ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Statistical seasonal forecasting of tropical cyclones over the western North Pacific

To cite this article: Kelvin T F Chan *et al* 2021 *Environ. Res. Lett.* **16** 074027

IOP Publishing Environ. Res. Lett. 16 (2021) 074027 https://doi.org/10.1088/1748-9326/ac05f1

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
22 March 2021

REVISED
25 May 2021

ACCEPTED FOR PUBLICATION
27 May 2021

PUBLISHED
2 July 2021

LETTER

Statistical seasonal forecasting of tropical cyclones over the western North Pacific

Kelvin T F Chan[1,2,3,*], **Zhenyuan Dong**[1,*] **and Minglin Zheng**[1,*]

[1] School of Atmospheric Sciences, and Guangdong Province Key Laboratory for Climate Change and Natural Disaster Studies, Sun Yat-sen University, Zhuhai, People's Republic of China
[2] Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, People's Republic of China
[3] Key Laboratory of Tropical Atmosphere-Ocean System (Sun Yat-sen University), Ministry of Education, Zhuhai, People's Republic of China
[*] Authors to whom any correspondence should be addressed.

E-mail: chenth25@mail.sysu.edu.cn, dongzhy5@mail2.sysu.edu.cn **and** zhengmlin@mail2.sysu.edu.cn

Keywords: tropical cyclone, seasonal forecasting model, western North Pacific

Abstract

Forecasting tropical cyclone (TC) activities has been a topic of great interest and research. Many studies and existing seasonal forecasting models have examined and predicted the number of TCs (including geneses and landfalls) mainly based on the environmental factors in the peak TC season. However, these predictions can be time-consuming, computationally expensive and uncertain, depending on the efficiency and predictability of the dynamical models. Therefore, here we propose an effective statistical seasonal forecasting model, namely the Sun Yat-sen University (SYSU) Model, for predicting the number of TCs (intensity at tropical storm or above) over the western North Pacific based on the environmental factors in the preseason. The nine categories comprising 103 candidate predictors in 1980–2015 (36 years) are systematically investigated. The best subset selection regression shows that the sea surface temperatures at the tropical North Atlantic and eastern North Pacific in April, the 500 hPa geopotential height difference between April and January at the open ocean southwest of Australia and the 700 hPa geopotential height at the North Pacific in April are the most significant predictors. The correlation coefficient between the modeled results and observations reaches 0.89. The model is successfully validated by leave-one-out, nine-fold cross-validations, and later 5 year (2016–2020) observations. The prediction of the SYSU Model exhibits a 95% hit rate in 1980–2020 (39 out of 41), suggesting an operational potential in the seasonal forecasting of TCs over the western North Pacific.

1. Introduction

A tropical cyclone (TC) is one of the most destructive natural phenomena in the world. It can lead to heavy precipitation, flooding, strong winds and storm surges to the coastal regions (Zhang *et al* 2018), causing great economic and human losses. On average, about 80 TCs form globally each year. Among the basins, the western North Pacific is the region with the most active tropical cyclogenesis. About one third of global TCs, 28 on annual average, form over the western North Pacific (Gray 1968), which could make landfall on Southeast Asia as well as East Asia. East Asia, including China, Korea, Japan, etc, accommodates around 25% of the world's total population with a large portion dwelling in the coastal regions. Thus, accurate seasonal forecasting on predicting the TC number in the western North Pacific is enlightening.

Camargo *et al* (2007), Zhan *et al* (2012) and Klotzbach *et al* (2019) reviewed all the existing seasonal forecasting models around the world. In general, there are three types of forecasting: statistical, dynamical and hybrid statistical-dynamical. Each has its advantages and disadvantages. Dynamical and hybrid forecastings comprise full model physics and a series of parameterizations so that they can provide physical reasoning. However, as a matter of fact, the model predictability decreases with simulation time because of the butterfly effect. Small changes in initial conditions, physics schemes and parameterizations can lead to large-scale and unpredictable variations in the future state of the system. Thus, the prediction

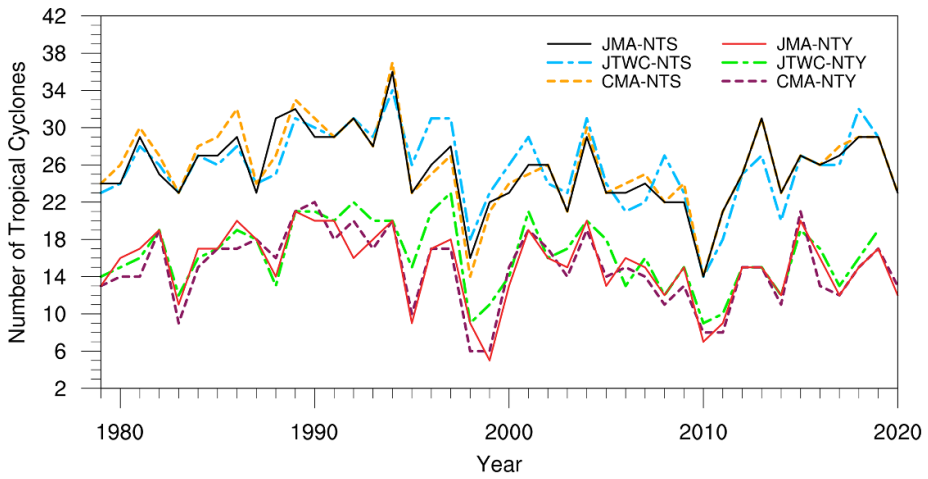

Figure 1. Time series of NTS and NTY from the JMA, JTWC and CMA best-track data.

performance highly depends on both the data and model predictability. Model ensembles are therefore often adopted to minimize the uncertainties. However, doing so would be computationally expensive and time-consuming. This is why statistical forecasting is still competitive nowadays although its physical reasoning is relatively weak.

Nonetheless, most of the existing statistical models for the seasonal forecasting of TCs over the western North Pacific are insufficient. They use many predictors to build their models, which could cause model overfitting and complexity (Chan *et al* 1998, Fan 2007a, Xia *et al* 2014). Moreover, many forecasts are closed-source, not documented, and/or for internal use only (Klotzbach *et al* 2019). Most of them do not provide model verifications or evaluations. In addition, it is noted that most of the existing dynamical or hybrid models make use of the environmental factors in the peak TC season to make predictions (e.g. Au-Yeung and Chan 2012, Wang *et al* 2012, Zhang *et al* 2017a, Camp *et al* 2019), while very few achieve this using the pre-season environmental factors (Chan *et al* 1998). A study by Tian and Fan (2019) utilized predictors from the preceding year (e.g. the preceding boreal summer SST) and adopted a year-to-year increment method to predict the number of landfalling TCs on China during June to August. In practice, some of the pre-season environmental factors could have a 'memory effect', suggesting that their seasonal or annual signals are relatively sustainable and transmittable through certain time periods. However, some of them may not be persistent (see figure 2), but they can store their signals in the ocean via air–sea interaction. Compared to the atmosphere, the memory effect of the ocean is more stable. Thereby, through selecting the particular factors in the preseason, it is plausible to predict the number of TCs in the upcoming TC season.

