STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, October 2026, pp 2957–2979.
Published online in International Academic Press (www.IAPress.org)

Regime-Aware Gated Ensemble for Monthly Temperature Forecasting in Nineveh Governorate

Ziadoon Mohand Khaleel[1,*], Safa Jawad Abed[2], Jalal Abdulkareem Sultan[2], Noor Marwan Ahmeed[3]

[1]*Department of Petroleum Reservoir Eng.,College of Petroleum and Mining Engineering, University of Mosul, Iraq*
[2]*Nineveh Agriculture Directorate, Iraq*
[3]*Department of Construction and Projects,University of Mosul, Iraq*

Abstract Monthly forecasting of maximum (MAX) and minimum (MIN) temperatures plays a vital role in scheduling agriculture activities, planning energy production, and heat risk management in semi-arid regions. In this work, we conduct an experimental study of the sequential expert weighting for monthly MAX and MIN temperature forecasting in Nineveh Governorate, Iraq based on historical time series recorded at Al-Hamdaniah, Baaj, Mosul, and Rabiaa sites. The modified procedure employs three forecasting experts: AIC-selected SARIMA, Exponential Smoothing (ETS), and reverse hybrid LSTM→SARIMA model, combining them through two adaptive combination procedures: prequential weight learning algorithm and gated ensemble method. To avoid temporal leakage in learning expert weights, all weights are calculated based solely on historical loss of rolling origin predictions rather than on current or future forecast errors. The evaluation utilizes chronological training/validation/test split, multi-origin rolling forecast predictions, 18 months prediction horizon, seasonal-naïve benchmarks correction, SARIMA model selection, and gate hyperparameters sensitivity analysis. The findings show that seasonal statistical models are still the most convenient models to be applied in this situation. For MAX temperature, SARIMA achieved the smallest average test error rate, with MAE values being between 1.473 and 1.489 °C for both fixed restricted and AIC-selected models. For MIN temperature, ETS proved to be the best performing model, having an MAE of 0.995 °C, closely followed by SARIMA. Adaptive approaches improved performance in some naive combination configurations but failed to beat best performing statistical benchmarks consistently. Diagnostics of regimes showed that seasonality was predominant in most cases, thus explaining the low additional value of adaptive gating. Overall, this paper presents a leakage-proof evaluation of adaptive ensembling in monthly temperature forecasting and shows that, in case of strongly seasonal regional climate time series, statistical baselines should be used operationally, and gated or prequential combinations should be regarded as conditional robustness tools.

Keywords monthly temperature forecasting; Nineveh Governorate; SARIMA model selection; LSTM-SARIMA hybrid; seasonal forecasting

AMS 2010 subject classifications Primary 62M20; Secondary 62M10, 62M45, 62P12.

DOI: 10.19139/soic-2310-5070-3852

1. Introduction

Temperature prediction on a monthly basis is a very important part of operation planning in semiarid regions, where agricultural, energy demand management, drought preparation and heat risk reduction require seasonal data. In semiarid regions, monthly temperature series usually have strong annual periodicity, however they might contain also local variability, trends and deviations from regular seasonality. It makes the problem of forecasting not trivial and the best suitable forecasting method might differ depending on location, variable and forecast horizon [1, 2].

Classic statistical models are still very important for time series forecasting. ARIMA and SARIMA models are well-known classical tools to model autocorrelation, trend and seasonality based on the Box-Jenkins methodology

*Correspondence to: ziadoon Mohand Khaleel (Email: ziadoon.khaleel@uomosul.edu.iq). Department of Petroleum Reservoir Eng., University of Mosul. Mosul, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

[3]-[6]. Seasonal statistical models usually perform quite good for monthly meteorological series with strong annual cycle, especially in case of linear and seasonal structure. Exponential Smoothing (ETS) models are still competitive for seasonal forecasting due to their ability to model changing level, trend and seasonality with relatively low computational costs [4, 7, 8]. Consequently, in any comparison of machine learning and hybrid approaches, SARIMA and ETS should be considered not just as weak benchmark models, but as solid competition.

At the same time, many climatic and environmental series might have nonlinear or time-varying structure which could not be modeled with statistical linear models. Deep learning models, in particular, Long Short-Term Memory (LSTM) models have been widely used for complex dependencies modeling in weather and temperature forecasting [10, 11, 15, 16]. However, the applicability of LSTM based models depends greatly on amount of data, presence of seasonality in the model, proper model tuning and stability of the process. For monthly temperature series with small samples size, LSTM models tend to overfit or cannot beat well-specified seasonal statistical baselines [1, 10, 11, 12, 13].

Hybrid forecasting models combine the advantages of statistical and neural approaches. One of the approaches is to use ARIMA or SARIMA to model linear and seasonal structure of the series and then to complement the model by ANN or LSTM part to deal with nonlinear residuals [12]-[14]. Another approach is the usage of reverse hybrid models like LSTM→ARIMA or LSTM→SARIMA, where recurrent model captures the nonlinear dynamics first and then the statistical model corrects the residuals [14]-[16]. Hybrids are beneficial when different components of the series are captured well by different assumptions. However, hybridization does not always lead to better forecasting performance, especially in case when the seasonal model already captures most of the predictable variability.

Forecast combinations and ensemble learning is another direction of investigation. Equal-weight averaging, stacking and mixture-of-experts models try to address the deficiencies of models through multiple model averaging [17]-[20]. Mixture-of-experts models combine several experts in one forecast through the gating mechanism, which assigns weights to the experts depending on the input. In theory, this approach allows the gating mechanism to adapt to the situation depending on the time point, forecast horizon and statistical regime. However, the practical value of the gating mechanism should be verified against strong baselines and simple combination rules [1, 2, 17, 18, 19, 20]. If the gate does not perform better than equal-weight averaging or the best statistical model, the additional complexity is not justified.

This problem becomes especially important for monthly temperature forecasting in Nineveh Governorate, Iraq. As the region is semiarid and sensitive to seasonality of the temperature, there is an open question whether adaptive model combination can bring any benefits beyond the robust seasonality-based baselines. Consequently, the current research does not attempt to demonstrate the superiority of gated or hybrid models in general. Instead, the goal is the evaluation of sequential expert weighting for monthly maximum (MAX) and minimum (MIN) temperature forecasting in Al-Hamdaniah, Baaj, Mosul and Rabiaa.

In the revised framework, the problem of forecasting is defined as the problem of expert combination where three major experts are SARIMA, chosen with AIC criterion, ETS and reverse hybrid LSTM→SARIMA models. Two adaptive combination strategies are investigated: the direct prequential weighting and the neural gating ensemble. Contrary to soft-oracle gate training based on current-origin forecast errors, in the revised framework the expert weights are estimated based on the past losses of the rolling origin. Thus, at each forecasting origin the adaptive mechanism is based only on the past information. It helps to avoid the temporal leakage and better reflects the reality of the out-of-sample forecasting [22]-[27].

The major contributions of the study are the following. First, it gives a leakage-aware sequential framework for monthly temperature forecasting with the chronological train/validation/test split and multi-origin rolling forecasts over 18-month period. Second, it evaluates adaptive expert weighting with robust and relevant experts, in particular, AIC-SARIMA, ETS and LSTM→SARIMA, instead of weak or artificially constrained baselines. Third, it reevaluates seasonal-naïve benchmarks and uses both restricted and selected SARIMA models to understand the impact of model identification. Fourth, it performs the tests of predictive accuracy and hyperparameter sensitivity analysis to understand the robustness of the adaptive weighting. Fifth, it gives an empirical evaluation of the limits of adaptive ensembling in the strongly seasonal regional temperature series.

The results of the study indicate that the major factor of the studied series is the seasonality. As a result, SARIMA remains the best practical model for MAX temperature, while ETS is the best or highly competitive model for MIN temperature. The adaptive weighting methods give the conditional and limited improvement over simple combination in some cases, but they cannot consistently beat the best individual statistical baselines. Consequently, the value of the framework developed is not the demonstration of universal superiority of the gated ensembling, but the transparent and leakage-aware evaluation of when the combination is beneficial and when the individual seasonal statistical models should be used.

The rest of the paper is organized as follows. Section 2 discusses the literature review on the topic. Section 3 describes the methodology of the study including preprocessing, expert models, prequential weighting, gating, and evaluation procedures. Section 4 presents the results and statistical comparison. Section 5 discusses the implications and limitations of the results. Section 6 concludes the paper.

2. Literature Review

2.1. Classical Seasonal Statistical Models for Climatic Time Series

Classical statistical models are still in use in climatic time series forecasting due to their clear description of trend, autocorrelation, and seasonality and well-developed techniques of model identification, estimation, and diagnostics [3]-[8]. Of these, the ARIMA and SARIMA models are particularly relevant to monthly meteorological time series because they account for non-seasonal and seasonal dependence in autoregressive, moving average and differencing terms. Also, the Box-Jenkins methodology implies stationarity assessment, seasonal differencing, ACF/PACF inspection, and residual diagnostics including, among others, Ljung-Box test [3]-[6].

In the context of monthly temperature forecasting, SARIMA models often act as strong practical benchmarks because temperature series usually contain an evident annual cycle. With the data generating process being mostly seasonal and nearly linear, a correctly specified SARIMA model is hard to beat even with advanced machine learning techniques. This is an important point since statistical models need to be treated as strong competitors in seasonal climate forecasting rather than weak baselines.

The Exponential Smoothing models with ETS formulations belong to another important class of classical forecasting models. They describe time series with the help of evolving level, trend, and seasonal components and are commonly used if the aim is to produce robust forecasts with minimum computational effort [4, 7, 8]. In climatic series with relatively smooth seasonal behavior, ETS can compete with SARIMA model, especially if gradual changes in level or seasonal amplitude are present. Hence, both SARIMA and ETS provide important baselines for evaluating whether hybrid and adaptive models offer any additional value.

