Comparative Analysis of Medium-Range (72-hour) Temperature Time Series Forecasting Using Korea Meteorological Administration Surface Observation Data

Hyun Koo Lee, Nak Hyun Jung

ABSTRACT

This study presents a comparative evaluation of medium-range (72-hour) air temperature forecasting performance among recent deep learning–based time series models—Neural Hierarchical Interpolation for Time Series Forecasting (NHITS), Temporal Fusion Transformer (TFT), Patch Time Series Transformer (PatchTST)—relative to a traditional statistical model (SARIMA) and a climatology-based baseline. The analysis is conducted using hourly surface observation data provided by the Korea Meteorological Administration, with a focus on a 72-hour multi-step forecasting task for the Seoul station.

Empirical results show that all three deep learning models substantially outperform the traditional baselines. Among them, the
TFT achieves the best overall performance, with a Mean Squared Error (MSE) of 4.39, a Root Mean Squared Error (RMSE) of 2.10, a Mean Absolute Error (MAE) of 1.65, and a coefficient of determination (R^2) of 0.775. This corresponds to an approximate 54% reduction in RMSE compared with the SARIMA model. NHITS also demonstrates a significant improvement over the statistical baseline, achieving an RMSE of 2.61.

These results indicate that the Temporal Fusion Transformer architecture, which combines a Long Short-Term Memory (LSTM) based sequence encoder with multi-head attention, is particularly effective in capturing the nonlinear dynamics and combined seasonal and diurnal structures inherent in air temperature time series. This study provides quantitative evidence supporting the applicability of advanced deep learning models to short- to medium-range temperature forecasting based on domestic meteorological observation data. The findings have practical implications for applications requiring accurate temperature prediction, including smart grid energy demand management and weather-related disaster prevention. Future work should extend this framework to multivariate and multi-station settings, as well as to probabilistic forecasting approaches for explicit quantification of predictive uncertainty

☞ Keyword: Air temperature forecasting, time series forecasting, deep learning, Temporal Fusion Transformer, medium range forecasting, meteorological observation data

1. Introduction

1.1 Research Background

Air temperature is a fundamental meteorological variable that plays a critical role across a wide range of societal domains, including agricultural productivity, energy supply and demand planning, disaster management, and public health. Temperature time series inherently exhibit complex seasonal structures driven by the Earth's rotation and revolution, characterized by daily (24-hour) and annual (365-day) cycles. Superimposed on these periodic components are nonlinear fluctuations arising from factors such as frontal passages, variations in cloud cover, and local atmospheric circulation, resulting in a highly complex dynamical system. Owing to these characteristics, temperature forecasting has been widely regarded as a representative yet challenging benchmark for evaluating the predictive performance and generalization capability of time series forecasting models.

For several decades, statistical approaches based on the ARIMA (AutoRegressive Integrated Moving Average) family constituted the dominant methodology for temperature forecasting. However, the recent intensification of meteorological variability associated with climate change has increasingly revealed the limitations of these linear models. As an alternative, deep learning–based time series forecasting methods have gained attention for their strong capacity to learn nonlinear patterns, leading to a gradual paradigm shift in forecasting methodology. In particular, the Transformer architecture proposed by Vaswani et al. (2017)[1], originally developed for natural language processing, has also had a significant impact on time series forecasting research. Following this development, a series of Transformer-based models—such as Informer[2], Autoformer[3], PatchTST[4], and the Temporal Fusion Transformer (TFT)[5]—have been proposed to better capture long-range dependencies and complex periodic structures in time series data. These recent models have demonstrated substantial performance improvements over traditional statistical models and early LSTM-based approaches on various global benchmark datasets, including electricity demand, traffic flow, and logistics indices.

Despite these advances, empirical studies that systematically apply and evaluate such state-of-the-art deep learning models using domestic meteorological data remain limited. In particular, research based on Korea Meteorological Administration (KMA) Automated Synoptic Observation System (ASOS)[6] data is still at an early stage. Much of the existing domestic literature continues to rely on statistical methods or early recurrent neural network architectures, underscoring the need for comprehensive validation of recent deep learning models under domestic meteorological conditions and for assessing their potential practical applicability.

1.2 Research Problem and Objectives

This study addresses the empirical validation gap observed in the domestic meteorological forecasting literature regarding recent deep learning–based time series models. Using hourly air temperature data from the Seoul region, the study aims to conduct a systematic and quantitative comparison of predictive performance across different forecasting approaches. Rather than focusing on short-term prediction, this research considers a 72-hour (3-day) multi-step forecasting setting in order to evaluate model robustness and accuracy as the prediction horizon increases. Based on this motivation, the following research questions are formulated.

