Dear Referee 1

We sincerely thank the reviewer for their time, positive evaluation, and highly constructive feedback on our manuscript. We have carefully reviewed all of your comments and are pleased to inform you that we have fully addressed and incorporated every suggestion into the revised version of our manuscript. To facilitate your re-evaluation process, all changes, additions, and modifications have been explicitly highlighted in **yellow** throughout the updated text. Below, we have provided clear, point-by-point responses detailing how each of your specific comments has been resolved.

1. The inclusion of 19 global teleconnection indices from major oceanic basins provides a robust and comprehensive physical basis for the forecasting models. However, the manuscript would be further strengthened if the authors could explicitly clarify the rationale behind selecting these specific 19 indices.

We sincerely thank the reviewer for this insightful and constructive comment. We completely agree that clarifying the rationale behind the selection of these specific 19 indices strengthens the physical foundation of our modeling framework.

The selection of these 19 indices was not arbitrary; rather, it was driven by an extensive and rigorous review of previous climatological studies focusing on the Middle East, and particularly Iran. While there are numerous global teleconnection patterns, our extensive literature review revealed that these specific 19 indices have been repeatedly confirmed by various researchers to have significant, albeit complex, impacts on the precipitation and temperature regimes of Iran across different temporal and spatial scales. We have now explicitly added this rationale to Section 3.1.2 of the revised manuscript to make the physical and empirical basis of our predictor selection completely transparent.

2. The classification table for SPI (Table 2) contains obvious errors. The authors are requested to carefully verify and revise.

We are very grateful to the reviewer for catching this oversight. The reviewer is completely correct. The errors in Table 2 were primarily due to a typographical and formatting glitch during the final manuscript preparation. We sincerely apologize for this confusion. We have now carefully verified the values and completely replaced Table 2 in the revised manuscript to accurately reflect the standard SPI classification defined by McKee et al. (1993).

3. Regarding the experimental design, the authors selected time lags of 1-3 months for teleconnection indices and 3, 6, 9, and12 months for autoregressive inputs,

while only 3-month and 6-month lags were utilized in the hybrid models. The authors should provide a brief justification for these specific choices and explain the logic behind the differing lag structures across scenarios.

We sincerely thank the reviewer for this precise and valuable comment. The reviewer rightly points out that the rationale for the differing lag structures needed explicit clarification.

Our specific choices were driven by a combination of climatological physical logic and machine learning optimization principles (specifically, avoiding the curse of dimensionality and overfitting).

For the teleconnection indices (S1), a 1-3 month lag was selected because the atmospheric response to large-scale oceanic forcing (the atmospheric bridge) typically manifests within a seasonal window; beyond this, the signal decays and turns into noise. For the autoregressive scenarios (S2), we tested a wide spectrum (3 to 12 months) to thoroughly evaluate the intrinsic hydrological memory of the region.

However, for the hybrid scenarios (S3), combining the 57 features from teleconnections (19 indices ×3 lags) with long-term SPI lags (9 or 12 months) led to severe feature space explosion. Our preliminary trials indicated that adding lags beyond 6 months in the hybrid models did not provide any new predictive value (information redundancy) but drastically increased computational cost and the risk of overfitting. Therefore, 3- and 6-month lags were chosen as the optimal balance for S3. We have now incorporated this justification into Section 3.3.1 of the revised manuscript.

4. The study employs a diverse suite of nine machine learning and deep learning algorithms to ensure robustness. To improve clarity and facilitate a better understanding of the performance differences discussed later, a summary table categorizing these algorithms is recommended.

We sincerely thank the reviewer for this excellent and constructive recommendation. We completely agree that a summary table significantly enhances the clarity of the methodology and provides a quick reference guide for readers to better interpret the performance differences discussed in the Results section.

Following this suggestion, we have synthesized the descriptions of the nine algorithms and added a new summary table (Table 5) to Section 3.3.2 of the revised manuscript. This table logically categorizes the algorithms into distinct families (Baseline, Classic ML, Tree-Based Ensembles, Probabilistic, and Deep Learning) and highlights their core characteristics and primary advantages.

