will evolve under a deterministic physics-based model [4,5] based on local thermodynamic states. While physics-based TC models have demonstrated great potential, their performance leaves room for improvement, mainly due to the bias from oversimplified physics [3]. Modifications are proposed [6,7], but their strategies, such as segmenting the basin for spatially heterogenous (regionalized) correction [8], will also face similar robust challenges in data-scarce regions or under changing climates, akin to the empirical track methods. In addition, basin segmentation might lack a consistent protocol to guarantee reliable results across basins or users. Beyond bias correction, another important challenge is to quantify the epistemic uncertainties in TC models stemming from their inadequacy, which will be essential to ensure adequate and effective communications to stakeholders.

Acknowledging the challenges, this paper aims to make the following novel contribution:

- Leveraging observation data, we aim to establish simultaneous bias correction and epistemic uncertainty quantification for physics-based TC models.
- Treating physics-based model as our prior knowledge, its simulated discrepancy versus TC observations will be used for data-driven discovery of stochastic governing equations.
- Empowered by symbolic regression, we identify stochastic, parsimonious yet interpretable equations, facilitating uncertainty-aware risk assessment under present and future climate.

2 DATA DESCRIPTION

In this study, two types of data for the Western North Pacific (WNP) basin are utilized: historical TC records and large-scale environmental variables, during the historical period of 1980–2014. For the historical TC record, track and intensity information at 6-hourly intervals were extracted from the best track dataset of the China Meteorological Administration (CMA). For the large-scale atmospheric and oceanic conditions, we utilized the reanalysis data from ERA5 (https://cds.climate.copernicus.eu/) for downscaling.

3 METHODOLOGIES

3.1 From Dynamical Systems to Stochastic Discrepancy

TC evolution can be expressed as a system of coupled nonlinear ordinary differential equations:

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{F}(\mathbf{x}(t), \mathbf{u}(t)) \tag{1}$$

where $\mathbf{x}$ is the system state (*e.g.*, TC position or intensity) at time t, $\mathbf{u}$ is the uncontrollable input

(large-scale environment) at time t, and $\mathbf{F}$ is the true unknown TC governing dynamics.

In physics-based TC models, we assume access to approximate governing dynamics $\tilde{\mathbf{F}}$, *e.g.*, the beta-and-advection track model and the fast intensity model [3,5]. Inevitably, there will be some missing physics in $\tilde{\mathbf{F}}$ due to its approximation nature, leading to *discrepancies* between simulation and observation. Mathematically, we represent such discrepancy as $\boldsymbol{\delta}$ in the dynamics:

$$\boldsymbol{\delta}(\mathbf{x}(t), \mathbf{u}(t)) = \frac{d}{dt}\mathbf{x} - \tilde{\mathbf{F}}(\mathbf{x}(t), \mathbf{u}(t)) \tag{2}$$

If we can model the discrepancies, they can be used to enhance physics-based TC models to drive simulations closely to reality. While there are various ways of constructing discrepancy models, we list two desiderata under our TC risk assessment context: 1) It must reflect the dependence on environment variables, inheriting such traits from the physics-based TC models for future climate applications; 2) it should be stochastic, to accommodate the random nature of unmodeled TC dynamics. Consequently, we choose to model the discrepancy with a stochastic differential equation illustrated in Figure 1:

$$\boldsymbol{\delta}(\mathbf{x}(t), \mathbf{u}(t)) \sim \mathbf{g}(\mathbf{x}(t), \mathbf{u}(t)) + \boldsymbol{\eta}(t) \tag{3}$$

where $\mathbf{g}$ is a deterministic *drift* function, dependent on both $\mathbf{x}$ and $\mathbf{u}$ for the environment dependency; $\boldsymbol{\eta}$ is the stochastic *diffusion* function to model the randomness.

