This manuscript proposes a dual-threshold double autoregressive framework for daily streamflow prediction that combines seasonal normalization, fractional differencing (long memory), a threshold structure (nonlinearity), and alternative residual distributions (Gaussian vs Student's t). The approach is evaluated at 15 gauges in the Yellow River Basin using the last year at each station as the test period. The topic is relevant to HESS and the manuscript is generally well structured. However, I have several concerns regarding the clarity and reproducibility of the methodological workflow, the interpretation of the diagnostic/statistical testing, and the robustness of the evaluation design. For these reasons, I recommend major revisions. The details of my concerns and suggested actions are provided in the Major comments below.

Mayor comments

1) Final paragraph of the Introduction The closing paragraph does not clearly state the study context and experimental setup. In particular, it does not specify the case-study region (Yellow River Basin), the number of gauges, the data period, and the evaluation design (training/testing split and forecasting protocol). I suggest restructuring the final paragraph to (i) state the objective, (ii) summarise the methodological contribution with consistent notation, (iii) explicitly mention the study area and dataset (15 stations in the Yellow River Basin), and (iv) briefly outline the benchmark models and verification metrics.
2) Many figure and table captions are not sufficiently descriptive and, in several cases, are not self-contained (some are also incomplete). Please revise all captions to clearly state (i) what is being shown, (ii) key definitions for acronyms/labels (e.g., M1–M4, AIW, CR), (iii) the evaluation setup (training/testing period and forecast horizon, where relevant), and (iv) what symbols/markers represent. Captions should allow readers to interpret figures and tables without relying heavily on the main text.
3) The manuscript reports and interprets diagnostic results (statistical tests applied to the original and transformed series, including after de-seasonal standardization and fractional differencing) within the data/characteristics section, i.e., prior to the Methods section. This sequencing is confusing and makes the analysis pipeline difficult to reproduce. I recommend keeping the Data section purely descriptive, and moving the test outcomes and interpretation to the Results section (or to a dedicated "diagnostic results" subsection). The Methods section should then clearly document the full workflow, including a brief description of each statistical test (null hypothesis, key settings such as lag/order/embedding dimension, and decision rule). Finally, please include a schematic/flowchart summarizing the sequence of preprocessing steps and diagnostic tests, and how their outcomes motivate the choice of the FDTDAR (and benchmark) model structures.
4) Because the evaluation uses the last year as a test period, it is essential that seasonal statistics (daily climatology) be computed using training data only, then applied to the test year. If the seasonal mean/variance are computed using the full record (including the test year), this introduces leakage and inflates apparent performance. Please clarify the implementation and, if needed, redo the evaluation using strictly training-only preprocessing.
5) The manuscript reports Hurst index values for all stations and concludes long-term memory based on $H > 0.5$, with stronger memory after deseasonalization. Please specify which estimator is used (R/S, DFA, Whittle, wavelet, etc.), and discuss the known sensitivity of Hurst estimation to trends, seasonality, regulation, and nonstationarities. Given the reliance on fractional differencing, it would also be valuable to report an uncertainty range for H or provide robustness checks.
6) The manuscript uses the last year of each station as the testing period. A single-year split can be highly sensitive to hydrologic conditions (wet vs dry year), regulation operations, and extremes, and it does not provide a stable estimate of generalization skill. I strongly recommend adding a more

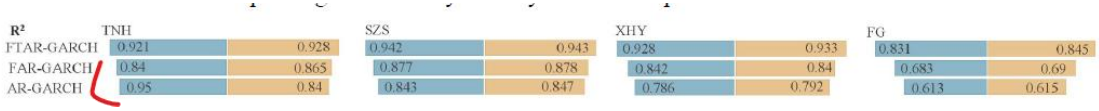
robust design, for example: multi-year rolling-origin evaluation, blocked cross-validation, or repeated splits that preserve temporal dependence. This is especially important because the record lengths differ across stations.

7) The manuscript is primarily a deterministic (point-forecast) time-series modeling study. If prediction intervals are intended (via Gaussian or Student's t residual assumptions), please make this explicit and describe how intervals are constructed. In that case, evaluating uncertainty using only AIW (interval width) and CR (coverage) is not sufficient; consider adding an interval skill metric (e.g., interval score/Winkler score) and reporting coverage at multiple nominal levels. If the study is not intended to provide uncertainty quantification, please remove or de-emphasize AIW/CR and focus on point-forecast metrics.
8) The evaluation is largely restricted to AR-GARCH and TAR-GARCH baselines. While these are appropriate time-series benchmarks, the lack of comparisons against more recent nonlinear data-driven approaches (e.g., Bayesian models or other modern machine-learning/deep-learning baselines commonly used for daily streamflow prediction) limits the strength and generality of the performance claims. At minimum, the manuscript should explicitly justify the choice of the benchmarking set and acknowledge this as a limitation. Ideally, the authors should include one or two representative modern nonlinear benchmarks, or provide a clear rationale for why the study is intentionally confined to the DAR/GARCH model family.

