Supplementary 1

Short-Term Drought Forecasting in Iran Using Multi-Source Machine Learning: An Assessment of Autoregressive, Teleconnection-Driven, and Hybrid Paradigms

Jun Jian[1, 2], Peyman Mahmoudi[3], Pouria Jafari[4], Alireza Ghaemi[3], Jing Yang[5], Fatemeh Firoozi[6]

[1]Navigation College, Dalian Maritime University, Dalian, China
[2]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, China
[3]Department of Physical Geography, Faculty of geography and environmental planning, University of Sistan and Baluchestan, Zahedan, Iran
[4]Department of Department of Electronic and Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Sistan and Baluchestan, Zahedan, Iran
[5]Faculty of Geographical Science, Key Laboratory of Environmental Change and Natural Disaster, Beijing Normal University, Beijing, China
[6]Department of Humanities and Social Science, Farhangyan University, Tehran, Iran

Correspondence to: Peyman Mahmoudi (p_mahmoudi@gep.usb.ac.ir)

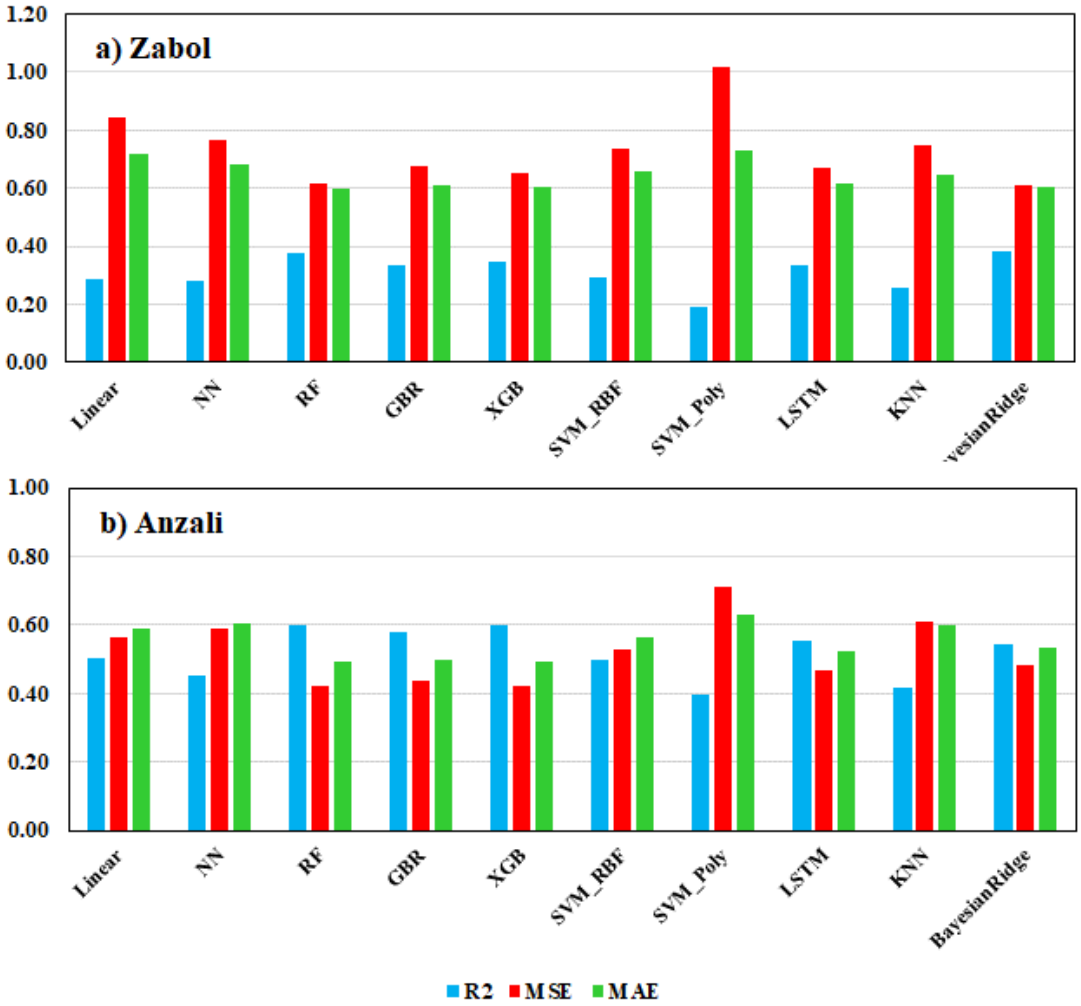

Figure S1 (a, b): Performance comparison of the nine machine learning algorithms in two-month-ahead forecasting of the drought index (SPI_{t+2}) under the teleconnection-based scenario (S1). The charts display the results for two stations with contrasting climates: Zabol (hyper-arid) and Anzali (hyper-humid). The metrics include the Coefficient of Determination (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE). These charts illustrate the general decrease in accuracy compared to the one-month forecast and the tighter competition among the top-performing models.

