Ocean Science Discussions

Ocean Sci. Discuss., referee comment RC1
https://doi.org/10.5194/os-2021-84-RC1, 2021

Comment on os-2021-84

Anonymous Referee #1

Referee comment on "Forecasting hurricane-forced significant wave heights using a long short-term memory network in the Caribbean Sea" by Brandon J. Bethel et al., Ocean Sci. Discuss., https://doi.org/10.5194/os-2021-84-RC1, 2021

General comments

This paper presents an interesting application of a recurrent neural network for nowcasting and forecasting of significant wave heights in the Caribbean Sea. The authors also provide a useful overview of related machine learning techniques in this field. The method appears to provide useful forecasts at up 12-hour leadtimes, although while maximum SWHs are reasonably well forecasted, the timing of peak SWH seems to be poorly predicted. However, there are several statements which do not appear to be supported by the current figures and there are several issues with the manuscript which require clarification:

- Representativeness of the training set.

  As noted by the authors, due to the focus on extreme conditions, the training set is relatively limited. However, I think the authors need to discuss further the expected effects of this limited training set.

  Additionally, from Table 2 the test date have generally deeper lows and faster wind speeds than the training set. Some discussion on the use of this method on hurricanes which fall at the extremes, or outside the range of conditions in the training set would be beneficial.

Line 77: Hurricane Humberto is explicitly excluded from the training set. However, the reasons for this removal are not fully explained: would swell contamination be known a priori to know that any forecast of this hurricane would be unreliable? How would such cases affect the use of this technique in real-time forecasting?

-

Difficulty in interpretation of figures and confusion of the colour used for observations and forecasts in the text and figure captions.

I have found it difficult to interpret Figures 3, 4, and 5. Neither the red nor blue lines in the figures have consistent values between all panels. Why do the observations at the same timesteps change value between the panels?

Line 133: It is not explicitly stated, but I assume that the forecasts are all of SWH at a point (the buoy locations which provide the validating observations). If not, this should be clarified. Also, an explanation of how and why the nowcast differs from the observations should be given. Are observations up to the current time used in the forecast?

Line 151: Referring to Fig3b, the stated observation and forecast values are the opposite of what is shown in the figure.

Figure 5: The caption states that the observations are as measured at two buoys. Are the values averaged? Are the forecasts then averaged between the two locations also?

Section 3.2: This section would be clearer if an explanation were included of what constitutes an event in the SWH histograms presented. Is each event a particular time period? How does this relate to the earlier time-series?

Figure 6: The legend and caption disagree on which colour is used for the observations and forecast.

Figure 6: There are some events with SWH<2m. Why are these not seen in the time-series in Figure 3? Also, the time-series in Fig3a peaks at ~7m while Fig6a shows events up to the 9-10m bin. Line 214 also states that Hurricane Dorian has a maximum SHW of 8m, but Figure 6a-d shows events in the 9-10m range. The differences between these figures requires some explanation.

- Lack of a metric to capture the timing of the peak SWH.

Figures 3, 4, and 5 all show (to different extents) a phase shift in the timing of the peak SWH between the observations and forecasts. This appears to increase as the forecast leadtime increases. This requires some consideration and discussion as it appears to be a major drawback from this method of forecasting.

I suggest including an additional metric to quantify the ability to capture the timing of the peak SWH. Perhaps a lagged correlation of the observations and forecasts would quantify how well the timing of the peak is captured at different forecast leadtimes.

Indeed, the lag between observations and forecast shown for some of the test data (e.g., Figure 5) appears to match the forecast leadtime.

- Lack of a comparison method.

The paper would be strengthened considerably from a comparison of this method to another. Even in broad terms, describing the approximate RMSE expected from wave model forecasts at similar leadtimes would be useful. This could help the reader understand the forecast leadtimes beyond which a wave model significantly outperforms this method.

Alternatively, since this paper includes both SWH and surface wind speeds as input to the LSTM neural network, a comparison against forecasts using only SWH as input would allow readers to judge the benefit of the developments made by the authors.