Despite their strengths, classical statistical models may face difficulties when dealing with nonlinearity, time-varying dependence, structural instabilities, and residual patterns not accounted for by linear seasonality [9, 12, 13]. In such cases, nonlinear and hybrid models may be used, although their applicability should be shown against properly specified statistical baselines and not assumed a priori.

2.2. Neural and Deep Learning Models

Various types of neural network-based forecasting models were developed for learning complex nonlinear relationships and temporal dependencies from the data. LSTM networks are often used among these because they are designed to learn sequential dependencies and avoid the vanishing gradient problem faced by regular recurrent neural networks [10]. LSTM-based models were used in weather forecasting, air-temperature prediction, and other applications related to time series in environmental field, especially in the cases of expected nonlinearity [1, 13, 14, 15, 16].

It should be kept in mind, however, that the forecasting accuracy of deep learning models largely depends on the sample size, network architecture, input representation, learning procedure, regularization, and other features of time series [1, 10, 11, 12, 13]. In the context of monthly climatic applications, the number of observations may be relatively small compared to high-frequency data series and hence lead to possible overfitting. Also, deep learning models might find it difficult to account for strong annual seasonality without proper learning or special design.

That is why the assumption that LSTM models will always outperform classical statistical ones in monthly temperature forecasting is not justified. Instead, their role should be seen as complementary: they may be used to capture nonlinearity that is missed by linear seasonal models, although they should be evaluated via out-of-sample criteria and compared to properly specified statistical models.

2.3. Hybrid Forecasting Models

Hybrid models try to combine the strengths of statistical and machine learning models. One common type of hybrid models is the residual hybrid model, which uses ARIMA or SARIMA model to model the linear and seasonal component of the series and then applies ANN or LSTM model to the residual component to capture remaining nonlinearity [12]-[14]. This approach seems to be attractive because of separating the forecasting process into seasonal-linear and nonlinear parts.

Residual hybrids can work efficiently when the statistical model accounts for the main seasonal pattern but there is some systematic structure in the residual part. Nevertheless, their efficiency depends on how much valuable residual information is contained in the model. When the statistical model already accounts for the most of predictable variability in the series, the neural residual model may add no additional value and even increase noise and instability.

Another design of hybrid models is the reverse hybrid model like LSTM→ARIMA or LSTM→SARIMA, which uses recurrent neural network to model the main pattern in time series and then fits the statistical model to the residuals produced by the neural model [14]-[16]. This approach may be useful if nonlinear dependence is supposed to dominate the process, but some linear or seasonal pattern in the residuals remains. As above, however, hybrid models should be evaluated empirically rather than be assumed better than others. The order of components in hybrid models is therefore important and reflects some assumption about whether the dominating structure is linear-seasonal, nonlinear, or mixed.

2.4. Forecast Combination, Mixture of Experts, and Gating

Forecast combinations are used to improve robustness by aggregating the forecasts of several models. The underlying idea of forecast combinations is that various forecasting methods may catch different aspects of the data and their forecasting errors may partially compensate each other [17]-[21]. The simplest way of forecast combination is equal-weight averaging, which acts as a strong benchmark because of its simplicity, transparency, and stability to overfitting [1, 2, 31, 32]. More flexible approaches like stacking and meta-learning estimate the combination rule from data by training a second-level model based on the outputs of base forecasters [19, 20].

One of the extensions of this idea is the use of mixture-of-experts models, which employs the gating mechanism for assigning observation- or horizon-specific weights to experts [17, 18]. From the forecasting perspective, such gate may in theory adapt weights of experts according to the characteristics of seasonality, persistence, volatility, trend, stationarity, or forecast horizon. That is why this approach is appealing if the model suitability changes in time and space.

However, the practical benefits of the gating mechanism depend on whether the gate helps to improve the result. If a gate behaves similarly to equal-weight averaging or cannot outperform the best individual model, the additional complexity of the approach becomes unjustified. Moreover, the gate trained with the monthly data may easily overfit. For this reason, the gate approach should be compared not only to individual experts but also to equal-weight averaging and best statistical baselines like SARIMA and ETS.

2.5. Leakage-Free and Rolling-Origin Evaluation

Proper evaluation of models is crucial in time series forecasting because temporal dependence makes random sampling inappropriate. The forecasting models should be evaluated by using chronological split of data to training, validation, and testing sets and using rolling-origin or expanding window approach [22]-[24]. In this case, all the preprocessing, modeling, feature computation, and hyperparameter selection are made based on information available up to the particular forecast origin.

This point is especially important for adaptive combination and gated ensembles. The gate model trained on forecast errors of the same origin or horizon as the predicted by the gate introduces implicit forward-looking information. Although the test set is not used in this case, the use of errors of the same origin or horizon to train the gate is not realistic because, in practice, the future value is not known at the forecast time. Thus, the correct approach would be to train the gate model on past rolling-origin errors and then use it for future origins.

Leakage-free evaluation of models is thus a stricter and more realistic test of the adaptive forecasting models. Also, it avoids excessive claims of adaptivity caused by the use of forecast errors. Hence, any regime-aware or gated forecasting model should be evaluated by a chronological and prequential protocol.

2.6. Synthesis and Research Gap

The literature review yields the following four conclusions. First, the classical seasonal statistical models, especially SARIMA and ETS, can serve as powerful benchmarks for monthly temperature forecasting in the case of dominating annual seasonality [3]-[8]. Second, neural and hybrid models can capture the nonlinearity but their performance strongly depends on the sample size, model design, and seasonal pattern in the data [10]-[16]. Third, forecast combinations and mixture-of-experts frameworks can increase the robustness, but their benefits should be tested against equal-weight averaging and best individual models [17]-[21]. Fourth, the adaptive weighting models should be trained and evaluated in a leakage-free way using only available information at the forecast origin [22]-[27].

These conclusions motivate the current research. Most of the existing literature on hybrid and ensemble forecasting focuses on whether the proposed architecture is able to outperform individual models, although not enough attention is paid to whether the adaptive part is really useful in the presence of strong seasonal benchmarks and leakage-free evaluation. In the context of monthly temperature forecasting in semi-arid regions, this problem is important because the main annual cycle is likely to be well-captured by SARIMA or ETS.

In the local scale, recent modeling works in Nineveh Governorate show that the statistical and machine learning approaches can be used in regional risk assessment in semi-arid areas [33]. However, the problem of leakage-free monthly MAX and MIN temperature forecasting with strong seasonal benchmarks and sequential expert weighting was not sufficiently addressed.

Hence, the current paper is aimed not at another claim of superiority of the hybrid models, but at the controlled evaluation of adaptive expert combination for monthly temperature forecasting in Nineveh Governorate. In the joint use of the AIC-selected SARIMA, ETS, and LSTM→SARIMA models and their comparison to the direct prequential weighting and gated ensemble model, the paper answers the following practical question: with the availability of strong seasonal statistical models, does the adaptive expert weighting provide any additional value for monthly temperature forecasting?

3. Methodology

3.1. Data Description and Forecasting Problem

The study considers maximum (MAX) and minimum (MIN) monthly temperature forecasting for four locations in Nineveh Governorate, Iraq: Al-Hamdaniah, Baaj, Mosul, and Rabiaa. Monthly temperature data were provided by the Iraqi General Authority for Meteorology. Series for each site-variable combination were analyzed independently in order to maintain local climatic specifics and prevent pooling of heterogeneous temperature behavior across stations.

Consider the monthly temperature series $y_1, y_2, \ldots, y_T$ at a given site and variable. At a forecast origin t, the task is to produce multi-step ahead forecasts (see Eq.1)

$$\hat{y}_{t+h|t}, \qquad h = 1, 2, \ldots, H, \tag{1}$$

where $H = 18$ months is the largest forecast horizon. Seasonal period m was set to 12, since we consider monthly temperature series. Lag period was set to $L = 12$ for recurrent and lagged components.

Adaptive combination was conceived as a controlled experiment on expert weighting rather than a proof of supremacy of a particular hybrid model. The basic set of experts included AIC-selected SARIMA, ETS, and LSTM→SARIMA. Next, two adaptive combination approaches were evaluated: prequential expert-weighting and neural gated ensemble.

3.2. Chronological Splitting and Rolling-Origin Design

As random splitting is invalid for time-series forecasting, all experiments used chronological splitting and rolling-origin evaluation [22]-[24]. Every time series was divided into three segments (see Eq.2) :

$$\mathcal{D}_{train} = \{y_1, \ldots, y_{T_{train}}\}, \quad \mathcal{D}_{val} = \{y_{T_{train}+1}, \ldots, y_{T_{val}}\}, \quad \mathcal{D}_{test} = \{y_{T_{val}+1}, \ldots, y_T\}. \tag{2}$$

The validation period consisted of 120 months and was used to train the neural gate and build prequential expert weighting information. The test period consisted of 60 months and was used only for evaluation purposes. Rolling origins were advanced with an annual step of 12 months. With a test length of 60 months and a forecasting horizon of 18 months, four test origins were available for every site-variable combination. Eighteen forecasts for every origin were produced; therefore, 72 test forecasts were produced for each site-variable combination, and for eight site-variable combinations 576 test forecast points altogether were generated.

3.3. Leakage-Free Preprocessing

All preprocessing steps were done in a leakage-free way. Normalization parameters were estimated on the training segment alone (see E.q.3,4):

$$\mu_{train} = \frac{1}{T_{train}} \sum_{t=1}^{T_{train}} y_t, \tag{3}$$

$$\sigma_{train} = \sqrt{\frac{1}{T_{train} - 1} \sum_{t=1}^{T_{train}} (y_t - \mu_{train})^2} \tag{4}$$

The standardized series was then computed as (see E.q.5)

$$z_t = \frac{y_t - \mu_{train}}{\sigma_{train}} \tag{5}$$

The same normalization parameters were used to standardize the validation and test periods without any modifications. Forecasts were computed in the standardized domain and transformed back to the original scale using (see E.q6)

$$\hat{y}_{t+h|t} = \mu_{train} + \sigma_{train} \hat{z}_{t+h|t} \tag{6}$$

For every rolling origin, all features, expert forecasts, expert losses, and adaptive weights were calculated using information up to that origin only. No future observations from the validation or test horizon were used to produce forecasts or calculate expert weights.