Hence, this study aims to propose an effective (no more than four predictors), open-source (free public disclosure, i.e. this paper) and competitive (at least 80% hit rate) operational statistical seasonal forecasting model that can predict the number of TCs over the western North Pacific in the upcoming TC season using pre-season environmental factors. The forecast is targeted to be issued by every middle of May. Section 2 describes the data and methodology. Section 3 introduces the statistical seasonal forecasting model with model validation. Section 4 summarizes the study with discussion.

2. Data and methodology

The best-track data from the Japan Meteorological Agency (JMA), China Meteorological Administration (CMA) and Joint Typhoon Warning Center (JTWC) are adopted. The TCs reaching tropical storm (TS) intensity or above (maximum sustained wind speed $\geqslant$34 kt) over the western North Pacific in 1980–2020 (41 years) are counted and mainly examined in this study. As the recorded number of TSs (NTS) is similar among the three sets of best-track data (figure 1), and the JMA is the World Meteorological Organization's official Regional Specialized Meteorological Center for the western North Pacific basin, the JMA best-track data are used for the major discussion hereafter. On a 41 year average (1980–2020), there are around 25–26 TSs appearing over the western North Pacific per year. The corresponding standard deviation is about 4.2. A slight but insignificant decrease in the trend of NTS is found (-0.78 decade^{-1}, $p = 0.17$). It is noted that the number of typhoons (NTY; TCs reaching typhoon intensity or above (maximum sustained wind speed $\geqslant$64 kt)) is not studied here because the benchmarks are divergent. They vary significantly among the three sets of best-track data. Although the absolute diversity of NTY among them seems not apparent, the proportion of diversity in NTY is obviously higher than that in NTS (figure 1).

As the forecast is targeted to be issued by every middle of May, the candidate predictors before May, that is, those from December of the preceding year to April of the following year are mainly examined. Based on previous studies, correlation analysis and data availability, nine categories, comprising 103 candidate predictors in total are selected in this study. The categories that have been suggested to be conducive to tropical cyclogenesis are put into a pool of predictors. For instance, the SST indices (Xiao and Li 2007, Zhou and Cui 2011, Zhan *et al* 2013, Ding *et al* 2015, Zhang and Villarini 2019, Hu *et al* 2020), meridional modes (Zhang *et al* 2016, 2017b, Gao *et al* 2018), atmospheric circulations (Zhou and Cui 2008, Guo and Tan 2018), climate indices (Zhou and Cui 2014, Chen *et al* 2018, Ding *et al* 2019), solar activities (Li *et al* 2019), cryosphere (Fan 2007b), ocean heat content (Tu *et al* 2011), stratosphere (Chan 1995) and atmospheric chemistry (Takahashi *et al* 2017). Table 1 summarizes all the categories, predictors and their corresponding data. In order to achieve the objective of this study, the data should cover from 1980 to the present and should be operationally ongoing. The data descriptions are tabulated in table 2. The monthly fifth generation of the European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric reanalysis of the global climate (ERA5) is primarily used in this study, including direct data retrievals and indices calculations. The predictors reaching correlation coefficients >0.32 with the NTS ($p < 0.05$) are chosen as the candidate predictors (figure 2).

The best subset selection regression is then adopted to build the model in this study. In this approach, the selection procedure is first performed. Based on the sequential replacement method, the predictors are sequentially replaced. Replacements that improve the performance are retained. Next, all combinations of predictors are examined to obtain the best combination (highest adjusted R squared). In order to reduce the instability brought by different categories of predictors, such as solar activity and atmospheric chemistry, we iterate the first two steps nine times by adding one category to the consideration each time so that more predictors would be considered. Last, we get 90 regression models (nine considerations and ten models each). After screening out the models having high collinearity or high squares of residuals in 2016–2020, the model that has the best prediction competence (highest R squared) is selected as the best subset.

Standardization of the predictors is applied before the regression analysis so that all the predictors are in the form of indices. The data before 2016 are used for model training. In order to further justify the model skills, the leave-one-out, nine-fold cross-validations and the recent 5 year (2016–2020) observations are performed for the model validation.

3. Statistical seasonal forecasting model

The correlation matrices between the 103 candidate predictors and the predictand NTS are shown in figure 2. Notably, about 60% of the predictors in the pool reached the peak of absolute correlation with the NTS in April. This makes sense because April is the month closest to the TC season among the five pre-season months (December of the preceding year to April of the following year) so that the signals acquired in April should have more influence on the later months. Meanwhile, some factors have high correlations over a long period of time, for example, SST, meridional modes and ocean heat content. As will be shown, although some candidate predictors have high correlations with the NTS independently, they are not selected as our final predictors because of their high mutual correlations and collinearity.

The best subset selection regression suggests that the SSTs at the tropical North Atlantic ($SST_{TNA,4}$; 0°–20° N, 30°–75° W) and eastern North Pacific ($SST_{ENP,4}$; 10°–25° N, 115°–140° W) in April, the 500 hPa geopotential height difference between April and January at the open ocean southwest of Australia ($Z500_{SWA,4-1}$; 40°–60° S, 90°–120° E) and the 700 hPa geopotential height at the North Pacific ($Z700_{NP,4}$; 25°–40° N, 170° E–160° W) in April are the best four significant predictors. Their combination is the most optimum one for forecasting the NTS among all 103 candidate predictors. The corresponding geophysical regions are shown in figure 3.

To avoid the uncertainties or arbitrary conclusions given by a single data source, another set of mainstream reanalysis data (the Climate Forecast System Reanalysis (CFSR) and the Extended Reconstructed Sea Surface Temperature version 5 (ERSSTv5) from the National Oceanic and Atmospheric Administration (NOAA), USA) are employed. Table 3 shows that the mutual correlations between the four selected predictors and the NTS are comparable between the two, ECMWF and NOAA, datasets. Although $Z700_{NP,4}$ and $Z500_{SWA,4-1}$ have a moderate correlation, the variance inflation factor values of $SST_{TNA,4}$, $SST_{ENP,4}$, $Z700_{NP,4}$ and $Z500_{SWA,4-1}$ are 1.11, 1.12, 1.15 and 1.15, respectively. Since they are all below 2, this means the collinearity between the four predictors is low.

The best subset selection regression model based on the ECMWF data (hereafter Model E),

$$NTS_E = 25.50 - 1.89SST_{TNA,4} - 1.89Z500_{SWA,4-1} - 1.33Z700_{NP,4} + 1.02SST_{ENP,4}$$

explains 83% of variance. The correlation coefficient reaches 0.91. The root-mean-square error (RMSE) and mean absolute error is 1.77 and 1.40, respectively.

