I read through this interesting manuscript by Xie et al. with great interest. Here, the authors have attempted to resolve the coarse temporal resolution of GRACE(-FO) data by downscaling to the daily time series by employing machine learning methods. Further, a new flood index, namely, the normalized daily flood potential index (NDPFI), is proposed to better characterize the extreme flood events in the Yangtze river basin (YRB).

I have some suggestions that authors may find beneficial while revising their manuscript.

**Major suggestions**

(1) **Selection of flood events:** I found the authors have selected only extreme flood event(s) to analyze and demonstrate the capability of NDFPI (Lines 395, 413). I think that this may not be the best case. Selecting other flood events that are not well captured in other indices may be a better choice to discern the outperformance of and additional insights by the proposed index over conventional indices, e.g., SPI, SPEI.

(2) **Terminology of wet and dry seasons:** If I am correct, the wet season in the basin spans from June to September, and this is also the same period when the flood of 2020 is observed (e.g., Figure 10). The newly proposed index is also shown better to detect the extreme events during the wet season (e.g., as stated in lines 408, 427). However, in a few places (e.g., lines 433, 551, 817), it is mentioned that the flood events in 2020 as the ‘summer 2020’. Is there something that I missed or confused between summer or wet season?

(3) **Handling the intermittent data gaps due to battery management (Line 111-112):** It is
not clear how the intermittent data gaps occurring about every six months in the GRACE and GRACE-FO TWSA time series were filled. Most probably, they were filled by linear interpolation or by the average of the bounding one-two months values. In either case, the filled values are likely (a) underestimating the actual (positive/negative peak) TWSA if the data gap happens to be in the peak of the wet or dry season (there are a lot of such times, please see footnote of Table 3 in Abhishek et al., 2022 ), or (b) overestimation or underestimation in case of the high short-term fluctuations in the TWSA. Furthermore, this overestimation or underestimation can be critical given the topical issue of daily monitoring dealt with herein and, subsequently, might lead to inappropriate inferences in the YRB, which is highly vulnerable to floods.

In my opinion, these data gaps can either be filled by running the three machine learning models (already used in the manuscript) for monthly TWSA, or alternatively, this can be stated as a likely source of uncertainty.

**Minor suggestions**

Line 32: ‘severe extreme’. Choosing one of these two words may be better.

Line 52: ‘evaluation’ may not be necessary.

Line 57: either ‘a limited attention has been paid or ‘very few studies have paid attention to’

Line 59: Since downscaling of the TWSA data is the aim (Line 60), the phrase ‘using the temporally downscaled GRACE data’ may be removed.

Line 81: Reference of Fig. 1 may be provided here.

Line 84: ‘temperature’ to ‘temperate’

Lines 106, 119, 131: ‘basins’ to ‘subbasins’

Line 131: How the ‘extreme flood events’ were extracted is not clear? Providing a reference to Section 4.4 may be better.

Line 142: ‘o’ to ‘to’

Line 146, 530: using the single model output for soil moisture may include the implicit biases. How about using the ensemble mean from multiple model outputs, maybe even within the GLDAS series, subject to the availability and consistency with the study period.

Line 157: ‘machine’ learning-based

Line 168: ‘served’ to ‘used’ or ‘employed’ or some other appropriate verb

Line 411: ‘rivers’ may not be the best word. Please see another suitable word, if possible.

**Reference**