3.4. Overall Workflow of the Revised Framework

Figure (1) provides the scheme of the revised forecasting framework. The workflow starts with chronological splitting of the data and train-only preprocessing. At every rolling origin, the three experts generate 18-month forecasts. The previous losses of the experts at rolling origins are used to build the prequential weights and to train or apply the neural gate. Finally, forecasts are evaluated using accuracy metrics, predictive accuracy testing, regime summarizing, gating weights, and sensitivity analysis.

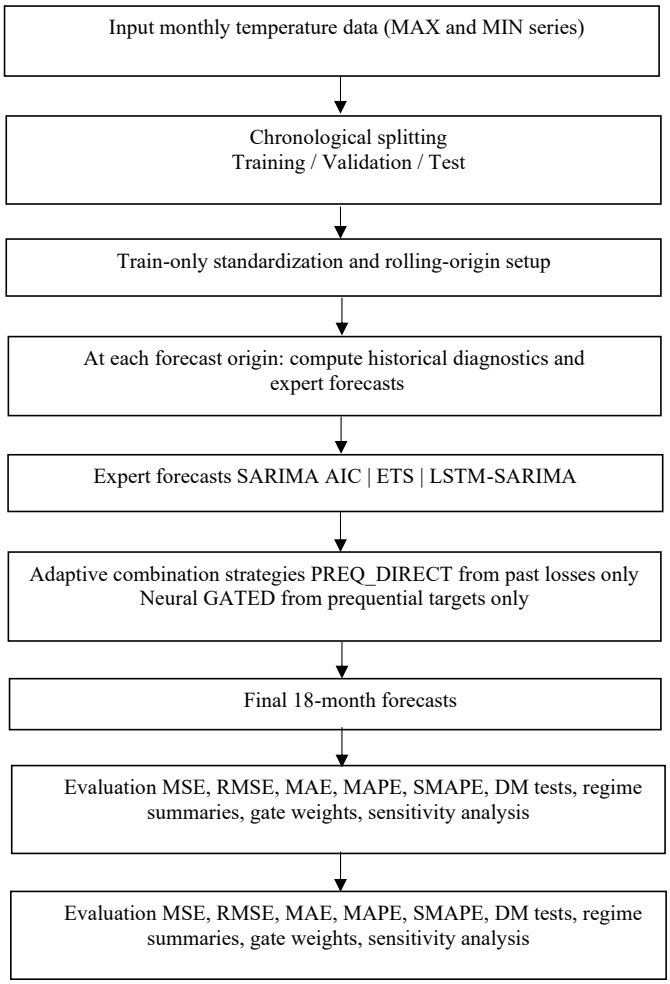

Figure 1. Revised leakage-aware workflow of the forecasting framework.

3.5. Forecasting Experts

3.5.1. AIC-Selected SARIMA Expert

The first expert is an AIC-selected seasonal autoregressive integrated moving-average model, SARIMA_AIC. For every site-variable combination and rolling origin, SARIMA models were fitted within a given range of orders and the model with minimal Akaike Information Criterion was chosen. The Akaike Information Criterion is calculated as (see E.q.7)

$$AIC = 2k - 2\ln\left(\hat{L}\right) \tag{7}$$

where k is the number of estimated parameters and L is the maximized likelihood.

The general SARIMA model can be represented as (see E.q.8)

$$\Phi_P\left(B^m\right)\phi_p\left(B\right)\left(1-B\right)^d\left(1-B^m\right)^D y_t = \Theta_Q\left(B^m\right)\theta_q\left(B\right)\varepsilon_t \tag{8}$$

where B is the backshift operator, $m = 12$ is the seasonal period, p, d, q are the nonseasonal orders, and P, D, Q are the seasonal orders [3]-[6]. The AIC-selected SARIMA expert was included to address the limitation of using a single fixed SARIMA order for all sites and variables [5].

For the sake of transparency, the fixed restricted model $(1,1,1)\,(1,1,1)_{12}$ was also kept as a separate baseline, denoted SARIMA_FIXED_RESTRICTED. This expert was not considered as the main SARIMA expert but as a statistical benchmark.

3.5.2. ETS Expert

The second expert is the exponential smoothing state-space model, ETS. ETS models describe a time series using a state vector consisting of level, trend, and seasonality terms [4, 7, 8]. In this study, additive seasonality specification was used if the historical sample was sufficiently large. The additive ETS model can be written as (see E.q 9)

$$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t \tag{9}$$

where ℓ_t is the level component, b_t is the trend component, s_t is the seasonal component, $m = 12$, and ε_t is the error term.

The ETS expert was introduced to capture the smooth seasonal pattern, which is commonly present in monthly temperature series and can be modeled with level, trend, and seasonal smoothing. If fitting of ETS failed or there were not enough data to estimate the parameters of ETS, a seasonal naïve forecast was made.

3.5.3. Reverse Hybrid LSTM→SARIMA Expert

The third expert is a reverse hybrid model in the form of LSTM→SARIMA. In the model, an LSTM network produces forecasts for the standardized temperature series. Then, the residuals are calculated, and a SARIMA model is fitted to the residual series and its forecast is added to the LSTM forecast.

Let $\hat{z}^{LSTM}_{t+h|t}$ denote the standardized LSTM forecast. The one-step residual sequence is defined as (see E.q 10)

$$r_t = z_t - \hat{z}^{LSTM}_{t|t-1} \tag{10}$$

A seasonal statistical model is then fitted to r_t, producing the residual forecast $\hat{r}_{t+h|t}$. The reverse hybrid forecast is (see E.q 11)

$$\hat{z}^{LSTM\to SARIMA}_{t+h|t} = \hat{z}^{LSTM}_{t+h|t} + \hat{r}_{t+h|t} \tag{11}$$

If the residual series was too short for reliable SARIMA fitting, the expert was simplified to the LSTM forecast. The LSTM→SARIMA expert was used to check if the proposed nonlinear-first modeling combined with subsequent residual seasonality correction can improve upon the predictions of SARIMA and ETS. The motivation behind the LSTM component is sequence learning [10, 11], whereas the reverse hybrid approach is related to the hybridization of ARIMA and neural forecasting [12]-[16].

3.6. Recurrent Forecasting Component

A fixed architecture was used for all site-variable combinations. Input sequences were created using 12-month look-back window (see E.q 12):

$$\mathbf{x}_t = (z_{t-L}, z_{t-L+1}, \ldots, z_{t-1}), \qquad L = 12 \tag{12}$$

The target is z_t. The model included two LSTM layers with 32 hidden units, dropout regularization after LSTM layers and a dense output layer. The model was trained using the mean squared error loss and the Adam optimizer [29] with the dropout regularization for avoiding overfitting [30].

Multistep forecasts were produced recursively. First, the one-step forecast was made, and the prediction was added to the input window to make the next step forecast until all $H = 18$ steps were completed. The LSTM was evaluated both as the stand-alone baseline and as a component of the LSTM→SARIMA expert.

3.7. Baseline and Reference Models

Several baselines and reference models were included for the fair comparison. First, SARIMA_FIXED_RESTRICTED was kept as the statistical benchmark. Second, stand-alone LSTM was used to evaluate the LSTM component without residual correction. Third, the corrected SEASONAL_NAIVE forecast was implemented as (see E.q 13)

$$\hat{y}^{SN}_{t+h|t} = y_{t+h-m} \tag{13}$$

where $m = 12$. For multi-step horizons larger than one seasonal cycle, the corresponding month of the last seasonal cycle was replicated accurately. Moreover, the SEASONAL_NAIVE drift was considered.

Fourth, the equal-weight ensemble was defined as (see E.q 14)

$$\hat{y}^{AVG}_{t+h|t} = \frac{1}{K}\sum_{k=1}^{K} \hat{y}^{(k)}_{t+h|t} \tag{14}$$

where $K = 3$ and the three experts are SARIMA_AIC, ETS, and LSTM→SARIMA.

Finally, the ORACLE performance was reported as an infeasible upper bound. It selects the best expert ex-post at every horizon (see E.q 15):

$$\hat{y}^{ORACLE}_{t+h|t} = \hat{y}^{(k^*)}_{t+h|t} \tag{15}$$

where :

$$k^* = \arg\min_{k} \left| y_{t+h} - \hat{y}^{(k)}_{t+h|t} \right| \tag{16}$$

The ORACLE model was not used as the deployable forecasting method. Its purpose was to show the maximum possible benefit of perfect expert selection.

3.8. Diagnostic and Regime Features

At every forecast origin, a set of diagnostic features were computed based on historical series only. The maximum diagnostic look-back period was 120 months. These features captured the properties of persistence, seasonality, variability, trend, stationarity, and residual dependence.

The diagnostic feature vector consisted of the lag-1 autocorrelation, seasonal autocorrelation at lag 12, standard deviation of the level series, standard deviation of first differences, ratio of differenced and level variabilities, linear trend slope, Augmented Dickey-Fuller p-value, Ljung-Box p-value, and the seasonal naïve error statistics. The Augmented Dickey-Fuller test was used as the stationarity diagnostic [28]; the Ljung-Box test was used to capture residual autocorrelation [6].

