Statistical Framework for Utilisation of Modelled Data for Tropical Cyclones

Problem presented by

Richard Sproson & Dave Sproson

Fugro GEOS

Report author

Study Group 1

Executive Summary

Fugro GEOS produce wind and wave design criteria e.g. estimates of 100-year or 10,000-year return period values for wind and wave heights, linked to tropical cyclones in various regions, including those where such events occur relatively infrequently. Modelling individual storms by multiple model runs with small differences in the input parameters and model physics allows to produce a collection of storm events, and to predict extremes based on the obtained data sample.

The task is to check the validity of the methods used, to estimate the confidence level that can be placed in the results of the analysis, and to suggest ways to improve this level. Two sets of data were presented. We perform statistical analysis of the data, and show that for NW Australia, where tropical cyclones are frequent and good hindcast databases are available, the data shows good agreement with the theoretical extreme value distribution, thus enabling one to get reliable prediction for significant wave height maxima. For Bay of Bengal data, where the number of historical tropical cyclones is very limited, we use quantile regression analysis to check the reliability of the estimates for maximum wave heights for individual cyclones.

A number of ways to improve the quality of the forecasts are proposed. First, we suggest a method for estimating the quantiles of the wind speeds (or wave heights) at several uniformly distributed points along the tracks trajectories, using the generated ensembles. This method is based on Harrell-Davis quantile estimation and allows to estimate the uncertainty associated with a given number of model runs, and to add more runs if required. Second, we discuss an optimal strategy for adding runs. Third, we propose a statistical model for the generation of synthetic tropical cyclone tracks, which is an inexpensive method to generate a large number of tropical cyclones in the basin of interest, where the hindcast data is insufficient.

Version 1.0
March 13, 2016
iv+13 pages

Contributors

Sergei Annenkov (Keele University)
Thomas P. Leahy (Imperial College London)
Emily Kawabata (University of Edinburgh)

Contents

1 Introduction and problem statement

Calculating return period values (e.g. estimates of 100-year or 10,000-year maximum wind or wave height) for tropical cyclones is a difficult task, especially in regions where such events occur relatively infrequently. Meanwhile, such estimates are crucial for risk analysis for offshore sites. Methods used are different for regions where tropical cyclones or hurricanes are frequent, and for those where they occur relatively infrequently. In the former case, the estimates are based on existing hindcast databases: data is taken from neighbouring grid points, and then the extreme values analysis is used to predict the extremes. In the latter case, there are no existing hindcast databases available, and the task becomes much more difficult.

The current operational research approach is to model individual storms, although getting both the storm track and intensity to match historical data is not easy. Currently one produces multiple model runs with small differences in the input parameters and model physics in order to produce a collection (ensemble) of plausible storm events. These "small differences" in input parameters and model physics can change both the track and intensity of the modelled storm, any of which could be considered as a possible representation of a cyclone in the region. However, these so-called small differences in the input parameters may have unknown effects to the track and intensity depending upon the model.

The current method for deriving extreme wave criteria in regions where cyclones are infrequent is:

1. Gathering the cyclone data for the region in question over the last 45 years. There is good satellite data since 1970, and there may have been, say, 6-10 severe storms in the area of interest during that period.
2. For each of those storms, running simulations of it with numbers of combinations of different microphysics models, different boundary layer models, different sea surface models etc. The simulations of the storm area take external conditions from a large-scale atmospheric model, interpolated to give values on the grid size of the storm simulation. The microphysics models are different ways of representing the homogenised effects of the physics at sub-grid scale. The boundary layer models are, similarly, different ways of representing the effects of the atmospheric boundary layers. They include, for instance, models of enthalpy transfer across the air-sea interface.
3. Using each of those simulated storms to predict wave height at the specified location. This is done using a standard wind-wave model.
4. Using that set of data to estimate the 1-in-100-year maximum wave height at the location.

First, we need to find out what confidence can be placed in the results of this analysis, and how that confidence level depends on the number and the type of variations in step 2. Two sets of data were presented. The

first set has maximal values of significant wave heights for 7 points off the northwestern coast of Australia, where a large hindcast database of historic tropical cyclones is available. The second set refers to Bay of Bengal, where tropical cyclones are rare (although they can be quite severe and dangerous). The data shows wave height evolution vs time at the target point in Bay of Bengal. In the next section, we perform statistical analysis of the two sets of data, in order to find out whether the extreme value estimates are statistically reliable. In section 3, we propose a number of solutions: a method to estimate whether the number of model runs is sufficient for a reliable statistical estimate, based on Harrell-Davis quantile analysis, a way to efficiently add model runs with optimal use of available computational resources, and a synthetic tropical cyclone generator. Section 4 contains conclusions.