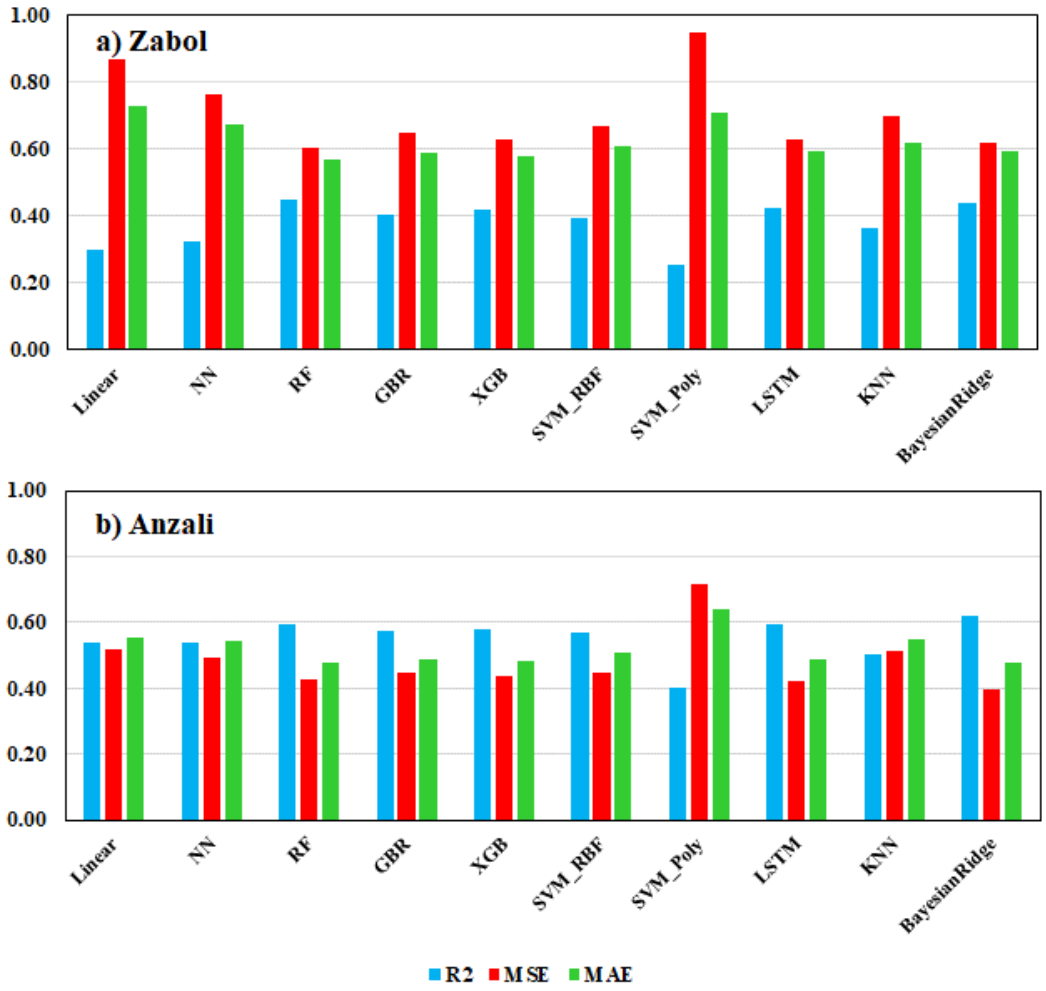

Figure S1 (c, d): Performance comparison of nine machine learning algorithms for three-month-ahead drought index forecasting (SPI_{t+3}) under the teleconnection-based scenario (S1). The charts display results for two stations with contrasting climates: Zabol (hyper-arid) and Anzali (hyper-humid). The metrics include the Coefficient of Determination (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE). These charts highlight the close competition between the RF and BRR models and the dependency of superior performance on climatic conditions.

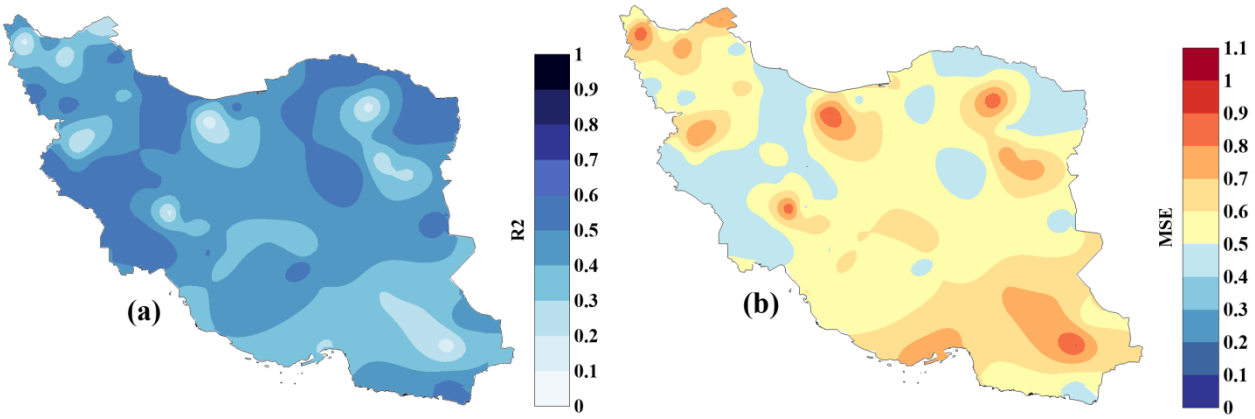

Figure S2 (a, b): Spatial distribution of the performance of the best-selected model (Random Forest, RF) for two-month-ahead drought forecasting (SPI_{t+2}) under scenario S1. Left panel: Coefficient of Determination (R^2). Right panel: Mean Squared Error (MSE). A comparison of these maps with Figure 5 indicates a significant decline in predictability across most regions of the country as the lead time increases, particularly the reduction in the extent of high-accuracy areas and the increase in error in the central and southern regions.

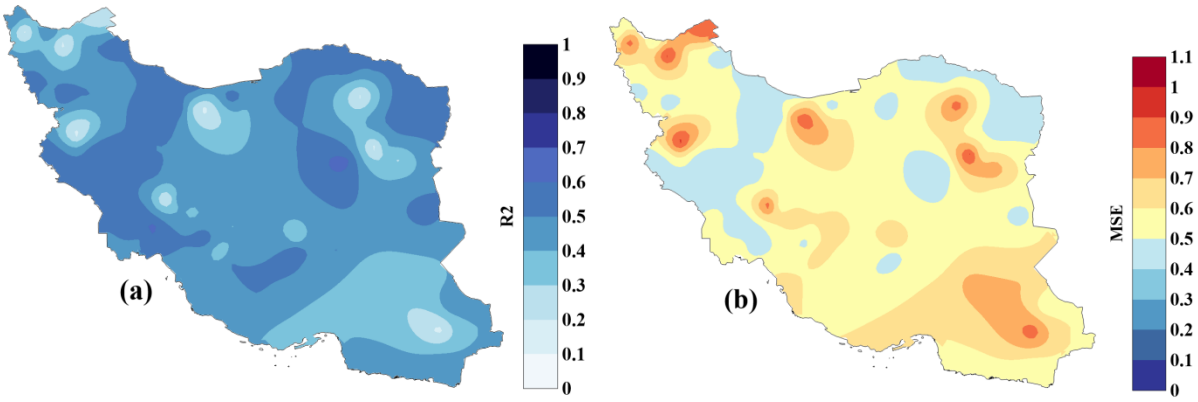

Figure S2 (c, d): Spatial distribution of the performance of the best-selected model (Random Forest, RF) for three-month-ahead drought forecasting (SPI_{t+3}) under Scenario S1. Left panel: Coefficient of Determination (R^2). Right panel: Mean Squared Error (MSE). A comparison of these maps with Figure 11 indicates a marked decrease in predictability across most parts of the country as the lead time increases, particularly the reduced extent of high-accuracy regions and increased error in central and southern areas.