For a given forecast origin t and horizon h, the diagnostic feature vector is denoted as (see E.q 17)

$$\mathbf{x}_{t,h} = [f_{1,t}, f_{2,t}, \ldots, f_{q,t}, h^*] \tag{17}$$

where $f_{1,t}, \ldots, f_{q,t}$ are historical diagnostic features and

$$h^* = \frac{h-1}{H-1} \tag{18}$$

is the normalized forecast-horizon indicator.

To get the intuitive meaning, every origin was labeled by broad diagnostics. Origins with high seasonality autocorrelation and stationary differenced variability were labeled as seasonality-dominant. Origins with weaker stationarity and higher variability were labeled as volatile/non-linear. Other cases were labeled as mixed. These labels were used for interpretation only; numerical diagnostic features were used in gating.

3.9. Direct Prequential Expert Weighting

The first adaptive combination approach is direct prequential weighting, PREQ_DIRECT. The method produces expert weights only based on previous rolling-origin losses. The future error of the current-origin forecast is not used for producing weights.

Let

$$e_{\tau,h}^{(k)} = \left| y_{\tau+h} - \hat{y}_{\tau+h|\tau}^{(k)} \right| \tag{19}$$

denote the absolute error of expert k at a previous origin $\tau < t$ and horizon h. The past average loss for expert k is (see E.q 20)

$$\bar{L}_{k,t,h}^{past} = \frac{1}{|\mathcal{P}_t|} \sum_{\tau \in \mathcal{P}_t} e_{\tau,h}^{(k)} \tag{20}$$

where $\mathcal{P}_t$ is the set of previous rolling origins available before origin t. Expert weights are then computed using an inverse-loss softmax rule (see E.q 21):

$$w_{k,t,h}^{PREQ} = \frac{\exp\left(-\bar{L}_{k,t,h}^{past}/\tau_g\right)}{\sum_{j=1}^{K} \exp\left(-\bar{L}_{j,t,h}^{past}/\tau_g\right)} \tag{21}$$

where τ_g is the temperature parameter controlling the sharpness of the weighting distribution. The final prequential forecast is (see E.q 22)

$$\hat{y}_{t+h|t}^{PREQ} = \sum_{k=1}^{K} w_{k,t,h}^{PREQ} \hat{y}_{t+h|t}^{(k)} \tag{22}$$

This method is straightforward, sequential, and fully deployable because it uses only losses that can be estimated prior to making the forecast at origin t. It also provides a strong benchmark for evaluating the added value of the neural gate.

3.10. Neural Gated Ensemble

The second adaptive combination method is neural gated ensemble, GATED. The gate gets the diagnostic feature vector and the vector of expert forecasts as inputs. Then, it produces horizon-dependent non-negative weights, which sum to one (see E.q 23):

$$\mathbf{w}_{t,h}^{G} = G_\theta\left(\mathbf{x}_{t,h}, \hat{\mathbf{y}}_{t,h}\right) \tag{23}$$

where :

$$\hat{\mathbf{y}}_{t,h} = \left[\hat{y}_{t+h|t}^{(1)}, \hat{y}_{t+h|t}^{(2)}, \ldots, \hat{y}_{t+h|t}^{(K)}\right] \tag{24}$$

The weights satisfy

$$w_{k,t,h}^{G} \geq 0, \qquad \sum_{k=1}^{K} w_{k,t,h}^{G} = 1 \tag{25}$$

The final gated forecast is:

$$\hat{y}_{t+h|t}^{GATED} = \sum_{k=1}^{K} w_{k,t,h}^{G} \hat{y}_{t+h|t}^{(k)} \tag{26}$$

The gate was implemented as the feedforward neural network with two hidden layers, dropout regularization, and softmax output layer. The use of gating network comes from the mixture-of-experts approach [17, 18]. However, in this study, the gate is evaluated carefully against equal averaging, PREQ_DIRECT, SARIMA, and ETS.

3.11. Prequential Gate Training Without Forward-Looking Targets

The original soft-oracle gate training design was replaced by the prequential training procedure. In the new design, the neural gate does not learn from the current-origin future errors. Instead, prequential targets are built using the past expert losses only. For every validation origin t, the prequential target vector is calculated as (see E.q 27)

$$q_{k,t,h} = \frac{\exp\left(-\bar{L}^{past}_{k,t,h}/\tau_g\right)}{\sum_{j=1}^{K}\exp\left(-\bar{L}^{past}_{j,t,h}/\tau_g\right)} \tag{27}$$

Then, the gate is trained to approximate these prequential targets using the cross-entropy loss with the mild entropy regularization (see E.q 28):

$$\mathcal{L}_{gate} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} q_{k,i}\log\left(w^{G}_{k,i}+\epsilon\right) - \lambda\frac{1}{N}\sum_{i=1}^{N}\left(-\sum_{k=1}^{K} w^{G}_{k,i}\log\left(w^{G}_{k,i}+\epsilon\right)\right) \tag{28}$$

where N is the number of gate-training samples, λ is the entropy regularization coefficient, and ϵ is a small numerical constant.

This procedure implies the following sequential evaluation logic: after the true value becomes known, its losses can be added to the loss record for the use in later origins, but not for the same forecast origin. Hence, the gate is trained using the information conditions closer to the real-world forecasting.

3.12. Hyperparameter Sensitivity Analysis

A sensitivity analysis was conducted for the main adaptive weighting hyperparameters. The temperature parameter τ_g was varied over a predefined grid. Also, the entropy regularization coefficient λ was varied. The goal of this analysis was not to tune the model on the test data, but to test stability of the conclusions under realistic changes of adaptive weighting behavior.

The sensitivity analysis was reported separately from the main results. The distinction is important because the direct selection of hyperparameters using the test set performance can cause data snooping. The main results should be interpreted together with, but not replaced by, the findings of the sensitivity analysis.

3.13. Forecast Accuracy Metrics

Forecast accuracy was evaluated using the mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and symmetric mean absolute percentage error (sMAPE). These metrics are defined as follows [27]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{29}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{30}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{31}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \tag{32}$$

$$sMAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{2\,|y_i-\hat{y}_i|}{|y_i|+|\hat{y}_i|} \tag{33}$$

MAE and RMSE were emphasized because they are directly interpretable in degrees Celsius. MAPE and sMAPE were reported as additional measures, but interpreted cautiously, since the percentage errors are sensitive to small denominators.

3.14. Predictive Accuracy Testing

Pairwise predictive accuracy testing was done using the Diebold-Mariano test [25]. The test was applied to the loss differentials based on the absolute-error loss. For two competing forecasts A and B, the loss differential is calculated as (see E.q 34)

$$d_i = \left|y_i - \hat{y}_i^A\right| - \left|y_i - \hat{y}_i^B\right| \tag{34}$$

The null hypothesis is :

$$H_0 : E\left(d_i\right) = 0 \tag{35}$$

which means that the two forecasts have equal predictive accuracy. The alternative hypothesis is that the expected losses are different. To account for dependence of multi-step forecast errors, heteroscedasticity and autocorrelation consistent variance estimates were used. Harvey-Leybourne-Newbold small-sample correction was also considered [26].

In addition to the Diebold-Mariano test, nonparametric paired tests, such as the Wilcoxon signed-rank test with sign-test fallback, were reported as robustness checks. The main comparisons were focused on whether PREQ_DIRECT and GATED outperformed AVG_ENSEMBLE and whether they were able to surpass the strongest statistical baselines, especially SARIMA_AIC and ETS.

3.15. Reproducibility and Output Artifacts

All experiments were run in Python using chronological splits, fixed random seeds, and saved configuration files. The code produced the required output tables, including test scores, mean performance per series, selected SARIMA orders, statistical tests, regime summaries, gating weights, hard expert selections, prequential targets, and sensitivity results. This setup supports reproducible and leakage-free evaluation of adaptive expert combinations for monthly temperature forecasting.

4. Experiments and Results

4.1. Experimental Setup

The revised experiments were conducted for monthly maximum (MAX) and minimum (MIN) temperature series at four sites in Nineveh Governorate: Al-Hamdaniah, Baaj, Mosul, and Rabiaa. Each site-series was modeled independently using the chronological splitting and rolling-origin protocol described in Section 3. The validation segment contained 120 months and was used for prequential expert-weight construction and gate training, while the final test segment contained 60 months and was reserved for out-of-sample evaluation only.

Forecasts were generated over an 18-month horizon using a rolling-origin protocol with an annual stride of 12 months. This resulted in four test origins per site-series and 72 forecast points per site-series. Across the eight site-series combinations, the final test evaluation therefore included 576 forecast points. This evaluation design follows the principle that time-series forecasting models should be tested using chronological and rolling-origin procedures rather than random splitting [22]-[24].

The main expert set consisted of SARIMA_AIC, ETS, and LSTM→SARIMA. In addition, SARIMA_FIXED_RESTRICTED, stand-alone LSTM, corrected SEASONAL_NAIVE, SEASONAL_NAIVE_DRIFT, AVG_ENSEMBLE, PREQ_DIRECT, GATED, BASE_AVG, and ORACLE were evaluated. SARIMA_FIXED_RESTRICTED corresponds to the fixed SARIMA$(1,1,1)\,(1,1,1)_{12}$ model and was retained as a restricted statistical benchmark. BASE_AVG was retained only as an ablation component and not as part of the final expert pool. ORACLE is an infeasible reference that selects the best expert ex post at each horizon and is used only to indicate the possible upper bound of expert selection. $(1,1,1)\,(1,1,1)_{12}$

All accuracy values were computed after transforming forecasts back to the original temperature scale. The main evaluation metrics were MSE, RMSE, MAE, MAPE, and sMAPE, with MAE and RMSE emphasized because they are directly interpretable in degrees Celsius [27].

4.2. Average Accuracy Across Sites

Table 1 presents the average test results for MAX and MIN temperature series across the four sites.