Table 1. Pool of predictors, corresponding descriptions and data examined in this study.

Category/Factor	Predictor	Description	Data
1. SST indices			
1.1 SSTG	$SSTG_3$ $SSTG_4$	SST gradient between the western South Pacific (40°–20° S, 160° E–170° W) and the western North Pacific warm pool (0–16° N, 125°–165° E) in March and April (Zhan *et al* 2013).	ERSSTv5, ERA5
1.2 SST_{EEP}	$SST_{EEP,12[-1]}$ $SST_{EEP,1}$ $SST_{EEP,2}$ $SST_{EEP,3}$ $SST_{EEP,4}$	SST at the equatorial East Pacific (5° S–5° N, 80°–110° W) in the preceding December and the following January, February, March and April.	ERSSTv5, ERA5
1.3 SST_{WUS}	$SST_{WUS,3}$ $SST_{WUS,4}$	SST to the west of the United States (30°–45° N, 130°–150° W) in March and April.	ERSSTv5, ERA5
1.4 SST_{TNA}	$SST_{TNA,1}$ $SST_{TNA,2}$ $SST_{TNA,3}$ $SST_{TNA,4}$	SST at the tropical North Atlantic (0°–20° N, 30°–75° W) in January, February, March and April.	ERSSTv5, ERA5
1.5 SST_{ENP}	$SST_{ENP,3}$ $SST_{ENP,4}$	SST at the eastern North Pacific (10°–25° N, 115°–140° W) in March and April.	ERSSTv5, ERA5
1.6 AWT	$AWT_{12[-1]}$ AWT_1 AWT_2 AWT_3 AWT_4	Atlantic warming tendency (AWT) index in the preceding December and the following January, February, March and April. AWT index is defined as the first principal component of SST in area 20°–40° N, 10°–40° W (Xiao and Li 2007).	ERSSTv5, ERA5
1.7 IOBM	$IOBM_1$ $IOBM_2$ $IOBM_3$ $IOBM_4$	Indian Ocean Basin Mode (IOBM) index in January, February, March and April.	ERSSTv5, ERA5
1.8 NPVM	$NPVM_4$	North Pacific Victoria Mode (NPVM) index in April (Ding *et al* 2015).	ERSSTv5, ERA5
1.9 SPDO	$SPDO_{6[-1]}$ $SPDO_{7[-1]}$	South Pacific Decadal Oscillation (SPDO) index in June and July of the preceding year. SPDO index is defined as the first principal component of SST in area 70°–20° S, 140° E–60° W.	ERSSTv5, ERA5
1.10 SIOD	$SIOD_3$	South Indian Ocean Dipole (SIOD) index in March. SIOD index is defined as the first principal component of SST in area 50°–10° S, 40° E–120° E (Hu *et al* 2020).	ERSSTv5, ERA5
2. Meridional modes			
2.1 PMM	$PMM_{12[-1]}$ PMM_1 PMM_4	Pacific Meridional Mode (PMM) index in the preceding December and the following January and April.	PSL
2.2 AMO	$AMO_{12[-1]}$ AMO_1 AMO_2 AMO_3 AMO_4	Non-filtered Atlantic Multidecadal Oscillation (AMO) Index in the preceding December and the following January, February, March and April.	PSL
2.3 AMM	AMM_2 AMM_3 AMM_4	Atlantic Meridional Mode (AMM) Index in February, March and April.	PSL
3. Atmospheric circulations			
3.1 $Z700_{NP}$	$Z700_{NP,4}$	700 hPa geopotential height at the North Pacific (25°–40° N, 170° E–160° W) in April.	CFSR, ERA5
3.2 $Z500_{SWA}$	$Z500_{SWA,4-1}$	500 hPa geopotential height difference between April and January at the open ocean southwest of Australia (60°–40° S, 90°–120° E).	CFSR, ERA5
3.3 $MSLP_{NP}$	$MSLP_{NP,4}$	Mean sea-level pressure at the North Pacific (55°–70° N, 170° E–165° W) in April.	CFSR, ERA5
3.4 $Z200_T$	$Z200_{T,12[-1]}$ $Z200_{T,1}$ $Z200_{T,2}$ $Z200_{T,3}$ $Z200_{T,4}$	200 hPa geopotential height at the tropics (20° S–20° N, 180° E–180° W) in the preceding December and the following January, February, March and April.	CFSR, ERA5

(Continued.)

Table 1. (Continued.)

Category/Factor	Predictor	Description	Data
3.5 $Z500_T$	$Z500_{T,12[-1]}$ $Z500_{T,1}$ $Z500_{T,2}$ $Z500_{T,3}$ $Z500_{T,4}$	500 hPa geopotential height at the tropics (20° S–20° N, 180° E–180° W) in the preceding December and the following January, February, March and April.	CFSR, ERA5
3.6 $U850_{EEP}$	$U850_{EEP,12[-1]}$ $U850_{EEP,1}$ $U850_{EEP,2}$ $U850_{EEP,3}$ $U850_{EEP,4}$	850 hPa zonal wind at the equatorial East Pacific (10° S–10° N, 90°–150° W) in the preceding December and the following January, February, March and April.	CFSR, ERA5
4. Climate indices			
4.1 AO	AO_1	Arctic Oscillation (AO) index in January.	PSL
4.2 NAO	NAO_2	North Atlantic Oscillation (NAO) index in February.	PSL
4.3 NAD	NAD_1 NAD_2 NAD_3	North America Dipole (NAD) index in January, February and March. NAD index is defined as the difference of normalized sea-level pressure anomalies between the southern (93°–58° W, 9°–29° N) and northern (77°–43° W, 53°–68° N) poles (Ding *et al* 2019).	CFSR, ERA5
4.4 NPO	NPO_4	North Pacific Oscillation (NPO) index in April. NPO index is defined as the second principal component of sea-level pressure anomalies in the region 10°–80° N, 120° E–60° W (Chen *et al* 2018).	CFSR, ERA5
5. Solar activities			
5.1 SN	SN_1	Sunspot number in January.	WDC-SILSO
5.2 NMS	$NMS_{12[-1]}$ NMS_2 NMS_3 NMS_4 NMS_y	Number of magnetic storms in the preceding December and the following February, March, April and all year round (Ivanov 2007).	OMNI
5.3 SR	SR_1 SR_2	F10.7 cm radio emissions in January and February.	OMNI
6. Cryosphere			
6.1 SI_A	$SI_{A,12[-1]}$ $SI_{A,1}$ $SI_{A,2}$ $SI_{A,3}$ $SI_{A,4}$	Sea ice at the Arctic (45°–75° N, 0°–90° W) in the preceding December and the following January, February, March and April.	HadISST, ERA5
6.2 SC_{NH}	$SC_{NH,2}$	Area of snow cover extent in the Northern Hemisphere (0°–90° N, 180° E–180° W) in February.	RUS, ERA5
7. Ocean heat content			
7.1 $OHCU_{ESP}$	$OHCU_{ESP,1}$ $OHCU_{ESP,2}$	Ocean heat content in the upper subsurface (⩽155 m) of the eastern South Pacific (65°–45° S, 70°–140° W) in January and February.	GODAS
7.2 $OHCD_{NIO}$	$OHCD_{NIO,1}$ $OHCD_{NIO,2}$ $OHCD_{NIO,3}$ $OHCD_{NIO,4}$	Ocean heat content in the deep (>155 m) North Indian Ocean (10° S–30° N, 40°–110° W) in January, February, March and April.	GODAS
7.3 $OHCD_{WSP}$	$OHCD_{WSP,1}$ $OHCD_{WSP,2}$ $OHCD_{WSP,3}$ $OHCD_{WSP,4}$	Ocean heat content in the deep (>155 m) western South Pacific (70°–40° S, 150° E–140° W) in January, February, March and April.	GODAS
8. Stratosphere			
8.1 $U10_{SH}$	$U10_{SH,1}$ $U10_{SH,3}$ $U10_{SH,4}$	10 hPa zonal wind in the Southern Hemisphere (53°–43° S, 145°–180° E) in January, March and April.	CFSR, ERA5
8.2 $U10_{NH}$	$U10_{NH,3}$ $U10_{NH,4}$	10 hPa zonal wind in the Northern Hemisphere (30°–45° N, 140°–180° E) in March and April.	CFSR, ERA5