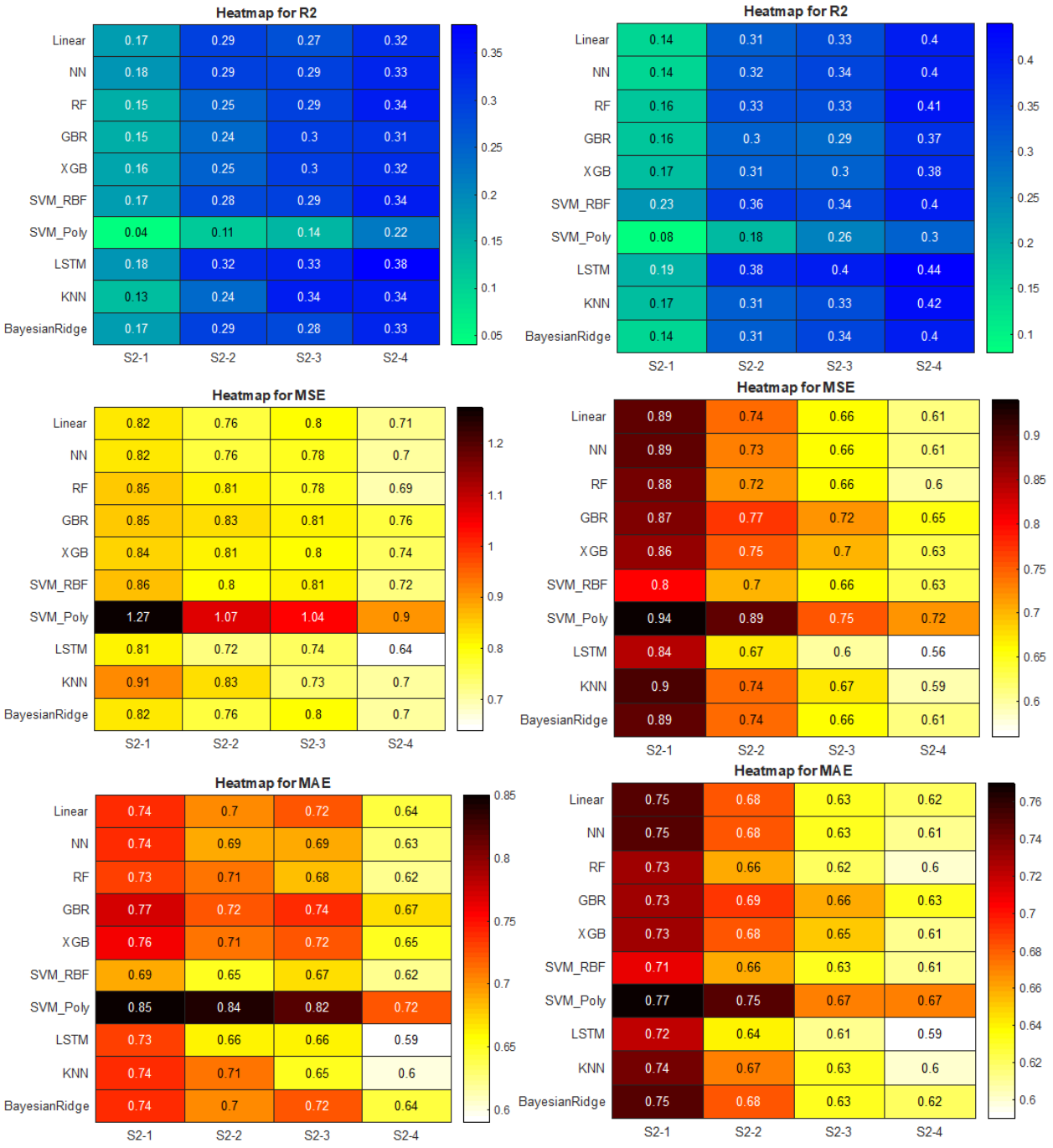

Figure S3 (a): Heatmaps for evaluating the performance of the 9 machine learning models in two-month-ahead drought forecasting (SPI_{t+2}) under the four temporal memory scenarios (S2-1 to S2-4) for the Anzali (hyper-humid, right columns) and Zabol (hyper-arid, left columns) stations. The metrics include R^2, MSE, and MAE. The results clearly demonstrate that increasing the input memory length from 3 months (S2-1) to 12 months (S2-4) significantly improves the performance of most models, particularly LSTM, emphasizing the importance of long-term temporal dependencies in forecasting with a longer lead time.