First, this study investigates whether state-of-the-art deep learning time series models—namely AutoTFT, AutoNHITS, and AutoPatchTST—exhibit statistically meaningful performance improvements over a traditional statistical model (SARIMA) and a climatology-based baseline in 72-hour multi-step temperature forecasting using KMA ASOS data.

Second, this study examines how architectural differences among deep learning models—specifically the MLP-based NHITS, the attention-based Temporal Fusion Transformer (TFT), and the patch-based PatchTST—affect their ability to capture seasonality and nonlinear variability in temperature time series, and seeks to identify the structural factors underlying these performance differences.

The primary objective of this study is to address these research questions by comparing the predictive performance of five models using hourly temperature observations from the Seoul station (STN 108), based on quantitative evaluation metrics such as MSE, RMSE, and MAE. Through this analysis, the study aims to provide empirical evidence supporting the adoption of deep learning models in the domestic meteorological domain and to establish a foundational data framework for future studies extending to multivariate inputs and multiple observation stations.

2. Literature Review

2.1 Development of Time Series Temperature Forecasting Methodologies

2.1.1 Traditional Statistical Model-Based Time Series Forecasting

The ARIMA model has long been regarded as a classical standard in time series analysis, providing a framework that mathematically models the autocorrelation structure of observed data to generate future predictions. For time series exhibiting pronounced periodic patterns, such as air temperature, the SARIMA

(Seasonal ARIMA) model—which explicitly incorporates seasonal components—has been widely employed.

Despite their extensive use, these linear statistical models suffer from inherent limitations. First, because they are fundamentally based on linear dependency assumptions, their ability to capture nonlinear meteorological dynamics—such as abrupt changes associated with frontal passages or variations in cloud cover—is limited. Second, as the forecast horizon extends, predicted values tend to converge toward the long-term mean, leading to a rapid increase in uncertainty and a corresponding decline in the informational value of long-range forecasts.

2.1.2 Transition to Deep Learning-Based Time Series Forecasting

To address the limitations of traditional statistical approaches, early deep learning–based forecasting methods primarily focused on recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) networks. LSTMs partially alleviated the long-term dependency problem by employing gated mechanisms that selectively retain and discard information within sequential data. However, difficulties related to limited parallelization and information degradation in very long sequences remained unresolved.

Since 2017, the dominant paradigm in time series forecasting has shifted toward Transformer-based architectures utilizing self-attention mechanisms. The three representative architectures considered in this study adopt distinct strategies for modeling temporal dependencies and complex patterns.

The TFT integrates an LSTM-based encoder with multi-head attention, enabling the simultaneous learning of local temporal patterns and long-range dependencies. In addition, it incorporates a Variable Selection Network and gating mechanisms to suppress irrelevant inputs while enhancing model interpretability.

NHITS decomposes time series into high-frequency and low-frequency components through multi-rate sampling and hierarchical interpolation, thereby achieving both computational efficiency and improved accuracy in long-horizon forecasting tasks.

PatchTST tokenizes time series by dividing them into fixed-length patches, preserving localized temporal information while substantially reducing the computational complexity of attention operations. This design allows the model to effectively leverage longer historical input sequences in forecasting.

2.2 Current Status and Limitations of Domestic Temperature Prediction Research

2.2.1 Current Status of Domestic Temperature Prediction Research

Previous domestic studies (see Table 1) have primarily employed early deep-learning architectures such as LSTM and have been conducted in short-term, region-/context-specific settings (e.g., single-city temperature, limited ASOS stations, urban maximum temperature, or summer-focused analyses).

Accordingly, it remains insufficiently examined whether recent long-horizon forecasting models such as PatchTST and N-HiTS, which have shown strong performance on international benchmarks, provide comparable gains under domestic meteorological conditions characterized by pronounced seasonality and complex topography.

Table 1. Summary of previous studies on domestic temperature forecasting

study	model	Target Variable	Forecast Horizon
Yun and Jeon (2017)	LSTM	Air temperature (Gwangju)	24 hours
Hong et al. (2021)[9]	LSTM	Meteorological variables at 9 ASOS stations	Short-term
Lim et al. (2020)[10]	Deep learning	Urban maximum temperature	Short-term
Cho et al. (2023)[11]	Deep learning + spatial interpolation	Summer temperature	Short-term

2.2.2 Limitations of Previous Studies

A review of prior domestic studies reveals several research gaps.