2 Data analysis

2.1 North-West Australia - extreme value analysis

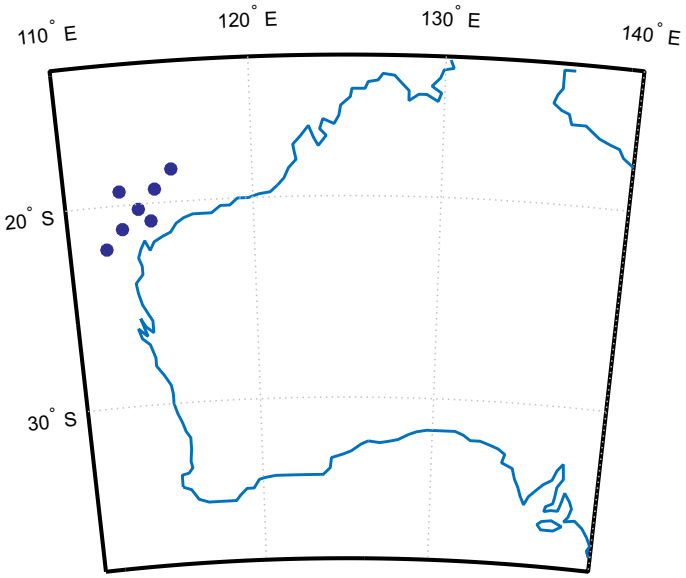

Figure 1: The data consists of maximal values of significant wave heights for 76 tropical cyclones at 7 points.

The first set of data contains maximal values of wind speed and significant wave height for 76 tropical cyclones. Data is pooled from 7 neighbouring grid points (figure 1). This data is used to predict the maximal values of wind speed and wave height with a specified probability, or return period.

In order to assess the quality of the data and the reliability of the prediction, in figure 2 we plot histograms of significant wave height maxima for all 7 points. Theoretically, the probability density function for the extreme value

