https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Transferable Hourly Ozone Forecasting with Transformers

Sindhu Vasireddy[1], Michael Langguth[2], and Martin Schultz[1]

[1]Jülich Supercomputing Centre, Forschungszentrum Jülich GmBH, Jülich, Germany
[2]Now at Quadra Energy, Dusseldorf, Germany, Previously Jülich Supercomputing Centre, Forschungszentrum Jülich GmBH, Jülich, Germany

Correspondence: Sindhu Vasireddy (s.vasireddy@fz-juelich.de)

Abstract. We investigate the suitability of a transformer-based approach for air-quality forecasting, focusing on 4-day ahead hourly predictions of surface ozone (O_3). The study employs Google's Temporal Fusion Transformer (TFT) to integrate meteorological predictors, historical pollutant observations, and static station metadata, using an open source implementation with minimal domain-specific preprocessing. The analysis addresses two questions: (1) how efficiently a transformer model can be deployed for regional air quality forecasting, and (2) how well the learned representations transfer across geophysically distinct regions.

Model performance is evaluated against state-of-the-art regional chemical transport model Copernicus Atmosphere Monitoring Service (CAMS) ensemble forecast using observations from Germany. The TFT consistently achieves lower bias and higher forecast skill across all lead times. Suburban monitoring sites exhibit the highest skill relative to CAMS based on RMSE and SMAPE-based metrics. Urban stations show moderate skill against CAMS baseline, while rural stations have reduced skill in comparison but remain positive across the full 96 h forecast, with the strongest improvements observed at shorter lead times. Post–day-1 results indicate a clear separation of performance by station type; suggesting increasing performance stratification by station type beyond day 1, with larger relative gains at urban and suburban sites and smaller but consistently positive skill at rural locations.

Geographic transferability is assessed by adapting a model trained over Germany to South Korea by retraining region-specific metadata embeddings while preserving learned temporal representations. Forecast errors increase by only 5-10%, indicating that the model captures meteorological drivers of O_3 variability that generalize across contrasting anthropogenic and climatic regimes. Ablation experiments further demonstrate the robustness of the chosen experimental configuration for both forecasting performance and cross region transferability.

Keywords: TFT, hourly forecasting, O_3 air quality, transfer learning, skill scores, meteorology, CAMS, ablation study

1 Introduction

Hourly surface ozone (O_3) forecasts are essential for environmental monitoring (Manisalidis et al., 2020), public health (Keswani et al., 2022), and policy planning (Gupta and Kumar, 2023). Operational Air Quality forecasting systems rely primarily on numerical chemical transport models (CTM), which are computationally expensive and typically run at relatively coarse spatial resolution. State-of-the-art systems, such as the CAMS (Flemming et al., 2022), provide continuous regional forecasts

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

using an ensemble of CTMs, yet these forecasts exhibit systemic biases, particularly at longer lead times and finer spatial scales, as documented in several evaluation studies (Bertrand et al., 2023; Riccio and Chianese, 2024; Shetty et al., 2025). Bias correction mechanisms (ECMWFCode4Earth, 2025) are planned to be included in CAMS final operational predictions (ECMWF), but in this study we use the publicly available uncorrected ensemble output[1] from Copernicus Climate Data Store (CDS) as our baseline for model comparison.

Despite their physically grounded formulation, operational CTM-based forecasts often show reduced skill beyond short lead times and limited sensitivity to exogenous anthropogenic drivers such as emissions, land use, and population density. Recent assessments by the Barcelona Supercomputing Centre (Guevara et al., 2025) highlight the importance of explicitly incorporating anthropogenic emission sources (e.g. wood combustion, road transport) for major air pollutants, including NO_2, O_3, $PM_{2.5}$, and PM_{10}, and recommend its inclusion in future CAMS ensemble developments to improve model sensitivity towards such aspects.

Motivated by these limitations, several studies (Bertrand et al., 2022; Mujtaba et al., 2025; Bodendorfer, 2025; Leufen et al., 2021a, 2022b, 2023b) have demonstrated successful air quality forecasting with machine learning (ML) models. These ML approaches have demonstrated that past pollutant concentrations contribute significantly to ozone forecast skill (Leufen et al., 2021a, 2022b, 2023b). Inclusion of past and future meteorological factors has also been proven to improve ozone forecasts in (Leufen et al., 2021a, 2022b), as meteorological variables are often more reliably predicted than pollutant concentrations themselves. Most existing ML-based air-quality models (Leufen et al., 2023a; Bodendorfer, 2025), however, rely on custom architectures, limited feature sets and most do not provide hourly resolution of forecasts. Their formulations also do not fully exploit cross-variable dependencies between meteorology, chemistry, and anthropogenic influences.

Recent advances in attention-based architectures, particularly transformer models, have enabled general-purpose time-series forecasting frameworks capable of learning complex temporal dependencies with minimal task-specific feature engineering (Vaswani et al., 2017; Brauwers and Frasincar, 2023; Choromanski et al., 2020). These models have shown robust performance across diverse application domains such as energy, finance, and healthcare, and are increasingly being explored for air-quality prediction. Existing transformer-based air-quality studies, focused on auto-regressive or uni-variate formulations (Bodendorfer, 2025) for short-range pollutant forecasts, have demonstrated ability of augmenting attention mechanisms to outperform other approaches like CTMs. However, they lack in systematic inclusion of anthropogenic factors (Zheng et al., 2022; Hickman et al., 2023), and evaluations of transferability across geophysically distinct regions on which they have been trained. The current study addresses this gap and also evaluates the trained model as a reliable source for probabilistic forecast of Ozone helpful in capturing uncertainty, needed for operation.

In this work, we assess whether a task-agnostic transformer architecture can serve as a practical and transferable alternative to physics-based operational models for hourly ozone forecasting. Specifically, we evaluate the TFT (Lim et al., 2020) for 4-day-ahead hourly surface ozone prediction using a multivariate time-series formulation that integrates meteorological inputs, historical pollutant concentrations, and static station-level metadata representing anthropogenic influences. The model is im-

[1] median forecast is considered best model forecast estimate from ensemble as per ECMWF Website (European Centre for Medium-Range Weather Forecasts and signed on 22/07/2021)., 2025)

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

plemented using an open-source framework (Beitner, 2020) with minimal domain-specific preprocessing and is trained using a probabilistic loss to quantify forecast uncertainty.

Observational ozone data are obtained from the Tropospheric Ozone Assessment Report Phase II (TOAR-II) database (Schultz et al., 2017), including all available measurements since 2014 and associated station metadata. TOAR-II provides accessible data from quality controlled data sources of tropospheric ozone measurement sites around the world and does not include any uncontrolled sources e.g. from OpenAQ. Static metadata from TOAR-II, such as population density, land-use characteristics, and emissions proxies, are incorporated to account for human influences on ozone variability. The forecasting system is formulated as a station-based time-series model rather than a three-dimensional geo-spatial model, allowing the use of mixed observational and simulated inputs while maintaining compatibility with operational forecasting workflows.

Model performance is evaluated against CAMS regional ensemble forecasts over Germany, following established benchmarking practices (Leufen et al., 2022a, 2023a), and is further assessed through a transfer-learning experiment in which the trained model is adapted to South Korea by retraining only region-specific metadata embeddings. This setting reflects a realistic operational scenario with limited observational history (less than 3 years) and incomplete metadata availability (also missing past NO_2 data). Additional ablation experiments investigate the role of static metadata, future meteorological inputs, and context window length in controlling forecast skill as detailed in Appendix D.

The main contributions of this study are as follows:

1. We present the first systematic benchmark of a general-purpose transformer model against an operational CTM-based ensemble - the regional Copernicus Atmosphere Monitoring Service (CAMS) system, for hourly surface ozone forecasting.
2. We formulate ozone prediction as a probabilistic forecasting task, producing distributions of air pollutant concentrations with multiple confidence intervals.
3. We demonstrate improved forecast skill through the joint integration of meteorological predictors and static station-level anthropogenic metadata.
4. We evaluate cross-regional transferability (Germany $\rightarrow$ South Korea) under data-sparse conditions, demonstrating robust generalization across contrasting climatic and emission regimes.

2 Related Works

Time-series forecasting methods are commonly categorized as univariate or multivariate. Many recent high-performing architectures, such as Lag-LLama (Rasul et al., 2024) and PatchTST (Nie et al., 2023), address multivariate problems through a *multi-univariate* formulation, in which variables are modeled independently while sharing parameters across channels. More generally, time-series models can be grouped into channel-independent, partially dependent, and fully dependent strategies (Qiu et al., 2025; Rasul et al., 2024; Liang et al., 2024; Brimos et al., 2024),implemented using architectures such as Transformers, MLPs, CNNs, GNNs, and recurrent networks (Leufen et al., 2021b; Lim et al., 2020; Brimos et al., 2024; Yunita et al.,

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

2025). Attention-based models have become dominant due to their ability to capture long-range temporal dependencies and to condition naturally on known future inputs.

Large-scale pretraining and transfer learning have further improved robustness and generalization in time-series forecasting, particularly in data-limited settings (Kottapalli et al., 2025). *Foundation models*, have demonstrated strong performance across domains such as traffic analysis (Brimos et al., 2024), finance, and healthcare. Transfer learning and parameter-efficient adaptation methods have further enabled cross-domain generalization in time-series settings (Nie et al., 2024). This paradigm has also been applied to spatiotemporal Earth system modeling, as illustrated by Aurora (Wu et al., 2025) , which leverages pretrained attention-based architectures for global weather and air-pollution prediction.

Within this context, Transformer models have increasingly been adopted for air-quality forecasting, including ozone prediction. In contrast to physics-based operational systems based on Numerical Weather Prediction (NWP) and chemical transport models (CTM)(Grell and Baklanov, 2011; Lang et al., 2023; Flemming et al., 2022) driven by the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS), data-driven approaches provide a flexible alternative for AQ prediction. Approaches such as AirFormer (Zheng et al., 2022) introduce spatiotemporal attention mechanisms for regional prediction, demonstrating competitive performance relative to traditional baselines. Other studies (Ibrahim, 2026; Hickman et al., 2023; Liang et al., 2022), explore attention-based time-series models for ozone forecasting across Europe, highlighting their potential compared to classical machine learning methods.

These developments motivate the use of attention-based, pretrained time-series models for probabilistic, station-level air-quality forecasting that integrate historical pollutant observations, known future meteorology, and static contextual information.

2.0.1 Established Baselines:

Previous work from our team, including MLAir (Leufen et al., 2021b), IntelliO3-ts (Kleinert et al., 2021), and O3ResNet (Leufen et al., 2021a, 2022a, 2023a), has established several empirical findings that guide the present study:

1. The inclusion of past meteorological predictors enables skillful AQ forecasts up to at least day 2.
2. Past ozone concentrations and temperature are dominant predictors of future ozone variability.
3. Conditioning on forecast meteorology, as known future inputs, substantially improves prediction.

2.0.2 Chosen Model:

We adopt a general-purpose TFT (Lim et al., 2020) as the forecasting model, in the current study, due to its suitability for interpretable, multivariate, multi-horizon time-series prediction.

This model integrates several architectural components such as variable selection networks, gated residual connections, and interpretable multi-head attention, enabling the model to learn both short- and long-range temporal dependencies while providing insights into variable importance. These components are particularly relevant for air-quality forecasting.[2].

[2]Feature importance is explained from the perspective of the AQ forecasting domain, rather than from general timeseries forecasting, as it is already done in the TFT publication (Lim et al., 2020)

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

The architecture (as explained in Appendix B) combines LSTM-based encoder decoder components for capturing local temporal dynamics with attention mechanisms that integrate information across the full forecast horizon. Importantly, probabilistic forecasting via quantile regression,is performed rather than point estimates, allowing an explicit representation of predictive uncertainty.

3 Data and Methodology

In our formulation, the model predicts future surface ozone (O_3) concentrations using historical pollutant observations (O_3, NO, NO_2), meteorological predictors (past and future), and static station-level anthropogenic metadata.

Unlike prior Transformer-based AQ studies (Hickman et al., 2023), we incorporate station-level anthropogenic metadata alongside meteorological and chemical predictors, benchmark performance directly against CAMS ensemble median forecasts, and assess robustness through cross-regional transfer from Germany to South Korea. This transfer setting, which involves distinct climatic regimes, emission patterns, and data availability constraints, remains underexplored in the existing literature.

The two central objectives addressed in this study are:

(i) evaluating whether a general-purpose transformer can be deployed for hourly ozone forecasting with minimal domain-specific tuning, and

(ii) assessing its ability to transfer across geophysically distinct regions—we designed two complementary experiments.

The first experiment focuses on 4-day-ahead hourly surface ozone forecasting over Germany using historical pollutant observations, past and future meteorological predictors, and static station-level anthropogenic metadata. The second experiment evaluates geographic transferability by adapting the model trained on Germany to South Korea through retraining of region-specific static metadata while preserving the learned meteorology–chemistry relationships. Feature engineering was intentionally restricted to basic meteorological covariates and static regional descriptors. This ensures that performance gains can be attributed to the task-agnostic architecture rather than domain-specific tuning.