First, there is a limitation in model coverage. Existing domestic temperature forecasting studies have largely concentrated on recurrent neural network–based approaches such as LSTM, GRU, and CNN-LSTM, which became prevalent between approximately 2015 and 2018. In contrast, domestic applications of more recent time series forecasting models—including TFT, NHITS, and PatchTST, which have attracted increasing attention since 2021—remain difficult to identify, particularly in studies using Korean meteorological data.

Second, limitations are observed in the forecasting horizon. Most previous studies have focused on short-term forecasts within 24 hours, and systematic evaluations of deep learning model performance in medium-range forecasting settings extending to 72 hours or longer are largely absent.

Third, there is a lack of systematic comparative analysis. While many studies examine the applicability of individual models, relatively few studies conduct quantitative comparisons of multiple state-of-the-art deep learning models alongside statistical baselines under identical datasets and experimental conditions.

2.2.3 Distinctiveness of This Study

In response to the research gaps identified above, this study differentiates itself from existing work in several respects.

First, this study presents an empirical comparative analysis applying recent deep learning time series models—specifically TFT, NHITS, and PatchTST, all proposed after 2021—to domestic Korea Meteorological Administration (KMA) ASOS data. Although these models have demonstrated strong performance on international benchmark datasets, their applicability to domestic temperature time series—characterized by combined diurnal and annual seasonality as well as abrupt fluctuations associated with frontal passages—has not yet been sufficiently examined.

Second, this study adopts a medium-range forecasting horizon of 72 hours (3 days), thereby extending beyond the 24-hour short-term prediction range that has been the primary focus of most domestic studies.

Third, this study systematically compares a climatology-based baseline, SARIMA, and three deep learning models under a unified dataset and experimental setting, enabling a quantitative assessment of performance improvements achieved by deep learning approaches relative to traditional statistical models.

Finally, this study enhances reproducibility by employing the Auto-Hyperparameter Optimization (Auto-HPO) functionality provided by the NeuralForecast library, thereby establishing a consistent experimental framework that can be utilized in future comparative studies.

3. Method (Research Methodology)

3.1 Data Collection and Preprocessing

This study used hourly data from the Automated Synoptic Observation System (ASOS) observed at the Seoul meteorological station (STN 108) for the period from January 1, 2015, to November 21, 2025. The raw data were refined and preprocessed according to the procedures described below.

Special values defined by Korea Meteorological Administration (KMA) specifications (e.g., −9, −9.0) were regarded as missing observations. The time series was then reconstructed using a continuous hourly datetime index. Missing values were imputed using linear interpolation based on temporal information (method = “time”). The linear interpolation formula is given as shown in Eq. (1):

$$y = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0} \tag{1}$$

where x denotes the missing time point, while x_0 and x_1 represent the valid observation time points adjacent to the missing interval. Missing values at the beginning and end of the time series, where interpolation is not applicable, were imputed using forward and backward filling (ffill/bfill). As a result, the final analysis dataset contained no remaining missing values.

Since the purpose of this experiment is to compare predictive performance across different model architectures, a univariate framework was maintained, with air temperature (TA) specified as the sole input and target variable. Data normalization for ensuring training stability of the deep learning models was not performed manually; instead, it was carried out automatically using scalers such as TemporalNorm provided within the NeuralForecast library.

Data partitioning was designed in consideration of the target 72-hour (3-day) multi-step forecasting structure. The period from January 1, 2015, to November 18, 2023, was primarily used for calculating the climatology baseline, deriving long-term statistical measures, and conducting exploratory analyses to identify seasonal and trend characteristics of the time series. The main model training period was limited to the interval from 00:00 on November 19, 2023, to 23:00 on November 18, 2025, during which the statistical model (SARIMA) and the deep learning–based models AutoNHITS, AutoTFT, and AutoPatchTST were trained. Finally, a 72-hour period from 00:00 on November 19, 2025, to 23:00 on November 21, 2025, was designated as the final test interval, and all models were configured to generate predictions for the same time window.

With respect to the training–validation split, a strategy was adopted in which the final 72 hours ($h = 72$) of the training period were used as the validation set for hyperparameter tuning of the deep learning Auto models. Specifically, from the data spanning November 19, 2023, to November 18, 2025, the last 72 observations were reserved for validation, and the preceding portion was used to train each trial with different parameter configurations. The trial that yielded the minimum loss on the validation set (based on MAE) was selected as the optimal hyperparameter configuration. Using the selected configuration, the model was retrained on the entire training period and subsequently used to generate 72-hour forecasts for the final test period.