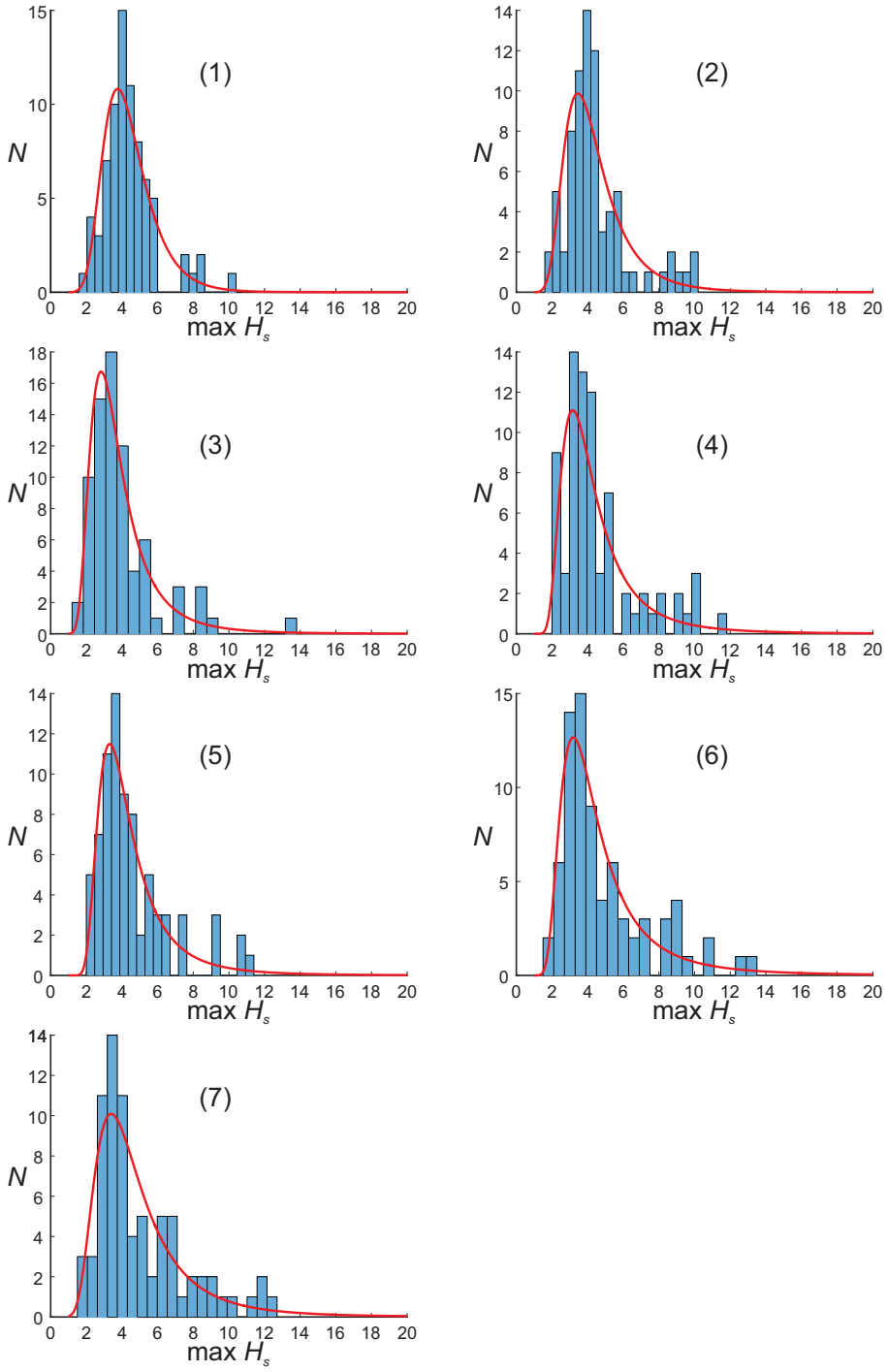

Figure 2: Frequency distributions of maximal wave heights at 7 grid points from North-West Australia hindcast database (histograms), and the theoretical probability density function for the extreme value distribution (red curves).

distribution with location parameter μ and scale parameter σ, suitable for

modelling the maximum value, has the form

$$y = f(x|\mu, \sigma) = \sigma^{-1} \exp(-\frac{x-\mu}{\sigma}) \exp(-\exp(-\frac{x-\mu}{\sigma}))$$

Given the distribution of maximum H_s from the data, we can fit it to the theoretical curve, test the quality of the data, and use the obtained continuous p.d.f. to model the probability of extreme H_s (say, the 100 year wave).

In figure 2, the curve for extreme value distribution p.d.f. is shown along with the histograms. The distribution functions have been normalized so that their integral corresponds to the total area of each histogram. We see that the distribution of maxima corresponds to the theoretical curves in each case rather well, so that a reliable prediction of the maxima with the specified probability can be made, based on the obtained p.d.f.

2.2 Bay of Bengal - basic data analysis

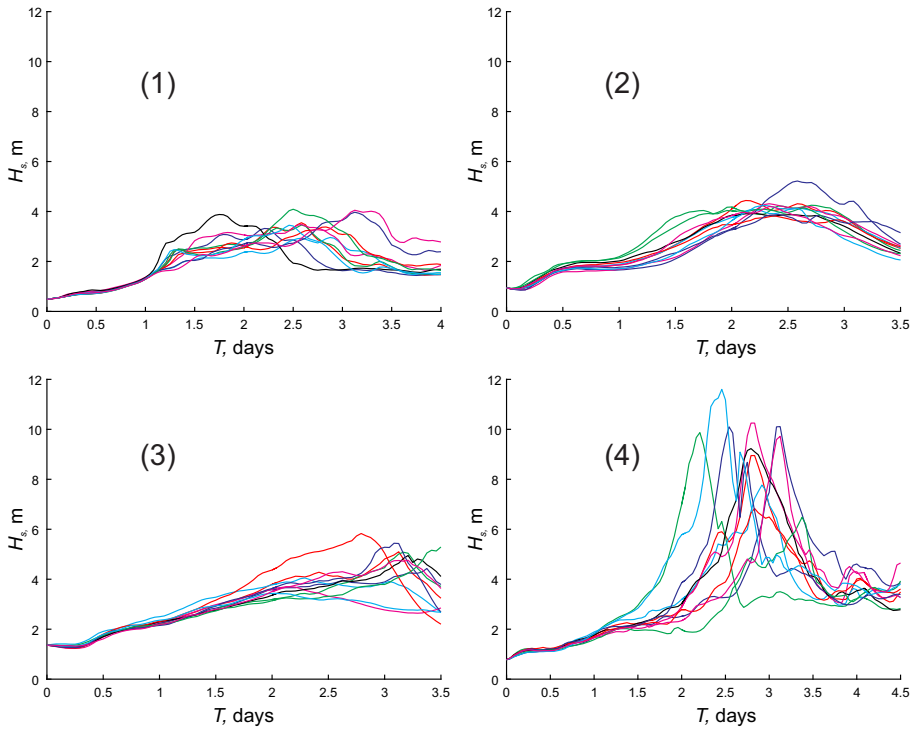

Figure 3: Wave height evolution vs time at the target point in Bay of Bengal (4 tropical cyclones, 11 realisations each).