3.1 Data Sources and Preprocessing

3.1.1 Data Acquisition

Hourly surface ozone observations were obtained from the Tropospheric Ozone Assessment Report Phase II (TOAR-II) database (Schultz et al., 2017), alongside collocated measurements of NO and NO_2. Meteorological predictors were sourced from the ERA5 reanalysis from 1990–2023 and interpolated to station locations using nearest-neighbor matching.

For benchmarking against operational forecasts, near-real-time ozone predictions were retrieved from the CAMS regional ensemble via the Copernicus Climate Data Store (CDS) API. The forecast horizon was set to 96 h, consistent with operational CAMS products and prior evaluation studies (Leufen et al., 2022a).

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

3.1.2 Preprocessing and Sample Construction

Data preparation followed a standardized and reproducible pipeline. Missing values in pollutant time series were forward-filled for gaps up to 6 h; samples with longer gaps were excluded. No interpolation was applied to meteorological inputs and static metadata.

Each station time series was segmented into overlapping samples comprising a fixed-length encoder window of 336 h (14 days), followed by a 96 h forecast horizon. [3] All variables were standardized using Z score normalization. Normalization was applied independently per variable to ensure numerical stability and comparability across stations during optimization.

3.1.3 Input-Output Configuration

The model ingested three categories of inputs:

(i) known past covariates, consisting of two weeks of historical pollutant concentrations (O_3, NO, NO_2) and meteorological variables (temperature, surface pressure, relative humidity , zonal and meridional winds, Planetary Boundary Layer height, and cloud cover);

(ii) known future covariates, comprising meteorological variables available for the subsequent 4 days; and

(iii) static covariates, including station-level anthropogenic metadata and spatiotemporal descriptors.

The model produced probabilistic 4-day-ahead hourly ozone forecasts, producing probabilistic forecasts, including the median and three prediction intervals of 96%, 80% and 50% (seven quantiles in total - 0.02, 0.10, 0.25, 0.50, 0.75, 0.90, 0.98), as well as the corresponding attention weights.

A complete list of predictor variables is provided in Appendix A.

3.2 Regional Datasets

3.2.1 German Dataset

The German dataset spans 1 January 1990 (00:00 UTC) to 31 December 2022 (23:00 UTC) for 493 monitoring stations. Out of the chosen stations 386 were used in training and rest unseen by the model is used for testing the trained model. The stations were chosen for the study to cover the full spatial region and also such that they are equally balanced across different types of locations such as rural, urban and sub-urban.

After preprocessing, the dataset comprised approximately 44.1 million training samples due to the use of overlapping sliding windows and 15.3 million validation samples. Samples were constructed using a total context length of 432 h (336 h history + 96 h forecast horizon) and trained with a batch size of 128.

Past data since 1990 till current date is made available in the TOAR-II database collected and updated continuously from observation stations. Temporal generalization was assessed using a time-based train-val-test split (75:22:3) as detailed below, with a limited inference presented for summer 2023 data to probe model robustness during high-variability photochemical

[3] Forecast horizon ablation study results are presented in Appendix D

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

regimes of O_3. From the 33-year period (1990–2022), 75% of the data were used for training and the remaining portion for validation. Specifically, the training set spans 25.25 years from 1 January 1990 to 24 March 2015[4], while the validation set covers 7.75 years from 1 April 2015 to 31 December 2022. The test set corresponds to summer 2023, from 11 June to 26 September[5].

The test set consisted exclusively of German monitoring stations unseen during training, enabling the evaluation of spatial generalization across rural, suburban, and urban site types.

To ensure a fair baseline comparison, CAMS ensemble forecasts were matched to TOAR stations using nearest-grid-point selection; the maximum horizontal displacement was less than 5 km.

3.2.2 Korean Dataset (Transfer Learning)

For cross-regional transfer experiments, a secondary dataset covering South Korea was constructed using TOAR. Only ozone and NO_2 observations were present for the region, no NO pollutant concentration data was available. This was merged with ERA5 meteorological covariates along with static metadata. The dataset spans January 2020 to August 2022, yielding approximately one million hourly samples.

Due to data completeness constraints, only 52 stations with fully aligned pollutant and meteorological time series were retained from available TOAR-II stations.

The dataset was partitioned into a training set covering January 2020 to December 2021, a validation set from January 2022 to April 2022, and a test set spanning May 2022 to August 2022.

The temporal configuration (336 h context+96 h forecast) and model architecture were kept identical to the German setup to enable controlled comparison. Evaluation, therefore, probes temporal generalization within Korea and cross-regional adaptation of the pretrained German model under limited observational data availability.

3.3 Model pipeline

No extensive domain-specific feature engineering or chemical transformations beyond below two spatiotemporal features are introduced, ensuring that performance gains can be attributed to the task-agnostic architecture rather than problem-specific tuning.

Two spatiotemporal features are :

(i) sample grouping by station using latitude, longitude, altitude, and station code to account for the concurrent availability of pollutant and meteorological observations across locations at each timestamp. This grouping ensures that samples belonging to the same station are treated jointly for relationship learning, alongside static covariates that are fixed at the station level.

[4]The last date indicated is 5 days before end of month based on the last available complete sample date

[5]Although trained model's Spring and winter months of 2023 performance is same as CAMS due to limited variability there is no significant variation to study or discuss in this paper.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

(ii) A relative time index (initialized to 0 at the first available timestamp for each station and incremented by 1 at each subsequent hourly timestep) is also introduced to explicitly indicate the timestep relative to the data available in that particular station sample and encourage cross-location, time awareness. Although this is implicitly covered by the recurrent nature of encoders and decoders, inclusion of relative index has been observed as an advantage in most of the multi-horizon sequence to sequence modeling cases.

3.3.1 Training Strategy

3.3.2 Data Sampling

We trained our TFT from scratch on a mixed-station dataset comprising 386 German monitoring stations, balanced across rural (143), suburban (131), and urban (112) sites rather than randomly using all available observations from a region. Training distribution was deliberately chosen to understand AQ pollutant characteristics for each of anthropogenic context of the area and to estimate model sensitivity towards them rather than random stations within a geospatial boundary.

Additionally 107 stations, unseen during training, were reserved exclusively for inference to evaluate out-of-sample spatial generalization across station types. Each sample was generated using a sliding window with a stride of 1 h, consisting of a 432 h (14-day/336h historical context and a 4-day/96h forecast horizon). This configuration aligns with synoptic-scale meteorological variability and enables direct comparison with operational 96 h CAMS forecasts and it is in agreement with established works such as (Leufen et al., 2023a). Additional ablation experiments with shorter context windows (168 h and 240 h) are described in Appendix D.

Static embeddings representing station-level metadata, including geographic attributes, land use, and emissions proxies were incorporated as auxiliary inputs to condition forecasts on the local anthropogenic context. All variables were standardized using statistics computed from the training set, and samples containing gaps were discarded as described in Section 3.1.

3.3.3 Optimization and Training Configuration

The model architecture was used as provided, with all model weights initialized randomly and trained from scratch, without any pretrained components. Training was implemented using the PyTorch Forecasting framework (Beitner, 2020) with a modified dataset handler that excludes incomplete samples rather than interpolating missing values. This modification is provided as part of the released source code as a custom wrapper package.

The model was trained using quantile loss with seven quantiles as detailed in Section 3.1.3, enabling probabilistic forecasting. Optimization employed the Ranger optimizer, combining RAdam and LookAhead (Tong et al., 2019), with early stopping based on validation loss (patience = 10, tolerance = 10^{-5}). Hyperparameters, including learning rate, were tuned using Optuna and presented in Appendix E.

Losses were monitored using TensorBoard to verify convergence stability and consistency across metrics. Distributed data-parallel training was used to accommodate the large dataset; implementation details and hardware specifications are provided in the Appendix C and Source code repository information is also provided for reproducibility under 6.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

3.3.4 Transfer Learning Strategy

To assess cross-regional transferability, the pretrained German model was adapted to South Korea using transfer learning. Model weights were initialized from the checkpoint, with gating layers selectively unfrozen to evaluate adaptation under limited data availability.

The Korean dataset was with limited pollutant concentrations for shorter duration as described earlier, hence, the corresponding pretrained inputs and parameters were discarded accordingly. All available stations were also classified as urban on TOAR-II database by the provider, resulting in limited categorical variability; class embeddings were retained only for architectural consistency[6].

Static metadata embeddings and associated gating layers were reinitialized and retrained to capture region-specific anthropogenic and geographic characteristics. In addition, attention output layers were unfrozen to allow adaptation of temporal dependencies and pollutant response dynamics. Other layers remained frozen to preserve learned meteorology and chemistry relationships.

Due to this limited station availability, evaluation was also focused on temporal generalization within the Korean domain rather than stations unseen during training.

Training employed the same loss function, optimizer, and early stopping criteria as the German experiment. Convergence was achieved rapidly, with early stopping triggered after 11 epochs, indicating efficient adaptation under constrained data conditions. As no regional CAMS forecasts are available for Korea, evaluation is reported using absolute error metrics only.

4 Results and Evaluation

Evaluation follows the benchmarking methodology of (Leufen et al., 2022a), ensuring consistency with established air-quality forecast assessment practices as described below, followed by results for the German forecasting experiment, and cross-regional transfer to Korea. Results are presented using hourly time series plots with uncertainty, stratified skill score plots, spatial maps, and interpretability analyses (variable importance and attention), with ablation studies reported in Appendix D.2.

4.1 Evaluation Metrics

We report deterministic error metrics, including root mean squared error (RMSE) and symmetric mean absolute percentage error (SMAPE), together with probabilistic performance assessed using quantile loss. Quantile loss is evaluated at seven quantile levels, including the median (0.50), and summarized using prediction intervals corresponding to 50% (0.25-0.75), 80% (0.10–0.90), and 95% (0.02–0.98) confidence levels.

[6]This design choice was made to preserve model compatibility for future cross-regional or mixed-domain experiments if full categories are available in such cases.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

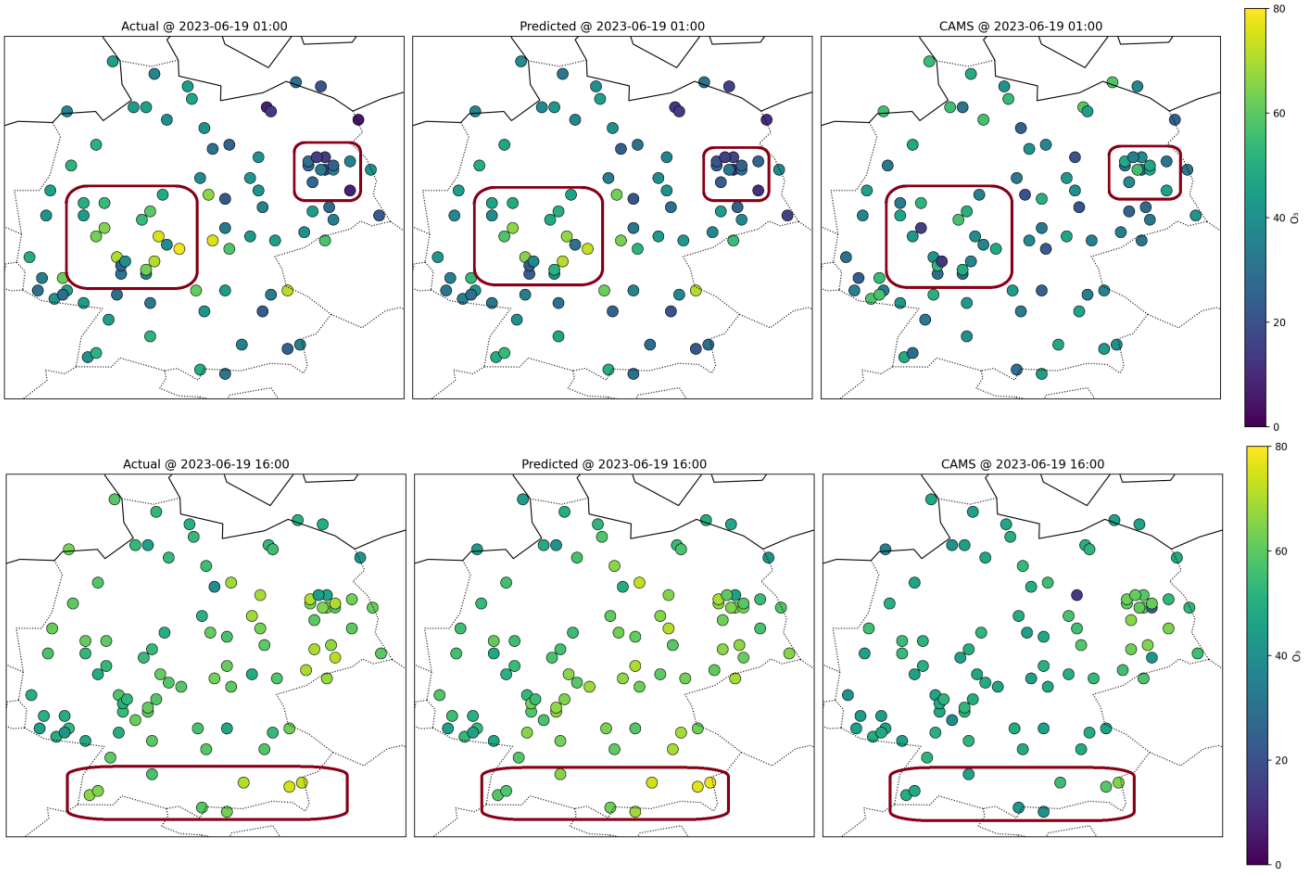

Figure 1. Map visualization of two selected timestamps comparing actual vs TFT vs CAMS prediction. A compilation of all time steps for the inference period is presented plots folder of the source code repository.

