Response to Reviewer 2:

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

- *I certainly agree that rescued observations are of great value for improving simulations of historic weather events. However, I feel the positive statements from comparisons and argumentations could be toned down a bit at times and more discussion is needed. For example, a short discussion about their trustworthiness, accuracy, error range of the rescued observations should be included.*

In the 20CRv3 system the land stations are all assigned an uncertainty of 1.2mb (surface pressure) or 1.6mb (sea level pressure), and ship observations are assigned an uncertainty of 2.0mb. This is unchanged in our experiments. The cited Craig & Hawkins (2020) paper discusses the collection and QC of most of the new pressure data, but we now include more details about the number of stations and their assumed uncertainty in Appendix A. It is likely that, as these stations were formal observatories that were regularly inspected by the Met Office, the data is of very high quality and perhaps the uncertainty assigned is too large. We have made some edits and deletions to slightly tone down the text in some places, such as Section 3.

- *The discussion is purely based on the ensemble mean (except for the sting jet precursor). I wonder if the missing wind jets in Figure 5 and that you discuss in l. 164ff are rather due to looking at the mean and are actually present in individual ensemble members. With a large spread, the maximum wind speeds of smaller features, such as the cold conveyor belt jet or sting jet, probably differ in location and, hence, are weaker in the ensemble mean. Of course, Figure 5 shows an improvement nonetheless, however, the argumentation why that is changes and you should state that the features "are not present in the 20CRv3 mean" (l. 172).*

We agree and have clarified that the features are not present in the ensemble mean of 20CRv3 (old line 172). None of the ensemble members of 20CRv3 have a wind jet.

- *Please consider using hPa instead of mb.*

Thanks – we have considered this and prefer to stick to mb.

- *How did you track the storm?*

The storm is tracked by interpolating the gridded pressure field to find the local minimum. This is added to the text.

- *Please be consistent with figure labels (e.g., "new data" vs. "new observations", colorbar labels).*

Thanks for spotting this - we have changed all the figure panel titles to say 'new observations'.

- *Some figures are not discussed to their full extent. You show interesting information in the figures, which are – sometimes – not even mentioned in the text (e.g., probabilities in Fig. 6)*

Thanks. Some additional text has been added discussing Figure 6.

- *l. 72f: As you state, 960mb is an estimate, so the comparison of the pressure minima in simulations with this value should be put in relation and not seen as the absolute truth.*

We agree with the reviewer, and have edited this sentence to say ‘is more consistent with’, rather than ‘better matches’.

- l. 85ff: Please add 1-2 sentences with more information about the added data and especially the improved data assimilation to the main text. It would be good to at least have an idea about the improvements without having to read the Appendix, which should then be for readers with further interest.

Noted. Some additional text on the number of locations added, and the assimilation improvements has been added.

- Figure 3 and elsewhere: Please consider putting the colorbar labels right next to the colorbar.

We have added labels to these colourbars.

- *l. 218: Please elaborate the “simpler grid point approach”. Do you mean you simply make the tool independent of neighbouring grid points, hence you could use it on every grid point independently? Please discuss shortcomings of this approach.*
- *l. 221ff: When do you define a member to show precursors: Is this already the case for only one grid point? How do these percentages compare to the probabilities in Fig. 6?*

The text has been edited to make this clearer. We simply count the number of ensemble members with at least one grid point indicating a DSCAPE precursor. The operational warning system has a more complex approach requiring a larger region (multiple grid points) to have a DSCAPE indicator present. There was also a mistake in the text meaning that the % given in the text did not match the figure – this has been fixed.

- *Figure 6: The difference between 20CRv3 and new data seems to be much smaller than new data and new data + improved assimilation. Can you comment on this? Could the improved assimilation be more important for the improvement than the new observations? However, this is not really the case in other figures.*

Unfortunately we do not have an experiment with the change in assimilation scheme applied to the original observations to test some of these issues. We have added: “For the DSCAPE precursor likelihood, there is a clear difference between the experiments that only differ due to the assimilation scheme changes (39% vs 55%). Such differences have been less clear in metrics presented earlier, and we suggest that this may be because DSCAPE is a non-linear threshold-based metric meaning that the reduction in ensemble spread has a larger effect.”

- *l. 257: “observed”: As in the caption, you should at least mention the HadUK-Grid, i.e., interpolated in-situ observations.*

Agreed – the text has been edited to mention this.

- *Figure 8: Please consider another colour scheme. Furthermore, what is the reasoning behind the 16-84% range?*

We have considered the colour scheme but retained the existing version. We have added the following text to the caption: The 16-84% range is roughly equivalent to showing the ensemble standard deviation but is more appropriate for a non-normally distributed variable such as high-frequency rainfall.