(Continued.)

Table 1. (Continued.)

Category/Factor	Predictor	Description	Data
9. Atmospheric chemistry			
9.1 BC_G	$BC_{G,1}$ $BC_{G,2}$ $BC_{G,3}$ $BC_{G,4}$	Average black carbon content in the globe in January, February, March and April.	MERRA-2
9.2 BC_{NP}	$BC_{NP,2}$ $BC_{NP,3}$ $BC_{NP,4}$	Average black carbon content at the North Pacific (10°–50° N, 120° E–120° W) in February, March and April.	MERRA-2
9.2 SO_{2NH}	$SO_{2NH,1}$ $SO_{2NH,2}$	Sulfur dioxide content at mid and high latitudes in the Northern Hemisphere (40°–90° N, 180° E–180° W) in January and February.	MERRA-2
9.3 DU_G	$DU_{G,4}$	Dust content in the globe in April.	MERRA-2

Table 2. Data descriptions.

Data	Center	Description	Availability
CFSR	National Oceanic and Atmospheric Administration (NOAA)	Coupled atmosphere–ocean reanalysis (Saha *et al* 2010, 2014)	1979–present
ERA5	European Centre for Medium-Range Weather Forecasts (ECMWF)	Coupled atmosphere–ocean reanalysis (Hersbach *et al* 2020)	1979–present
ERSSTv5	NOAA	SST analysis (Huang *et al* 2017)	1854–present
GODAS	NOAA	Ocean analysis and reanalysis (Derber and Rosati 1989)	1980–present
HadISST	Hadley Centre	A unique combination of monthly globally-complete fields of SST and sea ice concentration (Rayner *et al* 2003)	1871–present
MERRA-2	NOAA	Atmospheric reanalysis (Gelaro *et al* 2017)	1980–present
OMNI	National Aeronautics and Space Administration (NASA)	Spacecraft data (King and Papitashvili 2005)	1963–present
PSL	Physical Sciences Laboratory	Climate indices	1948–present
RUS	Rutgers University	Snow cover extent data (Estilow *et al* 2015)	1966–present
WDC-SILSO	Royal Observatory of Belgium	Sunspot number data	1749–present

Meanwhile, the best subset selection regression model based on the NOAA data (hereafter Model N),

$$\mathrm{NTS_N} = 25.50 - 1.85\mathrm{SST_{TNA,4}} - 1.91\mathrm{Z500_{SWA,4-1}} - 1.44\mathrm{Z700_{NP,4}} + 1.12\mathrm{SST_{ENP,4}}$$

explains 79% of variance. The correlation coefficient reaches 0.89. The RMSE and mean absolute error is 1.98 and 1.59, respectively. It is noted that the regression coefficients of the predictors among the two models are highly similar and all pass the significance test ($p < 0.01$). In addition, the weak mutual correlations among the predictors imply that the predictors are mostly independent of each other (table 3). These results suggest that both models are consistent with each other and consolidate the robustness of the predictors.

The models are validated by leave-one-out and nine-fold cross-validations. The correlations between the observations and Model E and Model N in the leave-one-out cross-validations reach 0.89 and 0.85, respectively. Those in the nine-fold cross-validations reach 0.89 and 0.86, respectively. In addition, the models are further validated by the recent 5 year (2016–2020) observations. The corresponding absolute errors are $\leqslant 3$ in both models. All these findings suggest that the models are effective.

To bring the results from research into practice and application, a new model, namely the Sun Yat-sen University (SYSU) Model is proposed. It is composed of Model E and Model N. The RMSEs of Model E ($\mathrm{RMSE_E}$) and Model N ($\mathrm{RMSE_N}$) are utilized to set the prediction range. The prediction range issued by the SYSU Model in a particular year i is given by,

$$\mathrm{NTS_{SYSU,i}} = \left[\left\lfloor \min\left(\mathrm{NTS_{E,i}} - \mathrm{RMSE_E}, \mathrm{NTS_{N,i}} - \mathrm{RMSE_N}\right)\right\rfloor, \left\lceil \max\left(\mathrm{NTS_{E,i}} + \mathrm{RMSE_E}, \mathrm{NTS_{N,i}} + \mathrm{RMSE_N}\right)\right\rceil\right].$$