To assess relative performance against the operational baseline, skill scores with respect to CAMS are calculated as per equation 1:

$$\text{SkillScore}(M) = 1 - \frac{M_{\text{TFT}}}{M_{\text{CAMS}}}, \tag{1}$$

where M denotes a given error metric. Positive skill score indicates improvement relative to CAMS, whereas negative values indicate degraded performance.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Metrics are further stratified by station category $C \in \{\text{rural}, \text{suburban}, \text{urban}\}$ and aggregated by forecast day. Hourly horizons are grouped into four 24 h blocks (0-23, 24-47, 48-71, and 72-95 h) to evaluate error growth with lead time:

$$\text{RMSE}_C(t) = \frac{1}{|I_C|} \sum_{i \in I_C} \text{RMSE}_i(t), \tag{2}$$

$$\text{RMSE}_C(d) = \frac{1}{|H_d|} \sum_{t \in H_d} \text{RMSE}_C(t), \tag{3}$$

$$\text{Skill}_C(d) = \frac{1}{|H_d|} \sum_{t \in H_d} \text{Skill}_C(t), \tag{4}$$

where H_d represents the hourly indices for day d (from $24d$ to $24d + 23$), and I_C denotes the group of stations in category C. The quantities $|H_d|$ and $|I_C|$ represent the number of hours per day and the number of stations in category C, respectively.

4.1.1 Weighted Interval Score (WIS)

To evaluate the probabilistic performance of the model, we employ the *Weighted Interval Score* (WIS), a proper scoring rule for quantile-based forecasts that jointly assesses prediction accuracy, sharpness, and calibration. WIS aggregates information from multiple central prediction intervals derived from the predicted quantiles (Bracher et al., 2022).

Let y denote the observed value and $\hat{q}_\tau$ the predicted quantile at probability level τ. For a central $(1-\alpha)$ prediction interval with lower bound l_α and upper bound u_α, the interval score is defined as

$$IS_\alpha(l_\alpha, u_\alpha; y) = (u_\alpha - l_\alpha) + \frac{2}{\alpha}(l_\alpha - y)\mathbf{1}(y < l_\alpha) + \frac{2}{\alpha}(y - u_\alpha)\mathbf{1}(y > u_\alpha), \tag{5}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. The first term rewards narrow prediction intervals, while the latter terms penalize observations that fall outside the predicted interval.

Given K central prediction intervals and the predictive median m, the Weighted Interval Score is computed as

$$WIS = \frac{1}{K + \frac{1}{2}} \left(|y - m| + \sum_{k=1}^{K} \frac{\alpha_k}{2} IS_{\alpha_k}(l_{\alpha_k}, u_{\alpha_k}; y) \right). \tag{6}$$

In this study, the probabilistic forecasts provide seven quantiles $\{0.02, 0.10, 0.25, 0.50, 0.75, 0.90, 0.98\}$, which correspond to three central prediction intervals with nominal coverages of 50%, 80%, and 96%.

The WIS is computed for each forecast lead time and averaged across stations belonging to the same station category (rural, sub-urban, and urban). The resulting WIS values are then plotted across the 96-hour prediction horizon to assess how the quality of the probabilistic forecasts evolves with lead time for each station category.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

4.2 Forecast Performance over Germany

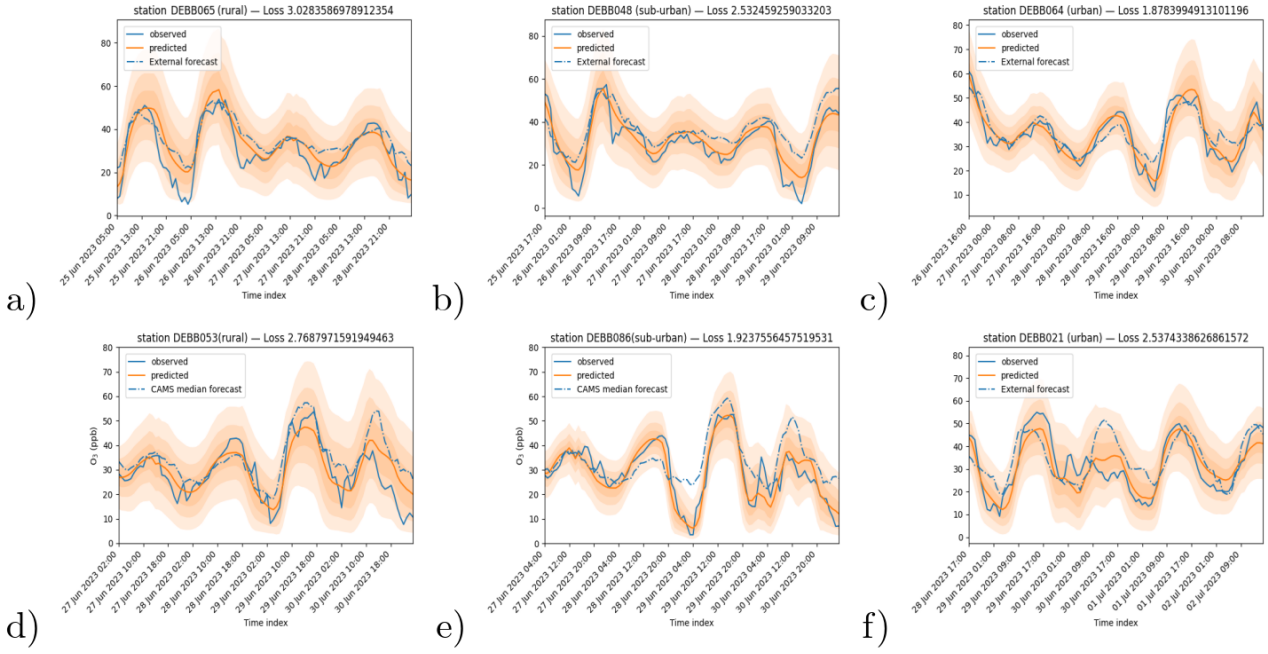

Figure 2. Plot of Actual vs TFT Prediction vs CAMS for each station type for a time period unseen by model during training or validation. subplots a,b,c are rural, suburban and urban station samples unseen during training or validation respectively subplots d,e,f are rural, suburban and urban stations seen apriori during training or validation respectively

Figure 1 shows example timesteps from summer 2023, comparing observed ozone fields with TFT predictions and CAMS forecasts. At the highlighted monitoring stations, TFT predictions more closely match observed values showing closer agreement between TFT and observations than CAMS, capturing local maxima and gradients more accurately.

Model performance varies across different station categories. Figure 2 presents hourly predictions, and their associated uncertainty. Peaks and troughs in this plot are where CAMS tends to overshoot relative to the transformer model; therefore, periods of high summer ozone variability during the inference window(2023) are shown to illustrate this behavior. This behavior is visible in Figure 2, where TFT remains closer to observations throughout the diurnal cycle rather than exhibiting performance peaks at specific hours (e.g., 00:00-03:00 UTC).

Figure 4 presents averaged skill and error metrics (RMSE and SMAPE-based) for urban, suburban, and rural stations. Averaging is performed across all stations in the test set within each category to assess generalized performance over the full summer 2023 period. Across the evaluation horizon, suburban stations exhibit the strongest performance, with RMSE

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

remaining below 2 ppb throughout. This is followed by urban stations, with a maximum RMSE of 2.6 ppb, while rural stations show smaller but generally positive skill scores relative to CAMS.

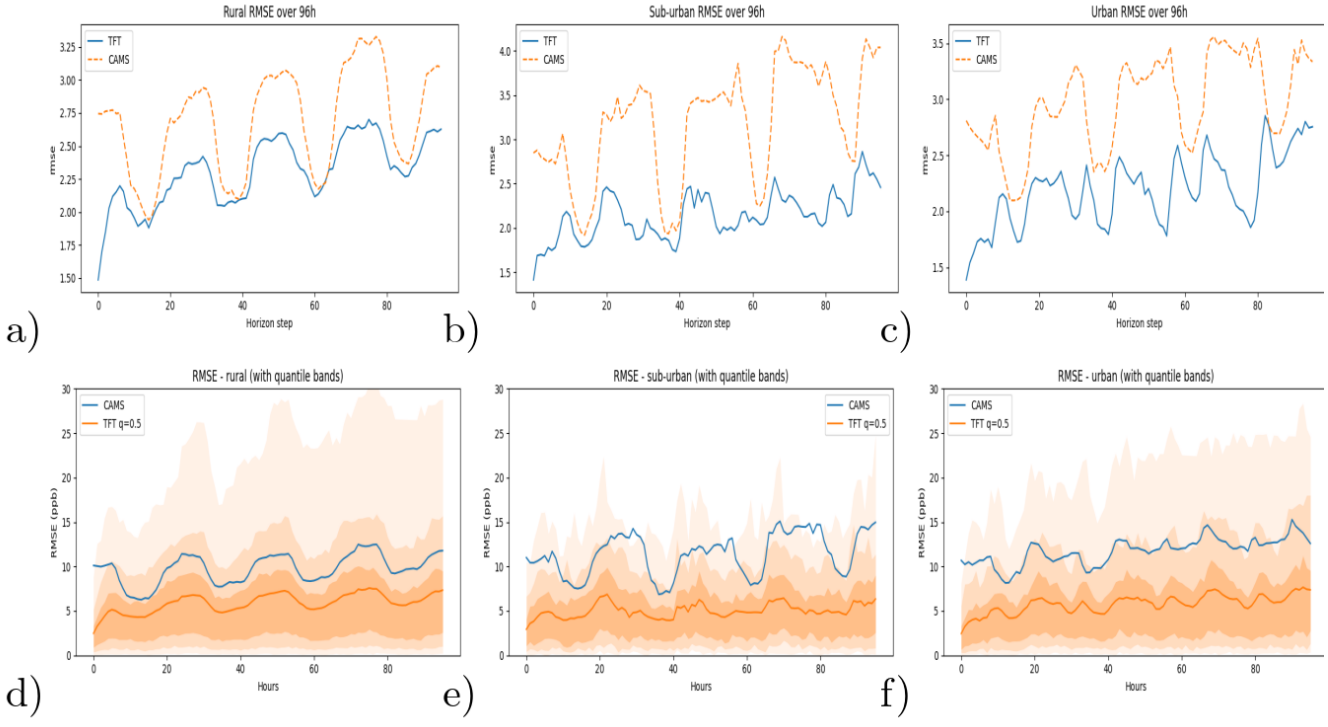

Figure 3. Subplots (a–c) show averaged RMSE across rural, suburban, and urban station categories, respectively, computed without considering the distributed forecast quantiles. Subplots (d–f) show the corresponding RMSE including errors evaluated against each forecast quantile, visualized as three shaded confidence intervals: 50% (darkest), 95%, and 98% (lightest).

Suburban and urban stations achieve approximately 35% improvement relative to CAMS, whereas rural stations show weaker gains. This behavior is consistent with the stronger influence of anthropogenic factors and static metadata at urbanized sites, which the TFT explicitly incorporates. [7] In rural regions, where ozone variability is more strongly driven by large-scale meteorology, the transformer model provides limited additional benefit over the physics-based baseline.

Day-averaged skill scores across station categories reveal clearer separation between station types than hourly skill curves. TFT maintains relatively stable error across the four forecast days, whereas CAMS exhibits lower error primarily for day-1, followed by a steady increase indicative of growing bias.