We propose to learn the discrepancy functions in Eq. (3) in a data-driven manner. We will exploit the TC observations, including $\mathbf{X} = [\mathbf{x}(t_1), \mathbf{x}(t_2), \cdots, \mathbf{x}(t_m)]$, the best tracks for historical TCs, the historical environment input $\mathbf{U}$ from reanalysis data, and $\dot{\mathbf{X}} = [\dot{\mathbf{x}}(t_1), \dot{\mathbf{x}}(t_2), \cdots, \dot{\mathbf{x}}(t_m)]$, the discretized TC dynamics observations. Plugging in $\mathbf{X}$, $\dot{\mathbf{X}}$ and $\mathbf{U}$ along with the physics-based model $\tilde{\mathbf{F}}$, we will obtain the observed discrepancies $\boldsymbol{\delta}$ in Eq. (2), for data-driven learning.

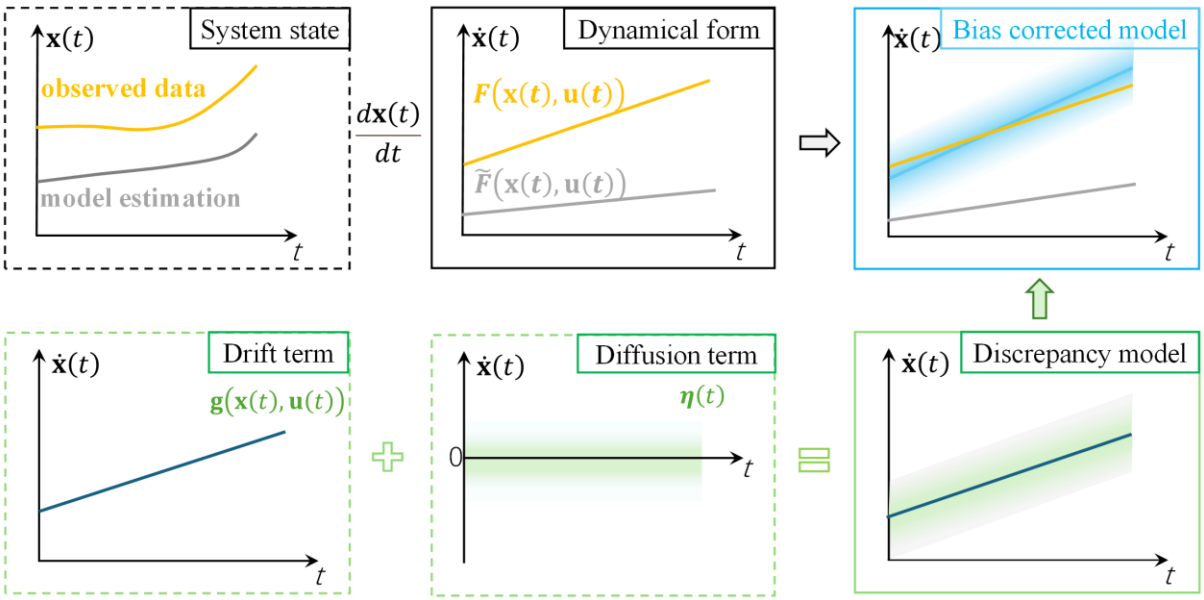


Figure 1: Correction using stochastic discrepancy model

For the drift function **g**, we assume that it can be approximated by a sparse combination of candidate functions dependent on TC states **x** and environment variables **u** [9]. Sparsity promotes interpretability and reduces overfitting, while the dependency on **u** enables compatibility with future environments. This is essentially the core task in various *symbolic regression* techniques. We adopt PySR [10], a multi-population evolutionary algorithm to search over the spaces of possible candidate functions to minimize the mismatch between drift and observed discrepancies.

Once the drift term is obtained, we will subtract drift from observed discrepancies to retain the residuals for diffusion discovery. We assume an Ornstein-Uhlenbeck (OU) process for the diffusion $\boldsymbol{\eta}$. For simulation convenience, we assume diffusion independent of TC states, allowing the Euler Maruyama approximation to be applied in the discrete timestep:

$$\eta(t) = \sigma\tau(t) = \sigma[\varepsilon\sqrt{1-\rho^2} + \tau(t_{-1})\rho\,] \tag{4}$$

where σ is the noise amplitude, $\tau(t)$ and $\tau(t_{-1})$ is the normalized OU process at successive time steps (6 hours, to be consistent with the best tracks), ε is a standard normal random variable, and ρ is the temporal correlation coefficient. A standard Maximum Likelihood Estimate (MLE) procedure will be applied to identify the OU parameters (σ and ρ) on the residuals.