Specific comments:

1) L12-14: Consider revising this sentence because 'non-stationarity' and 'time-varying fluctuations' largely refer to the same idea. You could either drop one term or specify what aspect varies in time (e.g., mean, variance, regime).
2) The use of DAR is potentially confusing and appears inconsistent with your notation, since the main proposed model is DTDAR. Please ensure acronym usage is consistent throughout the manuscript.
3) L29: Please revise 'The hydrological statistical method' (singular) to a clearer phrasing. This reads as a broad class of approaches, so 'hydrological statistical methods' or 'statistical methods in hydrology' would be more appropriate and less ambiguous.
4) L42-43: The sentence 'So, this study uses the seasonal standardization method…' feels out of place here, as this paragraph is still part of the literature review/background. I suggest moving this statement (or rephrasing it) to the end of the Introduction, where the manuscript clearly states its objectives and methodological contributions. The same comment applies to L72-74 and L80-84.
5) L86-87: A reference is needed to support this statement.
6) L98-99: "selected for their high quality and reliability" is too vague. Please state the specific selection criterion (e.g., minimum record length and/or data completeness threshold) and what it applies to (the streamflow record).
7) L110: The phrase "assess their modelling capabilities" is unclear and likely misphrased (ambiguous referent for "their" and redundant). Please reword to something specific, e.g., "assess model fit/performance on the training set" or remove this clause.
8) L149-152: "Figure 4 reports Ljung–Box p-values, but the manuscript also discusses ARCH LM test results (e.g., 'p-values … are all 0') without presenting them. Please report the LM test outputs (test statistic, p-value, and lag/order used), either in a table or in the Supplement.
9) L155-161: Please explicitly state the null hypothesis of the BDS test (typically i.i.d. behavior). Without defining H_0, it is difficult to interpret the reported p-values and the conclusion about nonlinearity.
10) In Eq. (1), σ_m is defined as the square root of the average squared deviations, i.e., the standard deviation, not the variance. Please correct the text accordingly (or change the formula if variance is intended) and ensure consistent terminology throughout. Check it in L218.

11) Section 3.2: Please define the acronym FDTDAR at first mention (spell out the full model name) and ensure consistent usage throughout the manuscript (avoid switching between FDTDAR, DTDAR, and DAR unless the distinction is explicitly stated).
12) Eq. (20): The penalty term includes "+ 8" inside the parameter count, but it is not explained where these 8 parameters come from. Please provide a clear accounting of all free parameters included in the AIC expression (e.g., intercepts/scale terms per regime, threshold parameters r_1, r_2, degrees of freedom for the t-distribution if applicable, etc.) and justify why this constant is 8.
13) L306: it should be: ", respectively."
14) Figure 6 reports MRE, but this metric is not defined or described in the Methods. Please define MRE explicitly (name, units/interpretation, and whether it is computed on the test period) and ensure the set of evaluation metrics is consistent between Section 3.6 and Figure 6.
15) Figure 9 is difficult to interpret without a clearer guide. Neither the caption nor the introductory text provides enough information to decode the acronyms and visual encoding. Please revise the caption/text to (i) define all model acronyms shown in the panels (e.g., DAR, FDAR, FTAR-GARCH, FAR-GARCH, etc.), (ii) explain what each block/panel represents (structure: integer differencing vs long memory vs long-memory threshold; residual distribution: Gaussian vs Student's t), (iii) clarify the meaning of colors/bars and how values should be compared across models and stations.
16) Figure 9. I think there is an error en length of the bar highlighted



17) The AIW and CR results are not interpretable without stating the nominal prediction interval level (for example 90% or 95%, equivalently the α used). Please revise the captions and/or figure annotations to explicitly report the chosen nominal coverage and briefly indicate how the prediction intervals were constructed (Gaussian vs Student's t residual assumption).
18) L426: There is an extra period at the end of the sentence ("time series.."). Please correct to a single full stop.
19) L470: it should be "on various". Please correct it.