Reviewer #1

I am very much satisfied with your responses to my comments. The paper has significantly improved its structure due to the incorporation of key flowcharts that describe the processes neatly. I have no further comments. I recommend the paper be accepted in its current form.

Minor correction:

Page 4, line 103: "ionic" is not a good fit here.

Response: Thank you very much for your thorough review and positive feedback on our revised manuscript. We greatly appreciate your recommendation for acceptance in its current form. This is very encouraging and motivating for us. And we have corrected the error.

Reviewer #2

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.

Response: We would like to thank your insightful and constructive feedback on our manuscript. We are pleased that you found the topic relevant to HESS and the manuscript well-structured. We take the concerns regarding methodological clarity, statistical interpretation, and evaluation robustness seriously. We have proposed a series of major revisions to address these points, which we believe will significantly enhance the quality and impact of the study.

Major 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.

Response: We thank the reviewer for this excellent suggestion. We agree that the final paragraph of the Introduction should provide a concise yet comprehensive summary of the study's scope and experimental design. Following the reviewer's suggestion, we have restructured the paragraph to state: (i) the research objective explicitly; (ii) the components of the threshold models' framework using consistent notation; (iii) the specifics of the 15 gauging stations in the Yellow River Basin (YRB) and the data split; and (iv) the benchmark comparison and evaluation metrics. The revised paragraph is presented below and has been updated in the revised manuscript.

"This study aims to improve the prediction accuracy of daily streamflow time series by constructing a novel model with high applicability that can simultaneously capture seasonality, non-stationarity, long-term memory, and nonlinearity of daily streamflow. We propose the FDTDAR framework that systematically integrates seasonal normalization, fractional differencing for long-memory modeling, and a dual-threshold structure to capture regime-specific nonlinearities. Furthermore, we evaluate alternative residual distributions—comparing the standard Gaussian distribution against the heavy-tailed Student's t-distribution—to improve error characterization during extreme events. The framework is applied to daily streamflow data from 15 gauging stations across the Yellow River Basin. For robust evaluation, the historical data are divided into a 70% calibration period and a 30% out-of-sample testing period to assess the one-day-ahead forecasting performance. The FDTDAR framework is rigorously evaluated against both the classical linear AR-GARCH model and the long short-term memory (LSTM) network. Evaluation metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of determination (R^2), Nash-Sutcliffe efficiency coefficient (NSE), and Absolute Maximum Error (AME), providing a multi-faceted assessment of model robustness."

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.

Response: We fully agree with the reviewer's assessment. We have revised all figure and table captions throughout the manuscript to ensure they are self-contained and descriptive. These revisions allow the figures and tables to be interpreted independently of 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.

Response: We agree with the reviewer that the logical flow of the analysis pipeline was initially suboptimal. In the revised manuscript, we have restructured the content as suggested:

Data Section: This section is now purely descriptive, focusing on the geographic location, hydrological characteristics of the Yellow River Basin, and a basic summary of the raw datasets.

Methods Section: We have expanded this section to include a dedicated subsection for the statistical tests used (ADF, Hurst, Ljung-Box, etc.). For each test, we now explicitly state the

null hypothesis (H_0), the parameter settings (e.g., lag/order), and the decision rules.
Results Section: We have moved the test outcomes and their interpretations to a new subsection titled "4.1 Preprocessing and Diagnostic Results."
Flowchart: Figure 3 illustrates the full analysis pipeline—from preprocessing to diagnostic tests—and clarifies how these steps justify the choice of the fractional-differenced dual-threshold DAR (FDTDAR) framework.