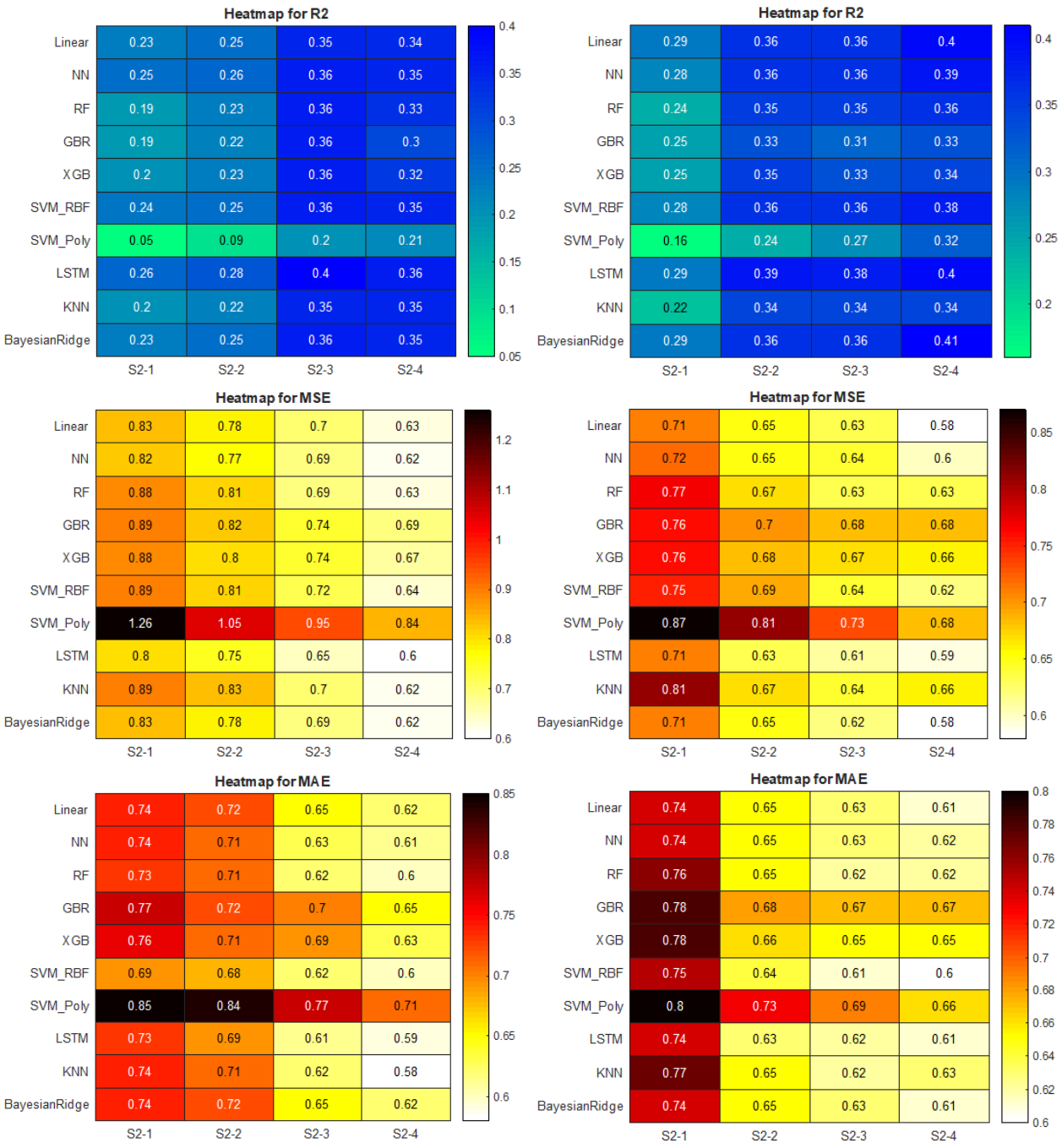

Figure S3 (b): Heatmap comparison of the performance of nine machine learning algorithms in three-month-ahead drought index forecasting (SPI_{t+3}) under the temporal memory-based scenario (S2). Each column represents an input structure with a different memory length (3, 6, 9, and 12 months). The results are displayed for two stations, Anzali (hyper-humid, right) and Zabol (hyper-arid, left), based on the R^2, MSE, and MAE metrics. The charts clearly demonstrate that increasing the memory length to 12 months (S2-4) dramatically improves the performance of most models, particularly LSTM.

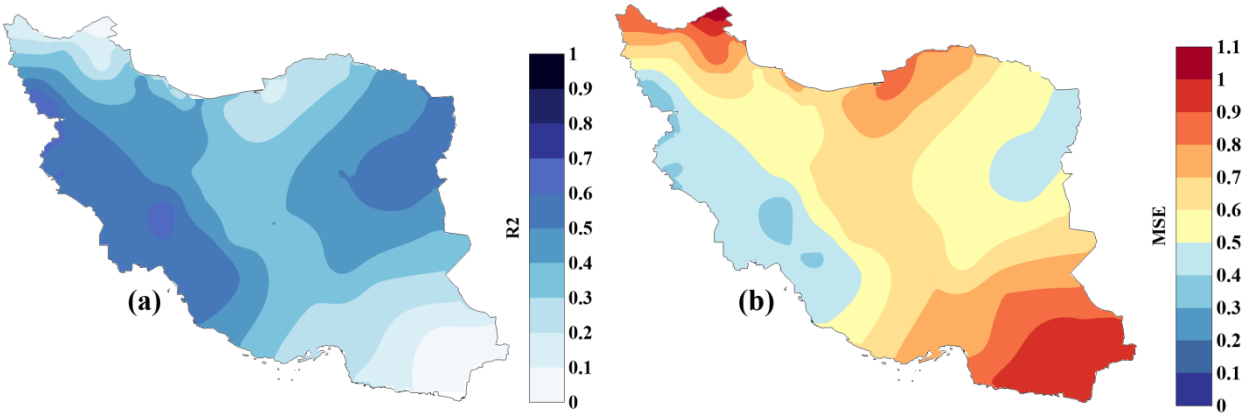

Figure S4 (a, b): Spatial distribution of the performance of the best-selected model (LSTM) with its optimal input configuration (12-month memory, S2-4) for two-month-ahead drought forecasting (SPI_{t+2}). Left panel: Coefficient of Determination (R^2). Right panel: Mean Squared Error (MSE). Despite the overall decrease in accuracy compared to the one-month forecast, the model still shows acceptable predictability (R^2>0.5) in the western and northwestern regions, while its performance in the southeastern areas and parts of the Caspian coasts is severely limited due to the irregular nature of precipitation at this lead time.

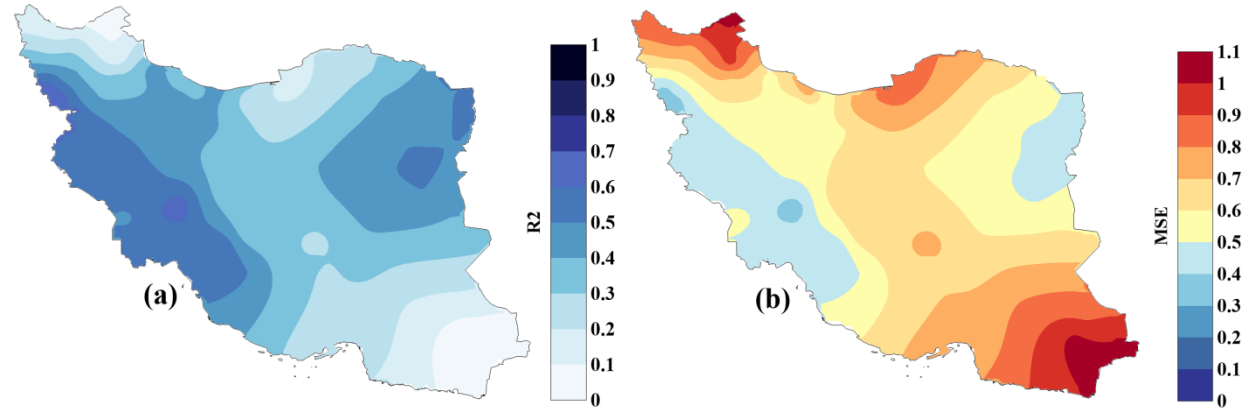

Figure S4 (c, d): Spatial distribution of the performance of the best-selected model (LSTM) with the optimal input structure (12-month memory, S2-4) for three-month-ahead drought forecasting (SPI_{t+3}). Left panel: Coefficient of Determination (R^2). Right panel: Mean Squared Error (MSE). Despite an overall decline in accuracy compared to shorter lead times, the model still exhibits acceptable predictability ($R^2 = 0.50$) in the western and northwestern regions, whereas its performance in the southeastern areas is severely limited due to the irregular nature of precipitation and increased uncertainty at this lead time.