Only rural station skill score show a pronounced reduction at day-1 before recovering and converging toward suburban and urban skill levels by day 3. At these daily transition points, both models occasionally show comparable error, corresponding

[7] Although CAMS implicitly incorporate anthropological forcing, as indicated earlier, works like (Guevara et al., 2025) establishes their sensitivity to such factors is less and needs to be more explicitly included in production deployments.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

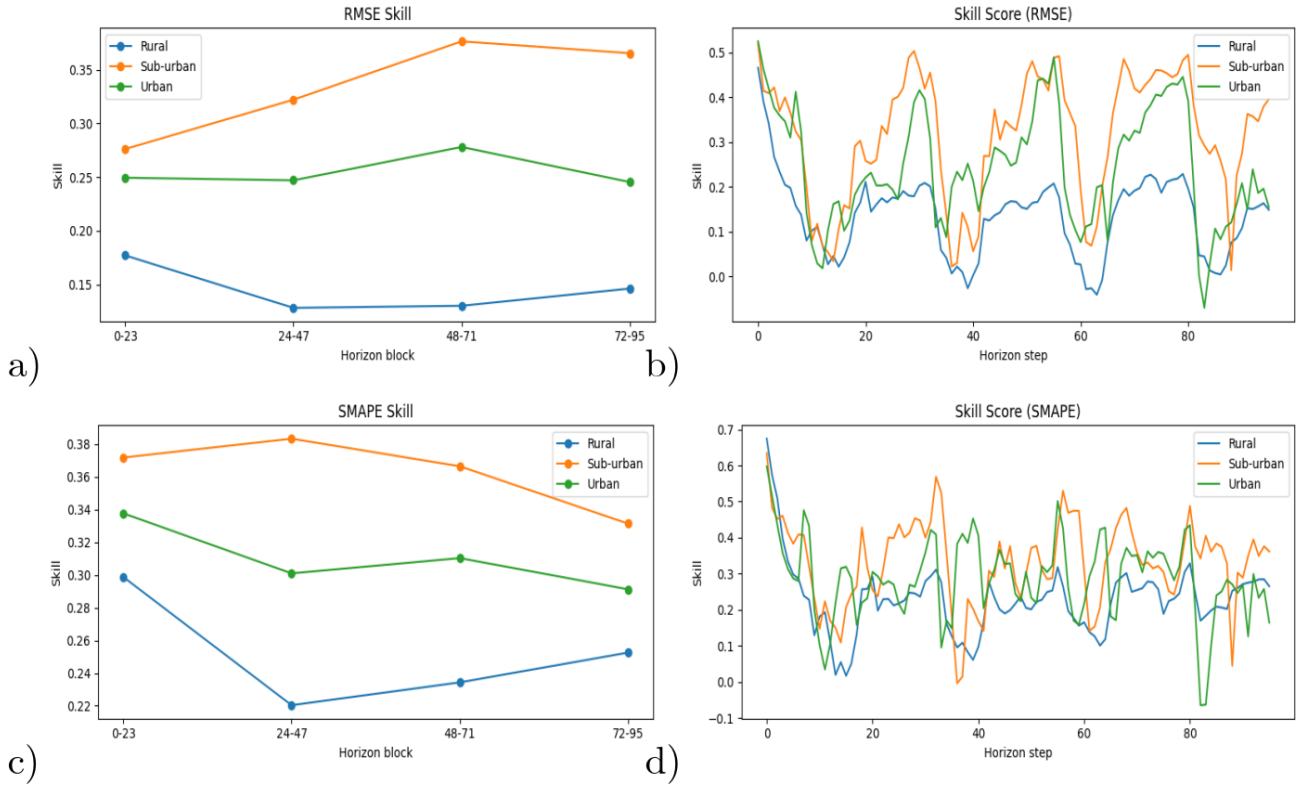

Figure 4. Forecast performance of the TFT relative to CAMS during summer 2023 (25 June–25 September). Sub plots (a-b) presents RMSE-based skill scores of TFT with respect to CAMS for stations unseen during training, shown as daily averages (a) and aggregated across the full period (b). Bottom row (c-d) shows corresponding SMAPE-based skill scores, reported as daily averages (c) and full-period aggregates (d). A positive skill score indicates improvement of TFT over CAMS.

to skill values near zero, indicating equivalent performance. Skill vs Lead time evaluation are all performed by considering forecasts with same shared initial time of 00:00 as per official CAMS 96h ensemble forecasts.

Figure 5 extends the deterministic evaluation by examining skill scores for all predicted quantiles of the TFT relative to CAMS using RMSE-based metrics. While the median prediction ($q = 0.5$) reflects the central forecast discussed previously, the additional quantiles provide insight into the behaviour of the predictive distribution. Across station categories, the central quantiles ($q = 0.25$–$q = 0.75$) exhibit the most consistent positive skill, indicating that the core predictive range of the model remains more accurate than the CAMS baseline across the forecast horizon. The outer quantiles ($q = 0.02$ and $q = 0.98$) exhibit higher variability and occasionally reduced skill, reflecting the increased uncertainty associated with extreme prediction intervals. Nevertheless, the overall structure of the quantile skill curves closely follows the behaviour of the median forecast, suggesting that the probabilistic forecasts remain coherent and that uncertainty grows in a physically consistent manner with

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

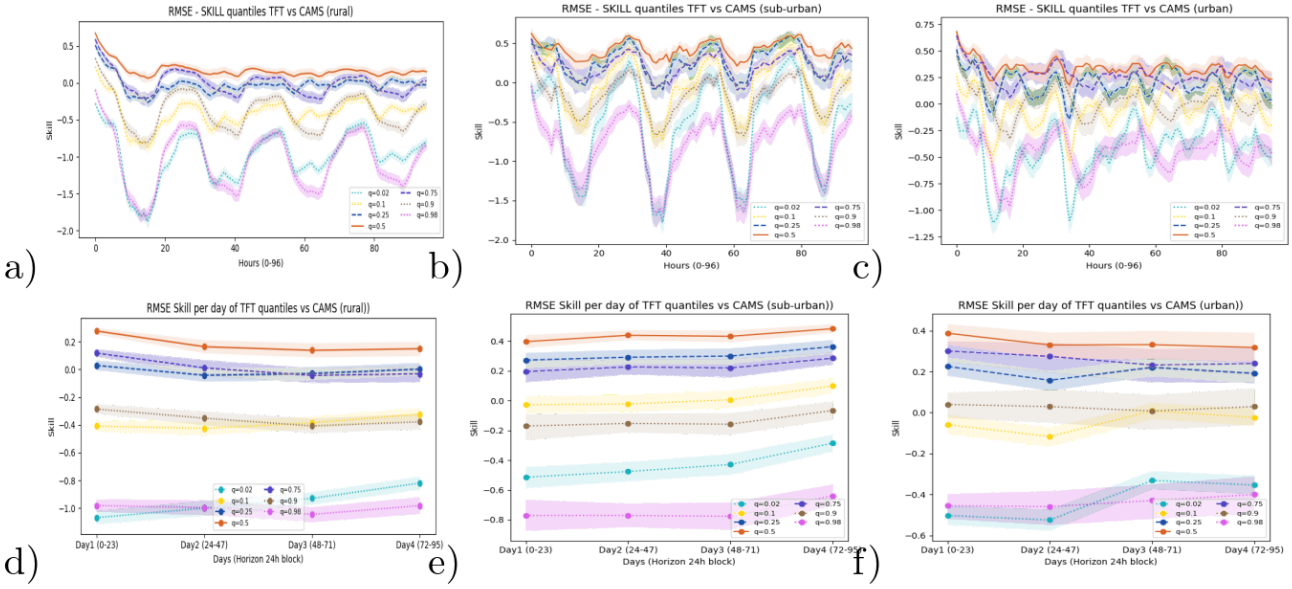

Figure 5. RMSE based Skill scores of TFT predictions for all 7 quantiles against CAMS baseline for rural, sub-urban and urban station categories are presented in subplots (a-c) respectively. Corresponding per day averaged skill scores are presented in subplots (d-f) with the most broadest confidence intervals of 95% and 98% indicated as dotted lines and closest CI of 50% presented as dashed line

lead time. When averaged by forecast day, the separation between quantiles becomes clearer, with central intervals maintaining stable skill while the outer intervals show greater spread, highlighting the increasing uncertainty envelope of the forecasts at longer lead times. Appendix D3 Section provides the corresponding distributed plots for SMAPE based skill scores and a discussion which is similar to the RMSE based skill score.

The category wise WIS, averaged over the 96-hour forecast horizon, is 4.17 ppb for rural stations, 3.46 ppb for sub-urban stations, and 4.00 ppb for urban stations. The overall mean WIS across all stations is 4.06 ppb. A horizon wide variation of WIS score across quantiles per category is also presented in Figure 8.

4.3 Transfer Learning to Korea

Transfer learning experiments were conducted by selectively freezing and retraining model components as described in Section 3.2.1. Figure 6 summarizes performance across the three transfer learning configurations.

Retraining static metadata embeddings alone captures the overall concentration range but fails to reproduce temporal dynamics. Joint retraining of metadata embeddings, attention layers, and output gating yields the best performance, as reflected in the averaged error metrics (Figures 6).

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

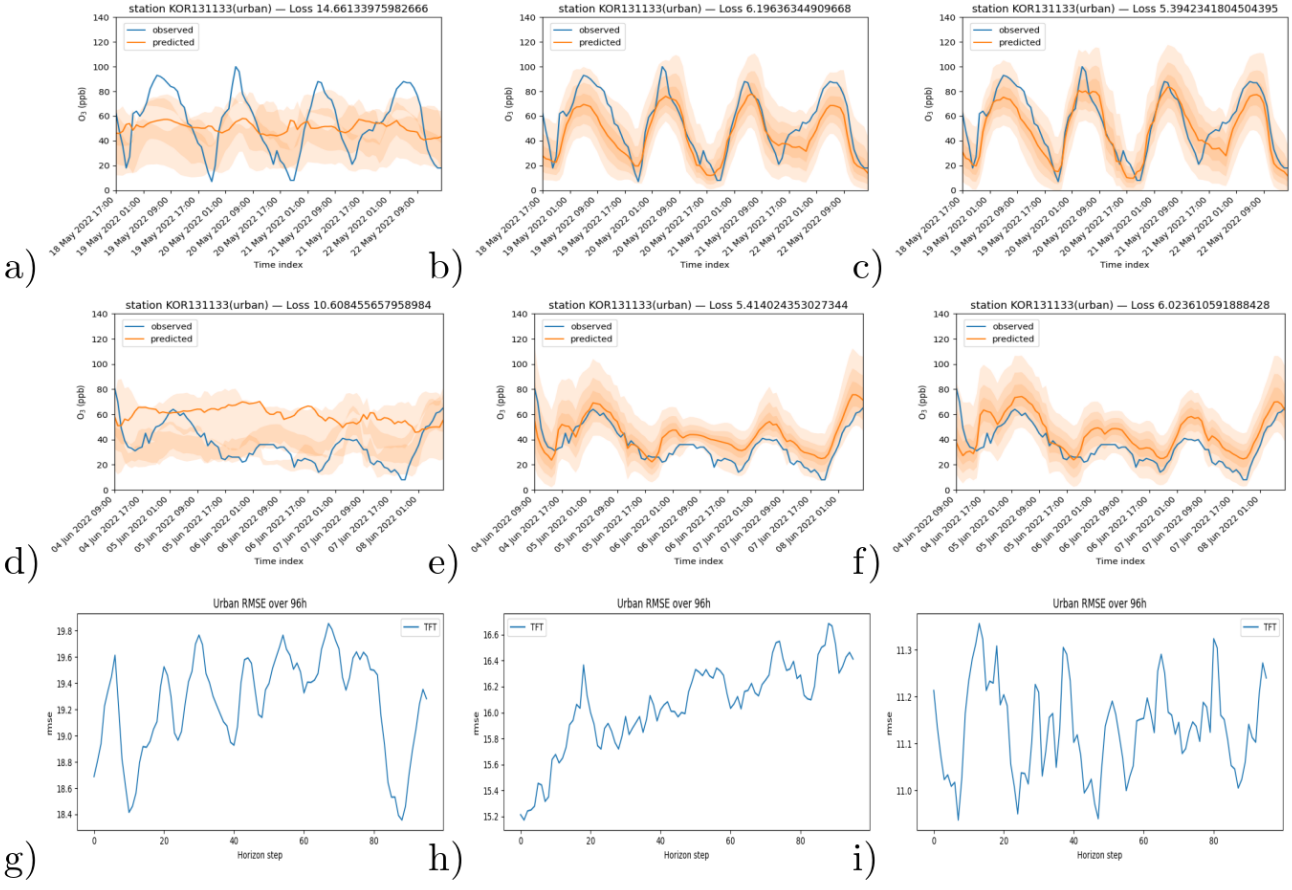

Figure 6. Transfer learning ablation results for the Korea domain adaptation experiment. Plots (a-f) show representative actual versus predicted time series for two unseen validation samples under three fine-tuning strategies: (left/plots a,d) retraining static anthropogenic and spatial metadata embeddings only, (center/plots b,e) retraining metadata embeddings and output layers with attention layers frozen, and (right/plots c,f) full fine-tuning including metadata embeddings, attention mechanisms, and output layers. Bottom row plots (g-i) reports the average RMSE corresponding to its column across all validation samples for each strategy. Results highlight the incremental benefit of attention fine-tuning for cross-regional generalization.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Although no Operational CAMS baseline is available for Korea, the transferred model achieves an RMSE of approximately 11 ppb across all stations when compared against CAMS deterministic global forecasts (Figure 6). Despite being trained on only two years of data, this indicates effective adaptation under data-limited conditions.

Figure 7 presents variable importance and attention weights for static, past, and future encoder–decoder inputs. Among static metadata, mean population density contributes the most (approximately 20%) to ozone prediction, followed by road distance and minimum relative topographic altitude (≈10% each). This indicates that a reduced subset of just these 3 static variables could retain more than 60% of the full model's predictive capacity under data-limited settings. Past and future temperature exhibit dominant influence, consistent with established ozone photochemistry. Similar patterns are observed for dynamic past inputs, with O_3 and NO concentrations contributing most strongly to future O_3 predictions, in agreement with previous findings like (Leufen et al., 2023b).

Attention weights peak at approximately 240 hours of historical context, suggesting the model finds the observation 240 hours ago highly informative for predicting the target output step. However, an ablation experiment using this reduced context window yields no improvement over the current two-week history. Details are provided in Appendix D.

These results demonstrate that re-learning attention mechanisms is essential for effective cross-regional adaptation, enabling the model to recalibrate temporal dependencies once region-specific static embeddings are updated.