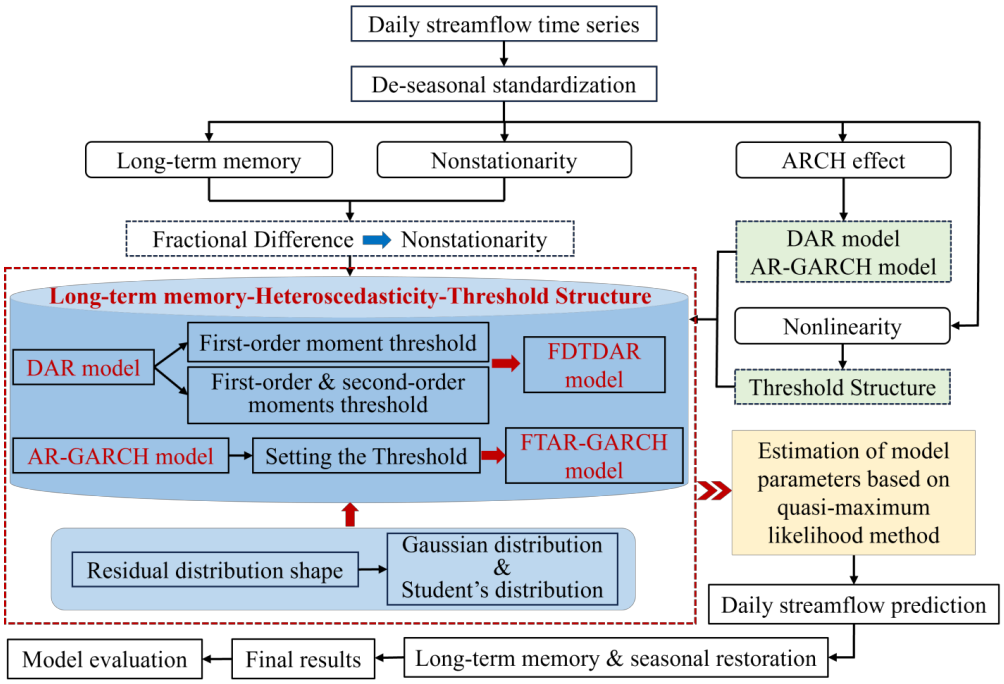

Figure 3: Methodological workflow of the Fractional-differenced dual-threshold double autoregressive (FDTDAR) framework for non-linear and long-memory streamflow prediction.

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.
Responses: The seasonal statistics (daily mean and standard deviation used for normalization) were computed exclusively from the calibration period (the initial 70% of the dataset). These calibration-derived seasonal parameters were then used to de-seasonalize the independent testing period (the remaining 30%).

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.
Responses: We have tested the significance of H values for 15 stations in the Supplementary file.
Significance test of Hurst values for 15 stations

To assess the statistical significance of the H estimated using the R/S method, a confidence interval for the test statistic was constructed based on the Monte Carlo simulation approach. Under the null hypothesis, the time series is assumed to exhibit no long-term memory, i.e., H=0.5, corresponding to a Gaussian white noise process. Specifically, for each hydrological station, 1,000 synthetic Gaussian white noise sequences of the same length as the observed daily streamflow series were generated. The R/S analysis was then applied to each simulated series to estimate the Hurst exponent, and the 95% confidence interval was derived from the resulting distribution. If the H value of the observed series lies outside this interval (at a significance level of p<0.05), the time series is considered to exhibit statistically significant long-term memory. As shown in Table S1, all 15 hydrological stations demonstrate significant long-term persistence in their daily streamflow time series.

Table S1. Significance test of H value of daily streamflow time series at 15 hydrological stations

Station	2.50%	97.50%	H	p	Station	2.50%	97.50%	H	p
TNH	0.49	0.60	0.82	<0.05	ZT	0.49	0.60	0.79	<0.05
SZS	0.49	0.60	0.86	<0.05	ZJS	0.48	0.61	0.69	<0.05
XHY	0.49	0.61	0.86	<0.05	LZ	0.49	0.61	0.94	<0.05
FG	0.48	0.60	0.83	<0.05	TDG	0.48	0.61	0.83	<0.05
TG	0.48	0.60	0.84	<0.05	LM	0.49	0.61	0.87	<0.05
SMX	0.48	0.61	0.84	<0.05	LT	0.48	0.59	0.76	<0.05
AS	0.48	0.60	0.84	<0.05	XY	0.48	0.58	0.80	<0.05
LJ	0.49	0.60	0.84	<0.05					

The daily streamflow time series exhibits seasonal characteristics due to the cyclical nature of the four seasons. Figure 4 and Table S1 examine the presence of long-term memory in the daily streamflow series based on the autocorrelation function (ACF) and Hurst exponent (H), respectively, and demonstrate whether seasonality influences the manifestation of long memory. The results show that, after deseasonalization (as detailed in Section 3.1), the ACF of the daily streamflow series decays more slowly, indicating stronger persistence. In addition, the H values of the daily streamflow series at 15 hydrological stations, calculated using the rescaled range (R/S) analysis method, are all greater than 0.5, suggesting the existence of long-term memory. The H values of the original series range from 0.69 to 0.94, which are lower than those of the deseasonalized series. The specific steps of the R/S method are presented in the supplementary manuscript, and Table S1 further shows the statistical significance of the H values of each station.

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.

Responses: We agree that a single-year testing period is insufficient to capture the inter-annual variability (e.g., wet vs. dry cycles) and may lead to an unstable estimation of the model's generalization capability, especially given the intensive human regulation in the Yellow River Basin. To address this concern and ensure a more robust evaluation, we have entirely revised

our experimental design and recalculated all results. Instead of a fixed one-year period, we now partition the data for each station chronologically, using the first 70% of the total record for model calibration and the remaining 30% as an independent testing period. Since our 15 gauging stations have different record lengths, this percentage-based approach provides a consistent and fair basis for performance comparison across all sites. We have updated all Tables and Figures in the revised manuscript to reflect these new results. The overall conclusions remain consistent, but the performance metrics now provide a more stable and reliable estimate of the DTDAR framework's prediction skill.