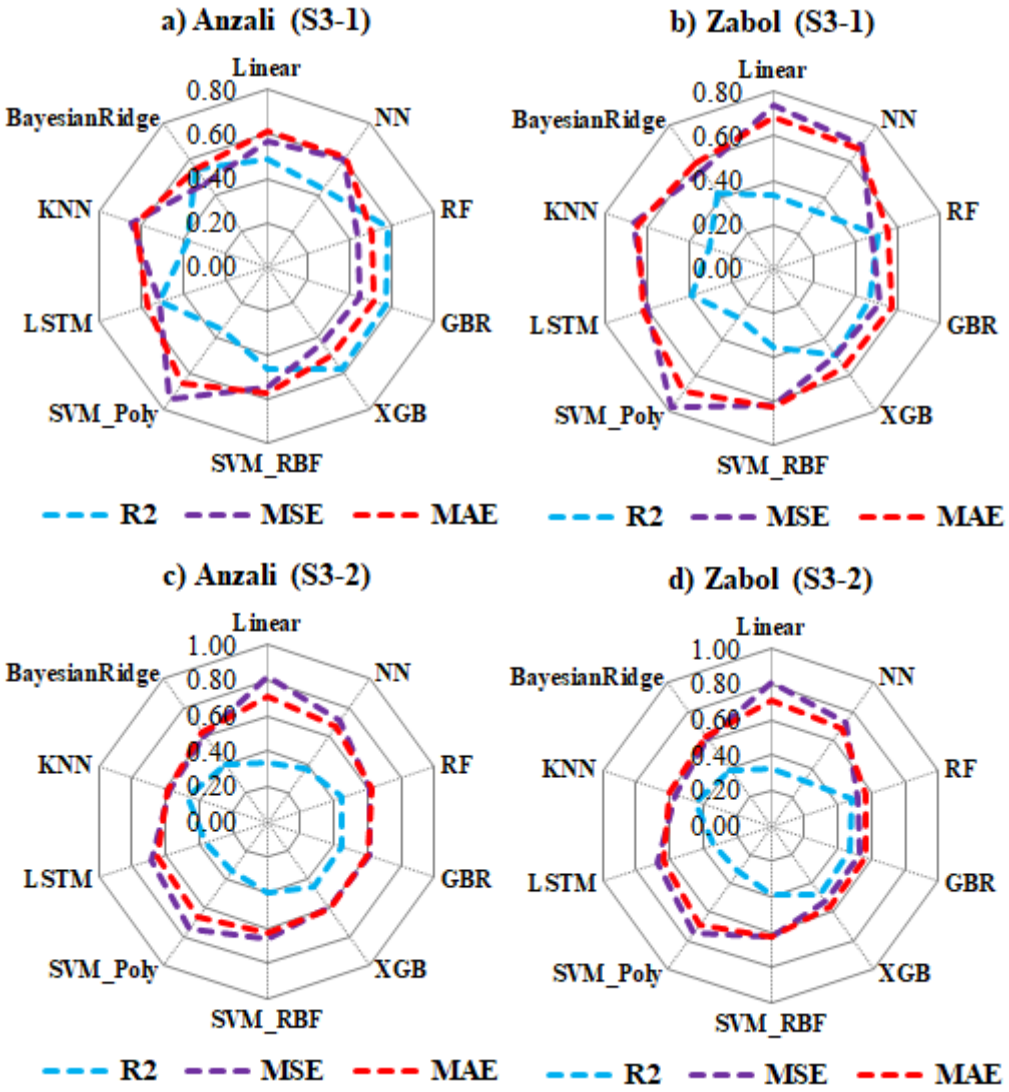

Figure S5 (a, b): Radar charts comparing the performance (R^2, MSE, and MAE) of the different machine learning models for two-month-ahead forecasting (SPI_{t+2}) at the Anzali (hyper-humid) and Zabol (hyper-arid) stations. This evaluation was conducted under the two input structures of the third scenario: S3-1 (a combination of teleconnections and 3-month memory) and S3-2 (a combination of teleconnections and 6-month memory). The charts illustrate the superiority of the Random Forest (RF) model and the varying impact of temporal memory length on model accuracy across different climates.

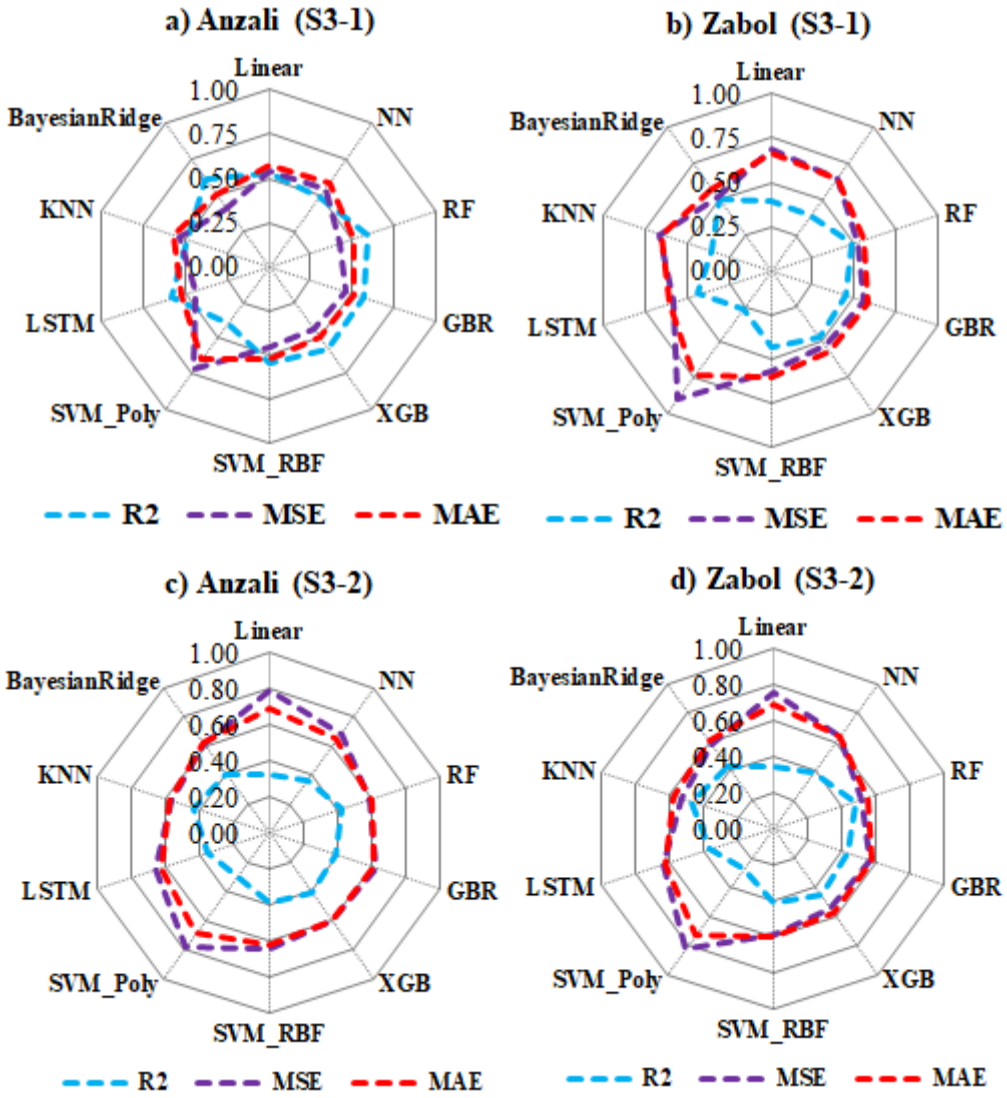

Figure S5 (c, d): Radar charts comparing the performance (including R^2, MSE, and MAE) for the nine machine learning models in three-month-ahead forecasting (SPI_{t+3}) at the Anzali (hyper-humid) and Zabol (hyper-arid) stations. The evaluation was conducted under the two hybrid structures of the third scenario: S3-1 (teleconnections + 3-month memory) and S3-2 (teleconnections + 6-month memory). The charts, while showing the overall superiority of the RF model, depict the contrasting impact of increasing temporal memory in the two different climates.