3.2 Prediction Problem Definition and Baseline Models

The objective of this study is to perform 72-hour-ahead (3-day) multi-step forecasting of hourly air temperature (TA) at the Seoul station using a univariate setting. Model performance is evaluated against two baseline approaches: a climatology lookup baseline and a statistical SARIMA baseline.

3.2.1 Climatology Baseline

As a minimal-reference baseline, a climatology model was constructed from historical hourly temperature observations spanning 2015 to November 18, 2023. For each timestamp, the day-of-year and hour-of-day were computed, and the climatological mean temperature was estimated for each $(day-of-year, hour-of-day)$ pair to form a lookup table. The climatology prediction at a test time point t is defined as shown in Eq. (2) and (3):

$$\hat{y}_t^{\text{clim}} = \mu_{\text{doy}(t),\,\text{hod}(t)} \tag{2}$$

$$\mu_{d,h} = \mathbb{E}[y_t \mid \text{doy}(t) = d,\ \text{hod}(t) = h] \tag{3}$$

This baseline requires little to no model training and serves as a minimum performance benchmark commonly used in operational meteorology.

3.2.2 SARIMA Baseline

As a statistical baseline, the Seasonal AutoRegressive Integrated Moving Average (SARIMA) model was employed. Since Seoul's hourly temperature series exhibits pronounced diurnal periodicity, the seasonal period was fixed to $s = 24$. Stationarity assessment (e.g., ADF test) and seasonal decomposition (details reported in the appendix) suggest the presence of trend and seasonal components; therefore, the differencing orders D were explored within a limited range focusing on small values.

Model orders were selected via a constrained grid search over candidate $\text{SARIMA}(p,d,q) \times (P,D,Q)_{24}$ specifications, based on model fitting results and diagnostics. The selected SARIMA specification is given in Eq. (4):

$$\text{SARIMA}(3,0,1) \times (0,0,1)_{24} \tag{4}$$

Using the chosen configuration, the SARIMA model was refitted for the full training period (November 19, 2023–November 18, 2025). Subsequently, 72-step-ahead (72-hour-ahead) forecasts $\{\hat{y}_{t+h|t}\}_{h=1}^{72}$ were generated for the test period.

3.3 Deep Learning–Based Auto Models

To benchmark modern deep learning approaches, this study employed the automatic hyperparameter optimization (Auto) variants of NHITS, Temporal Fusion Transformer (TFT), and PatchTST, implemented in the NeuralForecast library. All deep learning models were trained in a univariate setting using only past temperature values (i.e., no additional covariates), and the forecasting horizon was fixed to $h = 72$ steps (72 hours).

3.3.1 AutoNHITS, AutoTFT, and AutoPatchTST

AutoNHITS extends the N-BEATS family by modeling multiple temporal scales through stacked blocks. AutoTFT combines an LSTM-based sequence encoder with multi-head attention, while AutoPatchTST segments the series into fixed-length patches and applies a Transformer to learn long-range dependencies via patch-level representations.

Hyperparameter optimization was performed using the Ray Tune backend. The search space included the input window length, hidden dimension, dropout rate, learning rate, batch size, and window batch size. Each trial used a consistent training–validation protocol in which the last 72 hours of the training period were held out as the validation set. The hyperparameter configuration achieving the lowest validation MAE was selected. With the selected configuration, each model was retrained on the full training dataset and then used to generate 72-hour-ahead forecasts for the test period.

3.4 Training Setup and Evaluation Metrics

All models were evaluated and compared using the same test period, spanning a total of 72 hours from 00:00 on November 19, 2025, to 23:00 on November 21, 2025. Forecasting accuracy was assessed using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). The definitions of MSE, RMSE, and MAE are given in Eqs. (5)–(7), respectively.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

The coefficient of determination (R^2) was used as a measure of the proportion of variance in the observed values that is explained by the predictions.

Due to the characteristics of air temperature forecasting, conventional Mean Absolute Percentage Error (MAPE) can be distorted when temperatures are close to 0 °C, as the denominator approaches zero. Accordingly, for the deep learning models, temperature values were converted to Kelvin units prior to computing MAPE, and a Kelvin-based MAPE was employed in Eq. (8) .

$$\mathrm{MAPE}(K) = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{(y_i+273.15)-(\hat{y}_i+273.15)}{y_i+273.15}\right| \tag{8}$$

In contrast, at the initial implementation stage, MAPE values for the Climatology and SARIMA models were computed based on temperatures expressed in degrees Celsius, resulting in a substantial discrepancy in scale compared to the Kelvin-based MAPE used for the deep learning models. Owing to this inconsistency, the present study focuses primarily on MSE, RMSE, MAE, and R^2 for performance comparison across models. The accurate recalculation of MAPE(K) and the application of a unified definition across all models are identified as necessary tasks to be addressed in future work.