5. It is evident that the manuscript has been prepared with great care. However, to further enhance the overall quality of the manuscript, a thorough proofreading of the entire text is encouraged to remove minor but avoidable issues, such as inconsistent terminology, or slightly awkward phrasing that may have been overlooked.

We are very grateful to the reviewer for this final piece of constructive advice. We acknowledge that in a manuscript of this length, minor linguistic inconsistencies can sometimes be overlooked.

Following the reviewer's valuable recommendation, we have performed a comprehensive proofreading of the entire manuscript from start to finish. Special attention was paid to ensuring consistent terminology for all models, scenarios, and variables (e.g., consistently using abbreviations like "SPI", "RF", and scenario notations like "S1"). Furthermore, we have carefully reviewed and refined sentence structures throughout the document to improve readability and flow, and have corrected all minor typographical and grammatical errors found.

We are confident that these revisions have enhanced the overall quality and polish of the manuscript.

Dear Referee 2

We sincerely thank the reviewer for their time, thorough evaluation, and highly constructive feedback on our manuscript. We have carefully reviewed all of your insightful comments and are pleased to inform you that we have fully addressed and incorporated every suggestion into the revised version of our manuscript. To facilitate your re-evaluation process and to clearly distinguish these specific edits from those requested by the other reviewer, all changes, additions, and modifications made in response to your comments have been explicitly highlighted in **green** throughout the updated text. Below, we provide clear, point-by-point responses detailing how each of your specific concerns has been resolved.

1. The study clearly establishes its novelty. However, the discussion could be strengthened by briefly explaining why certain paradigms perform better in specific regions—for example, why teleconnections work well in coastal areas while autoregressive models perform better in arid inland regions. A short physical explanation would help readers understand the observed patterns.

We sincerely appreciate this insightful comment. We completely agree that providing a physical climatological rationale significantly strengthens the discussion of our empirical findings. To

address this, we have added a dedicated paragraph to the Discussion section. This new text explicitly explains the physical mechanisms behind the observed spatial patterns, detailing why coastal regions—being directly exposed to marine moisture fluxes and large-scale atmospheric circulations—benefit more from teleconnection drivers (S1). Conversely, it explains how the shielding effect of major mountain ranges (Zagros and Alborz) attenuates these large-scale signals in arid inland areas, where local land-atmosphere feedbacks, high evapotranspiration, and soil moisture deficits make internal temporal memory (autoregressive paradigms) critical for capturing drought persistence. The exact additions have been highlighted in yellow in the revised manuscript.

2. The manuscript uses 1–3 month lags for teleconnection indices. A brief explanation of how these lags were chosen (e.g., based on literature or correlation analysis) would improve transparency. The finding that longer memory (up to 12 months) improves LSTM performance is interesting. It would be worth noting whether this improvement continues with even longer memory, or if there is a point where adding more data no longer helps.

We sincerely thank the reviewer for this constructive comment. Interestingly, another reviewer also inquired about the logic behind our lag selection, which highlights the necessity of making this aspect entirely transparent.

Regarding the 1-3 month lags for teleconnections, we chose this window based on preliminary cross-correlation analyses and existing climatological literature. The atmospheric response to large-scale oceanic anomalies (the atmospheric bridge) typically manifests within a seasonal window (1 to 3 months). Beyond this period, the signal decays and essentially turns into noise.

Regarding the reviewer's excellent question about extending the memory beyond 12 months for LSTM: our preliminary exploratory analyses revealed a clear saturation point. Extending the autoregressive memory beyond an annual cycle (e.g., 18 or 24 months) provided no additional predictive skill. Instead, it introduced irrelevant noise and drastically increased the risk of overfitting, as the local hydrological memory required for a 1-3 month forecast horizon rarely exceeds one year. Adding more data beyond this point no longer helps and can even degrade model performance.

We have fully addressed both of these points by modifying and expanding Section 3.3.1 in the revised manuscript.

