<u>Response to Reviewer 2's comments</u>

We thank Reviewer 2 for the time to go through our manuscript in details. This manuscript describes a new and efficient method to produce a physical TC event set in the western North Pacific basin. In general, reviewers think after careful revision, the results of this study is of great interest and relevance, and it will be a useful contribution to the field of TC risk assessment. Here is our point-to-point response to Reviewer 2's comments.

<u>General comments:</u>
This manuscript describes a new and efficient method to produce TC event set in the Pacific basin. The TC events are detected from an ensemble data archive TIGGE, using an objective impact-oriented windstorm identification algorithm WiTRACK. This dataset contributes to existing synthetic datasets (mostly statistical basin-wide methods) as in this dataset the TCs are detected in GCMs, so that the complex physical processes of TCs are captured and hence TCs are physically realistic. More extreme TCs are found in the dataset and these data will help overcome the shortage in observational record. Overall, I think the results will be a nice contribution to the field of TC risk assessment. However, I have some basic questions about the methodology, the utility of certain results and I recommend major revisions prior to publication. Also, I would recommend careful editing of the manuscript. There are many terminologies in the manuscript and please make sure it is easy for readers to follow.

We thank the reviewer advice and we have carefully reviewed and edited the manuscript.

<u>Major comments:</u>
1) L191: The detection rates of historical TCs are reported here. However, will the detection algorithm produce more TCs? What fraction of TCs that the detection algorithm produce is real historical TCs?

We thank the reviewer for pointing this out and highlighting a section for which we can improve clarity. Strictly spoken, the detection algorithm we apply in this study (developed for TC detection in the West-Pacific in Befort et al., 2020) does not produce TCs, but enables us to detect them automatically in the large data set. As the data set we use is the output of operational NWP's forecast models, in a very narrow sense , none of the TCs that we detected is a one-to-one equivalent to a real historical TCs. However, events which satisfy the criteria in the MPES TC identifier (MTI, Section 3.2.3) (i.e. MEPS TC events) can be considered as events which are similar to the historical event. The percentage of TCs which are in this sense similar to real events that occurred in the TIGGE is ~60%. Thus, about 40% are pure ensemble predicted events that did not realise in the observed nature or do not have a very similar twin at the same time at the same location.

We have modified the text in L191 and the caption of Table 4 for more accurate description. Lines 225-228, "*A historical TC is said to be detected in a forecast model if there exists a TC counterpart in the forecast model, which is similar to the historical TC as identified by the MTI (Section 3.2.3). The detection rates of historical TCs which are detected in different* *forecast outputs, i.e. CMA, ECMWF, JMA, and NCEP, are 91.2%, 94.7%, 89.4%, and 90.7%, respectively, ...*"

2) The authors mention that one benefit of this dataset is "The TPEPS event set includes events which are unlikely but physically possible. This provides an important and unique advantage for typhoon risk assessment." Combined with Fig. 2, TC tracks in the detected dataset is very different with observations, and TPEPS tracks appear in locations with no historical tracks. If there is no historical track in some regions, are they supposed to be no storms or there can be storms but no storm has appeared in historical records due to the low probability? This needs to be explained.

We thank Reviewer 2 for pointing out this important issue. If there is no historical track in some regions, this does not mean storms cannot occur in those regions. The fact that we have not seen a TC during the time period of known observational records in those regions could be due to the observation period is too short and the sample size is not large enough to fully represent the distribution of the underlying basic population (i.e. all possible TCs in the given climate state). For example, if we follow the necessary but insufficient conditions of TC formation which are identified by Gray (1977) from historical observations, TC formation occurs away from the equator (> 5 deg). However, Tropical Storm Vamei (2001) formed close to the equator (~1.4 deg N). This shows storm can appear in the historically "storm-free" region.

Furthermore from the statistical perspective, we can view the JRA-55 event set as a subset which is randomly selected from the TPEPS event set. This means if we randomly sample the TPEPS event set, we can obtain a subset highly similar to the JRA-55 event set. For demonstration, we have conducted bootstrap resampling on the TPEPS event set to obtain 10,000 sets of subsample. Each set of subsamples has 668 events to mimic the number of events in the JRA-55 event set. For each set of subsamples, the track density is calculated, and used to calculate uncentred pattern correlation between the resampling set of subsamples and the JRA-55 event set. In order to focus on relevant entries, for a particular grid box, if the values of track density for a resampling set and the JRA-55 event set are both less than one, such grid box is not used in the pattern correlation calculation. The mean, standard deviation, minimum, and maximum of the uncentred pattern correlation of the 10,000 set of subsamples are 0.9380, 0.0107, 0.8961, and 0.9697, respectively. This suggests the spatial pattern of the JRA-55 event set is highly similar to a small random subset of the TPEPS event set. Consequently, the JRA-55 event set can be seen as a subset randomly selected from the TPEPS event set. On the other hand, it is not be possible to deduce the basic population (e.g. the TPEPS event set) from a small sample set (e.g. the JRA-55 event set). Although the spatial distribution of the small set sample is similar to the subsamples of the basic population and thus usable as one possible realisation of the basic population, the small sample set does not contain all of the information of the underlying population. Furthermore, the statistical estimate of extremes would also be different for the small sample set (e.g. JRA-55 event set) and the basic population (e.g. TPEPS event set).
We have included the above explanation in the revised manuscript (Lines 259-280).