4. Results (Empirical Analysis)

4.1 Data Characteristics and Exploratory Analysis

The hourly air temperature time series observed at the Seoul station (STN 108) reflects typical characteristics of a mid-latitude urban climate, exhibiting both long-term seasonal patterns associated with annual variability and pronounced diurnal patterns characterized by recurring daily maximum and minimum temperatures. The ASOS data used in this study were confirmed to capture these meteorological characteristics, underscoring the necessity of adopting model structures that explicitly account for seasonality and periodicity.

Although quantitative results from seasonal decomposition and stationarity tests have not yet been fully consolidated, the analysis was conducted under the assumption that the temperature time series is non-stationary and contains both trend and seasonal components. This assumption was also reflected in the design of the SARIMA model, in which the differencing orders were selected accordingly and a seasonal period of 24 was adopted to represent diurnal variation. More rigorous results from the Augmented Dickey–Fuller (ADF) test, as well as detailed statistics of the decomposed trend, seasonal, and residual components, may be presented in the main text or appendix in future revisions.

4.2 Comparison of 72-Hour Temperature Forecasting Performance

The forecasting performance of each model over the 72-hour test period, from November 19 to November 21, 2025, is summarized in Table 2.

Table 2. Performance comparison of models for 72-Hour Air Temperature Forecasting

model	MSE	RMSE	MAE	R^2	MAPE (K)
Climatology	23.04	4.80	4.15	0.216	-
SARIMA	20.44	4.52	3.70	0.304	-
AutoPatchTST	12.46	3.53	3.11	0.361	1.11%
AutoNHITS	6.80	2.61	2.10	0.651	0.75%
AutoTFT	4.39	2.10	1.65	0.775	0.59%

* MAPE(K) represents Kelvin-based Mean Absolute Percentage Error. For the Climatology and SARIMA models, MAPE values were not computed due to inconsistencies in definition.

The forecasting performance of each model over the 72-hour test period from November 19 to November 21, 2025, can be summarized as follows. The climatology model yielded an MSE of approximately 23.04, an RMSE of 4.80, an MAE of 4.15, and an R^2 value of 0.216. These results indicate that a simple climatology-based approach is able to explain a limited portion of temperature variability, but the associated prediction errors remain relatively large.

For the SARIMA model, the MSE, RMSE, MAE, and R^2 were approximately 20.44, 4.52, 3.70, and 0.304, respectively. Although this represents a modest improvement over the climatology baseline, the prediction errors are still substantial, and the explanatory power remains limited.

Among the deep learning–based Auto models, AutoNHITS achieved an MSE of approximately 6.80, an RMSE of 2.61, an MAE of 2.10, and an R^2 of 0.651. Compared with SARIMA, this corresponds to an error reduction of roughly 50% in terms of RMSE, indicating that AutoNHITS is able to capture a substantial proportion of the overall temperature variability. AutoTFT exhibited the best performance among the three deep learning models, with an MSE of approximately 4.39, an RMSE of 2.10, an MAE of 1.65, and an R^2 of 0.775. In particular, AutoTFT achieved the lowest RMSE and MAE values, outperforming not only the statistical baselines but also the other deep learning models, including AutoNHITS and AutoPatchTST.

AutoPatchTST recorded an MSE of approximately 12.46, an RMSE of 3.53, an MAE of 3.11, and an R^2 of 0.361. While its performance represents an improvement over the climatology and SARIMA models, it remains inferior to that of AutoNHITS and AutoTFT.

When comparing Kelvin-based MAPE values, AutoNHITS, AutoTFT, and AutoPatchTST yielded MAPE(K) values of approximately 0.75%, 0.59%, and 1.11%, respectively, indicating that absolute prediction errors over the 72-hour horizon are very small relative to the absolute temperature scale. In contrast, MAPE values for the climatology and SARIMA models were abnormally large due to differences in definition arising from the initial implementation, making direct comparison with the Kelvin-based MAPE of the deep learning models inappropriate. Accordingly, this study primarily discusses relative model performance based on MSE, RMSE, MAE, and R^2.