As an auxiliary diagnostic, we compare the model's predicted ozone variability against CAMS total-column ozone ($gtco_3$) in appendix D3. This variable is not directly comparable to surface observations; it mainly reflects stratospheric ozone. The goal is to verify that the model captures large-scale synoptic/seasonal ozone variability, not near-surface concentrations

4.4 Summary of Findings

The TFT consistently matches or outperforms CAMS across most lead times up to 96 h. Performance gains against operational models are strongest at suburban and urban stations, where anthropogenic factors are most influential, highlighting the advantages of using static data where available in data-driven approaches as a complement to physics-based systems. Improvements are particularly pronounced during elevated ozone episodes, where CAMS tends to underestimate peak concentrations.

Transfer experiments from Germany to Korea confirm the robustness of the TFT architecture, with only minor degradation relative to in-domain training. The model captures generalizable temporal structures despite the absence of explicit spatial modeling. This is partially attributed to inherent spatial modeling included in future meteorological covariates, along with the architectural scope to isolate and correlate the individual temporal tendencies along with the static spatial coordinate information.

Ablation results presented in Appendix D.2 validate the architectural design choices, demonstrating that adequate temporal context, future meteorological inputs, and static metadata contribute complementary information.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

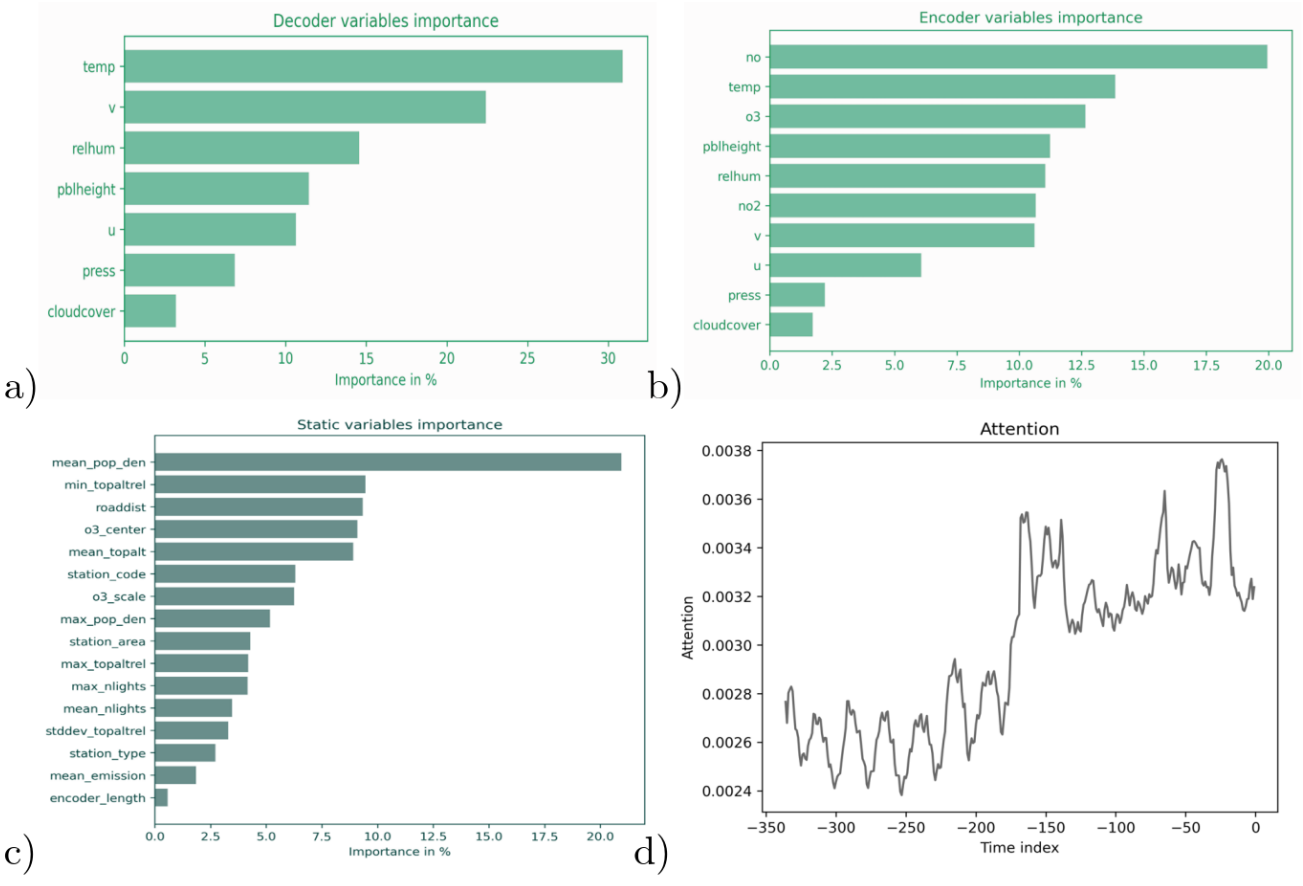

Figure 7. Model interpretability diagnostics for the Temporal Fusion Transformer. plots(a-c) shows normalized variable importance scores for decoder (future covariates), encoder (historical covariates), and static inputs, computed via gated residual network activations respectively. plot d visualizes temporal attention weights over the full 432 h context window (336 h encoder history plus 96 h decoder horizon), highlighting the relative contribution of past time steps to multi-day ozone forecasts.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

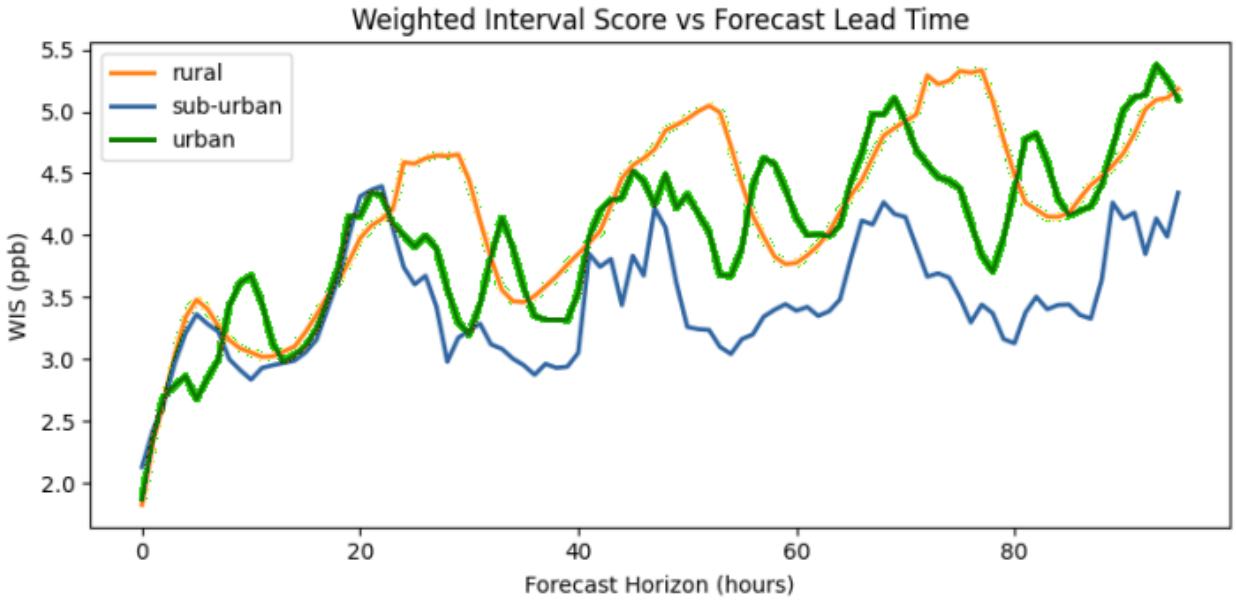

Figure 8. WIS Score per station category for the prediction horizon

5 Discussion

This study demonstrates that a task-agnostic transformer architecture can provide skillful, operationally relevant hourly ozone forecasts using a purely time-series formulation. When conditioned on meteorological inputs, historical pollutant concentrations, and static station-level metadata, the model outperforms the CAMS operational system across a 96 h forecast horizon.

The transformer maintains stable forecast skill and reduced bias growth across lead times, particularly at urban and suburban stations where anthropogenic influences dominate ozone variability. Conditioning on station-specific metadata together with known future meteorology enables improved alignment between local emission context and atmospheric forcing.

The Weighted Interval Score (WIS) averaged across prediction horizon as illustrated in Figure 8 indicates consistent probabilistic forecast performance across station categories, with values increasing gradually from approximately 2 ppb at short lead times to around 5 ppb at 96 hours. Sub-urban stations show the lowest WIS values, indicating comparatively higher predictive reliability in these environments. The observed periodic variations in WIS correspond to the diurnal cycle of ozone concentrations, suggesting that the probabilistic forecasts capture the temporal dynamics of the system. Overall, the TFT model produces stable and physically consistent uncertainty estimates across the forecast horizon.

Prior transformer-based ozone forecasting studies, most notably (Hickman et al., 2023), focus on aggregated targets such as daily maximum 8-hour ozone, use extended historical context windows of approximately three weeks, and train and evaluate on

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

the same monitoring stations within Europe. As a result, station-level generalization to unseen locations, broader geographic variability, and systematic differences in emission regimes are not explicitly assessed. For this reason, raw hourly CAMS ensemble forecasts for Germany and Global deterministic total Ozone for Korea, over a 4-day horizon are used as the primary operational benchmark in this study.

Generalization is evaluated through withheld-stations within Germany and cross-regional transfer to South Korea. Adaptation via selective retraining of static metadata embeddings and attention mechanisms yields limited performance degradation under sparse data conditions, to the extent that it captures the overall concentration range but fails to reproduce regional extremes, indicating that learned meteorology–chemistry relationships extend beyond region-specific correlations. Due to lack of hourly operational forecast for the Korean region as indicated earlier CAMS global total Ozone single level forecast is used as a benchmark; available operational alternatives are restricted to closed-source regional models or global reanalysis products.

Ablation experiments confirm that historical context length, future meteorological conditioning, and static anthropogenic metadata provide complementary information. Retraining attention layers is required when static embeddings are reinitialized, highlighting the coupling between attention alignment and contextual gating.

6 Conclusions

Together, these results indicate that task-agnostic transformer architectures provide a promising and operationally relevant complement to physics-based air-quality forecasting systems, particularly in settings where computational efficiency, adaptability, and robustness to data limitations are critical. This is coupled with an additional ability of generalizing geochemical relationships across regions in operational settings reliably in data-sparse or rapidly changing environments providing a unified assessment of operational readiness, spatial generalization, and geographic transferability.

Potential limitations should be acknowledged. Performance depends on the availability and quality of observational data and on the accuracy of meteorological inputs and forecasts. Since it is known that small-scale features, including small-scale orography variations, small-scale meteorological conditions and related dispersion regimes, local pollutant emissions not captured by large-scale emission inventories, and many others, are not explicitly accounted for in large-scale air-quality forecast models (Casciaro et al., 2022). As with other data-driven approaches, extrapolation beyond observed conditions remains challenging. Finally, while the station-based formulation enables flexible deployment, the absence of explicit spatial transport modeling may limit performance in regions dominated by long-range advection or complex topography. Although given the challenging nature of numerical modeling in cases of complex terrain, general timeseries forecasting can support efforts.

Overall, this study establishes that general-purpose transformer models constitute a scalable alternative for hourly air-quality forecasting, with strong potential for operational use and cross-regional deployment. By demonstrating competitive performance, robustness under transfer, and clear interpretability pathways, the work provides a foundation for the future development of transformer-based air-quality forecasting systems within the geoscience modeling community.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Future work could extend this framework to explore hybrid formulations that integrate spatial context, and further exploit the interpretability of transformer architecturally for uncertainity estimation including multi-staged training with existing foundational weather models.

Code and data availability. The source code of data pre-processing pipeline and model pipeline for training, inference, transfer learning, and visualization are available for direct replication via GitLab repositories: *TOAR data pipeline* and *AQ Forecasting pipeline*.

The exact versions of these repositories used to produce the results in this study are archived on Zenodo as *TOAR Ozone Data Processing Pipeline for Transformer-Based Forecasting* (Vasireddy, 2026c) and *Transformer-Based Framework for Transferable Hourly Ozone Forecasting* (Vasireddy, 2026b).

Pre-trained model checkpoints and inference outputs, and evaluation data used in this study are archived on Zenodo as *Model Checkpoints, Inference Outputs and Plots for Transferable Hourly Ozone Forecasting* (Vasireddy, 2026a).

Financial support. This research was supported by Horizon Europe Grant No. 101113400 (AQplus4).

Author contributions. S.V.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft; M.L.: Conceptualization, Validation, Resources, Supervision, Writing – review & editing; M.S.: Resources, Supervision, Project administration, Funding acquisition, Writing – review & editing. All authors have read and approved the published version of the manuscript.

Competing interests. The authors declare no conflicts of interest.