Specific comments

Figure 1: It would also be useful to highlight the test data tracks in particular to allow readers to judge the representativeness of the training data compared to the test data.

Figures 3, 4, & 5: It is difficult to judge the relative difference between observation and forecast between the panels. I suggest the authors consider keeping the top panel as it is (to show the range of SWHs), but for each forecast to show the observation-minus-forecast differences.

Line 115: "data partitioned into a 70/30 split". I'm confused by this statement as this doesn't correspond to the number of hurricanes used in the training and validation sets.

Line 126: This is not the standard Pearson correlation coefficient. Although a value of 1 is returned if x=x dot, a value of -1 is not returned when x= -1(x dot). In fact, the return values are not restricted to the range -1 to 1. As such, it is difficult to interpret the meaning of the correlation coefficients presented.

Line 141: The authors state that the LSTM nowcast is unable to capture the extremely fine details seen in observations. I would welcome some discussion as to why this is the case. Are the observations noisy, or is there something fundamental to the LSTM which smooths the output compared to the inputs?

Line 161: "but this was minor" The difference referred to is approximately 2m, I do not agree that this is minor.

Lines 197-199: The text here doesn't match what is shown in Figure 6a.

Line 198: “completely reproduce” While the correspondence is close, it is not exact. Please rephrase.

Line 201: “providing results for wave heights at the 8-9m range, though there are no observed occurrences”. This doesn’t agree with what is shown in Figure 3a which shows some events for both observations and forecast for all bins >3m.

Line 217: “At the 12hour horizon…completely missed”. From the figures, I do not see any clear difference in this respect between panels a-d.

Line 219-220: “model predicts wave heights that are approximately 1m higher than the total” Rather than “the total” I assume the authors mean “the model”. Also, this seems to be an important point, but it is not clearly demonstrated by the current figures. I would suggest considering how best to support this point with an additional figure.

Line 221: “This may indicate…inclusion of less powerful hurricanes”. Since the test data includes more powerful hurricanes than the training data, a conclusion that additional less powerful hurricanes should be added to the training set seems counter-intuitive. I would welcome some additional discussion on this.

Line 226: “previously not observed for either Hurricanes Dorian or Sandy”. It is not clear to me from Figures 6, 7, and 8 that the behaviour is markedly different.

Line 227: “identical to the previous hurricanes, the frequency of maximum wave height predictions are overestimated”. If the blue histogram shows the forecasts (unclear as caption and legend disagree) this is true for Fig8 and Fig7bcd, but not Fig7a or Fig6.

Figure 9: Although the text details the differences seen between the hurricanes in terms of R and RMSE (shown in Figure 9), I think some discussion of why the correlation for Sandy decreases faster than the other hurricanes is warranted.

Lines 252-253: Stating these values in the text doesn’t add anything, they are already shown clearly in the figures.

Lines 261-267: Again, this adds nothing to the paper. These values are all listed in a table and shown in a figure. It would be more useful to comment on the large difference in MAPE between Dorian and Sandy/Igor at long forecast leadtimes.

Technical comments

Figure 1 uses a rainbow colour scale which is generally discouraged as it is not easily interpreted by colour-blind readers. I would encourage a change to a more accessible colour-scale.

Figure 1: Triangles marking buoy locations are difficult to discern. Perhaps increase symbol size.

Line 213: “Fig5a” I think the authors mean to refer to Fig 4a here.

Lines 235-238: This text appears to refer to all 3 hurricanes, but the paragraph continues on to Sandy and Igor, so this section must be specific to Dorian. Please clarify. I think it would be clearer to focus on the broad results of all 3 hurricanes rather than the detailed values of R, RMSE, etc for each (as these are in the figures anyway).

Line 245: “decreased to 0.82” I think this is the value for Dorian, not Igor (which is being discussed at this point in the text).

Figure 9: It is difficult to distinguish the symbols used for Sandy and Dorian. Perhaps an open square for one would be better, or simply increasing the symbol size throughout.

Figure 9 caption refers to only one hurricane while 3 are shown in the figure.