Overall, all three Auto deep learning models substantially reduced prediction errors compared with the climatology and SARIMA baselines. AutoTFT achieved an error reduction of approximately 50% relative to the statistical model in terms of RMSE and MAE. From the perspective of R^2, AutoNHITS($R^2 \approx 0.65$), AutoTFT ($R^2 \approx 0.77$), and AutoPatchTST ($R^2 \approx 0.36$) all exhibited markedly higher explanatory power than the climatology ($R^2 \approx 0.22$) and SARIMA ($R^2 \approx 0.30$) models, confirming that the deep learning approaches are considerably more effective in capturing temperature variability.

4.3 Descriptive Summary of Visual Comparisons

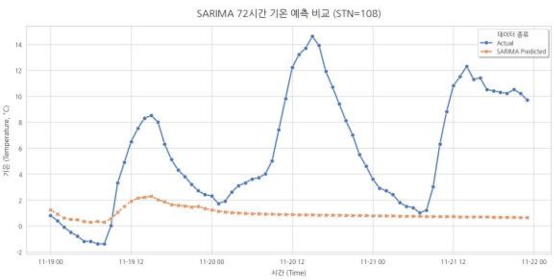

Figure 1. Comparison of Observed and SARIMA-Predicted Air Temperature over the 72-Hour Test Period

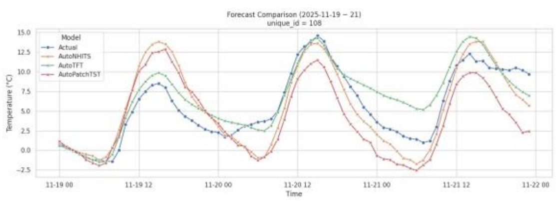

Figure 2. Comparison of Observed and Predicted Air Temperature Using AutoNHITS, AutoTFT, and AutoPatchTST over the 72-Hour Test Period

When the observed temperature time series and the model-predicted series are overlaid, the climatology and SARIMA models generally follow smooth temperature variation patterns (Figure 1). However, during periods of rapid temperature increases or decreases, these statistical baselines exhibit delayed responses and fail to accurately reproduce the timing and magnitude of peaks and troughs (Figure 1). In particular, over the 72-hour forecast horizon, SARIMA tends to regress toward the mean when abrupt temperature changes occur, leading to large errors in daily maximum and minimum temperatures (Figure 1).

In contrast, AutoTFT captures both the trend and diurnal patterns more effectively even under abrupt short-term changes, reproducing the timing and magnitude of peaks and troughs with relatively higher fidelity (Figure 2). AutoNHITS also tracks the overall pattern more smoothly and stably than SARIMA; although mild over-smoothing is observed in some intervals, its deviations from the observed series remain smaller overall than those of the statistical model (Figure 2). For AutoPatchTST, the predicted trajectories appear sensitive to the input window length, patch size, training steps, and the hyperparameter search range adopted in this study; in some trials, highly oscillatory prediction patterns are observed, which may help explain its comparatively weaker performance relative to the other deep learning models (Figure 2). Overall, these qualitative visual observations align with the tendencies indicated by the quantitative evaluation metrics (Figures 1–2).

5. Discussion

5.1 Interpretation of Model Performance

The empirical results demonstrate that the deep learning–based Auto models clearly outperform traditional baselines such as the climatology model and SARIMA. In particular, AutoTFT achieved an error reduction of more than approximately 50% in terms of RMSE compared with SARIMA, even under the relatively constrained setting of a single station and univariate input. This finding suggests that Transformer-based deep learning models can provide meaningful performance improvements for short- to medium-range forecasting problems, such as 72-hour air temperature prediction in the Seoul region.

These results can be naturally interpreted in light of the nonlinear characteristics and long-range dependencies inherent in temperature time series. In addition to seasonal and diurnal cycles, air temperature is influenced by a range of physical processes, including frontal passages, variations in cloud cover, and local wind fields, resulting in complex and nonlinear dynamics. Linear models such as SARIMA have limited capacity to represent such nonlinear structures and tend to regress toward the mean as the forecast horizon increases. In contrast, AutoNHITS and AutoTFT leverage deep neural network architectures to learn nonlinear relationships within the time series and to capture patterns across multiple temporal scales, enabling more stable predictions of trends and periodic components even over a relatively long horizon of three days.

Furthermore, AutoTFT integrates an LSTM-based sequence encoder with a multi-head attention–based fusion layer, allowing the model to assign greater weight to temporally important time points or intervals. Although this study considered only a univariate setting, such architectural characteristics appear to be well suited to temperature data with pronounced seasonal and diurnal patterns. Internal scaling and gating mechanisms may also have contributed to improved training stability and generalization performance.