Acknowledgements. This work was supported by the AQPlus4 project supported by Horizon Europe Grant No. 101113400, and we acknowledge the Jülich Supercomputing Centre (JSC) for providing computational resources under project DE002302.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

References

Beitner, J.: Pytorch forecasting package built on PyTorch Lightning to allow training on CPUs, single and multiple GPUs out-of-the-box, https://pytorch-forecasting.readthedocs.io/en/stable/index.html, 2020.

Bertrand, J.-M., Meleux, F., Ung, A., Descombes, G., and Colette, A.: Technical note: Improving the European air quality forecast of Copernicus Atmosphere Monitoring Service using machine learning techniques, https://doi.org/10.5194/acp-2022-767, 2022.

Bertrand, J.-M., Meleux, F., Ung, A., Descombes, G., and Colette, A.: Technical note: Improving the European air quality forecast of the Copernicus Atmosphere Monitoring Service using machine learning techniques, Atmospheric Chemistry and Physics, 23, 5317–5333, https://doi.org/10.5194/acp-23-5317-2023, 2023.

Bodendorfer, N.: A HEART for the environment: Transformer-Based Spatiotemporal Modeling for Air Quality Prediction, arXiv preprint arXiv:2502.19042, 2025.

Bracher, J., Ray, E. L., Gneiting, T., and Reich, N. G.: Correction: Evaluating epidemic forecasts in an interval format, PLOS Computational Biology, 18, e1010 592, https://doi.org/10.1371/journal.pcbi.1010592, 2022.

Brauwers, G. and Frasincar, F.: A General Survey on Attention Mechanisms in Deep Learning, IEEE Transactions on Knowledge and Data Engineering, 35, 3279–3298, https://doi.org/10.1109/TKDE.2021.3113017, 2023.

Brimos, A., Liu, K., Wang, C., et al.: Lag-Llama: Pre-trained Foundation Models for Traffic Flow Forecasting, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, https://dl.acm.org/doi/full/10.1145/3716554.3716619, 2024.

Casciaro, G., Cavaiola, M., and Mazzino, A.: Calibrating the CAMS European multi-model air quality forecasts for regional air pollution monitoring, Atmospheric Environment, 287, 119 259, https://doi.org/https://doi.org/10.1016/j.atmosenv.2022.119259, 2022.

Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., and Sun, J.: RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism, in: Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), pp. 3504–3512, https://papers.nips.cc/paper/6321-retain-an-interpretable-predictive-model-for-healthcare-using-reverse-time-attention-mechanism, 2016.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlós, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, , Belanger, D., Colwell, L. J., and Weller, A.: Rethinking Attention with Performers, arXiv preprint arXiv:2009.14794, abs/2009.14794, https://api.semanticscholar.org/CorpusID:222067132, 2020.

Dauphin, Y. N., Fan, A., Auli, M., and Grangier, D.: Language Modeling with Gated Convolutional Networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), vol. 70 of *Proceedings of Machine Learning Research*, pp. 933–941, PMLR, https://proceedings.mlr.press/v70/dauphin17a.html, 2017.

ECMWF: CAMS Global Reanalysis (EAC4), https://www.ecmwf.int/en/forecasts/dataset/cams-global-reanalysis, accessed 2025-08-18.

ECMWFCode4Earth: GitHub - ECMWFCode4Earth/aq-biascorrection: Bias correction of air quality CAMS model predictions by using OpenAQ observations., https://github.com/ECMWFCode4Earth/aq-biascorrection, 2025.

Elguindi, N., Granier, C., Stavrakou, T., Darras, S., Bauwens, M., Cao, H., Chen, C., Denier van der Gon, H. A. C., Dubovik, O., Fu, T. M., Henze, D. K., Jiang, Z., Kuenen, J. J. P., Kurokawa, J., Liousse, C., Miyazaki, K., Müller, J.-F., Qu, Z., Sekou, K., Solmon, F., and Zheng, B.: Analysis of recent anthropogenic surface emissions from bottom-up inventories and top-down estimates: are future emission scenarios valid for the recent past?, Earth's Future, 8, e2020EF001 520, https://doi.org/10.1029/2020EF001520, 2020.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

European Centre for Medium-Range Weather Forecasts, o. o. C. o. b. o. t. E. U. D. A. s. o. . and signed on 22/07/2021)., C. A.: CAMS Regional: European air quality analysis and forecast data documentation - Copernicus Knowledge Base - ECMWF Confluence Wiki, https://confluence.ecmwf.int/display/CKB/CAMS+Regional%3A+European+air+quality+analysis+and+forecast+data+documentation, 2025.

Flemming, J., Inness, A., et al.: The Copernicus Atmosphere Monitoring Service, Bulletin of the American Meteorological Society, 103, E3024–E3060, https://doi.org/10.1175/BAMS-D-21-0314.1, 2022.

Grell, G. and Baklanov, A.: Integrated modeling for forecasting weather and air quality: A call for fully coupled approaches, Atmospheric Environment, 45, 6845–6851, https://doi.org/https://doi.org/10.1016/j.atmosenv.2011.01.017, modeling of Air Quality Impacts, Forecasting and Interactions with Climate., 2011.

Guevara, M., Colette, A., Guion, A., Petiot, V., Adani, M., Arteta, J., Benedictow, A., Bergström, R., Bolignano, A., Camps, P., Carvalho, A. C., Christensen, J. H., Couvidat, F., D'Elia, I., Denier van der Gon, H., Descombes, G., Douros, J., Fagerli, H., Fatahi, Y., Friese, E., Frohn, L., Gauss, M., Geels, C., Hänninen, R., Hansen, K., Jorba, O., Kaminski, J. W., Kouznetsov, R., Kranenburg, R., Kuenen, J., Lannuque, V., Meleux, F., Nyíri, A., Palamarchuk, Y., Pérez García-Pando, C., Robertson, L., Russo, F., Segers, A., Sofiev, M., Struzewska, J., Timmermans, R., Uppstu, A., Valdebenito, A., and Ye, Z.: Technical note: Sensitivity of the CAMS regional air quality modelling system to anthropogenic emission temporal variability, EGUsphere, 2025, 1–47, https://doi.org/10.5194/egusphere-2025-1287, 2025.

Gupta, S. and Kumar, R.: Urban Areas and Air Pollution: Causes, Concerns, and Mitigation, in: Urban Air Quality: Monitoring, Modelling and Management, pp. 163–185, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-45300-7_7, 2023.

Hickman, S., Griffiths, P., Nowack, P., and Archibald, A.: Short-term forecasting of ozone air pollution across Europe with transformers, ResearchGate Preprint, https://doi.org/10.1017/eds.2023.37, 2023.

Ibrahim, N. J.: A Hybrid Transformer–BiLSTM–Attention Framework for High Accuracy Multivariate Air Quality Prediction, Academia Open, 11, 10.21 070/acopen.11.2026.13 837, https://doi.org/10.21070/acopen.11.2026.13837, 2026.

Jarvis, A., Reuter, H. I., Nelson, A., and Guevara, E.: Hole-Filled SRTM for the Globe, Version 4, http://srtm.csi.cgiar.org, data derived from NASA Shuttle Radar Topographic Mission (SRTM); available from the CGIAR-CSI SRTM 90m Database, 2008.

Keswani, A., Akselrod, H., and Anenberg, S. C.: Health and clinical impacts of air pollution and linkages with climate change, NEJM Evidence, 1, https://doi.org/10.1056/EVIDra2200068, 2022.

Kleinert, F., Leufen, L. H., and Schultz, M. G.: IntelliO3-ts v1.0: a neural network approach to predict near-surface ozone concentrations in Germany, Geoscientific Model Development, 14, 1–25, https://doi.org/10.5194/gmd-14-1-2021, 2021.

Kottapalli, A., Zhao, Z., Wu, H., et al.: Foundation Models for Time Series: A Survey, arXiv preprint arXiv:2504.04011, https://arxiv.org/abs/2504.04011, 2025.

Lang, S., Rodwell, M., and Schepers, D.: IFS upgrade brings many improvements and unifies medium-range resolutions, https://doi.org/10.21957/slk503fs2i, 2023.

Lawrence, M. G.: The Relationship between Relative Humidity and the Dewpoint Temperature in Moist Air: A Simple Conversion and Applications, Bulletin of the American Meteorological Society, 86, 225–233, https://doi.org/10.1175/BAMS-86-2-225, 2005.

Leufen, L., Kleinert, F., and Schultz, M.: O3ResNet: A Deep Learning–Based Forecast System to Predict Local Ground-Level Daily Maximum 8-Hour Average Ozone, Artificial Intelligence for the Earth Systems, https://juser.fz-juelich.de/record/1007047/files/aies-AIES-D-22-0085.1.pdf, 2022a.

Leufen, L., Kleinert, F., and Schultz, M.: O3ResNet: A Deep Learning–Based Forecast System to Predict Local Ground-Level Daily Maximum 8-Hour Average Ozone in Rural and Suburban Environments, Artificial Intelligence for the Earth Systems, 2, 1–42, https://doi.org/10.1175/AIES-D-22-0085.1, 2023a.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Leufen, L. H., Kleinert, F., and Schultz, M. G.: MLAir (v1.0) – a tool to enable fast and flexible machine learning on air data time series, Geoscientific Model Development, 14, 1553–1574, https://doi.org/10.5194/gmd-14-1553-2021, 2021a.

Leufen, L. H., Kleinert, F., and Schultz, M. G.: Exploring decomposition of temporal patterns to facilitate learning of neural networks for ground-level daily maximum 8-hour average ozone prediction, Environmental Data Science, 1, e10, https://doi.org/10.1017/eds.2022.9, 2022b.

Leufen, L. H., Kleinert, F., and Schultz, M. G.: O3ResNet: A Deep Learning–Based Forecast System to Predict Local Ground-Level Daily Maximum 8-Hour Average Ozone in Rural and Suburban Environments, Artificial Intelligence for the Earth Systems, 2, e220 085, https://doi.org/10.1175/AIES-D-22-0085.1, 2023b.

Leufen, L. H. et al.: MLAir (v1.0) – a tool to enable fast and flexible machine learning on air data, Geoscientific Model Development, 14, 1553–1573, https://doi.org/10.5194/gmd-14-1553-2021, 2021b.

Liang, X., Wang, Z., Liu, Y., et al.: Foundation Models for Time Series Analysis: A Tutorial and Survey, arXiv preprint arXiv:2403.14735, https://arxiv.org/abs/2403.14735, 2024.

Liang, Y., Xia, Y., Ke, S., Wang, Y., Wen, Q., Zhang, J., Zheng, Y., and Zimmermann, R.: AirFormer: Predicting Nationwide Air Quality in China with Transformers, https://arxiv.org/abs/2211.15979, 2022.

Lim, B., Arik, S. O., Loeff, N., and Pfister, T.: Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting, https://arxiv.org/abs/1912.09363, 2020.

Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NIPS) 2017, pp. 4768–4777, http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions, proceedings of NIPS 2017, 2017.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., and Bezirtzoglou, E.: Environmental and health impacts of air pollution: A review, Frontiers in Public Health, 8, https://doi.org/10.3389/fpubh.2020.00014, 2020.

Mujtaba, M., Munir, M. A., Ali, S., Petrů, J., Ansar, T., Akhlaq, W., Ahmad, M., Iqbal, H., Ali, F., Bashir, M. N., and Alexander, T.: Using machine learning for air quality prediction and sustainable urban planning, Sustainable Futures, 10, 100 981, https://doi.org/https://doi.org/10.1016/j.sftr.2025.100981, 2025.

Nie, T., Mei, Y., Qin, G., Sun, J., and Ma, W.: Channel-Aware Low-Rank Adaptation in Time Series Forecasting, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24, p. 3959–3963, ACM, https://doi.org/10.1145/3627673.3679884, 2024.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J.: A Time Series is Worth 64 Words: Long-term Forecasting with Transformers, https://arxiv.org/abs/2211.14730, 2023.

NOAA National Centers for Environmental Information (NCEI): DMSP Data — Image and Data Processing by NOAA's National Geophysical Data Center, https://www.ncei.noaa.gov/, dMSP data collected by the U.S. Air Force Weather Agency, n.d.

OpenStreetMap contributors: OpenStreetMap data retrieved via Overpass API, https://overpass-api.de/, https://www.openstreetmap.org/copyright, version 0.6, Overpass API 0.7.57.2 (build 48842a1b), 2022.

Qiu, X., Cheng, H., Wu, X., Hu, J., Guo, C., and Yang, B.: A Comprehensive Survey of Deep Learning for Multivariate Time Series Forecasting: A Channel Strategy Perspective, https://arxiv.org/abs/2502.10721, 2025.

Rasul, K., Ashok, A., Williams, A. R., Ghonia, H., Bhagwatkar, R., Khorasani, A., Bayazi, M. J. D., Adamopoulos, G., Riachi, R., Hassen, N., Biloš, M., Garg, S., Schneider, A., Chapados, N., Drouin, A., Zantedeschi, V., Nevmyvaka, Y., and Rish, I.: Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting, https://arxiv.org/abs/2310.08278, 2024.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Ribeiro, M. T., Singh, S., and Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, https://doi.org/10.1145/2939672.2939778, 2016.

Riccio, A. and Chianese, E.: Technical note: Accurate, reliable, and high-resolution air quality predictions by improving the Copernicus Atmosphere Monitoring Service using a novel statistical post-processing method, Atmospheric Chemistry and Physics, 24, 1673–1689, https://doi.org/10.5194/acp-24-1673-2024, 2024.