3. The abstract could be slightly shortened to focus on the most important findings.

We thank the reviewer for this helpful suggestion. We completely agree that a more concise abstract improves readability and impact. Accordingly, we have streamlined the introductory

sentences and the methodology description to focus more sharply on the core findings. The revised abstract has been shortened while carefully retaining all key results and messages.

4. Some figures (e.g., Figures 5, 7, 9) would benefit from clearer legends and labels.

We genuinely thank the reviewer for pointing this out. We agree that the readability of the figures is crucial for the final publication. Accordingly, we have thoroughly revised Figures 5, 7, and 9. Specifically, we have:

1. *Increased the font size and bolded the labels and numbers on all color bars (legends) to ensure clarity even if the images are scaled down.*
2. *Corrected the typographical formatting of the Coefficient of Determination in the legends (changed from R2 to R^2).*
3. *Added standard map elements (scale bar and north arrow) for geographical clarity.*
4. *Exported the revised figures at a higher resolution (300 DPI) to prevent any pixelation.*

We believe these modifications have significantly enhanced the visual quality and readability of the mentioned figures.

5. Ensure all abbreviations (SPI, RF, LSTM, etc.) are defined at first use in the main text.

We appreciate the reviewer's careful attention to detail. Following this comment, we have thoroughly proofread the entire manuscript from beginning to end. We have ensured that all abbreviations and acronyms (including SPI, RF, LSTM, among others) are explicitly defined at their very first occurrence in the abstract, the main text, and the figure/table captions. Any previously missed instances or inconsistencies have been entirely rectified.

6. The study period is 1993–2022—a brief note on whether this period is representative of longer-term climate patterns would be useful. Why not 1991-2020 (officially recommended by WMO)

We sincerely thank the reviewer for raising this insightful point regarding the WMO standard climatological normals.

The reviewer is absolutely correct that 1991–2020 is the WMO-recommended period for standard climate normals. However, our decision to use the 1993–2022 period was driven by the specific data requirements of machine learning-based forecasting models, rather than statistical climatology or climate change anomaly assessments.

First, our primary constraint was data quality; the year 1993 was the earliest point where we could ensure a continuous, gap-free, and homogenous dataset for all 96 selected synoptic stations (as detailed in our QA/QC protocol). Second, for short-term predictive modeling, it is

crucial to train the algorithms on the most recent atmospheric dynamics. Truncating the dataset at 2020 would have meant discarding two recent and highly valuable years of drought/precipitation data.

Nevertheless, a 30-year continuous period (360 months) fully satisfies the statistical threshold required to be representative of long-term climate variability, capturing decadal oscillations and major multi-year drought cycles in the region.

As accurately suggested by the reviewer, we have added a brief explanatory note to ***Section 3.1.1*** *of the revised manuscript to clarify the representativeness of this period and the rationale behind our selection.*

7. extract the table on the right side from Figure 2 as a seperated one.

We thank the reviewer for this excellent suggestion, which significantly improves the manuscript's clarity and professional presentation.

As suggested, we have extracted the list of station names from Figure 2 and presented it as a new, separate table (now ***Table 1****). Consequently, Figure 2 has been updated to only show the map, and its caption now refers to Table 1 for the station details. This change has also allowed us to present the map in a larger and clearer format.*

Dear Referee 3

We sincerely thank the reviewer for their time, thorough evaluation, and highly constructive feedback on our manuscript. We have carefully reviewed all of your insightful comments and are pleased to inform you that we have fully addressed and incorporated every suggestion into the revised version of our manuscript. To facilitate your re-evaluation process and to clearly distinguish these specific edits from those requested by the other reviewer, all changes, additions, and modifications made in response to your comments have been explicitly highlighted in **blue** throughout the updated text. Below, we provide clear, point-by-point responses detailing how each of your specific concerns has been resolved.

1. The Introduction currently lacks a clear narrative flow and moves abruptly between topics. It should be reorganized to highlight the critical research gap of this study (e.g., the lack of a systematic comparison of these paradigms). The specific advancement of this paper and the main objectives should be highlighted concisely at the end.