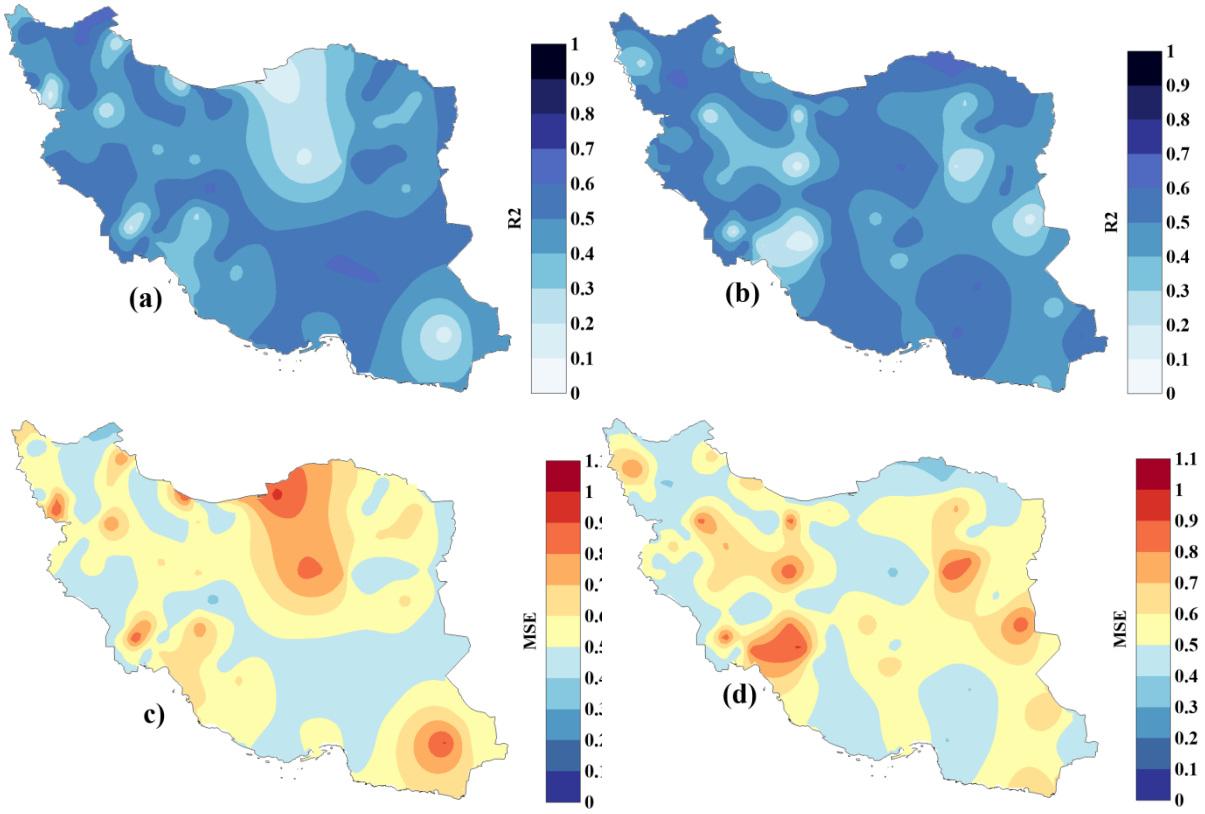

Figure S6 (a, b): Spatial distribution map of the Coefficient of Determination (R^2) and Mean Squared Error (MSE) values for the best model (RF) in forecasting SPI_{t+2} across Iran, based on the two scenarios S3-1 (left column) and S3-2 (right column). The maps reveal the spatial complementarity of the two scenarios; while S3-2 demonstrates superiority in the north and center of the country, S3-1 provides better performance in the eastern regions and parts of the south.