3.2 Track Discrepancy Model

The beta-and-advection model is a commonly used TC track model, describing the TC translation speed $\vec{V}_T$ by combining the surrounding winds (steering flow $\vec{V}_{st}$) and generally westward poleward deviations from non-linear advection (beta drift $\vec{V}_\beta$):

$$\vec{V}_T = \vec{V}_{st} + \vec{V}_\beta \tag{5}$$

For the steering flow $\vec{V}_{st}$, a vertical mean at levels of 850, 500, and 200 hPa across a radial band ranging from 5° to 7.5° from the TC center was adopted [7]. For the beta drift $\vec{V}_\beta$, we used the WNP beta drift formula by Shan and Yu (Model S20) [7].

We develop a track discrepancy model for Model S20 in the following form:

$$\delta_{x,y} \sim g_{x,y}\big(\mathbf{B}_T(t)\big)dt + \eta_{x,y}(t) \tag{6}$$

where $g_{x,y}$ will be the track discrepancy drift function, $\mathbf{B}_T$ represents the candidate variables, and $\eta_{x,y}$ will be the stochastic diffusion. Subscripts x and y represent zonal and meridional components.

For $g_{x,y}$, PySR [10] was applied to identify the functional form. For the candidate library $\mathbf{B}_T$, the TC center latitude and longitude (*lon, lat*) were selected. In [9], the TC movement was related to the horizontal shear $\left(\partial U_x/\partial y + \partial U_y/\partial x\right)$ and the maximum sustained wind speed v_m. These variables were also added to the library. Furthermore, the translation speed V_T and heading direction α_h derived from Model S20, along with the steering flow in two directions (U_x, U_y) were also considered. The final candidate library is: $\mathbf{B}_T = \{lon, lat, V_T, \alpha_h, \left(\partial U_x/\partial y + \partial U_y/\partial x\right), U_x, U_y, v_m\}$. For diffusion, we adopt the OU process:

$$\eta_{x,y}(t) = \sigma_{x,y}\tau_{x,y}(t) = \sigma_{x,y}[\varepsilon\sqrt{1-{\rho_{x,y}}^2} + \tau(t-1)\rho_{x,y}\,] \tag{7}$$

whose parameters include $\sigma_{x,y}$ (noise amplitude) and $\rho_{x,y}$ (correlation coefficient), which will be optimized to maximize the residual likelihood of the observations.

Under the 1980–2004 CMA historical data, we identified the discrepancy drift part by Eq. (8) – (11), and the diffusion term in Eq. (12). **Bold** indicates model parameters:

- Discrepancy drift for westward TCs:

$$g_x = -\mathbf{0.2446}U_x - \mathbf{0.2195}U_y + \mathbf{0.3601}\,cos(\mathbf{0.0384}v_m) - \mathbf{0.0868} \quad (8)$$

$$g_y = -\mathbf{0.5813}U_y - \mathbf{0.5813}\cos(\mathbf{0.1132}U_x + \mathbf{0.0566}{U_y}^2 + \mathbf{0.0566}v_m) \quad (9)$$

- Discrepancy drift for eastward TCs:

$$g_x = \mathbf{0.0818}lat - \mathbf{0.4112}U_x - \mathbf{0.0518}v_m \quad (10)$$

$$g_y = -\mathbf{0.5949}U_y + \cos(\mathbf{0.0703}lat + \mathbf{0.0703}v_m + \mathbf{1.8801}) \quad (11)$$

- Discrepancy diffusion:

$$\sigma_x = \mathbf{2.3405}, \rho_x = \mathbf{0.5588};\ \sigma_y = \mathbf{1.9342}, \rho_x = \mathbf{0.5263} \quad (12)$$

Empirically, track discrepancy is nonlinearly influenced by large-scale winds and sustained wind speeds. Latitude also plays a role, particularly eastwards, which may be associated with the Coriolis effect. The track discrepancy model has a total of 21 parameters for the whole basin (10 for westward TCs drift; 7 for eastward TCs drift; 4 for diffusion), orders of magnitude less than alternative track bias correction approaches.