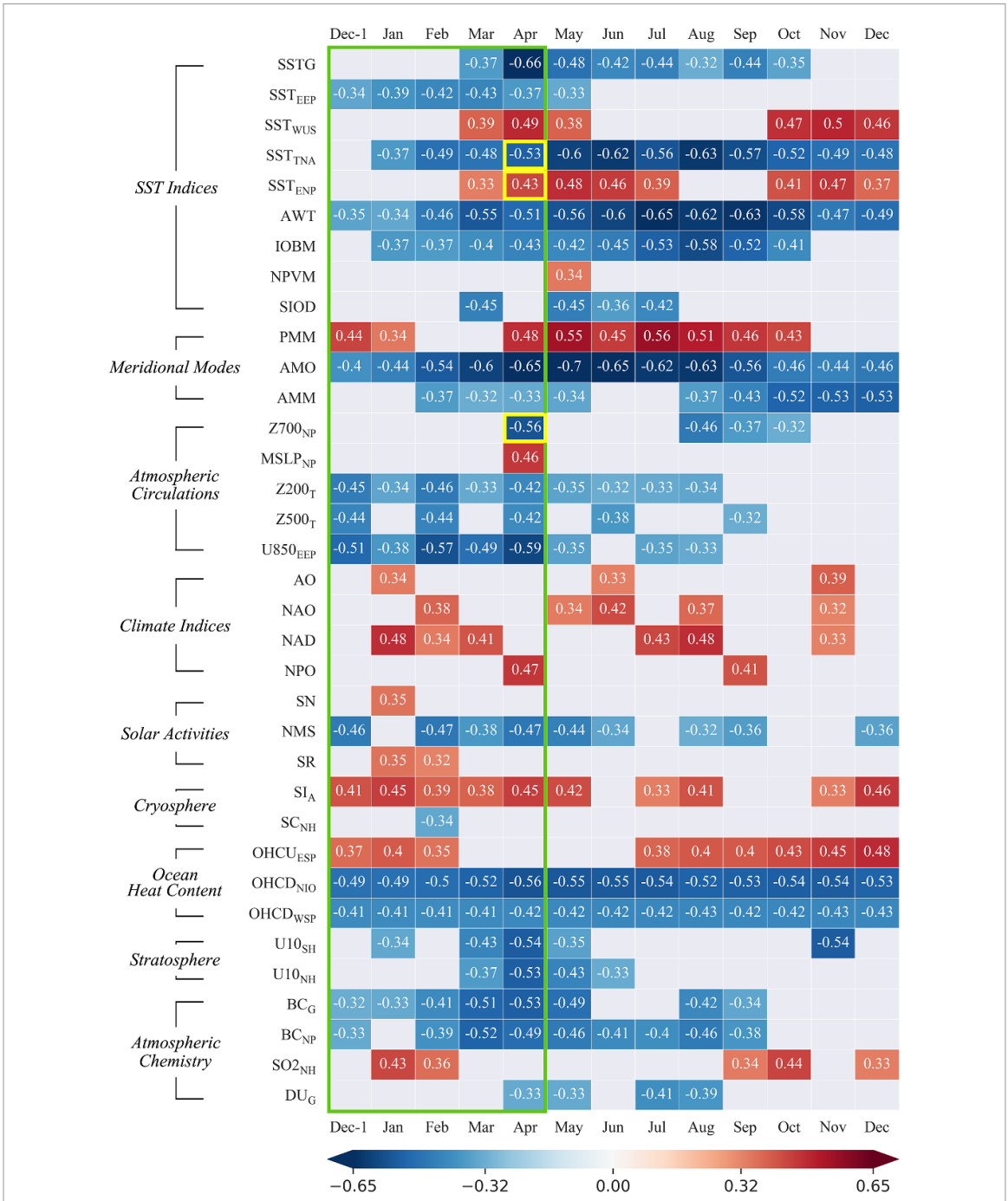

Figure 2. Correlation matrices between the various factors and the predictand NTS. Factors are in monthly mean starting from the preceding December to the concurrent December. Absolute correlations higher than 0.32 (passing 95% significance test) are shown. 99 candidate predictors picked in this study are bounded by the big green frame. The other four candidate predictors $SPDO_{6[-1]}$ ($r = -0.36$), $SPDO_{7[-1]}$ ($r = 0.34$), $Z500_{SWA,4-1}$ ($r = -0.64$) and NMS_{Y} ($r = -0.47$) are not shown. Model predictors $SST_{TNA,4}$, $SST_{ENP,4}$, and $Z700_{NP,4}$ that are employed in Model E and Model N are outlined by the small yellow frames.

In plain text, it is the maximum range of $NTS_{E,i} \pm RMSE_E$ and $NTS_{N,i} \pm RMSE_N$ by rounding down the minimum and rounding up the maximum to the nearest integers.

Figure 4 shows that the SYSU Model catches the highs and lows well. The range of the annual prediction (maximum prediction minus minimum prediction) in 1980–2020 is about 5–6 (the green shading shown in figure 4), which is operationally reasonable and meaningful. Most of the observations fall within the prediction. Once the observed NTS falls within the prediction range, we call the model hits. The hit rate is encouragingly promising since it reaches 95% in 1980–2020 (39 years of observations fall within the prediction range out of the total 41 years; figure 4).

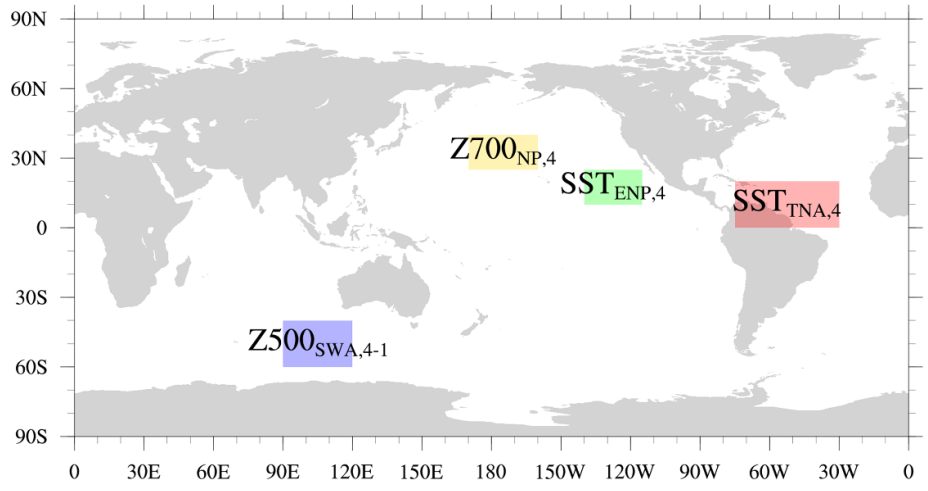

Figure 3. Model predictors. Shaded boxes denote their corresponding regions.

Table 3. Mutual correlation coefficients among the predictors and predictand NTS using the ECMWF and NOAA data. Single and double asterisk(s) denote the statistical significance at the 95% and 99% confidence levels, respectively.