3) The sensitivity and performance of four ensemble data archive are not well described. For example, in some dataset, the storms are much weaker than historical storms. And some models *have biases in simulating extratropical cyclone transition. More explanations and descriptions of the data archive needs to be added. Also, how these biases would have an impact on the detection algorithm?*

The four data sets selected from the TIGGE archive are the state-of-the-art NWP models as used by four leading synoptic weather forecast centres worldwide. Although a full assessment of their respective models' skill and potential biases is not in the scope of this study, **we added a section with information on the general performance of these four selected NWP models (Lines 106-116)**. For the dedicated purpose of this study, the reviewer is fully correct and we need to check for biases in the underlying climatological features as provided by a time- and ensemble-aggregated view of the data set (a task normally not necessarily done in forecast model evaluation departments for all levels of severe and rare extremes). This evaluation for extreme TC occurrence is what we did in section 4.1, showing respective results in Fig.1-7. **We included a paragraph to clarify which part of the study is model validation and which part is event set building (Lines 223-225)**.

TIGGE data's main difference to the operationally used NWP output is that TIGGE did archive a lower resolution. Nevertheless, all underlying processes and feedbacks are captured in the originally resolution of the NWP products and are thus fully included. Thus, we would expect the best possible representation of dynamical processes in those forecast simulations than compared to lower resolution AOGCM simulations, e.g. for transient climate experiments. Beyond this, model resolution is known to be a limiting fact of simulating TC intensity (Bengtsson et al., 2007). One of the advantages of using WiTRACK is that it does not use raw wind speeds, instead, it uses the 98th percentile relative exceedance for tracking. This means that even if the simulation wind speed of TC is systematically weaker than in historical observations, the 98th percentile climatological wind should also be lower than the actual 98th percentile climatological wind speed, a TC will still be tracked as long as there exists a 98th percentile exceedance wind cluster. It can be shown that, within the study area, the 98th percentile relative exceedance of the 4 models, which we used to construct the TIGGE event set, have similar behaviour (i.e. similar to Figure 2 of Osinski et al. (2016)). Befort et al. (2020) showed the applicability of such an approach to relate information from observations (i.e. IBTrACS data) to automatically detected TCs from a much coarser resolution reanalysis product (JRA-55). Consequently, a bias due to resolution does not have significant impact on WiTRACK as the tracking algorithm serves as a bias correction in this sense (detailed discussion on the impact of weaker wind speed in model outputs on WiTRACK can be found in Osinski et al. (2016)). **We included a paragraph to discuss this in more detail (Lines 126-128; 141-152)**.

4) The authors have compared the TIGGE PEPS TCs with JRA-55 in terms of track density, landfall frequency, etc. How about other characteristics? For example, landfall intensity along coastline?

The distribution of landfall intensity (wind speed in m/s) for TC, which made landfall with at least typhoon strength, are very similar for the JRA-55 event set and TPEPS event set. The table below shows some of the statistics of these two distributions. The two-sample Kolmogorov-Smirnov test show these two distributions belong to the same distribution significant at the 0.05 significance level.

	JRA-55	TPEPS
Mean	23.5899	23.4044
Standard deviation	3.44527	3.84537
Median	22.58	22.2
Number of events	184	23343

We have included the above discussion into revised manuscript [Lines 344-347].

5) Fig. 7 shows the difference between TIGGE PEPS event set and observation. In the text, you have mentioned possible reasons for these differences. Is there possible way to reduce these differences, for example in the detection algorithm, to also remove low-impact storms? Also, you mentioned the ESSI, is there a way to quantify this index?

Figure 7 shows some of the differences between the TPEPS event set and the JRA-55 event set. These differences are mainly due to the finite simulation time in forecast models. Some of these differences could be reduced based on additional assumptions that would depend on the specific application of the users. A more detailed analysis of the performance with respect to a data set not affected by a finite simulation time is a reanalysis product (e.g. JRA-55). We showed in Befort et al. (2020) in JRA-55 that our tracking already focusses on the most severe part of the TC severity distribution and thus does show some expected differences to e.g. IBTrACS data.