For Bay of Bengal, the data consists of the wave height evolution for 4 tropical cyclones, obtained with simulations of atmospheric and wave models with varying parameters to produce an ensemble of 11 realisations. All realisations for all 4 cases are shown in figure 3.

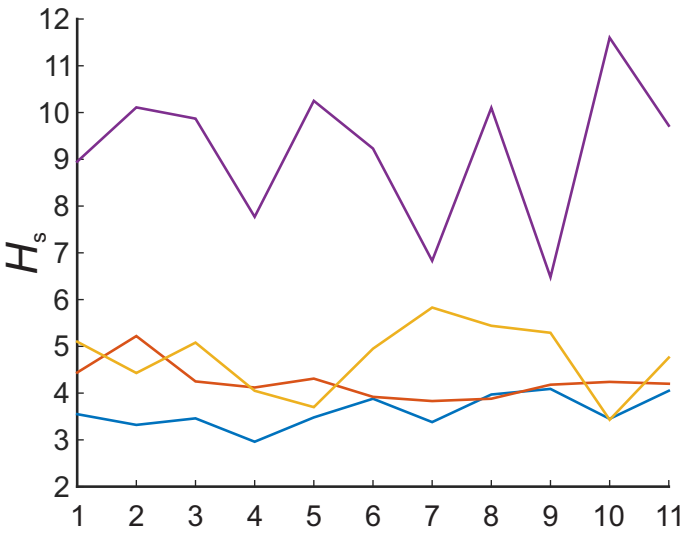

Figure 4: Maxima of wave heights in time vs realisation number.

We start the data analysis with looking at the maxima of all curves for all realisations (plotted in figure 4). These maxima give a rough indication of the maximal wave height that occurs at the target point. However, such basic analysis does not allow to get a measure of the statistical reliability of the maxima, or to produce a correct statistical estimate of return values for significant wave height. In fact, while working with random functions, it makes more sense to use quantiles instead of extremum values [1]. Quantile regression estimates are more robust against outliers in the response measurements than the maxima, and are often used for predicting relashionships between variables. The quantile regression analysis is performed in the next subsection.

2.3 Bay of Bengal - quantile regression analysis

In order to check the quality of the data, we use the quantile regression analysis, estimating the conditional quantile (in this case, 90%) of the response variable. This estimate is robust agains outliers and is useful for estimating the statistical envelope of response variables, together with confidence intervals.

To perform the analysis, we use the opensource statistics toolbox for Matlab and Octave by Anders Holtsberg
(http://www.maths.lth.se/matstat/stixbox/Contents.html)

In figure 5, we show the 90 % quantiles, obtained using the function "quantile" of the toolbox (note that the standard Matlab function "quantile" gives the same result), and the 95 % confidence intervals for the 90 % quantile, obtained using the function "ciquant". There are a few ways of estimating confidence intervals for quantiles; the toolbox used employs the binomial distribution method [2]. This result demonstrates that although we have