For MAX temperature, SARIMA_FIXED_RESTRICTED delivered the lowest mean MAE of $1.473\ ^\circ\mathrm{C}$, followed closely by SARIMA_AIC with an MAE of $1.489\ ^\circ\mathrm{C}$. Among the deployable combined models, PREQ_DIRECT ranked next with an MAE of $1.608\ ^\circ\mathrm{C}$, followed by AVG_ENSEMBLE with an MAE of $1.629\ ^\circ\mathrm{C}$. The neural GATED model produced an MAE of $1.695\ ^\circ\mathrm{C}$. It was better than ETS, LSTM, LSTM→SARIMA, and the corrected seasonal-naive baselines, but it was not better than SARIMA.

For MIN temperature, ETS showed the best average MAE of $0.995\ ^\circ\mathrm{C}$. SARIMA_FIXED_RESTRICTED and SARIMA_AIC also performed well, with MAE values of $0.999\ ^\circ\mathrm{C}$ and $1.010\ ^\circ\mathrm{C}$, respectively. PREQ_DIRECT and GATED yielded MAE values of $1.085\ ^\circ\mathrm{C}$ and $1.104\ ^\circ\mathrm{C}$, respectively. These values were slightly lower than AVG_ENSEMBLE but still higher than ETS and SARIMA. Therefore, adaptive weighting improved some combination baselines but did not replace the best statistical models.

Table 1. Average test performance across sites

Series	Model	MSE	RMSE	MAE	MAPE	sMAPE
MAX	Oracle	2.312	1.485	1.074	4.544	4.489
MAX	Fixed	3.476	1.841	1.473	6.268	6.137
MAX	AIC	3.544	1.856	1.489	6.131	6.129
MAX	PREQ	4.269	2.030	1.608	6.916	6.736
MAX	AVG	4.422	2.070	1.629	7.036	6.834
MAX	Gate	5.212	2.206	1.695	7.415	7.095
MAX	ETS	4.781	2.135	1.744	7.357	7.161
MAX	SN	6.626	2.526	2.036	8.698	8.492
MAX	SN-Drift	6.809	2.560	2.056	8.798	8.560
MAX	L-S	8.157	2.782	2.096	9.159	8.615
MAX	LSTM	8.902	2.921	2.365	9.131	9.132
MAX	Base Avg	74.110	8.600	7.186	37.070	27.597
MIN	Oracle	1.014	0.997	0.717	6.608	6.542
MIN	ETS	1.648	1.277	0.995	9.406	9.045
MIN	Fixed	1.717	1.300	0.999	9.302	8.881
MIN	AIC	1.737	1.308	1.010	9.342	9.232
MIN	PREQ	1.844	1.346	1.085	10.034	10.012
MIN	Gate	1.901	1.365	1.104	10.212	10.213
MIN	AVG	1.921	1.373	1.109	10.272	10.272
MIN	SN	3.252	1.782	1.421	13.402	13.221
MIN	L-S	3.543	1.844	1.488	13.978	14.894
MIN	LSTM	4.314	2.019	1.667	14.934	16.898
MIN	SN-Drift	4.816	2.170	1.806	17.259	16.734
MIN	Base Avg	39.005	6.148	5.324	68.565	40.608

Notes: Fixed = SARIMA_FIXED_RESTRICTED; AIC = SARIMA_AIC; PREQ = PREQ_DIRECT; AVG = AVG_ENSEMBLE; Gate = GATED; SN = SEASONAL_NAIVE; L-S = LSTM→SARIMA.

The ORACLE result shows that there is still a potential gap between the feasible models and ideal expert selection. However, the adaptive methods did not close this gap enough to outperform SARIMA or ETS. This

means that having several experts is not sufficient by itself; adaptive weighting remains challenging when one or two seasonal statistical models are already highly competitive.

4.3. Site-Level Results

Table (2a,2b) provides the MAE values of the main models at the site level. The best deployable model depended on the site and variable, but the general pattern was still clear. SARIMA_FIXED_RESTRICTED was the best model for MAX temperature at Al-Hamdaniah, Baaj, and Rabiaa, while SARIMA_AIC was the best for Mosul. For MIN temperature, ETS was best for Baaj and Mosul, SARIMA_FIXED_RESTRICTED was best for Al-Hamdaniah, and SARIMA_AIC was best for Rabiaa.

Table 2a. MAX temperature

Site	Fixed	AIC	ETS	L-S	AVG	PREQ	Gate	SN	Oracle
AHD	1.481	1.492	1.788	1.831	1.523	1.516	1.544	2.000	1.063
BAJ	1.665	1.783	2.269	2.787	2.040	2.032	2.274	2.507	1.346
MOS	1.692	1.596	1.729	1.736	1.616	1.609	1.614	2.244	1.196
RAB	1.056	1.084	1.190	2.028	1.339	1.275	1.347	1.392	0.691

Table 2b. Site-level MAE values for MIN temperature series

Site	Fixed	AIC	ETS	L-S	AVG	PREQ	Gate	SN	Oracle
AHD	1.019	1.026	1.088	1.996	1.295	1.226	1.258	1.418	0.847
BAJ	1.148	1.089	1.069	1.369	1.117	1.115	1.109	1.623	0.751
MOS	1.005	1.113	0.935	1.565	1.154	1.139	1.201	1.577	0.702
RAB	0.826	0.814	0.889	1.023	0.869	0.860	0.848	1.067	0.566

Notes: AHD = Al-Hamdaniah, BAJ = Baaj, MOS = Mosul, RAB = Rabiaa. Fixed = SARIMA_FIXED_RESTRICTED; AIC = SARIMA_AIC; L-S = LSTM→SARIMA; AVG = AVG_ENSEMBLE; PREQ = PREQ_DIRECT; Gate = GATED; SN = SEASONAL_NAIVE.

The site-level results confirm that the revised SARIMA and ETS baselines are strong and must be treated as operationally meaningful models. The adaptive methods performed reasonably in some cases, especially relative to AVG_ENSEMBLE, but their gains were not large enough to change the overall model recommendation.

The corrected SEASONAL_NAIVE baseline also produced credible values. Its average MAE was $2.036\ ^\circ\mathrm{C}$ for MAX and $1.421\ ^\circ\mathrm{C}$ for MIN. Although it did not outperform SARIMA or ETS, it provides a useful low-complexity seasonal reference.

4.4. Statistical Comparison of Predictive Accuracy

Pairwise Diebold-Mariano tests were used to compare predictive accuracy using absolute-error loss [25]. The Harvey-Leybourne-Newbold correction and nonparametric paired tests were also considered as robustness checks [26]. Table 3 summarizes the most important comparisons.

The comparison between GATED and AVG_ENSEMBLE showed mixed evidence. GATED had lower MAE than AVG_ENSEMBLE in four of the eight site-series cases. However, the Diebold-Mariano test indicated a statistically significant GATED improvement only for Al-Hamdaniah MIN. In several MAX cases, GATED was similar to or slightly worse than AVG_ENSEMBLE. Therefore, the neural gate cannot be described as consistently superior to equal-weight averaging.

PREQ_DIRECT provided a more stable adaptive result. It achieved lower MAE than AVG_ENSEMBLE in all eight site-series cases. However, the Diebold-Mariano test showed a statistically significant improvement only for Al-Hamdaniah MIN, while the remaining improvements were generally small. This suggests that direct prequential weighting is a useful and straightforward adaptive benchmark, but its practical improvement over equal weighting is limited.

In comparison with the best-performing statistical models, the adaptive approaches did not dominate. For GATED, there was no site-series case where the MAE was lower than SARIMA_AIC. There were three site-series cases where GATED had lower MAE than ETS: Al-Hamdaniah MAX, Mosul MAX, and Rabiaa MIN. Likewise, PREQ_DIRECT did not achieve a lower MAE than SARIMA_AIC in any site-series case.

Table 3. Summary of key predictive-accuracy comparisons.

Comparison	Lower MAE for A	Sig. wins for A	Conclusion
Gate vs AVG	4/8	1/8	Mixed evidence.
PREQ vs AVG	8/8	1/8	Stable but modest gains.
Gate vs AIC	0/8	0/8	No improvement over SARIMA_AIC.
Gate vs ETS	3/8	0/8	Only selected-case improvement.
PREQ vs AIC	0/8	0/8	No improvement over SARIMA_AIC.
AIC vs Fixed	3/8	0/8	Model selection did not uniformly improve.

Notes: Gate = GATED; AVG = AVG_ENSEMBLE; PREQ = PREQ_DIRECT; AIC = SARIMA_AIC; Fixed = SARIMA_FIXED_RESTRICTED; Sig. = statistically significant.

This supports the interpretation that adaptive weighting can help in some simple combination scenarios, but it is not necessarily better than well-specified seasonal statistical methods in this dataset.

4.5. Regime Diagnostics and Gate Behavior

Based on the diagnostic regime classification, all rolling origins in the validation and test sets were seasonality-dominant. Across 112 validation and test rolling origins, no Mixed or Volatile/nonlinear regime was identified by the diagnostic rule used in this study. This is important because it explains the strong performance of SARIMA and ETS and the limited added value of adaptive gating.

The average weights of the neural gate showed that the gate did not collapse into one expert. For MAX temperature, the average gate weights were 0.293 for SARIMA_AIC, 0.365 for ETS, and 0.342 for LSTM→SARIMA. For MIN temperature, the weights were 0.336, 0.387, and 0.277, respectively. Thus, the gate distributed weights among experts and assigned the highest average weight to ETS in both series groups.

Table 4 shows the average gate weights and hard expert selections across all test forecasts. The most frequently selected expert was ETS, followed by LSTM→SARIMA and SARIMA_AIC. However, these results should be interpreted carefully because all origins were classified as seasonality-dominant.

Table 4. Gate behavior across test forecasts.