We apologise we did not include the text associated with the SSI (and ESSI). Leckebusch et al. (2008) introduced this objective severity measure for gridded datasets of extreme storms in the North-Atlantic and the method was applied for TCs in the North-West Pacific in Befort et al. (2020). **The relevant text is included in revised manuscript [Lines 215-219]**.

Minor comments:

L41-42: more recent papers should be added. Such as the following two recent models:
- Lee, C.-Y., M. K. Tippett, A. H. Sobel, and S. J. Camargo, 2018: An environmentally forced tropical cyclone hazard model. Journal of Advances in Modeling Earth Systems, 10 (1), 223–241.
- Jing, R., and N. Lin, 2020: An environment-dependent probabilistic tropical cyclone model. Journal of Advances in Modeling Earth Systems, 12 (3), e2019MS001 975.

We thank the reviewer's suggestions. **We have included these references in the revised manuscript.**

L45: I didn't understand the sentence 'the typhoon event set might not be physically consistent'. What is 'physically' consistent?

It means event sets created by stochastic perturbations will create TC events that (with respect to their inner dynamical structure) are not necessarily physically consistent anymore. As just surface footprints are stochastically modelled from existing tracks, there is no check whether those events (in the stochastically modelled from) are physically possible and how they could be realised in a fully dynamical consistent view, thus fulfilling all known physical relations and derived constraints by the means of physical laws. Consequently, the amount of unrealistic physical properties due to the oversimplified stochastic simulation is unknown and laws of physical interactions are potentially ignored. **We have modified the sentence in the revised manuscript to clarify this point [see lines 46-52]**.

L79: "The domain of this study covers the Western North Pacific (WNP), east and south-east Asia spanning from 85 E to 195E and 15 S to 75 N." Why data around equator is also used? There is no TCs forming around equator.

We thank Reviewer 2 for pointing this out and we apologise for the confusion. The domain stated in the manuscript is part of the parameter set up for WiTRACK. However, the true domain which is used for tracking is 90-180° E, and 0-70° N and **we have made this correction in the revised manuscript.**

We included regions close to the equator although TCs rarely form around the equator, it is still possible for TCs to form close to the equator, for example Tropical storm Vamei (2001). Furthermore, while the core pressure centre of the TCs might be away from the equator, the damaging wind field, as identified by the 98th percentile relative exceedance, could be quite large, impacting potentially regions close to the equator.

L102-104: Is there a reason why an old version of IBTrACS is used?

IBTrACS v03r10 was the most up-to-date official version of IBTrACS when this study was first started. Furthermore, for our study period (with 6-hourly observations), the data in v04 and v03r10 are the same.

L152: "the accuracy of the LRC is about 90%" What is the fraction of TC (or positive samples?) Does there exist issue of imbalanced data?

No, the validation set is not imbalanced. In the validation set, 49 out of 96 tracks are TCs (~51% of the validation set). **We have included a more detailed description in the revised manuscript. Lines 180-181** "*Validation using JRA-55 event set (2015-2017), which has 49 TC events and 47 non-TC events…*"

L197: "Percentage of total TC windstorms as PEPS TCs can be treated as a proxy to quantify the forecast skill of the model." In Table 5, NCEP is almost twice of that in JMA, what does this percentage mean?

This indicates the NCEP model generates more "wrong" forecasts than JMA yet these wrong forecasts are physically possible. **We included a clarifying sentence to a respective possible interpretation at lines 235-238**: *"For example, NCEP has 47.1% of TC windstorms as PEPS TCs whereas JMA has 26.5%. This indicates the NCEP model generates more "wrong" forecast than JMA however these wrong forecasts are physically possible. Yet, examining the* *forecast skill of models is not the focus of this study and the rest of the discussion focuses on the TPEPS TC event set.*"

L203: do you mean Fig 3? Also, more explanations should be added in the text. I can't understand this figure.

We thank Reviewer 2 for identifying this error. The reviewer is correct that we are referring to Fig 3. Fig. 3 shows the feature scaled times series of number of TCs which are first identified in each day from May to December. The core message of Fig 3 is that the temporal variability of the TPEPS event set and the JRA-55 event set are largely similar (except for the earlier years). **We have modified the text in the revised manuscript. Lines 239-240** "*Figures 2 and 3 show the spatial pattern and temporal variability of the number of TC which are first detected for each day, …*"

L203: In Fig. 2, all TPEPS are much more similar with each other, comparing with JRA-55. How to explain this?

