Referee Report

Evaluation of Extreme Sea-Levels and Flood Return Period using Tidal Day Maxima at Coastal Locations in the United Kingdom

General comments

The preprint proposes "TMAX," a workflow based on tidal-day maxima and linear regression on Gumbel reduced variates, positioned as an automated alternative to SSJPM and other established approaches. The topic sits squarely within the scope of *Ocean Science*, and the paper is motivated by an operational need for transparent and automatable estimation of return levels. In its current form, however, the scientific framing and validation are not yet sufficient for design-grade use: key assumptions (tail type, independence/declustering, stationarity) are not stress-tested; uncertainty is not fully quantified at the site level; and reproducibility is limited by brief methodological descriptions (peak selection, QC, and parameter tuning).

Alignment with OS evaluation aspects (brief assessment).

- **Scientific significance.** Addresses a relevant question and aims to deliver operational simplification; the novelty is primarily in workflow/automation rather than in new statistical theory. Substantial impact would require stronger validation and generalization beyond the UK testbed.
- **Scientific quality.** Core assumptions (Gumbel tail, independence of selected peaks, uniform de-trending) need diagnostics and comparisons (e.g., GEV/POT baselines). Results are compared against SSJPM but without per-site uncertainty bands or sensitivity analyses.
- **Presentation quality.** Generally clear narrative, but several essential methodological details are under-specified (peak selection, independence checks, QC). Figures would benefit from confidence bands and clearer indication of the fitted range.
- **Reproducibility/traceability.** The text should provide enough algorithmic detail (and QC rules) to enable replication; parameters selected by heuristics (e.g., number of peaks per year) should be justified or selected by diagnostics.

Specific comments

S1. Automation as primary justification vs. reliability for design. The manuscript motivates TMAX mainly by its potential for full automation as an alternative to SSJPM, which is presented as difficult to operationalize. If design decisions based on return levels may have severe consequences, methodological reliability should come first and automation second. Automation is valuable for broad regional screening, but for site-specific design one would typically prefer the best-supported method (SSJPM or POT/GEV with explicit diagnostics), even if harder to automate. Please either (i) provide stronger evidence that automated TMAX delivers design-grade performance (diagnostics, uncertainty, sensitivity, external validation) or (ii) clearly position TMAX as a rapid, approximate tool for comparative assessments rather than as a standalone basis for design.

S2. Geographical scope and transferability beyond the UK. The paper frames TMAX as broadly automatable, but all evidence and tuning are limited to United Kingdom gauges. Please clarify whether the method is intended as UK-specific or globally transferable. If transferable, add an external validation or stress test at a few non-UK sites spanning different tidal regimes (micro/meso/macro; mixed/diurnal/semidiurnal), sampling cadences (15-min vs. hourly), and surge climatologies. Otherwise, state clearly that conclusions are restricted to the UK, and discuss what would be required to extend elsewhere (re-tuning of n_A, declustering windows, trend handling, tail-shape checks).

S3. Assumption of Gumbel vs. GEV tail form (and relation to prior practice). By fitting a straight line on probability paper using the reduced variate, the analysis *assumes* a Gumbel (Type I) tail. The manuscript should test whether a GEV model with freely estimated shape ξ provides a better description (Gumbel, Fréchet, or Weibull). A POT (Pareto–Poisson/GPD) baseline is asymptotically consistent with GEV and would allow estimation of ξ. From a physical perspective, after de-trending, extreme still-water levels are plausibly bounded, which argues for exploring Weibull-type behavior ($\xi < 0$) and could explain curvature in low-rank portions of the plots. The paper contrasts with work where heavy tails were mitigated via spatial averaging; here the tail difficulty is sidestepped by *assuming* $\xi = 0$. Please justify this choice with tail diagnostics (GEV and POT/GPD fits, QQ/PP/return-level plots) or relax the assumption.

S4. Peak identification and independence of events. The peak-picking algorithm ("scan above MSL, store the largest value before returning below MSL") and the selection of the top $n = n_A n_y$ peaks are described briefly, but independence (declustering) is not demonstrated. Please report extremal index estimates and/or runs-declustering sensitivity, and show ACF/dispersion checks to confirm that the retained peaks are approximately independent across sites.

S5. Choice of sample vs. threshold modeling (link to r-largest/POT and threshold selection). Selecting the "top n_A per year" resembles r-largest approaches, and the choice of n_A plays a role analogous to *threshold selection* in POT. A single global variance minimum is unlikely to be optimal everywhere. Please provide sitewise sensitivity of return levels (with uncertainty) to $n_A \in \{3, \ldots, 10\}$, and outline diagnostics for automatic selection grounded in stability/predictive criteria (in line with established threshold-selection practice).

S6. Stationarity and de-trending. All series are de-trended using a fixed $2\,\mathrm{mm\,yr^{-1}}$ with a specified origin date. This uniform correction may not match local relative sea-level change, and no diagnostic is provided for residual nonstationarity in extremes. Please add simple diagnostics (moving-window location/scale of daily maxima, change-point tests) or adopt a basic nonstationary model if indicated.

S7. Uncertainty quantification for return levels. Formulas for σ_m and σ_p are given, but site-level return curves lack confidence or prediction bands. Please add per-site intervals for key return periods (e.g., 20, 100, 200 years) and discuss extrapolation uncertainty for longer periods.

S8. Data curation and sampling resolution. Combining 15-min and hourly records is pragmatic, but potential peak under-sampling is asserted rather than quantified, and the QC procedure is described as manual. Please specify reproducible QC criteria and include a down-sampling experiment (e.g., degrade the 15-min periods to hourly and re-estimate return levels) to demonstrate that conclusions based on hourly data are consistent with those from 15-min data.

S9. Relationship to JPM/SSJPM and tide–surge dependence. TMAX sidesteps explicit modeling of conditional dependence between surge and tide. Please show whether selected peaks cluster near springs and whether dependence structures vary by site. If dependence persists after conditioning, consider incorporating it (e.g., via covariates in exceedance rate/parameters or via an extremal index correction).