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.

Responses: Following the reviewer's suggestion, we have decided to focus exclusively on point-forecasting performance to ensure a clearer and more concise research narrative. We have removed the Average Interval Width (AIW) and Coverage Rate (CR) from all tables and text.

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.

Responses: We agree that comparing our proposed framework with modern machine learning/deep learning approaches is essential to demonstrate its competitiveness and robustness in the current hydrological modeling landscape. We have included a Long Short-Term Memory (LSTM) network as a representative modern nonlinear data-driven baseline. LSTM was selected due to its widespread success in daily streamflow prediction and its inherent ability to handle sequential dependencies. All 15 stations were re-evaluated using the expanded set of benchmarks based on the updated 70%/30% calibration-testing split. In the revised manuscript, we now explicitly discuss the rationale for our benchmarking set. While the LSTM model offers strong predictive power, our DT-DAR framework provides a unique balance between high accuracy and statistical interpretability—specifically by explicitly quantifying long-memory effects, threshold-based regime switching, and error distribution characteristics (Gaussian vs. Student's t).

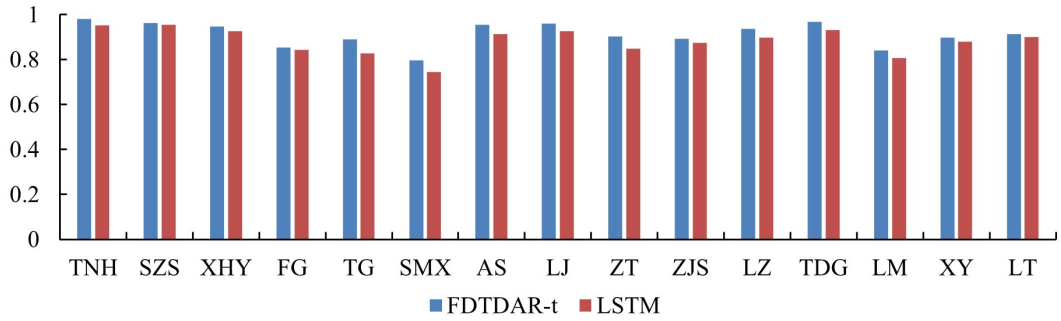

Figure 10: Comparison of prediction accuracy (NSE) during the validation period between LSTM and FDTDAR-t models

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).

Responses: We have changed the sentence to "The non-stationarity and non-linearity of streamflow have increased with changes in the environment……".

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.

Responses: We have unified the notation throughout the text.

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.

Responses: We have changed 'The hydrological statistical method' to 'time series analysis method'.

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.

Responses: We have removed the forward-looking statements from L42-43, L72-74, and L80-84. These sections now focus exclusively on the background, the progression of previous research, and the identification of the existing knowledge gaps. All specific methodological contributions have been consolidated into the final paragraph of the Introduction:

"This study aims to improve the prediction accuracy of daily streamflow time series by constructing a novel model with high applicability that can simultaneously capture seasonality, non-stationarity, long-term memory, and nonlinearity of daily streamflow. We propose the FDTDAR framework that systematically integrates seasonal normalization, fractional differencing for long-memory modeling, and a dual-threshold structure to capture regime-specific nonlinearities. Furthermore, we evaluate alternative residual distributions—

comparing the standard Gaussian distribution against the heavy-tailed Student's t-distribution—to improve error characterization during extreme events. The framework is applied to daily streamflow data from 15 gauging stations across the Yellow River Basin. For robust evaluation, the historical data are divided into a 70% calibration period and a 30% out-of-sample testing period to assess the one-day-ahead forecasting performance. The DTDAR framework is rigorously evaluated against both the classical linear AR-GARCH model and a deep learning-based LSTM network. Evaluation metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of determination (R^2), Nash-Sutcliffe efficiency coefficient (NSE), and Absolute Maximum Error (AME), providing a multi-faceted assessment of model robustness."

5) L86-87: A reference is needed to support this statement.
Responses: Thanks. We have added references to support this statement.
"However, the applicability of the TAR-GARCH model to hydrological time series is limited by narrow parameter constraints (Guo et al., 2021b; Li et al., 2016)."

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).
Responses: Yes. We have restated the specific selection criteria: "This study utilizes daily streamflow records from 15 hydrological stations in the Yellow River Basin. These stations were selected based on rigorous criteria to ensure statistical robustness: (i) a minimum continuous record length of 10 years; (ii) a data completeness requirement of 100%, with zero missing daily values throughout the selected study period to ensure continuous time-series modeling; and (iii) spatial representativeness covering the upper, middle, and lower reaches of the basin."