Schiavina, M., Freire, S., and MacManus, K. J.: GHS-POP R2019A — GHS population grid multitemporal (1975-1990-2000-2015), https://doi.org/10.2905/0C6B9751-A71F-4062-830B-43C9F432370F, 2019.

Schultz, M. G., Schröder, S., Lyapina, O., Cooper, O. R., Galbally, I., Petropavlovskikh, I., Hamad, S. H., Helmig, D., Henriques, D., Hermansen, O., Holla, R., Hueber, J., Im, U., Jaffe, D. A., Komala, N., Kubistin, D., Lam, K.-S., Laurila, T., Lee, H., Levy, I., Mazzoleni, C., Mazzoleni, L., McClure-Begley, A., Mohamad, M., Murovic, M., Navarro-Comas, M., Nicodim, F., Parrish, K. A., Read, N., Reid, N., Ries, L., Saxena, P., Schwab, J. J., Scorgie, Y., Senik, I., Simmonds, P., Sinha, V., Skorokhod, A. I., Spain, G., Spangl, W., Spoor, R., Springston, S. R., Steer, K., Steinbacher, M., Suharguniyawan, E., Torre, P., Trickl, T., Weili, L., Weller, R., Xiaobin, X., Xue, L., and Zhiqiang, M.: Tropospheric Ozone Assessment Report: Database and metrics data of global surface ozone observations, Elementa: Science of the Anthropocene, 5, 58, https://doi.org/10.1525/elementa.244, 2017.

Shetty, S., Hamer, P. D., Stebel, K., Kylling, A., Hassani, A., Berntsen, T. K., and Schneider, P.: Daily high-resolution surface PM2.5 estimation over Europe by ML-based downscaling of the CAMS regional forecast, Environmental Research, 264, 120 363, https://doi.org/https://doi.org/10.1016/j.envres.2024.120363, 2025.

Tetens, O.: Über einige meteorologische Begriffe, Zeitschrift für Geophysik, 6, 207–309, 1930.

Tong, Q., Liang, G., and Bi, J.: Calibrating the Adaptive Learning Rate to Improve Convergence of ADAM, https://arxiv.org/abs/1908.00700, 2019.

Vasireddy, S.: Model Checkpoints, Inference Outputs and Plots for Transferable Hourly Ozone Forecasting, https://doi.org/10.5281/zenodo.19151740, access restricted; available upon reasonable request, 2026a.

Vasireddy, S.: Transformer-Based Framework for Transferable Hourly Ozone Forecasting, https://doi.org/10.5281/zenodo.19151703, 2026b.

Vasireddy, S.: TOAR Ozone Data Processing Pipeline for Transformer-Based Forecasting, https://doi.org/10.5281/zenodo.19151435, 2026c.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I.: Attention Is All You Need, in: Advances in Neural Information Processing Systems (NeurIPS), https://api.semanticscholar.org/CorpusID:13756489, 2017.

Wen, R., Torkkola, K., Narayanaswamy, B., and Madeka, D.: A Multi-Horizon Quantile Recurrent Forecaster, in: 31st Conference on Neural Information Processing Systems (NIPS) 2017 — Time Series Workshop, https://arxiv.org/abs/1711.11053, 2017.

Wu, X., Jin, J., Qiu, W., Chen, P., Shu, Y., Yang, B., and Guo, C.: Aurora: Towards Universal Generative Multimodal Time Series Forecasting, https://arxiv.org/abs/2509.22295, 2025.

Yunita, A., Pratama, M. I., Almuzakki, M. Z., Ramadhan, H., Akhir, E. A. P., Firdausiah Mansur, A. B., and Basori, A. H.: Performance analysis of neural network architectures for time series forecasting: A comparative study of RNN, LSTM, GRU, and hybrid models, MethodsX, 15, 103 462, https://doi.org/https://doi.org/10.1016/j.mex.2025.103462, 2025.

Zheng, H. et al.: AirFormer: Predicting Nationwide Air Quality in China with Transformers, arXiv preprint arXiv:2211.15979, https://arxiv.org/abs/2211.15979, 2022.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Appendix A: Data Inputs and Metadata Overview

A1 Input Variables

Table A1. Input and Target Variables

Category	Variables	Description
Past observations	Air Pollutants (TOAR II)	O_3, NO, NO_2
	Meteorological factors (ERA5)	Zonal and meridional winds (U, V), temperature, surface pressure, relative humidity[1], Planetary Boundary Layer (PBL) height, and cloud cover. This data was obtained from ERA5_Single_PressureLevel dataset in the Copernicus Climate Data Store (CDS) API.
Known future covariates	ERA5 meteorology	Same variables as past meteorology[2]. Used as deterministic drivers.
Target	Surface O_3	Hourly O_3 concentration for the forecast horizon.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

A2 Static Metadata

A2.1 Anthropological Factors (TOAR II)

Table A2. Anthropological Metadata

Variable	Type / Units	Description
station_type	Categorical: Type of the locality of weather station that collected the TOAR observations	Locality type: [[0, Rural], [1, Sub-Urban], [2, Urban]]
station_area	Categorical: A controlled vocabulary category to indicate type of station with 4 classes	Station class: [[0, Unknown], [1, Background], [2, Traffic], 3, Industrial]][3]
mean_emission	Real	Mean value of NOX emission data within a radius of 10 km around station location averaged over years 2000 and 2015, from CAMS global and regional emissions data (Elguindi et al., 2020).
roaddist	Real: unit - km	Distance to major roads from the station location computed based on (OpenStreetMap contributors, 2022) for year 2020.

A2.2 Population Density

Population density around station location collected in radii of 250m, 5000m and 25000m for years 1990 and 2015 from European Commission, Joint Research Center (Schiavina et al., 2019) were used and central tendencies such as Mean and Max per year were further aggregated below and used for training.

[1]Relative humidity was calculated from the dew point temperature T_{dew} and the air temperature T using the ratio of saturation vapor pressures:

$$\mathrm{RH} = 100 \times \frac{e_s(T_{\mathrm{dew}})}{e_s(T)}, \tag{A1}$$

where $e_s(T)$ is the saturation vapor pressure at temperature T. A widely used approximation for $e_s(T)$ is the Magnus-Tetens formula (Tetens, 1930):

$$e_s(T) = 6.112 \exp\left(\frac{17.67T}{T + 243.5}\right) \quad [\mathrm{hPa}], \tag{A2}$$

where T is in degrees Celsius. This formulation allows a simple conversion between dew point and relative humidity (Lawrence, 2005).

[2]ERA5 meteorological data was used instead of observational weather data also available in TOAR since ERA5 is operational data with better continuity and also a gold standard in the global Earth science community with near real precision

[3]Additional categorical metadata on type of landcover and type of ecoregion, have been discarded delibrately although available in TOAR to avoid over-learning of these relationships as they are limited in either the year the metadata was calculated in the source or the distance from the station location considered

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Table A3. Population Density Metadata

Variable	Description
mean_pop_den	Average of population densities computed at 250m and 5000m radii are averaged per years 1990 and 2015 before averaging together and treated as mean of the population density for the station: $\bar{P}_{\text{mean\_pop\_den}} = \frac{1}{2}\left(\frac{P_{1990}^{(250)} + P_{1990}^{(5000)}}{2} + \frac{P_{2015}^{(250)} + P_{2015}^{(5000)}}{2}\right)$ where $P_y^{(r)}$ population density for year y at radius r (in meters).
max_pop_den	Highest value amongst population densities computed at radius of 25000m for years 2015 and 1990 was used as the maximum population density for the station at any given time as the training data ranges from 1990 to 2016 as elaborated in the next section. $P_{\text{max\_pop\_den}} = \max\left(P_{1990}^{(25000)}, P_{2015}^{(25000)}\right),$ where $P_y^{(r)}$ denotes the population density at radius r (in m) for year y.

A2.3 Nightlight Density

Nightlight intensity around station location collected in radii of 1,5 and 25 Kms for years 2013 and 1992 by US Air Force Weather agency (NOAA National Centers for Environmental Information (NCEI), n.d.) were used and central tendencies as below were used for training.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Table A4. Night Lights Metadata

Variable	Description
mean_nlights	Average of nightlight intensities computed at 1 and 5Km radii for years 2013 and treated as mean of the nightlight intensities for the station. $\bar{P}_{\text{mean\_nlights}} = \frac{P^{(1)}_{2013} + P^{(5)}_{2013}}{2}$ where $P^{(r)}_{y}$ Nightlight intensity for year y at radius r (in Km).
max_nlights	Highest value amongst nightlight intensities computed at radius of 25Km for years 2013 and 1992 was used as the maximum for the station at any given time for the training range. $P_{\text{max\_nlights}} = \max\left(P^{(25)}_{1992},\ P^{(25)}_{2015}\right),$ where $P^{(r)}_{y}$ denotes the Nightlight intensity at radius r(in Km) for year y.

A2.4 Topographic Metadata

Topographic landscape altitude range around station location collected in radii of 90m, 1Km and 5Km for year 1994 by NASA SRTM (Jarvis et al., 2008) were used and central tendencies as below were used for training.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Table A5. Topographic Metadata

Variable	Description
mean_topalt	Average of altitudes of Topographic landscape computed at 90m and 1Km radii for years 1994 and treated as mean of the topograhic altitude around the station. $\bar{P}_{\text{mean\_nlights}} = \frac{P_{1994}^{(1)} + P_{1994}^{(0.009)}}{2}$ where $P_y^{(r)}$ Altitude variation for year y at radius r (in Km).
min_topaltrel	Due to structuring of data the range of relative altitude to the station against the topographic landscape was only computed at a radius of 5Km from the station location for the year 1994 and the minimum of these values was taken as relative topographic altitude of landscape.
max_topaltrel	Same reason as above the TOAR only had maximum of relative altitude in the range computed only for year 1994 at a radius of 5Km. Hence only that was retained for the experiments.
stddev_topaltrel	Same reason as above the TOAR only had standard deviation of relative altitude in the range computed only for year 1994 at a radius of 5Km. Hence only that was retained for the experiments.

A2.5 Spatiotemporal Metadata

Table A6. Spatiotemporal Metadata

Variable	Description
lat	Geographic Latitude coordinate of the observation station location.
long	Geographic Longitude coordinate of the observation station location.
alt	Geographic Altitude coordinate of the observation station location. [4].
station_code	Internal TOAR Station code/category string in the data base - format "CNSTXXX" where CN is the country, ST is the state and XXX is a 3 digit integer ID for the station.[5].
relative_time_idx	Within each sampled sequence, this index will range from $-encoder\_length$ (context window e.g., -336) to $+prediction\_length$ (forecast horizon e.g., +96) to indicate explicitly the lag and lead times in data for forecasting.
encoder_length	encoder length is also provided as a static covariate to indicate the relationship to the look back time used.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

A2.6 Topography

When temporal context fades, static features like mean/scale give an anchor, preventing model drifts and help maintain realistic amplitude in forecasts. This becomes more relevant as the forecast horizon increases for medium and long range forecasting, although not significant in short range forecasts. Hence the absolute magnitude of unnormalized series is passed to the model as static real features.

Table A7. Target scales Metadata

Variable	Description
O_3_center	The mean absolute value of O_3 for the corresponding station to which sample belongs to.
O_3_scale	The standard deviation across range of absolute values of O_3 for the corresponding station to which sample belongs to.

In other words for each time series $y_t^{(i)}$, is standard normalization as

$$\tilde{y}_t^{(i)} = \frac{y_t^{(i)} - \mu_i}{\sigma_i}$$

adding μ_i and σ_i. Hence including the unnormalized data center and scale of target variables as additional inputs to the model helps in generalizing better for any range.

[4]It is worth noting that since TOAR is observational while meteorological data from ERA5 is more uniformly gridded hence nearest lat, long of ERA5 for the corresponding station was fetched and the coordinates of TOAR observation stations are retained and not the grid points of meteorological data, although since farthest point was 6.3Km the significance is not lost.