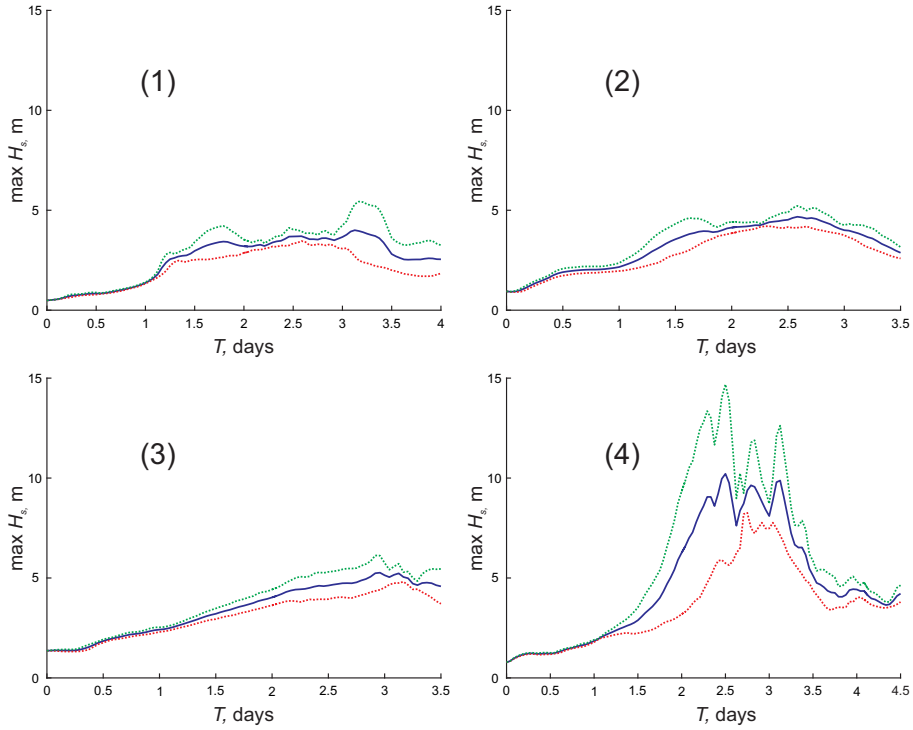

Figure 5: 90 % percentile fits for wave height vs time (blue solid curves) and 95 % confidence intervals for the percentile (dotted curves).

only 11 realisations of the model, the statistical estimate of maximal wave heights *for one individual tropical cyclone* appears to be quite reliable (although case 4, where there is considerable uncertainty in the ensemble of 11 realisations, somewhat stands out). In the next section, we develop a systematic approach, based on Harrell-Davis quantile estimation, to find the number of realisations required to keep statistical errors within specified limits. Then, in 3.2, we suggest an optimal strategy for increasing the number of realisations using limited computational resources. The main difficulty, however, lies in the insuffient data on tropical cyclones in the region. It appears reasonable to build a generator of synthetic cyclonic tracks to facilitate the statistical analysis. This is discussed in section 3.3.

3 Proposed solutions

3.1 Applying Harrell-Davis quantile estimation

We first wish to note that for the purposes of our research we did not have access to the various models and a description of how the microphysics variables were varied in order to generate the ensemble members. However, given the current ensemble member data that we have, we suggest to apply the Harrell-Davis estimator as a potential solution for having a measure of

confidence on the number of ensembles to run given how certain one would like to be about a certain location.

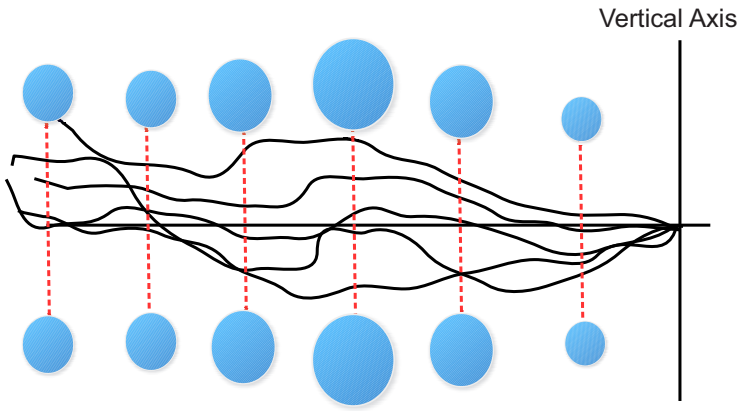

Figure 6: Illustration of the Harrell-Davis estimation process.

Here we describe a method for estimating the quantiles of the wind speeds at several uniformly distributed points along the tracks trajectories, using the generated ensembles. The first step in this method is, for each ensemble, to draw a straight line from the common genesis point to each of the corresponding termination points of the ensembles. Next, one calculates the angle θ between each of the plotted line and the y-axis centred on the longitude of the starting common genesis so we have a vector of angles $\theta = (\theta_1, \ldots, \theta_n)$, where n is the number of ensembles. The next step is to take the mean of this vector, $\hat{\theta}$. The we rotate the frame of reference by the angle $180° - \hat{\theta}$ by using the traditional 2D rotation matrix, ($R(\theta)$ below).

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

Next, we divide uniformly the reference line, i.e. the horizontal line, at each uniformly spaced point an orthogonal line is draw. This will give a wind speed for each ensemble at that point in space. One must note that the implementation will require an interpolation method in order to have an estimate of the wind speed at the intersection with each orthogonal line. With each of these samples of wind speeds $X_i = (X_{i,1}, \ldots, X_{i,n})$, one can perform a Harrell-Davis quantile estimation at each uniformly distributed point using the `Hmisc` package in R. This gives an estimate of the quantile and a standard error with very few points of data.