By contrast, the relatively weaker performance of AutoPatchTST is likely attributable not to inherent limitations of the model itself, but rather to the specific choices of input window length, patch size, and hyperparameter search range adopted in this study, which may not have been fully optimized for the given data and forecasting objective ($h = 72$). Patch-based time series Transformers are known to be sensitive to patch design and training configurations, and there remains potential for performance improvement through more refined architectural tuning and extended training. Nevertheless, such improvements require further experimentation and validation. At the current stage, the results indicate that AutoTFT and AutoNHITS represent more stable and interpretable options from a practical application perspective.

5.2 Academic and Practical Implications

Although this study focuses on a single station and a univariate temperature forecasting setting, it quantitatively demonstrates that recent deep learning–based time series models can significantly outperform traditional statistical approaches. This finding indicates that there is sufficient motivation to adopt Transformer-based models even at the single-station level, prior to extending the framework to more complex spatial models or multivariate configurations. From an academic perspective, the results provide empirical evidence that nonlinear deep learning architectures can yield substantial improvements in predictive performance over conventional ARIMA-family models, even for climate time series characterized by strong seasonality and trend components.

From a practical standpoint, national meteorological service agencies such as the Korea Meteorological Administration already provide high-resolution forecasts through numerical weather prediction (NWP) models. Nevertheless, discrepancies between forecasts and actual observations frequently arise due to local microclimate effects, complex terrain, and urban heat island phenomena. Data-driven deep learning forecasters such as those examined in this study have the potential to serve as post-processing modules that correct or refine NWP-based forecasts. For example, hybrid architectures that jointly incorporate ASOS observations at specific locations and outputs from NWP models could be designed to provide short-term correction models for multiple meteorological variables, including temperature, precipitation, and wind speed. Such research directions are closely linked to practical applications in urban-scale energy management, heatwave and cold-wave warning systems, and weather-sensitive services in the health and welfare sectors.

In addition, the use of Auto-HPO pipelines such as NeuralForecast contributes to improved research productivity and reproducibility. In this study, AutoNHITS, AutoTFT, and AutoPatchTST enabled the exploration of diverse hyperparameter configurations with relatively minor code modifications, allowing researchers to achieve competitive performance without excessive effort devoted to model implementation details. This approach lowers the barrier to conducting empirical experiments in domestic meteorological data science and provides a shared experimental framework through which results can be systematically compared and accumulated across different institutions and research groups.

5.3 Limitations and Future Research Directions

Despite its contributions, this study has several limitations that should be acknowledged, along with corresponding directions for future research. First, the analysis is limited to a single observation site, namely the Seoul station (STN 108), and the generalizability of the findings to other locations or climatic regimes has not yet been verified. Future studies should replicate the same experimental design across stations representing diverse climatic characteristics, such as inland cities, coastal regions, and mountainous areas, to examine whether models such as AutoTFT maintain consistent performance advantages. Furthermore, a natural extension of this work would involve designing models that incorporate spatial information across nationwide ASOS stations, including combinations with Graph Neural Networks (GNNs), ConvLSTM architectures, or spatiotemporal Transformers.

Second, although a univariate forecasting framework was adopted in this study to ensure clarity in structural model comparison, air temperature in reality is influenced by a wide range of factors, including humidity, wind speed, atmospheric pressure, surface temperature, upper-level pressure systems, and large-scale climate indices. Since AutoTFT is inherently designed to handle multivariate and multi-source inputs, future research should extend the current framework to multivariate forecasting settings that incorporate such potential covariates. This would allow for an assessment of whether the performance gains observed under the univariate setting are preserved or further enhanced in more realistic multivariate environments.

Third, the results presented in this study are based on a single 72-hour forecast window spanning November 19 to November 21, 2025. To draw more robust conclusions, future work should adopt rolling backtesting strategies that cover multiple seasons and a wider range of meteorological conditions after 2023. For example, repeated forecasts over multiple sliding 72-hour windows could be used to analyze average performance and variability across time. In particular, evaluating model performance under extreme events—such as heatwaves, cold spells, and abrupt temperature transitions—would constitute a critical validation step for potential operational deployment in meteorological services.

Fourth, all models considered in this study provide point forecasts only and do not quantify predictive uncertainty. Given the nature of meteorological forecasting, probabilistic predictions that provide forecast intervals or full predictive distributions are often more informative than single-point estimates. Future research should therefore explore extensions incorporating quantile-based loss functions, ensemble-based uncertainty estimation, or Bayesian deep learning approaches, with the aim of quantifying uncertainty in temperature forecasts and leveraging such information for weather warning systems and decision support.