Series	Expert	Average weight	Count selected
MAX	AIC	0.293	41
MAX	ETS	0.365	161
MAX	L-S	0.342	86
MIN	AIC	0.336	65
MIN	ETS	0.387	172
MIN	L-S	0.277	51
Total	AIC	0.314	106
Total	ETS	0.376	333
Total	L-S	0.310	137

Notes: AIC = SARIMA_AIC; L-S = LSTM→SARIMA.

The prevalence of seasonality-dominant labels indicates that there was not enough statistical evidence for strong regime switching under the current diagnostic criteria. Thus, the revised framework should not be interpreted as proving superiority under multiple regimes. Rather, it provides a way to test whether adaptive expert weighting is valuable under strong seasonality conditions.

4.6. Hyperparameter Sensitivity Analysis

A sensitivity analysis was conducted for the adaptive weighting hyperparameters. The temperature parameter τ_g was varied over $\{0.10, 0.25, 0.50, 1.00\}$, while the entropy regularization coefficient λ was varied over $\{0, 0.01, 0.02, 0.05\}$. The results showed that lower values of τ_g, especially $\tau_g = 0.10$, produced slightly better adaptive-combination performance. The effect of entropy regularization was small$\tau_g \in \{0.10, 0.25, 0.50, 1.00\}$, $\lambda \in \{0, 0.01, 0.02, 0.05\}$, and $\tau_g = 0.10$.

Table 5 reports the best sensitivity results for GATED and PREQ_DIRECT. For MAX temperature, PREQ_DIRECT improved from an MAE of 1.608 under the main setting to 1.590 when $\tau_g = 0.10$. GATED improved from 1.695 to 1.677. For MIN temperature, PREQ_DIRECT improved from 1.085 to 1.063, while GATED improved from 1.104 to 1.093.

Table 5. Best sensitivity results for adaptive methods.

Series	Method	τ_g	λ	MSE	RMSE	MAE	MAPE	sMAPE
MAX	PREQ	0.10	0.00	4.129	1.993	1.590	6.796	6.641
MAX	Gate	0.10	0.00	5.123	2.182	1.677	7.331	7.023
MIN	PREQ	0.10	0.00	1.783	1.325	1.063	9.801	9.757
MIN	Gate	0.10	0.01	1.875	1.357	1.093	10.149	10.087

Notes: PREQ = PREQ_DIRECT; Gate = GATED.

Even though the sensitivity analysis led to small improvements in adaptive approaches, it did not affect the core empirical finding. The best statistical models were still SARIMA for MAX and ETS for MIN temperature. Hence, the conclusions were not based on one particular hyperparameter setting and remained stable.

4.7. Visual Forecast Evaluation

The 18-month forecast paths from the last test origin are shown in Figure 2. The figure illustrates the observed data and the predictions from SARIMA_AIC, ETS, LSTM→SARIMA, AVG_ENSEMBLE, PREQ_DIRECT, GATED, and seasonal baselines. The graphical evaluation is consistent with the numerical results. For most MAX time series, SARIMA followed the observed seasonal pattern well. For MIN time series, ETS frequently captured the level and seasonality. The adaptive forecasting techniques did not dominate the strong experts but stayed within their range.

Figure 3 presents the gate-weight dynamics across test origins. The weights varied across origins and horizons, but the diagnostic classification remained seasonality-dominant throughout. Consequently, the gate behavior is best interpreted as adaptive weighting within a predominantly seasonal environment rather than as evidence of clear switching among distinct climatic regimes.

4.8. Summary of Empirical Findings

In total, the improved experiments give rise to five main conclusions. Firstly, SARIMA becomes the best practical forecasting model for MAX temperature, where both the fixed-restricted and the AIC-selected versions show good results. Secondly, ETS is the best practical forecasting model for MIN temperature, while SARIMA is its closest rival. Thirdly, the corrected seasonal naïve benchmark does not appear noticeably weak anymore, meaning that the improved experiment provides more reliable benchmarking. Fourthly, PREQ_DIRECT becomes the best adaptive combination model, consistently outperforming AVG_ENSEMBLE at all site-series setups, even though the advantages are usually small. Fifthly, GATED neural network model is unable to consistently beat AVG_ENSEMBLE or the best statistical baselines.

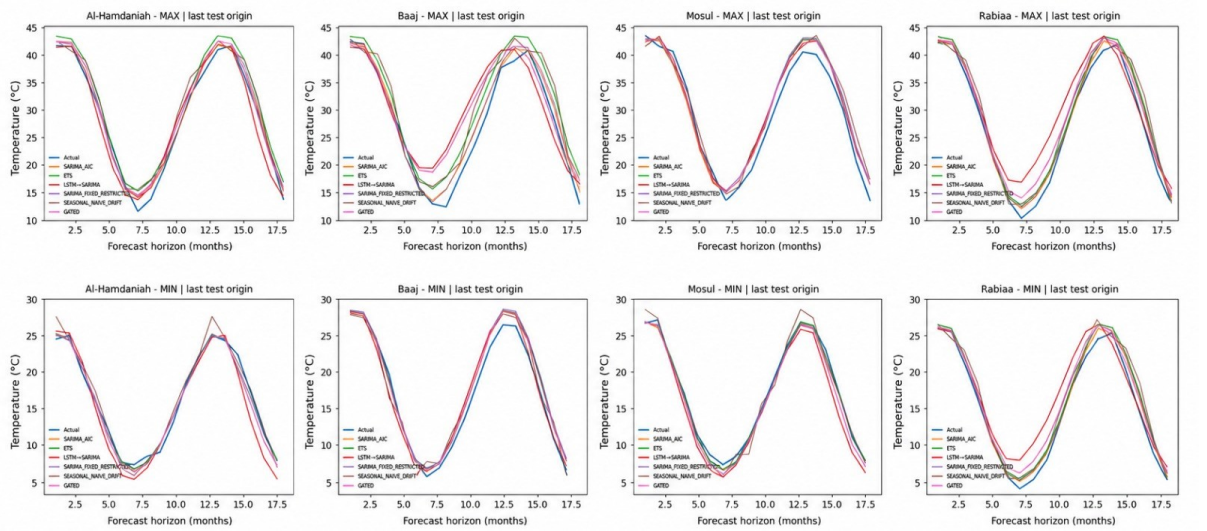

Figure 2. Eighteen-month forecasts for the last test origin across all site-series pairs. Panels compare actual observations against SARIMA_AIC, ETS, LSTM→SARIMA, AVG_ENSEMBLE, PREQ_DIRECT, GATED, LSTM, and seasonal baselines.

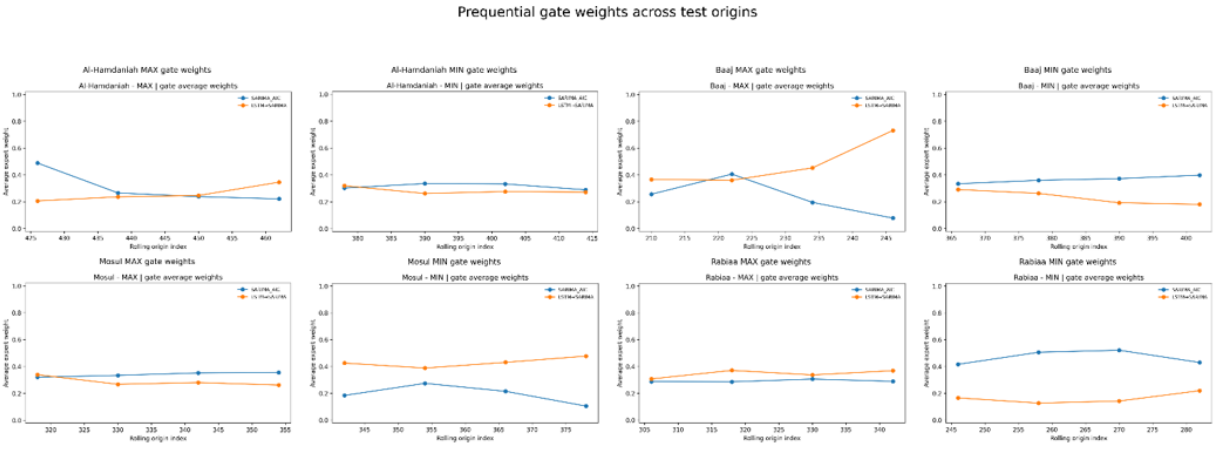

Figure 3. Average neural gate weights assigned to SARIMA_AIC, ETS, and LSTM→SARIMA across rolling test origins. The weights show expert preference within a predominantly seasonality-dominant setting.

In essence, these results imply a pragmatic approach to forecasting: seasonal statistical models should still be used as defaults in forecasting of strongly seasonal monthly temperatures in Nineveh Governorate. Adaptive weighting and gated ensembling can be viewed as helpful techniques of robustness checking, but they cannot be considered superior forecasting models by default.

5. Discussion

5.1. Main Findings

Thus, the revision yields clear and practically applicable results: among the statistical models for monthly temperature series considered in this study, seasonal models are the most resilient options for forecasting operationally. For MAX temperature, SARIMA yielded the best average performance in terms of test errors, both fixed restricted and AIC-selected specification demonstrated a lower error level than any of the adaptive ensemble approaches. For MIN temperature, ETS showed the lowest error rate, which was very close to SARIMA performance. Hence, the predictable structure is mostly of seasonal nature and can be captured by classical statistical models. While adaptive methods, PREQ_DIRECT and GATED, provided valuable information about the expert combination, their advantage was somewhat limited. The former showed itself as a more stable adaptive approach, as it consistently outperformed the equal-weighted approach in all site-series cases. However, the gain from using it was rather small. The neural GATED approach sometimes exceeded some baselines, but it did not outperform AVG_ENSEMBLE and did not exceed the performance of the strongest individual statistical models. Thus, the contribution of the adaptive framework is limited to the evaluation of the adaptive combination under strong seasonal baselines.