If one plots this method and plots the analysis as in figure 6, and the location of interest falls within one of the 'fuzzy balls' or the standard error of the estimate of the quantile, it indicates that if the client wishes to be more certain, more ensemble runs are required to reduce the error in the estimation.

3.2 Optimal strategy for adding realisations

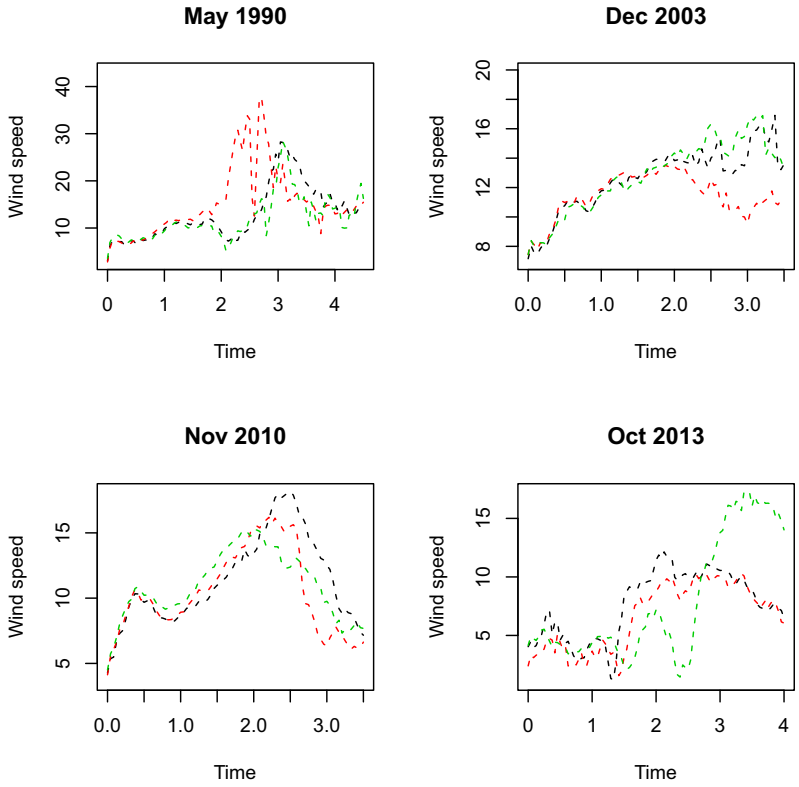

Figure 7: Simulated wind speeds from four observed cyclones (four plots) for three selected initial settings (three colours).

The Harrell-Davis estimator approached in this way gives an estimate of the uncertainty surrounding the number ensembles that is required. However, these ensembles are computationally expensive and running more ensembles takes time. But since our interest is only in the maximum, the question becomes how do we identify a pair of an observed cyclone and an initial setting that will produce the maximum wind speed, or wave height? If this is possible we need only run one simulation and not waste our resources on the ones that will not.

Let us now discuss this problem using wind speed as an example. The simulated wind speed is a function of the wind speed of the observed cyclone and the initial setting used to simulate it. The simulated wind speed is the speed at the eye of the cyclone so it then needs to be extrapolated using an appropriate method to estimate the speed at point P.