While recognizing these limitations, this study nonetheless holds both academic and practical significance in demonstrating that state-of-the-art deep learning time series models, including AutoTFT, can deliver substantial performance improvements over traditional statistical models and climatology-based baselines for the 72-hour air temperature forecasting problem in Seoul using KMA ASOS data.

6. Conclusion

This study compared and evaluated the temperature forecasting performance of recent deep learning–based time series models—AutoNHITS, AutoTFT, and AutoPatchTST—using hourly meteorological observation data provided by the Korea Meteorological Administration, against a traditional statistical model (SARIMA) and a climatology-based baseline.

The results show that AutoTFT achieved the best overall performance, with an MSE of 4.39, an RMSE of 2.10, an MAE of 1.65, and an R^2 of 0.775, corresponding to an error reduction of approximately 54% in terms of RMSE compared with SARIMA. AutoNHITS also significantly outperformed the statistical model, recording an RMSE of 2.61. These findings demonstrate that the Temporal Fusion Transformer architecture, which combines an LSTM-based sequence encoder with multi-head attention, is capable of effectively learning the nonlinear relationships and seasonal structures inherent in air temperature time series.

The results of this study provide practical insights for various application domains that require accurate meteorological forecasting, such as smart grid energy demand management and the prevention of cold damage to agricultural crops. Future research is expected to extend this work toward multivariate forecasting models that integrate multiple meteorological variables and multi-station data, as well as toward probabilistic forecasting approaches that enable quantitative uncertainty estimation.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008, 2017.

[2] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," *arXiv* preprint arXiv:2012.07436 [cs.LG], submitted Dec. 14, 2020, last revised Mar. 28, 2021. doi: 10.48550/arXiv.2012.07436.

[3] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting," arXiv preprint arXiv:2106.13008 [cs.LG], submitted Jun. 24, 2021, last revised Jan. 7, 2022. doi: 10.48550/arXiv.2106.13008.

[4] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers (PatchTST)," arXiv preprint arXiv:2211.14730 [cs.LG], submitted Nov. 27, 2022, last revised Mar. 5, 2023. doi: 10.48550/arXiv.2211.14730.

[5] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting," International Journal of Forecasting, vol. 37, no. 4, pp. 1748–1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.

[6] Korea Meteorological Administration (KMA), "Surface Synoptic Observation Data (ASOS)," dataset.

[7] C. Challu, K. G. Olivares, B. N. Oreshkin, F. Garza, M. Mergenthaler-Canseco, and A. Dubrawski, "N-HiTS: Neural Hierarchical Interpolation for Time Series Forecasting," *arXiv* preprint arXiv:2201.12886 [cs.LG], submitted Jan. 30, 2022, last revised Nov. 29, 2022 (v6), (accepted at AAAI-23). doi: 10.48550/arXiv.2201.12886.

[8] Yoon, J., & Jeon, M. (2017-11-24). Temperature Forecasting Model by using Deep Learning Technology based on LSTM. Conference of the Institute of Electronics and Information Engineers (IEIE), Incheon, Republic of Korea.

[9] Hong, S., Kim, J. H., Choi, D. S., & Baek, K. (2021). Development of Surface Weather Forecast Model by using LSTM Machine Learning Method. Atmosphere, 31(1), 73–83. https://doi.org/10.14191/ATMOS.2021.31.1.073

[10] Lim, J., Cho, D., Yoo, C., & Lee, Y. (2020, October 28). Improving next-day urban maximum temperature forecasts in a deep learning–based local forecasting model. Proceedings of the Korean Meteorological Society Conference, Seoul, Korea.

[11] Cho, D., Lim, J., & Jeong, S. (2023, November 1). Spatial refinement of summer short-term

temperature forecasts by integrating deep learning and spatial interpolation methods. Proceedings of the Korean Meteorological Society Conference, Busan, Korea.

Author Biography

Hyun Koo Lee

Master of Engineering in AI Big Data, aSSIST (Seoul School of Integrated Sciences & Technologies)

RESEARCH INTERESTS
- Large Language Models (LLMs)
- Transformers
- Time Series Analysis

E-mail: hyunkoo.lee@stud.assist.ac.kr

Nak Hyun Jung

Visting Professor, Seoul School of Integrated Sciences & Technologies (aSSIST)
Research Professor, The Institute for Industrial Policy Studies (IPS)
Head of Process Management Team, Mirae Asset Securities
Ph.D. in Business Administration, aSSIST University
Research Interests: AI, Finance, Time Series, Business Strategy
E-mail: nhjung@assist.ac.kr