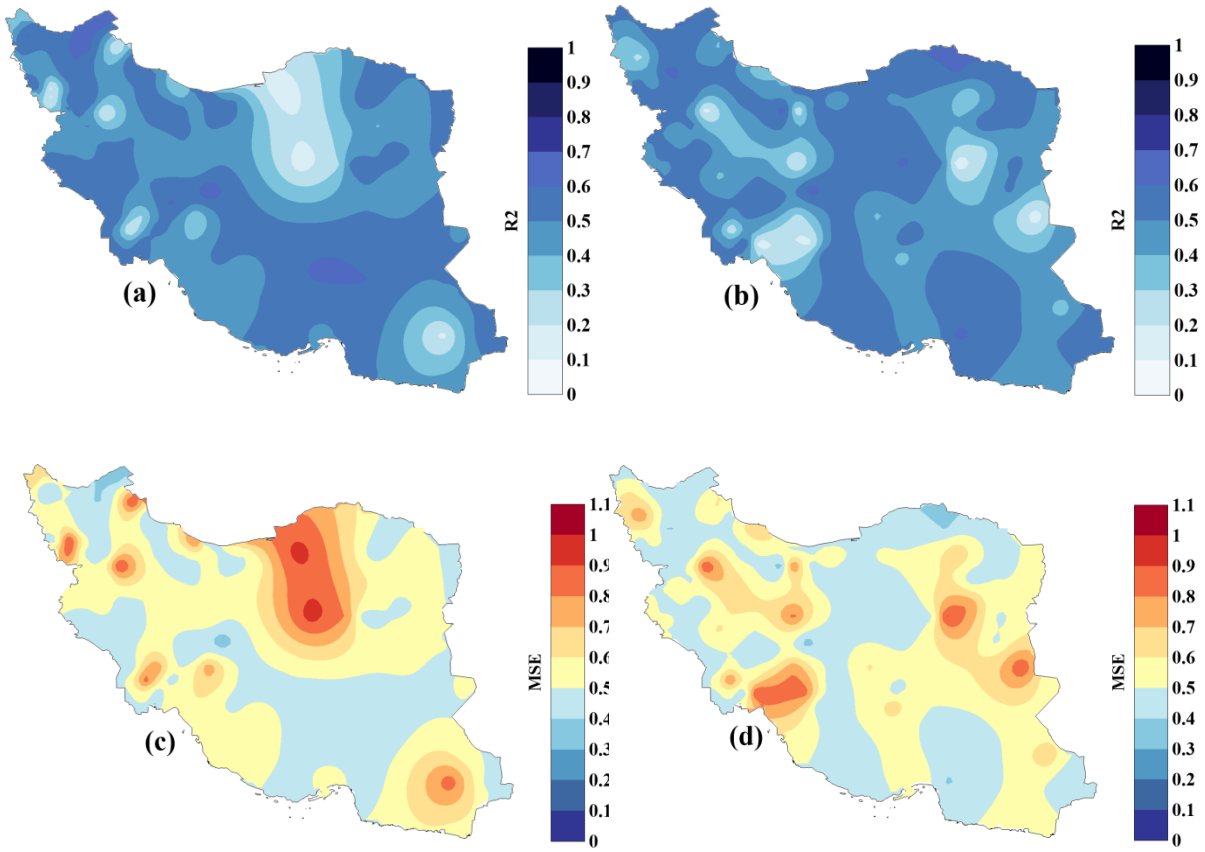

Figure S6 (c, d): Spatial distribution map of the Coefficient of Determination (R^2) and Mean Squared Error (MSE) values for the best model (RF) in forecasting SPI_{t+3} across Iran, based on the two scenarios S3-1 (left column) and S3-2 (right column). The maps reveal the spatial complementarity of the two scenarios; while S3-2 demonstrates superiority in the west and northeast, S3-1 provides better performance in the southern and eastern regions.

Table S7: Comparative performance evaluation of the best-selected models from each of the three main scenarios (S1, S2, S3) for two-month-ahead forecasting of the SPI_{t+2} index at sample stations. This table presents the statistical metrics R^2, MSE, and MAE to facilitate the identification of the optimal input structure in different climatic regions of Iran and highlights the geographical dependency of the models' performance on the chosen scenario.

Station	Model	Scenario	R2	MSE	MAE
Zabol	RF	1	0.38	0.61	**0.60**
	LSTM	S2-4	0.38	0.64	**0.59**
	RF	**S3-1**	**0.50**	**0.47**	0.55
Anzali	**RF**	**1**	**0.60**	**0.42**	0.49
	LSTM	S2-4	0.44	0.56	**0.59**
	RF	S3-1	0.58	0.43	**0.50**
Mashhad	**RF**	**1**	**0.60**	**0.41**	0.49
	LSTM	S2-4	0.48	0.52	**0.57**
	RF	S3-2	0.57	0.44	**0.50**
Esfahan	RF	1	0.50	0.48	**0.55**
	LSTM	S2-4	0.39	0.63	**0.62**
	RF	**S3-1**	**0.65**	**0.35**	0.47
Ahwaz	RF	1	0.56	0.45	**0.50**
	LSTM	**S2-4**	**0.55**	**0.44**	0.46
	RF	S3-2	0.13	0.92	**0.72**
Chabahar	**RF**	**1**	**0.56**	**0.44**	0.50
	LSTM	S2-4	0.07	0.92	**0.77**
	RF	S3-1	0.40	0.63	**0.66**
Bandar	RF	1	0.29	0.74	**0.67**
Abbas	LSTM	S2-4	0.23	0.79	**0.68**
	RF	**S3-2**	**0.45**	**0.61**	0.62

Table S8: **Comparative performance evaluation of the best-selected models from each of the three main scenarios (S1, S2, S3) for the three-month-ahead forecast of the SPI index (SPI_{t+3}) at sample stations. This table presents the statistical metrics R^2, MSE, and MAE to identify the optimal input structure in different climatic regions of Iran, highlighting the geographical dependence of model performance on the chosen scenario.**

Station	Model	Scenario	R2	MSE	MAE
Zabol	RF	S1	0.45	0.60	**0.57**
	LSTM	S2-4	0.36	0.60	**0.59**
	RF	**S3-1**	**0.48**	**0.52**	**0.56**
Anzali	**RF**	**S1**	**0.60**	**0.43**	**0.48**
	LSTM	S2-4	0.49	0.55	**0.58**
	RF	S3-1	0.59	0.43	**0.50**
Mashhad	**RF**	**S1**	**0.60**	**0.43**	**0.48**
	RF	S2-4	0.50	0.48	**0.55**
	RF	S3-2	0.56	0.44	**0.50**
Esfahan	RF	S1	0.48	0.52	**0.56**
	LSTM	S2-4	0.34	0.66	**0.64**
	RF	**S3-1**	**0.68**	**0.33**	**0.43**
Ahwaz	RF	S1	0.54	0.49	**0.50**
	LSTM	**S2-4**	**0.58**	**0.42**	**0.46**
	RF	S3-2	0.18	0.84	**0.68**
Chabahar	**RF**	**S1**	**0.54**	**0.49**	**0.50**
	LSTM	S2-4	0.06	0.99	**0.81**
	RF	S3-1	0.44	0.57	**0.60**
Bandar Abbas	RF	S1	0.34	0.67	**0.63**
	LSTM	S2-4	0.22	0.76	**0.68**
	RF	**S3-1**	**0.45**	**0.56**	**0.59**