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.
Responses: We have rewritten the sentence: "The daily streamflow record for each station is partitioned chronologically: the initial 70% of the data serves as the calibration period for identifying statistical characteristics and estimating model parameters, while the remaining 30% is reserved as an independent testing set. This setup allows for assessing the model fit during calibration and conducting a robust, out-of-sample evaluation of prediction accuracy across the various models."

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.
Responses: We have added Table S4 to display the LM test results in the Supplement.

Table S4: The LM test of daily streamflow time series

Stations	Original		De-seasonal standardization		Fractional difference	
	Statistics	*p*	Statistics	*p*	Statistics	*p*

TNH	4298.98	< 0.01	3580.84	< 0.01	3712.7	< 0.01
SZS	4287.97	< 0.01	3493.27	< 0.01	2518.44	< 0.01
XHY	3705.25	< 0.01	1769.77	< 0.01	1647.48	< 0.01
FG	3614.65	< 0.01	2441.83	< 0.01	881.492	< 0.01
TG	3528.39	< 0.01	2411.24	< 0.01	881.567	< 0.01
SMX	2941.85	< 0.01	1530.06	< 0.01	542.472	< 0.01
AS	4008.39	< 0.01	3628.33	< 0.01	2925.07	< 0.01
LJ	4087.1	< 0.01	3562.17	< 0.01	3076.79	< 0.01
ZT	2146.79	< 0.01	2721.1	< 0.01	819.449	< 0.01
ZJS	470.058	< 0.01	1736.57	< 0.01	173.742	< 0.01
LZ	3274.57	< 0.01	1682.35	< 0.01	392.955	< 0.01
TDG	3559.98	< 0.01	3067.69	< 0.01	1963.78	< 0.01
LM	1900.84	< 0.01	1903.24	< 0.01	22.1043	< 0.01
LT	1936.85	< 0.01	3692.69	< 0.01	384.945	< 0.01
XY	3619.99	< 0.01	6059.19	< 0.01	963.709	< 0.01

9) L155-161: Please explicitly state the null hypothesis of the BDS test (typically i.i.d. behavior). Without defining $H0$, it is difficult to interpret the reported p-values and the conclusion about nonlinearity.

Responses: Yes. The null hypothesis of the BDS test is that the time series is independently and identically distributed (i.i.d.). We have added several sentences to explain the BDS test.

"The null hypothesis (H0) of the BDS test is that the time series is independently and identically distributed (i.i.d.). If the null hypothesis is rejected ($p < 0.05$), it indicates that the series is not i.i.d. and is a nonlinear stochastic process."

10) In Eq. (1), σm 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.

Yes. The σm is the standard deviation, and we have changed the text throughout the text.

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).

Responses: We thank the reviewer for pointing out the inconsistency in our model terminology. We have implemented the following corrections. At the first mention in Section 3.2 (and throughout the Introduction), we now explicitly define the acronym as the "Fractional-differenced Dual-Threshold Double Autoregressive (FDTDAR)" model. We have conducted a full-text audit to ensure consistent usage. We now use FDTDAR to refer to our proposed final framework. To avoid confusion, we have added a brief sentence in the Methods section to clarify the distinction between the baseline DAR (Double Autoregressive), the intermediate DT-DAR (Dual-Threshold DAR), and our proposed FDTDAR (which incorporates fractional differencing).

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 $r1$, $r2$, degrees of freedom for the t-distribution if applicable, etc.) and justify why this constant is 8.

Responses: To ensure a more robust evaluation, we have re-evaluated all results based on a 70/30 training-testing partitioning strategy. This involved retraining all models and identifying the optimal architectures through the Bayesian Information Criterion (BIC). We believe this updated approach provides a more stable estimate of model performance.

$$BIC = kln(n) - 2ln(\hat{L})$$

where n is the sample size, k is the number of parameters, and $\hat{L}$ is the maximized value of the likelihood function of the model.

13) L306: it should be: ", respectively."

We have changed.

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.

Responses: We have added the 'Mean Relative Error (MRE)' in Section 3.8.

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.

Responses: We have replotted Fig. 9 according to the new results and given a clear introduction in the caption.

16) Figure 9. I think there is an error en length of the bar highlighted

Responses: We have replotted Fig. 9 according to the new results.

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).

Responses: We have removed the results about interval prediction.

18) L426: There is an extra period at the end of the sentence ("time series.."). Please correct to a single full stop.

Responses: We apologize for any oversight. While we were unable to pinpoint the specific error mentioned, we have conducted a comprehensive review of the entire manuscript to ensure linguistic accuracy and technical consistency.

19) L470: it should be “on various”. Please correct it.

Responses: Yes. We have corrected it.