We sincerely thank the reviewer for this constructive feedback. We completely agree that the core research gap and the main objectives must be explicitly clear and concisely highlighted for the reader.

The Introduction was originally structured to build a comprehensive step-by-step foundation: starting from the general problem of drought, introducing the SPI index, reviewing the evolution of ML models, discussing the physical drivers (teleconnections), analyzing the limitations of previous local studies, and finally arriving at the research gap. Because we wanted to preserve this rich literature and the foundational context, we addressed your valuable comment by refining the visual and structural organization of the final paragraphs rather than removing existing content.

To ensure a smoother narrative flow and to strictly follow your recommendation to "highlight concisely at the end," we have restructured the final section of the Introduction. Specifically, we explicitly emphasized the research gap and transformed our main objectives (the three predictor scenarios) into a concise, bulleted, and bolded list at the very end of the section. This reorganization makes the study's specific advancements immediately visible to the reader without losing the depth of the literature review. These structural modifications have been highlighted in ***blue*** *in the revised manuscript.*

2. The study uses 19 global indices but provides no clear rationale for their selection. Given the regional focus on Iran, it is unclear if all 19 are physically relevant or if this introduces noise. The authors must justify why these specific indices were chosen (e.g., based on literature documenting their influence on Middle East climate) and discuss whether a feature selection step was considered to retain only the most predictive drivers, thereby improving model parsimony and interpretability.

We sincerely thank the reviewer for this highly insightful comment, which touches upon a critical challenge in ML-based climate modeling: balancing a comprehensive physical input space with model parsimony to avoid overfitting (noise).

Regarding the rationale for selecting these 19 indices, another reviewer raised a similar point, and we had accordingly expanded Section 3.1.2. As explained in the revised text, the selection was driven by an extensive review of previous climatological literature focused on the Middle East and Iran. These specific 19 indices were chosen because prior empirical and dynamical studies have repeatedly confirmed their significant influence on the region's climate.

Regarding your excellent point about potential "noise" and the ***feature selection*** *step: Rather than applying an* a priori *explicit feature selection method (such as filtering or PCA) which might prematurely discard indices with complex, non-linear, or geographically localized*

*interactions, we deliberately relied on the **embedded feature selection and regularization capabilities** of our advanced ML/DL models.*

For instance, tree-based ensembles (RF, XGBoost, CatBoost, LightGBM) inherently perform feature selection during training by heavily penalizing or ignoring variables that do not maximize information gain. Similarly, the LSTM model utilizes its internal gating mechanisms (like the forget gate) to suppress irrelevant signals over time, while regularization techniques (e.g., L2 penalties in BRR and XGBoost) shrink the coefficients of uninformative predictors to zero. This approach ensures model parsimony and filters out noise without losing potential secondary drivers.

*To clarify this methodological decision for the readers, we have added a new paragraph at the end of Section 3.1.2 explicitly addressing how the algorithms handle potential noise from the 19 indices. This addition is highlighted in **blue** in the revised manuscript.*

3. 3. While the results thoroughly document which models/scenarios perform best where, they offer limited explanation of why. The Results/Discussion should be expanded to link observed performance patterns to known physical mechanisms. Incorporating insights from atmospheric dynamics (e.g., persistence of soil moisture, role of specific teleconnections) would significantly elevate the paper's scientific contribution.

We deeply appreciate this highly constructive and insightful feedback. We completely agree that moving beyond statistical performance documentation to provide robust physical and climatological interpretations is essential for elevating the scientific contribution of this study.