3.3 Intensity discrepancy model

For intensity, we aim to enhance the seminal physics-based fast TC intensity simulator in Emanuel [5] (denoted as Model E17):

$$\frac{dV_m}{dt} = \frac{1}{2}\frac{C_D}{h}\left[\alpha\beta V_p^2 m^3 - (1-\gamma m^3)V_m^2\right] \quad (13)$$

where V_m is the maximum tangential wind; C_D is the coefficient for surface drag; α is an ocean interaction parameter; β = 1-ε-κ and γ=ε+$\alpha\kappa$ are dimensionless parameters, ε is the thermodynamic efficiency; κ is the coefficient. m is a nondimensional inner core moisture variable estimated by:

$$\frac{dm}{dt} = \frac{1}{2}\frac{C_D}{h}[(1-m)V_m - 2.2Sm] \quad (14)$$

where C_k is the surface exchange coefficient for enthalpy; T_s is the sea surface temperature; L_v is the latent heat of vaporization; q_0^* is the surface saturation specific humidity; R_d is the gas constant.

We build a stochastic intensity discrepancy from Model E17 as follows:

$$\delta_{int} = g_{int}\big(\mathbf{B}_{int}(t)\big)dt + \eta_{int}(t) \quad (15)$$

where g_{int} is the intensity discrepancy drift term; $\mathbf{B}_{int}(t)$ is the candidate library; η_{int} is diffusion.

For drift g_{int}, PySR is adopted to identify the function. Apart from the intensity variable

itself (V_m), the candidate function library $\mathbf{B}_{int}$ includes the longitude *lon*, latitude *lat*, wind shear *S*, and nondimensional inner core moisture *m* recommended in [9]. The timespan from TC genesis is introduced as *t*, normalized with the interval (6 hours). The candidate library for the intensity discrepancy is: $\mathbf{B}_{int}(t) = [V_m, lon, lat, t, S, m]$. Fo diffusion, the OU process is adopted:

$$\eta_{int}(t) = \sigma_{int}\tau(t) = \sigma_{int}[\varepsilon\sqrt{1-\rho_{int}{}^2} + \tau(t-1)\rho_{int}\,] \tag{16}$$

where σ_{int} is the noise amplitude and ρ_{int} is the correlation coefficient, to be identified according to the Maximum Likelihood principle.

Finally, based on CMA TCs during 1980-2004, our TC intensity discrepancy model is expressed as (**bold** indicates parameters):

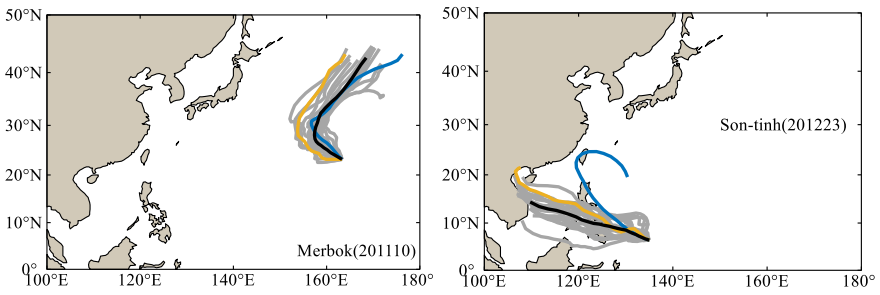
$$g_{int} = \cos\left(\frac{V_m}{t/6 + \dfrac{V_m}{\mathbf{0.8973}^{t/6} + t/6}}\right) + \mathbf{0.5330} - \mathbf{149.3893}\frac{m^2}{lat \cdot S};$$
$$\sigma_{int} = \mathbf{0.6826}, \rho_{int} = \mathbf{0.6390}; \tag{17}$$

Empirically, the intensity discrepancy is linked to the TC intensity levels and durations. Apart from environmental factors like humidity and wind shear, the TC geographical latitude also plays a role. Our intensity discrepancy model is parsimonious with a total of five parameters.