	ECMWF (1980–2015)				NOAA (1980–2015)			
	$SST_{TNA,4}$	$Z500_{SWA,4-1}$	$Z700_{NP,4}$	$SST_{ENP,4}$	$SST_{TNA,4}$	$Z500_{SWA,4-1}$	$Z700_{NP,4}$	$SST_{ENP,4}$
NTS	-0.61^{**}	-0.64^{**}	-0.56^{**}	0.48^{**}	-0.53^{**}	-0.61^{**}	-0.55^{**}	0.43^{**}
$SST_{TNA,4}$	1	0.13	0.15	−0.30	1	0.08	0.05	−0.19
$Z500_{SWA,4-1}$		1	0.34^{*}	−0.16		1	0.34^{*}	−0.08
$Z700_{NP,4}$			1	−0.16			1	−0.17
$SST_{ENP,4}$				1				1

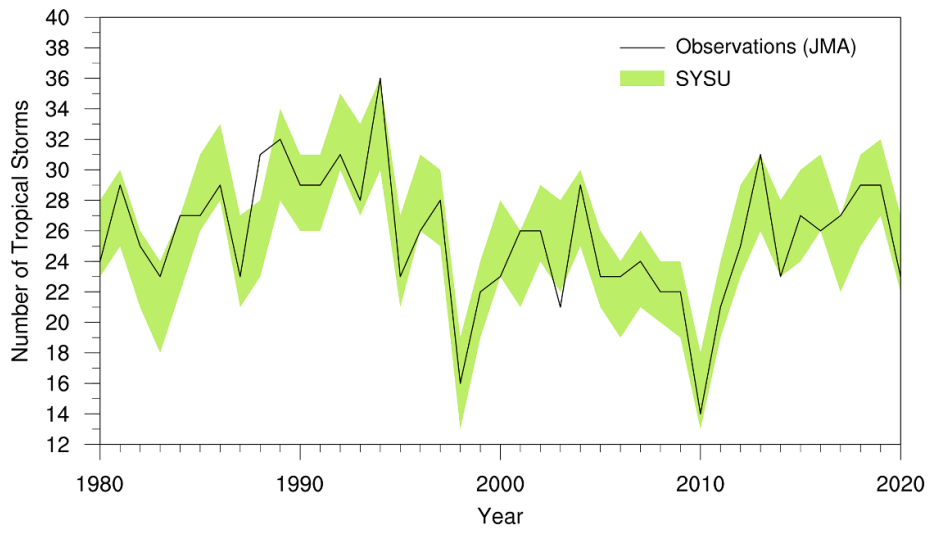

Figure 4. Model performance. Green shading indicates the prediction range of the SYSU Model. Note that the modeled results before 2016 are based on the model training, while those after 2016 are for the model validation.

4. Summary and discussion

In this study, an effective, open-source and competitive statistical seasonal forecasting model for predicting the NTS over the western North Pacific is proposed, namely the SYSU Model. Different to most of the existing seasonal forecasting models that consider the peak-season environment, the pre-season environmental factors are utilized to build the model. Four predictors are optimally selected from the 103 candidate predictors using the best subset selection regression. They are the SSTs at the tropical North Atlantic ($SST_{TNA,4}$) and eastern North Pacific ($SST_{ENP,4}$) in April, the 500 hPa geopotential height difference between April and January in the open ocean southwest of Australia ($Z500_{SWA,4-1}$) and the

700 hPa geopotential height in the North Pacific in April ($Z700_{NP,4}$). The correlation reaches 0.89 above in 1980–2015. The regression model is successfully validated by the leave-one-out, nine-fold cross-validations, and later 5 year (2016–2020) observations. The SYSU Model exhibits a 95% hit rate in 1980–2020, which is satisfactorily promising for a statistical model that is only composed of four predictors. It is effectively superior to many existing dynamical and statistical-dynamical models, in terms of time and computational power.

These four predictors show notably high and significant ($p < 0.05$) correlations with some climatic signals. The contemporaneous correlation coefficients between $SST_{TNA,4}$ and Atlantic Meridional Mode (AMM_4), $SST_{ENP,4}$ and Pacific Meridional Mode (PMM_4), $Z700_{NP,4}$ and North Pacific Oscillation (NPO_4), and $Z500_{SWA,4-1}$ and the Southern Annular Mode difference between April and January (SAM_{4-1}) are 0.812, 0.832, −0.635 and 0.409, respectively. These suggest the SYSU Model is implicitly physical, and could possess relevant climate dynamics mechanisms.

The high correlations between the two oceanic predictors, $SST_{TNA,4}$ and $SST_{ENP,4}$, and the meridional modes imply that the mechanisms of the meridional modes could be highly related to the TC activity in the western North Pacific. Zhang *et al* (2017b) found that the positive phase of AMM leads to a stronger descending branch in the tropical eastern and central Pacific, which intensifies the ascending branch of the Walker circulation in the western North Pacific. The stronger westerly in the upper troposphere and the stronger easterly in the lower troposphere associated with the enhanced Walker circulation result in stronger zonal shear, which suppresses the TC genesis. In addition, Wang *et al* (2019), Wu *et al* (2021) and Zhang *et al* (2020) investigated the roles of the PMM and the El Niño–Southern Oscillation (ENSO; in terms of the anomalous Walker circulation) in TC activity. They found that when the PMM is associated with the in-phase ENSO event, the positive relationship between the PMM and NTS becomes stronger.

Compared with the oceanic predictors, the mechanisms of the other two atmospheric predictors, $Z700_{NP,4}$ and $Z500_{SWA,4-1}$, are relatively uncertain. Newman *et al* (2016) suggested that the NPO signal in the atmosphere can be stored as the oceanic Pacific decadal oscillation (PDO) signal which can last for a few months. When the $Z700_{NP,4}$ is positive, the PDO signal tends to be in negative phase (not shown). The negative PDO signal in the spring can be characterized by the horseshoe SST distribution in the North Pacific (anomalous cooling at high (∼50° N) and low latitudes (∼10° N) and anomalous warming at middle latitudes (20°–30° N)) (Chen *et al* 2015). This pre-season oceanic signal can last into the peak season and then initiate three climate correspondences. First, it enhances the Pacific Walker circulation, and thus, strengthens the zonal shear, which is detrimental to the TC genesis. Second, the anomalous easterly in the lower troposphere weakens the monsoon trough activity. Finally, the anomalously warm SST at middle latitudes corresponds to the stronger subtropical high, which also suppresses the TC activity. This agrees with Zhao *et al* (2018) that the negative phase of PDO can result in a decrease in TC activity in the western North Pacific. On the other hand, Wang *et al* (2007) proposed a mechanism without the oceanic correspondence. They suggested that when the NPO is in negative phase (i.e. positive $Z700_{NP,4}$), the mid-latitude westerly jet stream is weaker, which leads to an anomalous cyclonic circulation in the tropical western North Pacific, and thus, strengthens the vertical wind shear and suppresses the TC activity.

Finally, the high correlation between the $Z500_{SWA,4-1}$ and SAM could suggest that the Southern Hemisphere Rossby wave at middle latitudes modulates the Madden–Julian Oscillation (MJO) activity with the upper-tropospheric zonal wind (Lin 2019), and consequently affects the TC activity (Nakazawa 1988, Liebmann *et al* 1994). However, it is uncertain how these signals can exist, propagate and/or transform in the atmosphere for a few months. Air-sea interaction should be involved in between. Meanwhile, Ho *et al* (2005), Wang and Fan (2007) and Choi *et al* (2010) gave disputing arguments about the relationship between the SAM and TC activity over the western North Pacific. Further study is urged on the linkage between the $Z500_{SWA,4-1}$, SAM, Rossby wave and MJO.