The major difference between the track density of TPEPS and JRA-55 is that there is an eastward bias in the TPEPS. There are several reasons that could contribute to this. The eastward bias in the track density appears to be a common feature in many GCMs (e.g. Camargo et al., 2005; Bell et al., 2013; Roberts et al., 2020), this has also been observed in seasonal forecast output (Camp et al., 2015). Finite simulation time has also contributed to this bias as TC that forms in the region east of 150 °E would not have the time to move into the western part of WNP before the end of simulation time. Differences in the amount of tracks could also contribute to the differences as more diverse tracks would be captured. **We have added a respective explanatory comment at lines 252-258**.

Fig4: The tracks in black are very easily messed up with the map. Probably change the color of coastline.

We thank Reviewer 2 for pointing this out. **We have changed the colour of the plot**.

Fig5: The y-axis is not clear to me, please add more explanation.

Fig 5 shows the climatological seasonal cycle of TC activity for the TPEPS TC event set and the JRA-55 event set. The daily number distribution, p_i, is calculated as follows:

$$p_i = \frac{n_i}{\sum_i n_i} \times 100\%$$

where n_i is the number of TC first detected on day i for the individual event set. As such, it is the probability of TC being first detected at a given day**. We have added more explanation in the caption of Figure 5.**

Fig8: The colored dots for single center are too light to see. If this figure is to show distribution, I would recommend not using same color bar for single model and for TIGGE total.

We have changed the colour scale of this figure in the revised manuscript.

Fig9: It's hard to see the distributions are in good agreement, probably can change to annual frequency instead of total number of landfall events. Also, the correlation coefficients could be used to show the landfall frequency in all TIGGE dataset is positively correlated with JRA-55.

We thank Reviewer 2 for these suggestions. **We have included pattern correlation between the spatial distribution for JRA-55 and TPEPS event sets in the revised manuscript. Line 338** "*…with uncentred pattern correlation of 0.8345*."

Fig12: I can see your points in showing the grey dashed lines. But the lower bound curves cannot show the trend properly. I would recommend add 75% or 80% confidence interval to show that the trends are same, but TIGGE PEPS event set has much narrower bounds.

We are not certain what Reviewer 2 refers to as the trend of the lower bound curves. There are two separate factors that determine the "shape" of the curve of the lower and upper bound of uncertainty. First, the return level-return period estimate has asymptotic behaviour. This means the return level estimate approaches to a certain value as the return period increases. Second, the uncertainty of the estimation increases with increasing return period. Combining these two factors we can see that the so-called "trend" in the lower bound grey curve does not exist. To show the 95% confidence interval reflects a typical setting for assessing statistical estimates uncertainty for GPD fitted return-level plots and the authors would prefer to stay with this representation.

References

Bell, R., Strachan, J., Vidale, P. L., Hodges, K., and Roberts, M.: Response of Tropical Cyclones to Idealized Climate Change Experiments in a Global High-Resolution Coupled General Circulation Model, J Climate, 26, 7966-7980, 10.1175/JCLI-D-12-00749.1, 2013.

Bengtsson, L., Hodges, K. I., and Esch, M.: Tropical cyclones in a T159 resolution global climate model: Comparison with observations and re-analyses, Tellus A, 59, 396-416, 2007.

Camargo, S. J., Barnston, A. G., and Zebiak, S. E.: A statistical assessment of tropical cyclone activity in atmospheric general circulation models, Tellus A, 57, 589-604, 10.1111/j.1600-0870.2005.00117.x, 2005.

Camp, J., Roberts, M., MacLachlan, C., Wallace, E., Hermanson, L., Brookshaw, A., Arribas, A., and Scaife, A. A.: Seasonal forecasting of tropical storms using the Met Office GloSea5 seasonal forecast system, Q J Roy Meteor Soc, 141, 2206-2219, 10.1002/qj.2516, 2015.

Gray, W. M.: Tropical Cyclone Genesis in the Western North Pacific, J Meteorol Soc Jpn, 55, 465-482, 1977.

Osinski, R., Lorenz, P., Kruschke, T., Voigt, M., Ulbrich, U., Leckebusch, G. C., Faust, E., Hofherr, T., and Majewski, D.: An approach to build an event set of European windstorms based on ECMWF EPS, Nat. Hazards Earth Syst. Sci., 16, 255-268, 10.5194/nhess-16-255-2016, 2016.

Roberts, M. J., Camp, J., Seddon, J., Vidale, P. L., Hodges, K., Vanniere, B., Mecking, J., Haarsma, R., Bellucci, A., Scoccimarro, E., Caron, L.-P., Chauvin, F., Terray, L., Valcke, S., Moine, M.-P., Putrasahan, D., Roberts, C., Senan, R., Zarzycki, C., and Ullrich, P.: Impact of Model Resolution on Tropical Cyclone Simulation Using the HighResMIP–PRIMAVERA Multimodel Ensemble, J Climate, 33, 2557-2583, 10.1175/JCLI-D-19-0639.1, 2020.