To fully address your recommendation, we have significantly expanded the Discussion section (Section 5) to explicitly link the observed spatial performance patterns to known atmospheric dynamics and physical geographic features. Specifically, we have added detailed explanations regarding:

1. ***The role of specific atmospheric dynamics:** We explained how coastal regions (e.g., Anzali and Chabahar) are directly modulated by marine moisture fluxes and unhindered exposure to large-scale atmospheric circulation patterns, such as the Siberian High and the Indian Summer Monsoon, which makes the teleconnection-driven scenario (S1) highly effective there.*
2. ***Topographic blocking effects:** We discussed how external oceanic signals are significantly attenuated in the inland and Central Plateau regions due to the blocking effect of the Zagros and Alborz mountain ranges.*
3. ***Persistence of soil moisture (as you aptly suggested):** We detailed how the local climate in arid inland regions is governed by land-atmosphere feedbacks. In these environments, the physical inertia of the system—driven by slow-responding variables exactly like **soil moisture deficits** and groundwater depletion—makes internal temporal memory*

(autoregressive models) and hybrid approaches critical for capturing drought persistence.

These comprehensive physical interpretations have transformed the Discussion section from a mere comparison of algorithms into an analysis of drought predictability dynamics. We believe these additions perfectly align with your suggestion and heavily enrich the manuscript. These extensive additions can be found in Section 5 of the revised manuscript.

4. 4. The design of input scenarios, particularly the lag structures for teleconnections (Tele t-1, t-2, t-3) and the autoregressive memory (up to 12 months), needs better justification. Why was 3-month maximum lag chosen for climate indices? Is this sufficient to capture known lagged responses of Iran's climate to phenomena like ENSO? Providing a brief explanation based on physical understanding or preliminary correlation analysis would strengthen the assessment.

We sincerely thank the reviewer for this highly insightful and critical question. Interestingly, the necessity of justifying the lag structures was a shared consensus among the reviewers, which highlighted the importance of making this methodological choice entirely transparent.

*We have comprehensively revised **Section 3.3.1** to provide a solid physical and statistical justification for our lag design. To directly answer your question regarding whether a 3-month lag is sufficient for phenomena like ENSO: Yes, based on both our preliminary cross-correlation analyses and the climatological literature for the Middle East. The delayed response of Iran's climate to large-scale oceanic forcing like ENSO occurs via the "atmospheric bridge" (e.g., the propagation of Rossby waves). This atmospheric manifestation typically reaches the region within a seasonal window (1 to 3 months). Extending the lag for teleconnections beyond 3 months does not capture more of the ENSO signal; rather, the external oceanic signal decays and becomes entirely overshadowed by local and regional atmospheric noise, drastically increasing the risk of overfitting in machine learning models.*

Regarding the autoregressive scenarios, our exploratory analyses revealed a clear saturation point at 12 months. Extending the internal hydrological memory beyond an annual cycle provided no additional predictive skill and merely introduced irrelevant noise.

*We have fully incorporated these physical understandings and statistical constraints into the newly expanded **Section 3.3.1**, explicitly addressing the atmospheric response to anomalies like ENSO. We believe this addition perfectly addresses your concern and significantly strengthens the methodological rationale of the paper.*

5. 5. The discussion could be enriched by contextualizing the findings within the broader literature on climate predictability.

We sincerely thank the reviewer for this excellent overarching suggestion. We completely agree that framing our empirical findings within the broader, fundamental theories of climate predictability significantly elevates the scientific narrative of the paper.

In response to your valuable comment, we have enriched the ***Discussion (Section 5)*** *by explicitly connecting our scenario outcomes (S1, S2, and S3) to the broader theoretical frameworks of Sub-seasonal to Seasonal (S2S) climate predictability. Specifically, we added new context (with key literature citations) emphasizing how our findings empirically validate the theoretical consensus that predictive skill is a delicate balance between:*

1. *Slowly varying oceanic boundary conditions (external sources of predictability, mirrored in our S1 scenario).*
2. *Local land-surface memory (internal sources of predictability, mirrored in our S2 scenario).*

Furthermore, we explicitly tied the observed performance drop in the teleconnection models to the broader literature on the transition from boundary-forced predictable states to weather-noise dominated states, acknowledging the well-known "signal-to-noise paradox."

These new contextualizations, supported by foundational S2S literature (e.g., Vitart et al., 2017; Scaife and Smith, 2018; Merryfield et al., 2020), have been integrated into the first and fourth paragraphs of the Discussion section. We believe these additions successfully address your comment and provide a much richer, globally relevant context for our findings.