Response to Reviewer 1:

We thank the reviewer for their careful reading of the manuscript and their thoughtful comments which we discuss below.

I believe a broader introduction is required and motivation of performing these reruns is required. This is noted clearly in the abstract that a transformation of our understanding of historical variability is possible, but the authors do not note this until L141-146. No real aim or scientific purpose of the study is given until the results are discussed and this is something that needs to be rectified.

We agree with the reviewer and have added some additional discussion in the opening section.

1. *For all figures I would recommend labelling of panels via a, b, c, d, etc. This would make it easier to know which panels are being referred to rather than 'bottom row', etc.*

We have added panel labels in some of the figures.

2. *L125, please include figure reference.*

We have added an extra reference to Figure 4 in this paragraph.

3. *L125, Fig 4. How does the assimilation density of 20CRv3 compare to the assimilation performed by the authors using the new data? Will this affect the ranking that you have done in Fig. 4 (d-f)? By improving the density of Ulysses this may mean its winds are not representative of the 1950-2015 reanalysis and so the ranking may not be correct. Please clarify this.*

We understand the reviewer's concern and is the reason why we included Figure B2 in the original manuscript and the discussion in L467-473. The density of observations assimilated in the experiments for 1903 is much smaller than for the modern period (1960s onwards), even after the addition of extra data. When compared to the 1950s, the experiments with added observations for 1903 have more locations over the UK, although the observation frequency is often lower. Figure B2 and the associated text highlights that the ensemble spread appears roughly reliable in the modern era when using one set of available independent observations and two example years. Although this is a simple test, it appears as though the assimilation scheme change might not be required for the 1950-2015 period and so the comparison of wind ranks is considered to be fair.

4. *Fig 4 (and throughout). It may be useful to show the plots of new data (and improved DA) as difference plots to highlight exactly where the re-imagined storms have strengthened relative to 20CRv3.*

We have considered this issue and decided to retain the figures as before as the main differences are clearly visible and it is the absolute values that are being assessed with the independent data.

5. *L221-224, how do the quoted percentages of ensemble members (49% and 22%) relate to the probabilities quoted in Fig 6. These values are different and I find it hard to understand why or how the authors have computed them to be different. This needs clarifying.*

Thanks. This was an error in the text which had not been fixed. The percentages now match between the figure and the text.

6. *L226-228, this feels like repetition of two paragraphs prior. Please consider rephrasing.*

The text has been edited to be less repetitive.

7. *L231-232, Fig 6, it would be good to also quote the windspeeds of the non-precursor members. Furthermore, are there any statistical differences in the distribution of windspeeds simulated between the precursor and non-precursor members? If not it needs to be stated that even though the ensemble mean is higher, there is no statistical increase in simulated windspeed with sting jet precursors.*

The wind speeds in the members without a precursor are significantly lower than the members with a precursor and this is now added.

8. *L316-319, Fig 8, Is the gauge data used in Fig. 8 a point estimate? If so I would not expect the output of the coastal surge model to match that of a point estimate as it has resolution of 12km. It may just be that the coarse nature of the reanalysis is unable to simulate such wave heights. This section is stated as if the storm is still not simulated to the correct strength is the driving factor of this, whereas it should be restated (in my opinion) that the difference in resolution of the two datasets is the leading driver of the difference and that an underestimation in intensity may be another reason why.*

Experience with present day records and operational storm modelling suggests that the underprediction of the storm intensity, and precise timing and direction of wind fields, is most likely to be the cause of the discrepancy, which is quite small. Whilst the data are point estimates, the gauges were designed to physically smooth local wave action over several minutes, as is done (slightly differently) by a modern stilling well. The signals at Liverpool and Hilbre, which are about 10 miles apart along the coast, are closer to each other than to the model at the peak surge. However, L320 has been edited to say that improved resolution of the storm surge model might help resolve remaining discrepancies.