No model is perfect. Although the physics and underlying mechanisms could not all be clearly corroborated, which is indeed an intrinsic deficiency of the statistical model, the performance of the SYSU Model demonstrates high potential in forecasting and merits consideration. The discussed mechanisms are the auxiliary, which suggests possible physical reasoning for the model. Further research is encouraged. Last but not least, in the near future the SYSU Model will be officially implemented and kept up-to-date, targeting the issuing of an open access forecast by every middle of May.

Data availability statement

The JMA best-track data were obtained from www.jma.go.jp. The CMA best-track data were extracted from http://tcdata.typhoon.org.cn. The JTWC best-track data were downloaded from www.metoc.navy.mil/jtwc/. The ERA5 reanalysis data were retrieved from www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5. The NCEP-CFSR reanalysis data were downloaded from https://rda.ucar.edu/datasets/ds093.0 and https://rda.ucar.edu/datasets/ds094.0. The GODAS reanalysis data were available at www.esrl.noaa.gov/psd/data/gridded/data.godas.html. The MERRA-2

reanalysis data were obtained from https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/. The ERSSTv5 SST analysis data were downloaded from www.esrl.noaa.gov/psd/data/gridded/data.noaa.ersst.v5.html. The HadISST sea ice data were retrieved from www.metoffice.gov.uk/hadobs/hadisst/. The RUS snowfall data were extracted from https://climate.rutgers.edu/snowcover/. The solar activity parameters were obtained from https://omniweb.sci.gsfc.nasa.gov/html/ow_data.html. The sunspot data were downloaded from www.sidc.be/silso/datafiles. The PMM index was retrieved from https://psl.noaa.gov/data/timeseries/monthly/PMM/. The AMO index was available at https://psl.noaa.gov/data/timeseries/AMO/. The AMM index was extracted from https://psl.noaa.gov/data/timeseries/monthly/AMM/. The AO index was obtained from https://psl.noaa.gov/gcos_wgsp/Timeseries/AO/. The NAO index was retrieved from https://psl.noaa.gov/gcos_wgsp/Timeseries/NAO/.

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

The authors sincerely thank O Y W Cheung, X Dong, J Liang and two anonymous reviewers for the discussion. This study was supported by the National Key R&D Program of China (Grant No. 2019YFC1510400), the National Natural Science Foundation of China (Grant No. 41975052), the National Natural Science Foundation of China and Macau Science and Technology Development Joint Fund (NSFC-FDCT) (Grant No. 41861164027), the Guangdong Province Key Laboratory for Climate Change and Natural Disaster Studies (Grant No. 2020B1212060025) and the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (Grant No. 2021qntd29).

ORCID iD

Kelvin T F Chan https://orcid.org/0000-0001-6150-7612

References

Au-Yeung A Y M and Chan J C L 2012 Potential use of a regional climate model in seasonal tropical cyclone activity predictions in the western North Pacific *Clim. Dyn.* **39** 783–94

Camargo S J, Barnston A G, Klotzbach P J and Landsea C W 2007 Seasonal tropical cyclone forecasts *WMO Bull.* **56** 297–309

Camp J, Robert M J, Comer R E, Wu P, MacLachlan C, Bett P E, Golding N, Toumi R and Chan J C L 2019 The western Pacific subtropical high and tropical cyclone landfall: seasonal forecasts using the met office GloSea5 system *Q. J. R. Meteorol. Soc.* **145** 105–16

Chan J C L 1995 Tropical cyclone activity in the western North Pacific in relation to the stratospheric quasi-biennial oscillation *Mon. Weather Rev.* **123** 2567–71

Chan J C L, Shi J E and Lam C M 1998 Seasonal forecasting of tropical cyclone activity over the western North Pacific and the South China Sea *Weather Forecast.* **13** 997–1004

Chen D, Wang H, Liu J and Li G 2015 Why the spring north pacific oscillation is a predictor of typhoon activity over the western north pacific *Int. J. Climatol.* **35** 3353–61

Chen Z, Gan B, Wu L and Jia F 2018 Pacific-North American teleconnection and North Pacific Oscillation: historical simulation and future projection in CMIP5 models *Clim. Dyn.* **50** 4379–403

Choi K S, Moon J Y, Kim D W and Chu P S 2010 Seasonal prediction of tropical cyclone genesis frequency over the western North Pacific using teleconnection patterns *Theor. Appl. Climatol.* **100** 191–206

Derber J and Rosati A 1989 A global oceanic data assimilation system *J. Phys. Oceanogr.* **19** 1333–47

Ding R, Li J, Tseng Y H, Sun C and Guo Y 2015 The Victoria mode in the North Pacific linking extratropical sea level pressure variations to ENSO *J. Geophys. Res. Atmos.* **120** 27–45

Ding R, Li J, Tseng Y H, Sun C, Li Y, Xing N and Li X 2019 Linking the North American dipole to the Pacific Meridional mode *J. Geophys. Res. Atmos.* **124** 3020–34

Estilow T W, Young A H and Robinson D A 2015 A long-term Northern Hemisphere snow cover extent data record for climate studies and monitoring *Earth Syst. Sci. Data* **7** 137–42

Fan K 2007a New predictors and a new prediction model for the typhoon frequency over western North Pacific *Sci. China Earth Sci.* **50** 1417–23

Fan K 2007b North Pacific sea ice cover, a predictor for the western North Pacific typhoon frequency? *Sci. China Earth Sci.* **50** 1251–7

Gao S, Zhu L, Zhang W and Chen Z 2018 Strong modulation of the Pacific Meridional mode on the occurrence of intense tropical cyclones over the western North Pacific *J. Clim.* **31** 7739–49

Gelaro R *et al* 2017 The modern-era retrospective analysis for research and applications, version 2 (MERRA-2) *J. Clim.* **30** 5419–54

Gray W M 1968 Global view of the origin of tropical disturbances and storms *Mon. Weather Rev.* **96** 669–700

Guo Y and Tan Z 2018 Impacts of the boreal spring Indo-Pacific warm pool Hadley circulation on tropical cyclone activity over the western North Pacific *J. Clim.* **31** 1361–75

Hersbach H *et al* 2020 The ERA5 global reanalysis *Q. J. R. Meteorol. Soc.* **146** 1999–2049

Ho C H, Kim J H, Kim H S, Sui C H and Gong D Y 2005 Possible influence of the Antarctic oscillation on tropical cyclone activity in the western North Pacific *J. Geophys. Res. Atmos.* **110** D19104