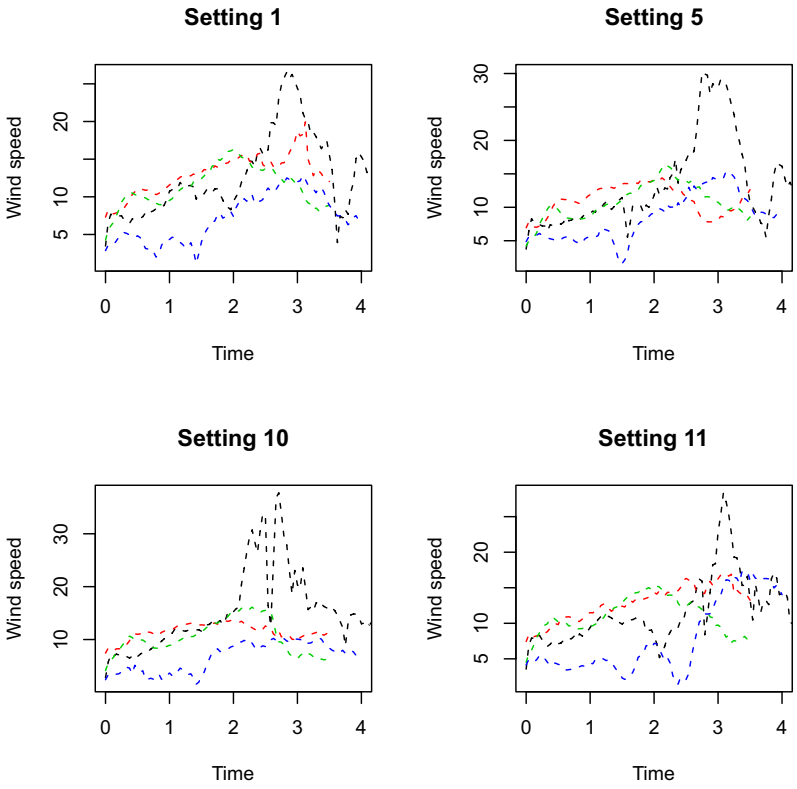

Figure 8: Simulated wind speeds from the same four observed cyclones (four colours) for each of four inital settings (four plots).

To identify an observed cyclone and an initial setting that will produce the maximum wind speed at P requires the knowledge of their effects on the simulated wind speed.

Figure 7 shows the wind speeds of three selected initial settings, ranging in three different colours, simulated from four selected cyclones. It is clear that the maximum wind speed in each cyclone is produced by a different coloured line (initial setting). Figure 8 shows the wind speeds simulated from the four cyclones, ranging in four different colours, for four selected initial settings. Even though the maximum wind speed is always achieved by the black event, the second largest comes from different cyclones depending on the initial setting chosen. So if we had used cyclones of similar sizes for simulation the maximum wind speed may have come from different events. Hence the cyclones and initial settings are interdependent. This means it is impossible to identify one pair of an observed cyclone and an initial setting to simulate the maximum wind speed.

This means we need to simulate as many as possible, starting with the maximum observed cyclone (after adjusted for P) and so on until the available resources are exhausted. It may not be necessary to simulate from all initial settings if it turns out that some of them always underestimate the speed.

3.3 Synthetic tropical cyclone generator

The main difficulty with the methods currently used in practice lies in the fact that the number of tropical cyclones is insufficient in some areas and does not allow to build reliable statistics. One may wish to have a return period distribution (for wind speed for example) for any spatial location in the observation window.

One way to do this is to use a statistical model for the generation of synthetic tropical cyclone tracks across the basin of interest, here the Bay of Bengal. This method is inexpensive and can be used to estimate a wind distribution at a particular location by reconstructing the wind profiles.

A general resampling synthetic tropical cyclone method involves three main models. The first model is the genesis model. This will generate the starting points of the synthetic tropical cyclones. The second propagates the synthetic tracks from genesis to lysis. This is generally constructed by looking at particular characteristics, for example direction, translation speed and maximum wind speed. At the initial step, one would sample from local distributions of the characteristics to gain an initial value. Then one can calculate the next point of the track and then sample from the local changes in those characteristics and so on for the next step. At each of these steps the model must decide whether the track terminates or not, this is the third model. A method for the termination of the tracks can be modelled using various parameters, for example, wind speed, an indicator if the cyclone is over land or not etc.. Below we outline one possible approach to simulate synthetic tropical cyclone tracks.

A method for the generation of synthetic tracks and calculation of synthetic tracks based on model in [4].

1. Identify the target distribution, for example the wave height distribution at a particular location.
2. Data analysis. Collecting the appropriate data for the proposed model and manipulate in order to get it into a usable format for the model.
3. Initiate the model by calculating appropriate estimators, for example the intensity estimator for the spatial Poisson process.
4. Select a generated cyclogenesis point. Construct local empirical cdfs for initial characteristics (direction, translation speed, maximum wind speed). Then, using a sampling method, for example inverse transform sampling, to sample a realisation of each characteristic.
5. The next location of the tropical cyclone can be computed by the sampled changes in the values of the characteristics.
6. Apply the termination model to determine if the track terminates at this point or not. If the track terminates then retern to step 4 if any cyclogenesis points remain, if not the algorithm terminates. Otherwise return to step 5.