4 RESULTS & DISCUSSION

4.1 Track and Intensity Validation

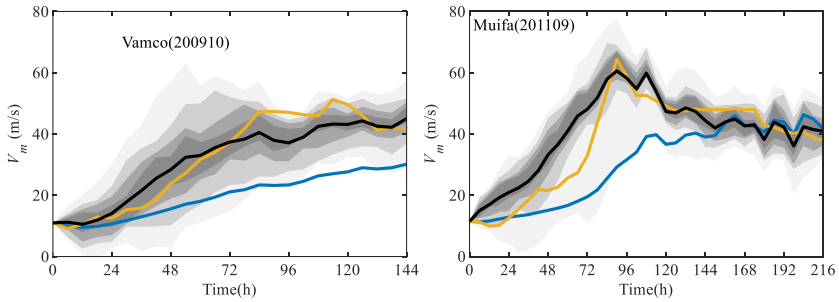
Two historical TCs are taken to demonstrate track model performance, including Severe Tropical Storm Merbok and Son-tinh. Historical genesis is taken from CMA and subsequently propagated by the track model. For each TC event, we conducted 20 realizations from the stochastic discrepancy model to reflect track uncertainty. We only focus on the duration before landfalling. Results are shown in Figure 2 along with historical records from the CMA and the reference simulation by Model S20. Across various tracks, our generated ensemble covers the realistic moving directions and lengths very well, demonstrating that they can offer a reliable range of potential landfall locations beyond a point estimate from Model S20.



(a) (b)

Figure 2: The observed (yellow line) and the simulated (grey line) track for (a) Merbok and (b) Son-tinh. Blue lines show results from the Model S20, while gray lines are realizations.

Furthermore, two historical TCs are taken to validate, including Super Typhoon Vamco and Muifa. Historical genesis and tracks were utilized to eliminate errors from non-intensity components, and the proposed stochastic intensity model estimated the intensity evolution for 20 realizations under each TC event. Four confidence intervals (30%, 50%, 70%, 90%) were illustrated in Figure 3, along with CMA observation and Model E17 simulations. From the hindcast, the TC intensity evolution mostly falls within the 90% interval of our TC model. Compared to Model E17, our model demonstrated a much better fitting to the actual intensity evolution.



(a) (b)

Figure 3: Observed (yellow line) and simulated (grey line) intensity for (a) Vamco (b) Muifa. Blue lines show results from Model E17. Shadings are confidence intervals of 30, 50, 70, and 90 percentiles.

4.2 Long-term wind speed in representative China coastline cities

Estimation of TC-induced long-term wind speeds is crucial for disaster protection and risk assessment. To this end, we simulate synthetic TCs which are initiated by the historical genesis and propagated using the discrepancy-enhanced track and intensity module. Historical genesis of TCs from 1980 to 2014 were considered, and each event was repeated 100 times. We then simulate the spatial distributions of wind speed during synthetic TC events with a parameterized wind field model, assuming the surface wind speed c_s at a particular site to be the combination of the translation component c_m and the vortex component c_g. To assess the long-term wind speeds at specific locations, we selected TCs that intersect the 500 km radius circle centered at the reference point. Subsequently, sampled values were fed into the following formula to calculate the wind speed with a specified return period T based on the maximum wind speed:

$$T = \frac{1}{1 - F(v)} \tag{18}$$

where $F(v)$ is the marginal distribution of extreme wind speed v.

We selected seven representative China coastline cities and presented the mean wind speeds and the 5th-95th intervals for the 100-year return periods in Table 1. Our estimated mean wind speeds closely align with the Chinese design code, with an absolute difference of less than 4 m/s across all cities. The code values consistently fall within our 5th -95th intervals, suggesting well-calibrated coverage. We maintain consistent agreement with other studies, and only differ from [8] by at most 3 m/s. Furthermore, if we consider the wind speed estimations from different existing models as an alternative reflection of epistemic uncertainties, the results from most existing studies also fall into our estimated confidence interval, another piece of evidence

on the high quality of our uncertainty estimations. Overall, one can observe that the influence of TC model uncertainties is indeed non-negligible for wind hazard assessments, with a 95% confidence interval consistently spanning across 10 m/s. Future skillful TC simulation models are needed to further reduce this margin of epistemic uncertainties.

Table 1: 100-yr wind speed (m/s) at seven cities.