Hu C, Lian T, Cheung H-N, Qiao S, Li Z, Deng K, Yang S and Chen D 2020 Mixed diversity of shifting IOD and El Niño dominates the location of Maritime Continent autumn drought *Natl. Sci. Rev.* **7** 1150–3

Huang B *et al* 2017 Extended reconstructed sea surface temperature version 5 (ERSSTv5): upgrades, validations, and intercomparisons *J. Clim.* **30** 8179–205

Ivanov K G 2007 Correlation between tropical cyclones and magnetic storms during cycle 23 of solar activity *Geomagn. Aeron.* **47** 371–4

King J H and Papitashvili N E 2005 Solar wind spatial scales in and comparisons of hourly wind and ACE plasma and magnetic field data *J. Geophys. Res. Space Phys.* **110** A02209

Klotzbach P *et al* 2019 Seasonal tropical cyclone forecasting *Trop. Cycl. Res. Rev.* **8** 134–49

Li H, Wang C, He S, Wang H, Tu C, Xu J, Li F and Guo X 2019 Plausible modulation of solar wind energy flux input on global tropical cyclone activity *J. Atmos. Sol. Terr. Phys.* **192** 104775

Liebmann B, Hendon H H and Glick J D 1994 The relationship between tropical cyclones of the western Pacific and Indian oceans and the Madden Julian Oscillation *J. Meteorol. Soc. Jpn.* **72** 401–11

Lin Z 2019 The South Atlantic–South Indian Ocean pattern: a zonally oriented teleconnection along the Southern Hemisphere westerly jet in austral summer *Atmosphere* **10** 259

Nakazawa T 1988 Tropical super clusters within intraseasonal variations over the western Pacific *J. Meteorol. Soc. Jpn.* **64** 17–34

Newman M *et al* 2016 The Pacific decadal oscillation, revisited *J. Clim.* **29** 4399–427

Rayner N A, Parker D E, Horton E B, Folland C K, Alexander L V, Rowell D P, Kent E C and Kaplan A 2003 Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century *J. Geophys. Res. Atmos.* **108** 4407

Saha S *et al* 2010 The NCEP climate forecast system reanalysis *Bull. Am. Meteorol. Soc.* **91** 1015–57

Saha S *et al* 2014 The NCEP climate forecast system version 2 *J. Clim.* **27** 2185–208

Takahashi C, Watanabe M and Mori M 2017 Significant aerosol influence on the recent decadal decrease in tropical cyclone activity over the western North Pacific *Geophys. Res. Lett.* **44** 9496–504

Tian B and Fan K 2019 Seasonal climate prediction models for the number of landfalling tropical cyclones in China *Meteorol. Res.* **33** 837–50

Tu J Y, Chou C, Huang P and Huang R 2011 An abrupt increase of intense typhoons over the western North Pacific in early summer *Environ. Res. Lett.* **6** 034013

Wang C, Wang B and Cao J 2019 Unprecedented Northern Hemisphere tropical cyclone genesis in 2018 shaped by subtropical warming in the North Pacific and the North Atlantic *Geophys. Res. Lett.* **46** 13327–37

Wang H and Fan K 2007 Relationship between the Antarctic Oscillation in the western North Pacific typhoon frequency *Chin. Sci. Bull.* **52** 561–5

Wang H, Sun J and Fan K 2007 Relationships between the North Pacific Oscillation and typhoon/hurricane frequencies *Sci. China Earth Sci.* **50** 1409–16

Wang Y, Song T, Liang J and Pan W 2012 Simulation of seasonal tropical cyclone activity over the western North Pacific by using the WRF model *Trans. Atmos. Sci.* **35** 24–31

Wu Q, Zhao J, Zhan R and Gao J 2021 Revisiting the interannual impact of the Pacific Meridional Mode on tropical cyclone genesis frequency in the western North Pacific *Clim. Dyn.* **56** 1003–15

Xia L, Zhu Y, Cheng C and Chen X 2014 Interannual to interdecadal variations and prediction of tropical cyclone frequency over the western North Pacific *Meteorol. Environ. Sci.* **37** 1–7

Xiao D and Li J 2007 Main decadal abrupt changes and decadal modes in global sea surface temperature field *Chin. J. Atmos. Sci.* **31** 839–54

Zhan R, Wang Y and Wen M 2013 The SST gradient between the southwestern Pacific and the western Pacific warm pool: a new factor controlling the northwestern Pacific tropical cyclone genesis frequency *J. Clim.* **26** 2408–15

Zhan R, Wang Y and Ying M 2012 Seasonal forecasts of tropical cyclone activity over the western North Pacific: a review *Trop. Cyclone Res. Rev.* **1** 307–24

Zhang H, Wu L, Huang R, Chen J-M and Feng T 2020 Does the Pacific Meridional Mode dominantly affect tropical cyclogenesis in the western North Pacific? *Clim. Dyn.* **55** 3469–83

Zhang W, Vecchi G A, Murakami H, Villarini G and Jia L 2016 The Pacific Meridional Mode and the occurrence of tropical cyclones in the western North Pacific *J. Clim.* **29** 381–98

Zhang W, Vecchi G A, Villarini G, Murakami H, Gudgel R and Yang X 2017a Statistical–dynamical seasonal forecast of western North Pacific and East Asia landfalling tropical cyclones using the GFDL FLOR coupled climate model *J. Clim.* **30** 2209–32

Zhang W, Vecchi G A, Villarini G, Murakami H, Rosati A, Yang X, Jia L and Zeng F 2017b Modulation of western North Pacific tropical cyclone activity by the Atlantic Meridional Mode *Clim. Dyn.* **48** 631–47

Zhang W and Villarini G 2019 Seasonal forecasting of western North Pacific tropical cyclone frequency using the North American multi-model ensemble *Clim. Dyn.* **52** 5985–97

Zhang W, Villarini G, Vecchi G A and Smith J A 2018 Urbanization exacerbated the rainfall and flooding caused by hurricane harvey in houston *Nature* **563** 384–8

Zhao J, Zhan R, Wang Y and Xu H 2018 Contribution of interdecadal Pacific oscillation to the recent abrupt decrease in tropical cyclone genesis frequency over the western North Pacific since 1998 *J. Clim.* **31** 8211–24

Zhou B and Cui X 2008 Hadley circulation signal in the tropical cyclone frequency over the western North Pacific *J. Geophys. Res. Atmos.* **113** D16107

Zhou B and Cui X 2011 Sea surface temperature east of Australia: a predictor of tropical cyclone frequency over the western North Pacific? *Chin. Sci. Bull.* **56** 196–201

Zhou B and Cui X 2014 Interdecadal change of the linkage between the North Atlantic Oscillation and the tropical cyclone frequency over the western North Pacific *Sci. China Earth Sci.* **57** 2148–55