5.2. Why SARIMA Performs Strongly for MAX Temperature

The strong performance of SARIMA model for MAX temperature is consistent with the nature of the monthly maximum temperature series in the study region. As MAX temperature shows a strong annual cycle, with higher readings in the summer months and lower in winter months, this cycle along with possible linear or seasonal dependence is a good fit for SARIMA model, since it incorporates both nonseasonal and seasonal autocorrelation, differencing, and moving averages.

Results show that both SARIMA_FIXED_RESTRICTED and SARIMA_AIC approaches are performing admirably. The fixed restricted SARIMA yielded the lowest average MAX MAE, which is equal to $1.473\ ^{\circ}\mathrm{C}$, while SARIMA_AIC performed almost as well with the MAE equal to $1.489\ ^{\circ}\mathrm{C}$. Therefore, it is safe to state that the original fixed SARIMA specification is not an artificially weak benchmark; actually, it is a strong seasonal model for these data. At the same time, it is crucial that the model structure varies across sites and series, which was achieved through SARIMA_AIC inclusion.

Thus, the outperformance of SARIMA after model selection based on AIC proves the practical recommendation: in the case of strongly seasonal monthly temperature data, good seasonal statistical modeling can outperform more complicated neural or ensemble structures. Consequently, in the current dataset, SARIMA should be used as the default forecasting model for MAX temperature.

5.3. Why ETS Performs Strongly for MIN Temperature

For MIN temperature, ETS proved to be the best performing model with the average MAE of $0.995\ ^{\circ}\mathrm{C}$. This result means that the MIN temperature series have a smooth level and seasonal component that is captured by the exponential smoothing. ETS models work best when the time series can be described with evolving level, trend, and seasonal components.

Hence, the close performance of ETS and SARIMA for MIN temperature indicates that the predictable structure is still mostly seasonal. However, in some site-level cases ETS performs better than SARIMA as it captures the level and seasonality of MIN temperature more accurately. From the perspective of forecasting practice, it is an important aspect since ETS is an easy-to-implement and fast model. Thus, for monthly MIN temperature forecasting in Nineveh Governorate, ETS should be recommended as the default option with SARIMA as an alternative.

5.4. Interpretation of the Adaptive Weighting Results

The main goal of adaptive weighting is testing the ability of expert combination to add value when different models capture different aspects of the time series. The results of the analysis show some validity of the assumption, yet its practical value in the current dataset is limited. PREQ_DIRECT approach outperforms AVG_ENSEMBLE in

all site-series cases, which proves that the past expert performance contains valuable information for building the weight vector. Still, the improvement obtained is relatively small and does not allow to exceed SARIMA in MAX case and ETS in MIN case.

At the same time, the results of GATED are even more subtle. Despite the usage of diagnostic features, horizon information, and expert forecasts in building the weights, the GATED model is not always better than AVG_ENSEMBLE. This means that the added flexibility of neural gate is not really helpful when there is not many monthly observations available and the data structure is already captured by strong statistical models. In this situation, the simpler prequential weight assignment rule works better.

Thus, the comparison of PREQ_DIRECT and GATED models reveals that the former is simpler, more interpretable, and more stable, while the latter is more flexible but also more prone to overfitting and unstable weight learning. This study thus proves the recommendation of using direct sequential weighting as the baseline for the proposed neural gating approach; it is consistent with the general forecasting literature recommendations for evaluation of combinations and adaptive model selection.

5.5. Regime Diagnostics and the Limited Added Value of Gating

Regime diagnostics revealed that all validation and test rolling origins are seasonality-dominant. This is the crucial result for interpretation. The goal of regime-aware weighting is adaptation to changing conditions of model suitability. However, if the diagnostic result shows that the data remain predominantly seasonal across origins, there is not much room for gate to exploit regime variety.

This explains the strong performance of SARIMA and ETS models and the modest improvement of the adaptive methods. The gate does not operate in the setting with transitions between different regimes (seasonal, mixed, volatile/nonlinear); instead, it operates in the regime where there is mostly seasonality-dominance. In this regime, it is quite hard to outperform the well-specified seasonal statistical model.

Thus, the regime results should not be viewed as evidence of regime switching in the data; instead, they prove the utility of the diagnostic framework by showing the absence of regime diversity under the current regime classification rule. Therefore, the regime framework should be interpreted carefully, as adaptive gating is more beneficial in datasets with regime heterogeneity, while in the current data set there is mostly the seasonal structure. This result is in line with the concept of the mixture-of-experts, according to which the added value of the gate depends on the regime diversity in input data.

5.6. Importance of Corrected Baselines

An important improvement in the revised study is the correction of the seasonal-naïve benchmark. In the original version of the framework, the seasonal-naïve model showed unexpectedly low performance, suggesting that the seasonal indexing did not work correctly. The revised framework uses the repetition of the corresponding month of the recent seasonal cycle and also the drift variant.

After this correction, the seasonal-naïve model showed much more plausible errors, equal to $2.036\ ^{\circ}\mathrm{C}$ for MAX and $1.421\ ^{\circ}\mathrm{C}$ for MIN. However, it still does not outperform SARIMA and ETS models; still, it gives a realistic low-complexity baseline. This is important because the implausible performance of the baseline model may inflate the performance of the more complex model. The correction of the seasonal-naïve model helps to produce a fairer comparison; it is important since forecast evaluation should be conducted against realistic benchmarks.

5.7. Practical Implications

The practical implication is simple: SARIMA should be adopted as the default operational model for monthly MAX temperature forecasting in Nineveh Governorate. For MIN temperature forecasting, ETS should be used as the default model with SARIMA as an alternative.

Adaptive combination models should not be used instead of the mentioned statistical baselines in the current dataset. They may be employed as an additional tool for monitoring the performance of experts when the forecaster expects regime changes or wants to follow the performance of experts over time. In this situation, PREQ_DIRECT will be especially useful due to its simplicity and reliance only on the past losses of the experts. GATED can also

be helpful in this situation to understand the dynamics of experts' weights, but the added complexity should be justified.

For the decision-makers working in the semi-arid regions, the current results mean that the simple and interpretable models have a lot of value. More complex machine-learning and gated ensemble methods should be used only in case of the evident superiority over robust seasonal statistical methods.

5.8. Methodological Implications

From the methodological point of view, the revised study emphasizes the importance of the leakage-free evaluation in adaptive forecasting. The gate that uses forecast errors from the future at the same origin appears adaptive, but this does not reflect the actual conditions of forecasting; thus, the prequential design is employed in the revised framework, allowing to use only the past rolling-origin losses to build expert weights and providing stricter evaluation.

The study also shows the importance of choosing right experts: the replacement of weak and unstable models with strong and relevant models, such as ETS, makes the ensemble setting more interesting, as the ensemble may be easily better simply by down-weighting the poor models. The fairer test of adaptive combination requires that the expert pool will contain strong and relevant models.

Finally, the sensitivity analysis revealed that the low values for the temperature of the adaptive softmax weighting produced slightly better performance, while entropy regularization had a smaller effect. Still, this did not change the main conclusion: SARIMA and ETS models are the strongest options. This shows the robustness of the main empirical result, which is not dependent on a particular setting of hyperparameters. These methodological choices are consistent with the general principle that time-series models should be evaluated under chronological, out-of-sample, and rolling-origin conditions.

5.9. Limitations

Several limitations should be discussed. First, the study uses only four sites within a single governorate. While these sites are important for the regional forecasting in Nineveh, the conclusion cannot be generalized for other climatic regions without additional research. Second, the monthly frequency limits the number of observations for training neural models and reduces their effectiveness compared to statistical models, which are more suitable for short periodic series. Third, the diagnostic regime classification showed seasonality-dominant origins only under the current classification rule, limiting the possibilities for gate evaluation under different regimes. The future research should use richer features of regime classification and apply the regime framework to datasets with structural changes, volatility switches, and nonlinear episodes. Fourth, the study focuses on the point forecasting; probabilistic forecasting and prediction intervals can be more valuable for decision-making under uncertainty. The extension of the framework to probabilistic forecasting will give a more useful decision-making tool. Finally, AIC-selected SARIMA was used, but the model search was limited due to the computational constraints. The larger model search or different automatic forecasting procedures may provide different statistical baselines.

5.10. Future Work

Future work can be conducted in several directions. First, the regime framework should be applied to additional stations and neighboring regions in order to understand the value of adaptive weighting in more diverse climatic regimes. Second, richer regime features, such as seasonal strength measure, spectral features, rolling volatility indicators, and decomposition-based trends, should be used in order to enhance the regime classification. Third, the regime framework can be extended to the probabilistic forecasting by combining predictive distributions instead of the point forecasts. Fourth, the expert pool can be enriched with other strong statistical and machine-learning models, provided that each expert is evaluated under the leakage-free rolling-origin protocol. Finally, the practical decision-level evaluation should be conducted, for example, for agricultural scheduling, energy demand planning, and heat-risk preparedness.

Overall, the study shows that the regime-adaptive ensembling is useful only when there is the variability in model suitability depending on the regime. In the current monthly temperature data, however, the predominance of the

seasonal structure makes SARIMA and ETS difficult to beat. This is not a weakness of the study; on the contrary, it is a valuable empirical lesson on the limitations of adaptive gating in strongly seasonal climate forecasting.

6. Conclusion

This paper presents an improved, leakage-controlled analysis of adaptive expert combination for the task of predicting the monthly maximum and minimum temperature in Nineveh Governorate, Iraq. This study examines the framework in four locations – Al-Hamdaniah, Baaj, Mosul, and Rabiaa using a chronological train-validation-test split, multi-origin roll forecasts on an 18-month horizon. In this case, the final set of experts comprises AIC-chosen SARIMA, ETS, and LSTM→SARIMA. The two adaptive combination strategies are prequential direct weighting and neural gated ensemble.