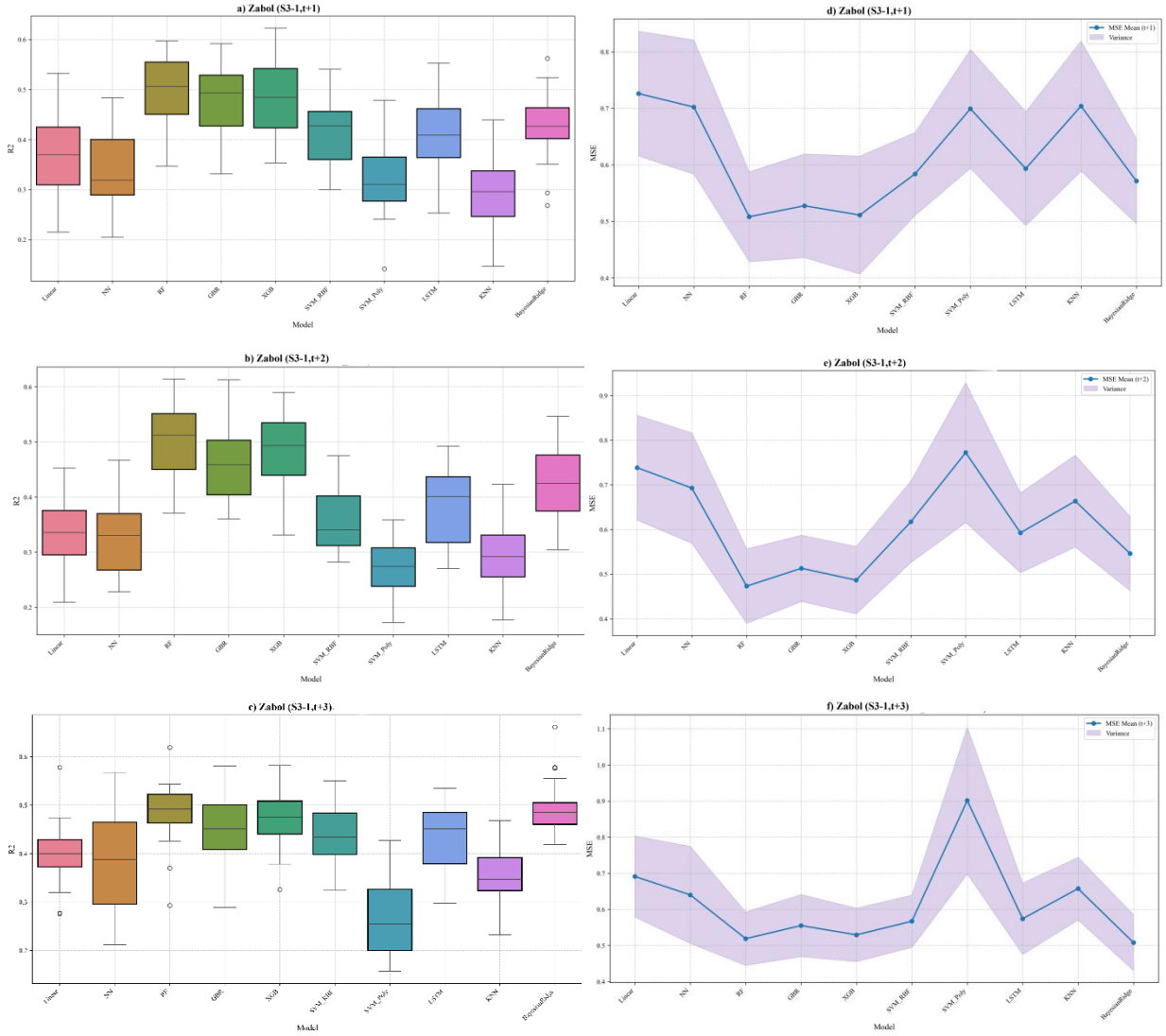

Figure S7: Evaluation of model stability and performance at the Zabol station under the optimal scenario S3-1 for one-, two-, and three-month forecast horizons (t+1, t+2, t+3). The left panels (a, b, c) display the distribution of the coefficient of determination (R^2) values in a box plot format for 20 independent runs. The right panels (d, e, f) show the mean (solid line) and the 95% confidence interval (shaded area) of the Mean Squared Error (MSE) for the same 20 runs. The results highlight the superior stability and accuracy of the RF model in an arid region.

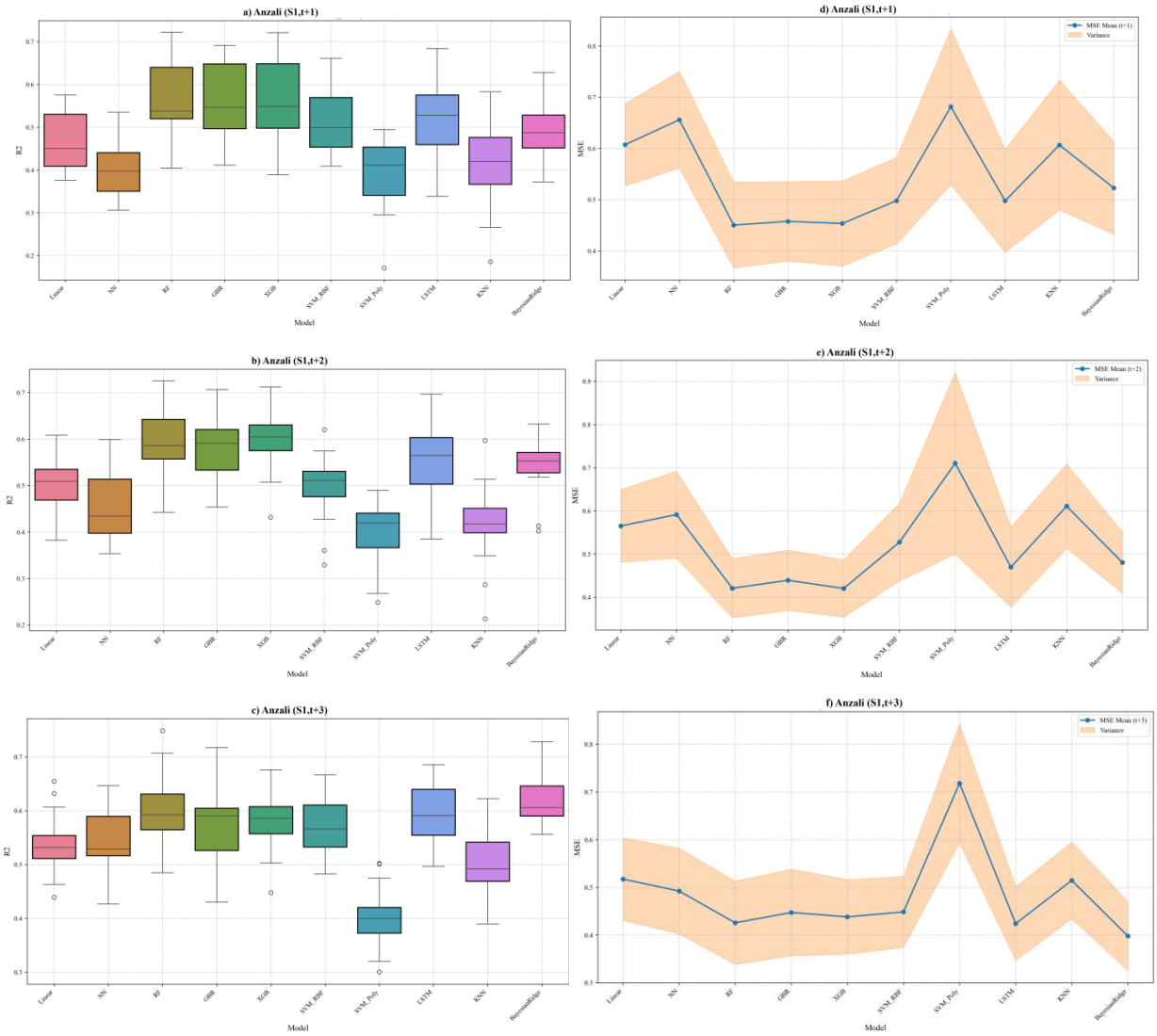

Figure S8: Evaluation of model stability and performance at the Anzali station under the optimal scenario S1 for one-, two-, and three-month forecast horizons (t+1, t+2, t+3). The left panels (a, b, c) show the distribution of the coefficient of determination (R^2), and the right panels (d, e, f) show the mean and variance of the Mean Squared Error (MSE) based on 20 independent runs. These plots are designed for a comparative analysis of the robustness and uncertainty of the models in a hyper-humid region with an irregular precipitation regime.