City	Proposed Mean (5^{th},95^{th})	Code	[11]	[12]	[8]
Zhanjiang	35.3 (30.0 , 40.5)	39.0	42.4	37.4	37.5
Shenzhen	37.8 (32.5 , 43.3)	37.9	43.8	36.8	36.4
Xiamen	36.3 (31.5 , 41.9)	39.0	46.7	38.9	37.7
Fuzhou	35.6 (31.3 , 41.2)	36.9	48.5	35.1	33.6
Wenzhou	34.9 (30.0 , 40.5)	33.5	48.8	36.5	34.9
Ningbo	33.8 (26.7 , 39.4)	31.0	44.9	33.0	34.5
Shanghai	33.1 (28.0 , 39.3)	31.0	48.3	31.7	31.7

5 CONCLUSION

This paper introduces a stochastic discrepancy framework to enhance physics-based TC models. Its symbolic regression-based drift component and an Ornstein-Uhlenbeck (OU) process-based diffusion component effectively reduces bias and quantify the inherent epistemic uncertainty in existing physics-based models. As a proof-of-concept, we establish two discrepancy models for commonly used track and intensity models in the WNP basin. We applied the developed stochastic discrepancy models for track and intensity simulation for historical TCs using ERA5 reanalysis data and CMA best tracks. Hindcast demonstrates that the proposed model can effectively capture realistic TC track and intensity evolution. In addition, we extended our application to typhoon wind hazard along the China southeast coastline. The proposed model demonstrates a good agreement with the design code recommendations, and our derived confidence intervals actually capture the diverse wind speed estimations from alternative models.

REFERENCES

[1] Meiler, S., A. Ciullo, C. M. Kropf, K. Emanuel, and D. N. Bresch. Uncertainties and sensitivities in the quantification of future tropical cyclone risk. *Commun. Earth Environ*. (2023). 4 (1): 371.

[2] Vickery, P. J., P. F. Skerlj, and L. A. Twisdale. Simulation of Hurricane Risk in the U.S. Using Empirical Track Model. *J. Struct. Eng*. (2000) 126 (10): 1222–1237.

[3] Emanuel, K., S. Ravela, E. Vivant, and C. Risi. A Statistical Deterministic Approach to Hurricane Risk Assessment. *Bull. Am. Meteorol. Soc.* (2006) 87 (3): 299–314.

[4] Emanuel, K., C. DesAutels, C. Holloway, and R. Korty. Environmental Control of Tropical Cyclone Intensity. *J. Atmospheric Sci*. (2004) 61 (7): 843–858.

[5] Emanuel, K. A fast intensity simulator for tropical cyclone risk analysis. *Nat. Hazards.* (2017) 88 (2): 779–796.

[6] Jing, R., and N. Lin. An Environment-Dependent Probabilistic Tropical Cyclone Model. J. Adv. Model. *Earth Syst.* (2020) 12 (3): e2019MS001975.

[7] Shan, K., and X. Yu. A Simple Trajectory Model for Climatological Study of Tropical Cyclones. *J. Clim.* (2020) 33 (18): 7777–7786.
[8] Chen, Y., and Z. Duan. A statistical dynamics track model of tropical cyclones for assessing typhoon wind hazard in the coast of southeast China. *J. Wind Eng. Ind. Aerodyn.* (2018) 172: 325–340.
[9] Zhong, X., W. Jiang, and J. Zhang. TC-SINDy: Improving physics-based deterministic tropical cyclone track and intensity model via data-driven sparse identification of Nonlinear Dynamics. *J. Wind Eng. Ind. Aerodyn*. (2024) 250: 105758.
[10] Cranmer, M. Interpretable Machine Learning for Science with PySR and Symbolic Regression.jl. *arXiv* (2023).
[11] Xiao, Y.F., Duan, Z.D., Xiao, Y.Q., Ou, J.P., Chang, L., Li, Q.S. Typhoon wind hazard analysis for southeast China coastal regions. *Struct. Saf.* (2011) 33, 286–295.
[12] Li, S.H., Hong, H.P. Typhoon wind hazard estimation for China using an empirical track model. *Nat. Hazards.* (2016) 82, 1009–1029