One can generate a large number of tropical cyclones in the basin of interest and reconstruct the wind profiles of those tropical cyclones in order to

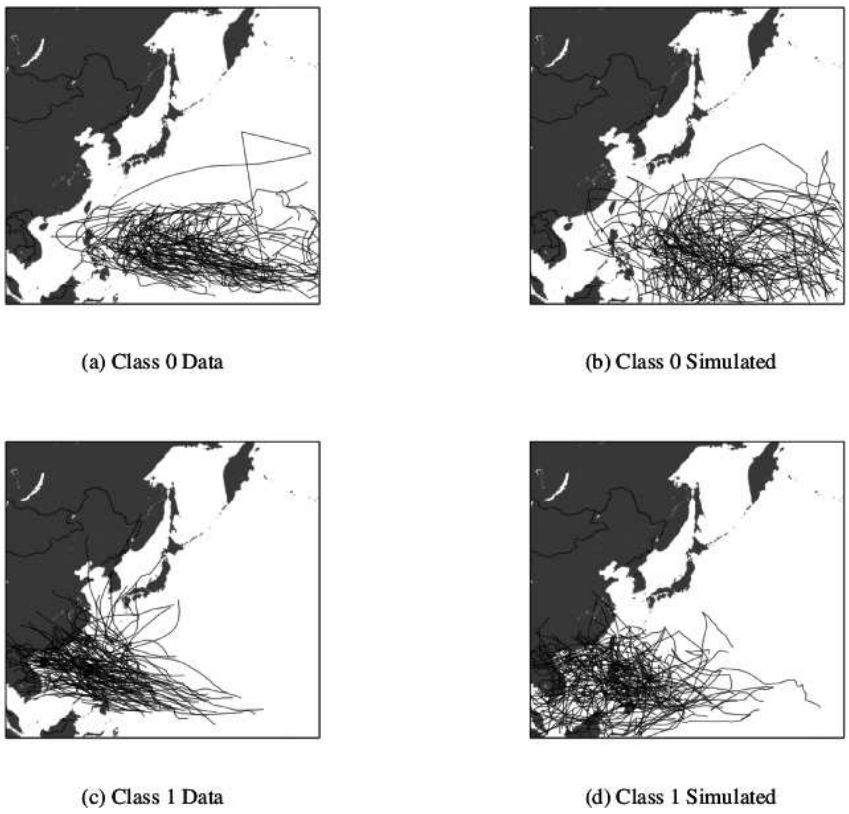

Figure 9: Illustration of historical vs. synthetic tracks in the western North Pacific [3].

construct a wind distribution at any location of interest. One can use this distribution to analyse potential extreme wind and associated wave events. Wind profiles can be reconstructed in numerous ways, for example via a modified Rankine vortex. However the type of model used to reconstruct the wind profile can have an affect on the resulting return period values [3].

4 Conclusions

- The quantile regression analysis shows that the maxima of significant wave height for individual tropical cyclone events can be estimated with a certain degree of reliability, even though we have got only 11 realizations.
- We propose a method for estimating the sufficiency of the number of realisations, based on the Harrell-Davis quantile estimation
- However, since the analysis is based on hindcast models, the quality of the extreme wave height forecast is dependent on the available historical data on the tracks and intensities of tropical cyclones affecting

the region of interest.

- Reliable and complete data on tropical cyclones tracks are insufficient for reliable statistical conclusions. This is especially true for Bay of Bengal, where tropical cyclones are relatively rare but can be quite severe.
- An alternative approach is to enlarge the sample of cyclones that can be used to drive the numerical models. This involves generating synthetic cyclones whose statistical characteristics simulate those of the population of real cyclones.
- At present, stochastic models for the simulation of tropical cyclone tracks are available for various oceanic regions. These models rely on the historical track data available. Complex meteorological aspects of tropical cyclone movement are simplified, creating the possibility for the simulation of large numbers of synthetic cyclone tracks. These models are general enough to be transferred to other ocean basins with only minor adjustments [4].
- We suggest a synthetic tropical cyclone generator, which can be used to generate a large number of tropical cyclones in the basin of interest and reconstruct the wind profiles of those tropical cyclones in order to construct a wind distribution at any location of interest.

Bibliography

[1] Gilchrist, W. G. *Statistical Modelling with Quantile Functions.* Chapman & Hall, 2000.

[2] Conover, W. J. *Practical Nonparametric Statistics.* Wiley, 1980.

[3] Leahy, Thomas P. Stochastic and Statistical Modelling of Extreme Meteorological Events: Tropical Cyclones. Imperial College London, University of Reading, MRes thesis, 2015.

[4] Rumpf J., Weindl H., Höppe P., Rauch E. and Schmidt V. Stochastic modelling of tropical cyclone tracks. Mathematical Methods of Operations Research, **66**, 475–490, 2007.