From the empirical results, we conclude that seasonal statistical models still represent the most practical choice for forecasting the temperature. As regards MAX series, the best average test performance shows SARIMA with an MAE of $1.473\ ^{\circ}\mathrm{C}$ (SARIMA_FIXED_RESTRICTED) and $1.489\ ^{\circ}\mathrm{C}$ (SARIMA_AIC). In case of MIN series, ETS demonstrates the lowest average error with an MAE of $0.995\ ^{\circ}\mathrm{C}$. These results imply that the prevailing and predictable pattern of the examined monthly temperature series is highly seasonal.

In this particular example, the use of adaptive approach provided limited and conditional improvements. Specifically, the PREQ_DIRECT technique outperformed the equal-weighted ensembling in all cases; however, the obtained improvements are small. The neural GATED model provided valuable insights about the expert weighting but it did not significantly outperform equal-weighted averaging and was inferior to the strongest statistical baseline. Therefore, our conclusion from this analysis is that adaptive gating should not be considered as an optimal forecasting model in this case. Instead, it should be viewed as a controlled, leakage-free tool to determine if sequential expert combination is really beneficial after robust statistical baselines.

This revised paper solves a number of methodological problems. First, the SARIMA model selection was done through AIC-based algorithm, and the fixed SARIMA was used just as a restricted benchmark. Second, the seasonal-naive baseline was corrected and evaluated again. Third, the forward looking soft-oracle gate training technique was substituted with a prequential one, in which the expert weights are calculated based on past roll losses only. Fourth, the sensitivity analyses were performed with regard to the hyperparameters of the primary adaptive weighting strategy.

Thus, we have concluded that this study provides a practical empirical lesson for the monthly temperature forecasting problem in semi-arid regions. It can be seen that if time series show high seasonality, simple and well-interpreted statistical models like SARIMA and ETS can be very hard to beat, even with the help of hybrid and adaptive ensemble techniques. Therefore, the value of adaptive combination is conditional, since it can facilitate robustness analysis and expert tracking but should be adopted operationally only in case of performance superiority compared to robust baselines.

In further research, the analysis should be extended to other stations and neighboring regions, where the heterogeneity of regimes is expected to be higher. Moreover, more elaborate diagnostics features, seasonality and trend strength measures using time series decomposition, spectral/volatility-based regimes and probabilistic forecasting with prediction intervals should be applied. Finally, the decision-level evaluation should be studied with respect to agricultural scheduling, energy planning, and heat risk preparedness tasks.

Data Availability

The monthly temperature data used in this study were obtained from the Iraqi General Authority for Meteorology. The data are not publicly archived by the authors because they are subject to institutional permission. The data may be made available from the corresponding author upon reasonable request and subject to approval by the data-providing authority.

Code Availability

The Python code used in this study is available at:
https://github.com/ziadoonkhaleel-stat/Code-for-Regime_Aware-for-Temperature-Forecasting

The repository includes the main Python script and the required files for running the experiments. The raw monthly temperature data are not publicly redistributed because they are subject to institutional permission from the Iraqi General Authority for Meteorology. The analysis can be reproduced when the approved data file is available.

REFERENCES

1. F. Petropoulos, D. Apiletti, V. Assimakopoulos, M.Z. Babai, D.K. Barrow, S. Ben Taieb, C. Bergmeir, R.J. Bessa, J. Bijak, J.E. Boylan, et al., *Forecasting: theory and practice*, International Journal of Forecasting 38(3) (2022) 705-871. doi:10.1016/j.ijforecast.2021.11.001
2. P. Montero-Manso, G. Athanasopoulos, R.J. Hyndman, T.S. Talagala, *FFORMA: Feature-based forecast model averaging*, International Journal of Forecasting 36(1) (2020) 86-92. doi:10.1016/j.ijforecast.2019.02.011
3. G.E.P. Box, G.M. Jenkins, G.C. Reinsel, G.M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed., Wiley, Hoboken, NJ, 2015. doi: 10.1111/jtsa.12194
4. R.J. Hyndman, G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed., OTexts, Melbourne, Australia, 2021.
5. R.J. Hyndman, Y. Khandakar, *Automatic time series forecasting: The forecast package for R*, Journal of Statistical Software 27(3) (2008) 1-22. doi:10.18637/jss.v027.i03
6. G.M. Ljung, G.E.P. Box, *On a measure of lack of fit in time series models*, Biometrika 65(2) (1978) 297-303. doi:10.1093/biomet/65.2.297
7. R.J. Hyndman, A.B. Koehler, J.K. Ord, R.D. Snyder, *Forecasting with Exponential Smoothing: The State Space Approach*, Springer, Berlin, 2008.
8. E.S. Gardner Jr., *Exponential smoothing: The state of the art–Part II*, International Journal of Forecasting 22(4) (2006) 637-666. doi:10.1016/j.ijforecast.2006.03.005
9. J.D. Hamilton, *A new approach to the economic analysis of nonstationary time series and the business cycle*, Econometrica 57(2) (1989) 357-384. doi:10.2307/1912559
10. S. Hochreiter, J. Schmidhuber, *Long short-term memory*, Neural Computation 9(8) (1997) 1735-1780. doi:10.1162/neco.1997.9.8.1735
11. F. Chollet, *Deep Learning with Python*, 2nd ed., Manning Publications, Shelter Island, NY, 2021.
12. G.P. Zhang, *Time series forecasting using a hybrid ARIMA and neural network model*, Neurocomputing 50 (2003) 159-175. doi:10.1016/S0925-2312(01)00702-0
13. M. Khashei, M. Bijari, *A novel hybridization of artificial neural networks and ARIMA models for time series forecasting*, Applied Soft Computing 11(2) (2011) 2664-2675. doi:10.1016/j.asoc.2010.10.015
14. S. Smyl, *A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting*, International Journal of Forecasting 36(1) (2020) 75-85. doi:10.1016/j.ijforecast.2019.03.017
15. M.A. Zaytar, C. El Amrani, *Sequence to sequence weather forecasting with long short-term memory recurrent neural networks*, International Journal of Computer Applications 143(11) (2016) 7-11. doi: 10.5120/ijca2016910497
16. P. Lara-Benítez, M. Carranza-García, J.C. Riquelme, *Temporal convolutional networks applied to energy-related time series forecasting*, Applied Sciences 10(7) (2020) 2322. doi:10.3390/app10072322
17. R.A. Jacobs, M.I. Jordan, S.J. Nowlan, G.E. Hinton, *Adaptive mixtures of local experts*, Neural Computation 3(1) (1991) 79-87. doi:10.1162/neco.1991.3.1.79
18. M.I. Jordan, R.A. Jacobs, *Hierarchical mixtures of experts and the EM algorithm*, Neural Computation 6(2) (1994) 181-214. doi:10.1162/neco.1994.6.2.181
19. D.H. Wolpert, *Stacked generalization*, Neural Networks 5(2) (1992) 241-259. doi:10.1016/S0893-6080(05)80023-1
20. L. Breiman, *Stacked regressions*, Machine Learning 24 (1996) 49-64. doi:10.1007/BF00117832
21. X. Wang, K. Smith-Miles, R.J. Hyndman, *Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series*, Neurocomputing 72(10-12) (2009) 2581-2594. doi:10.1016/j.neucom.2008.10.017
22. C. Bergmeir, J.M. Benítez, *On the use of cross-validation for time series predictor evaluation*, Information Sciences 191 (2012) 192-213. doi:10.1016/j.ins.2011.12.028
23. C. Bergmeir, R.J. Hyndman, B. Koo, *A note on the validity of cross-validation for evaluating autoregressive time series prediction*, Computational Statistics & Data Analysis 120 (2018) 70-83. doi:10.1016/j.csda.2017.11.003
24. L.J. Tashman, *Out-of-sample tests of forecasting accuracy: An analysis and review*, International Journal of Forecasting 16(4) (2000) 437-450. doi:10.1016/S0169-2070(00)00065-0
25. F.X. Diebold, R.S. Mariano, *Comparing predictive accuracy*, Journal of Business & Economic Statistics 13(3) (1995) 253-263. doi:10.1080/07350015.1995.10524599
26. D. Harvey, S. Leybourne, P. Newbold, *Testing the equality of prediction mean squared errors*, International Journal of Forecasting 13(2) (1997) 281-291. doi:10.1016/S0169-2070(96)00719-4
27. R.J. Hyndman, A.B. Koehler, *Another look at measures of forecast accuracy*, International Journal of Forecasting 22(4) (2006) 679-688. doi:10.1016/j.ijforecast.2006.03.001

28. D.A. Dickey, W.A. Fuller, *Distribution of the estimators for autoregressive time series with a unit root*, Journal of the American Statistical Association 74(366) (1979) 427-431. doi:10.1080/01621459.1979.10482531
29. D.P. Kingma, J. Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980 (2014). doi:10.48550/arXiv.1412.6980
30. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, *Dropout: A simple way to prevent neural networks from overfitting*, Journal of Machine Learning Research 15 (2014) 1929-1958.
31. S. Makridakis, E. Spiliotis, V. Assimakopoulos, *The M4 Competition: 100,000 time series and 61 forecasting methods*, International Journal of Forecasting 36(1) (2020) 54-74. doi:10.1016/j.ijforecast.2019.04.014
32. S. Makridakis, E. Spiliotis, V. Assimakopoulos, *M5 accuracy competition: Results, findings, and conclusions*, International Journal of Forecasting 38(4) (2022) 1346-1364. doi:10.1016/j.ijforecast.2021.11.013
33. Z.M. Khaleel, S.J. Abed, J.A. Sultan, N.M. Ahmeed, *Statistical and ANN-based modeling for desertification risk prediction in semi-arid regions: A case study of Nineveh Governorate*, Statistics, Optimization & Information Computing 15(1) (2025) 198-211. doi:10.19139/soic-2310-5070-2985