[5]Purpose of including an internal categorical variable like this that does not add any additional information to the model training than the above geographic coordinates. It is merely used for easy analysis of which station the sample belongs to during evaluation for better user friendly analytics.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Appendix B: Model significance for AQ Forecasting

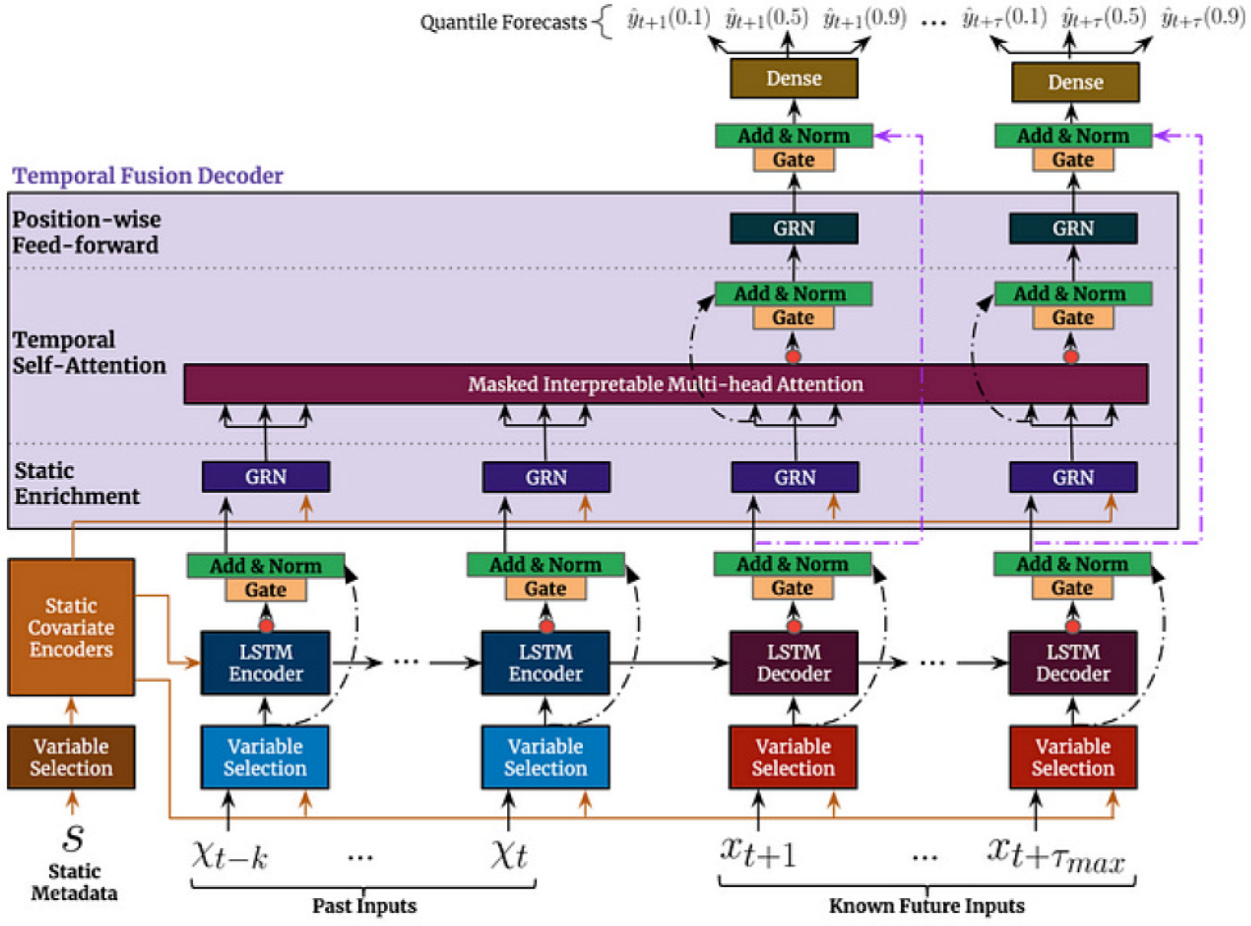

Figure B1. Model architecture of TFT, along with its main components (Source:[(Lim et al., 2020)

B1 Interpretable Multi-Head Attention (IMHA)

While post-hoc explainability techniques such as LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017) are widely used, they only provide localized, sample-specific interpretations. LIME constructs independent surrogate models for each input instance, and SHAP estimates feature influence based on perturbations of neighboring timesteps, which makes both approaches less suitable for interpreting temporally entangled, sequence-to-sequence architectures. However, multi-horizon multivariate transformer models consist of "complex nonlinear interactions between many parameters" (Lim et al., 2020). Hence, these post-hoc methods become less useful, as explained in detail in the Introduction section of (Lim et al., 2020) by the authors of TFT.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

This limitation is overcome in TFT by introducing a interpretable multi-head attention layer natively within its architecture, providing intrinsic transparency in how temporal dependencies and static covariates jointly influence predictions through a global importance weighting across all features.

$$InterpretableMultiHead(Q,K,V) = \frac{1}{H}\sum_{h=1}^{H} \text{Attn}(QW_Q^{(h)}, KW_K^{(h)}, VW_V)W_H,$$

where Q, K, V and H denote the Query, Key, Value and the number of heads respectively, with W_x corresponding to their weights and $W_x^{(h)}$ representing head-specific weights.

This importance weighting is applied in addition to the attention weights, and the attention layer is deliberately placed after the computation of variable-wise importance by selection networks, persistent temporal patterns from recurrent encoders, and significant outliers and regime changes identified through gating mechanisms. The significance of each of these three components is explained below. This placement ensures that the explainability obtained from TFT is neither limited to the attention layer inputs nor to the encoder and decoder networks locally, but instead provides a collective global importance for each component influencing the target, which is necessary for complex variable interactions such as AQ pollutant concentrations.

B2 Recurrent encoder & decoder

TFT was not the first model to use multi-horizon recurrent architectures for sequence-to-sequence modeling (Wen et al., 2017). The authors of TFT benchmarked against such works that use only recurrent networks, showing an average 30–40% improvement across quantiles with the combination of transformers, LSTMs, and gating mechanisms, bringing together the strengths of these techniques for sequence-to-sequence modeling. This is because the context vectors are used for attention computation after gating, rather than the raw sequences themselves. This introduces learning efficiency for models that work better with timestep-specific encoding, i.e., what happens at every t_{th} timestep is important in a chemical process that is time-driven, with different inputs acting with varying effects at different timesteps.

B3 Time specific Variable co-variates

Leveraging variable selection networks for temporal forecasting with interdependence is not new; works such as (Choi et al., 2016) have previously presented this. However, the ingenuity introduced by the TFT architecture lies in the placement of a carefully designed variable selection network, which does not compute only generic variable importance but rather an importance score for each t^{th} timestep of the horizon, which in turn incorporates a global flattened importance of all past timesteps up to that point. Algorithmically the variable selection weight is

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

$$v_{\chi t} = Softmax(GRN_{v_\chi}([\xi_t^{(1)^T}, ..., \xi_t^{(j)^T} t_i]^T, c_{variable}))$$

where GRN_{v_χ} is a gated residual network, $\xi^{(j)}t$ is the input entity embedding of the j^{th} variable at time t, and $cvariable$ is the context vector for the $variable$ corresponding to static, past, and future inputs respectively. Because this variable selection is placed before the encoders and decoders, an importance score for interdependence is captured at each timestep for each variable. This works effectively together with the gating mechanism explained below to determine when and where each variable contributes more during training.

This is done for all variables, including static metadata. The variable selection weights of static metadata are gated at each timestep for each variable across contexts and are chosen per sample and per instance, making the learned parameters more adaptive to extremities.

B4 Gating mechanisms

Normally, when all contributing factors are endogenous, the process is simple and all the above components can be retained. Inclusion of all is advantageous in this scenario. However, when dealing with exogenous variables such as in AQ forecasting, where complex covariations exist, not all of these combinations are necessarily required. However, the challenge is that until the appropriate combination of variable, timestep, and embedding dimension is reached, it is difficult to determine whether that layer should be considered for the next step in the sequence.

The simplest way to achieve this is by placing a gating mechanism at all output layers, including those following the attention layer, to ensure that only those with high importance are allowed to pass through to higher layers. TFT gating layer is based on Gated Linear unit from (Dauphin et al., 2017). It is important to note that gating alone would not work effectively without the pre-computed importance of covariation among all variables across all timestep combinations, which is why this architecture is particularly well suited for complex multivariate multi-horizon forecasting.

B5 Probabilistic forecasts

Alongside the above features, TFT produces probabilistic forecasts in place of point forecasts. This is done using a linear interpolation of the decoder output layer for the desired number of quantiles, which helps in capturing the uncertainty around the prediction. This is a desirable aspect for AQ forecasts and is not present in most models prevalent in the domain. There have been efforts to compute the uncertainty with other post processing or deliberate output layer modifications but not as distributed forecasts. This is now becoming more prevalent in upcoming architectures, and helps determine the confidence interval for AQ forecasts, which have inherent aleatoric uncertainty.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Appendix C: Implementation Details and Configuration

C1 Training Infrastructure

Model training was performed using distributed data-parallelism across multiple GPUs. Specifically, training employed 16 NVIDIA A100 GPUs (40 GB HBM2e each) distributed across four compute nodes. Each node was equipped with two AMD EPYC Rome 7402 CPUs (24 cores per socket, 2.8 GHz), resulting in a total of 384 logical CPU threads enabled via simultaneous multi-threading.

Distributed training relied on GPU-based data parallelism, while CPU resources were primarily used for data loading and input–output (I/O) operations. This configuration enabled efficient throughput when training on large spatiotemporal datasets with overlapping samples.

C2 Runtime Characteristics

Under the above configuration, the average runtime per training epoch was approximately 2 h 15 min, including validation. Overall training time scaled primarily with the degree of GPU data parallelism, while data ingestion and preprocessing performance depended on CPU I/O throughput.

No model-specific hardware optimizations or custom CUDA kernels were employed beyond standard distributed training capabilities provided by PyTorch and the PyTorch Forecasting framework. It was achievable leveraging the general timeseries forecasting support provided by the frameworks for both CPU and GPU workloads. This ensures that reported runtimes are representative of commonly available high-performance computing environments.

C3 Reproducibility Considerations

All training hyperparameters, preprocessing steps, and framework-level modifications are described in the main text and appendices. The hardware configuration documented here reflects the environment used for the experiments reported in this study but is not a requirement for reproducing the results. Equivalent experiments can be conducted on smaller-scale systems with fewer GPUs, at the cost of increased training time.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Appendix D: Ablation and Additional Studies

D1 Ablation Strategies

Ablation experiments were conducted to assess model sensitivity to key architectural and input-design choices and to quantify the contribution of individual components to forecast skill. Each ablation model was retrained from scratch using the same training data and protocol as the full configuration, with only the specified component removed or modified. The objective was to evaluate how reducing available information or architectural capacity affects the model's ability to capture ozone variability relative to operational baselines.

D1.1 Forecasting Ablations (Germany)

Reduced context window: Context lengths of 168 h (1 week) and 240 h (10 days), commonly used in prior studies, were compared against the default 336 h (2-week) context window. Models trained with shorter contexts retained basic forecast structure but exhibited limited improvement over CAMS, with MAE values typically in the range of 7-9 $\mu g\,m^{-3}$. In contrast, the 336 h configuration achieved consistently lower errors (approximately 5-6 $\mu g\,m^{-3}$) across the 96 h forecast horizon, while CAMS errors remained between 9-11 $\mu g\,m^{-3}$ as detailed in appendix. This indicates that longer context windows enable the model to better capture synoptic scale variability and diurnal cycles relevant for hourly ozone prediction.

No future meteorological covariates: To assess the role of decoder-side future inputs, all future meteorological variables were removed, forcing the model to rely solely on historical context and static metadata. This configuration resulted in pronounced performance degradation across all station types, with MAE increasing to approximately 9-11 $\mu g\,m^{-3}$, particularly at longer lead times (72-96 h) compared with 5–6 $\mu g m^{-3}$ in the full-feature configuration as detailed in appendix. Skill scores relative to CAMS became predominantly negative. The degradation was most pronounced at rural sites, where ozone variability is strongly driven by meteorological forcing. These results highlight the importance of conditioning on forecast meteorology to align pollutant evolution with anticipated atmospheric transitions rather than extrapolating solely from persistence.

No static anthropogenic and spatiotemporal metadata. In this ablation, static station-level embeddings were removed while retaining both encoder and decoder meteorological inputs. Forecast errors increased substantially across all horizons, and skill relative to CAMS was strongly negative. Although diurnal structure remained better aligned than in the no-future-meteorology case, forecasts exhibited persistent bias and miscalibrated amplitudes. This indicates that static metadata provide critical priors linking meteorological forcing to local emission intensity and land-use characteristics, without which the model fails to translate atmospheric drivers into station-specific concentration responses.

Across these 3 experiments, a clear hierarchy of importance emerged. Reduced context length led to moderate degradation, removal of future meteorological inputs caused larger errors due to loss of forward atmospheric conditioning, and removal of static metadata resulted in the strongest deterioration by eliminating station-specific calibration. These results confirm that adequate temporal context, future exogenous inputs, and static anthropogenic information contribute complementary and non-redundant information to hourly ozone forecasting.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

D1.2 Transfer Learning Ablation (Korea)

Frozen attention layers. To evaluate the necessity of retraining attention mechanisms during transfer learning, an ablation was conducted in which attention layers were frozen while static metadata embeddings and gating layers were reinitialized and retrained. This configuration consistently underperformed relative to the default transfer setup in which attention, metadata embeddings, and output layers were jointly retrained.

In the TFT architecture, attention weights do not encode geochemical relationships directly, but instead represent alignment between encoder and decoder context vectors modulated by static embeddings and gating mechanisms. When static metadata are reinitialized for a new region, the previously learned attention alignment becomes inconsistent with the updated gating structure. Empirically, freezing attention layers led to increased RMSE, damped amplitudes, and phase-shifted diurnal cycles. Joint retraining of attention restored both amplitude and phase fidelity, substantially reducing forecast error.

D2 Ablation Study Results

Results from the ablation experiments described in Appendix D are summarized in Figure D1.

Removing future meteorological inputs or static metadata leads to substantial performance degradation, with errors approaching or exceeding those of CAMS for much of the forecast horizon.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

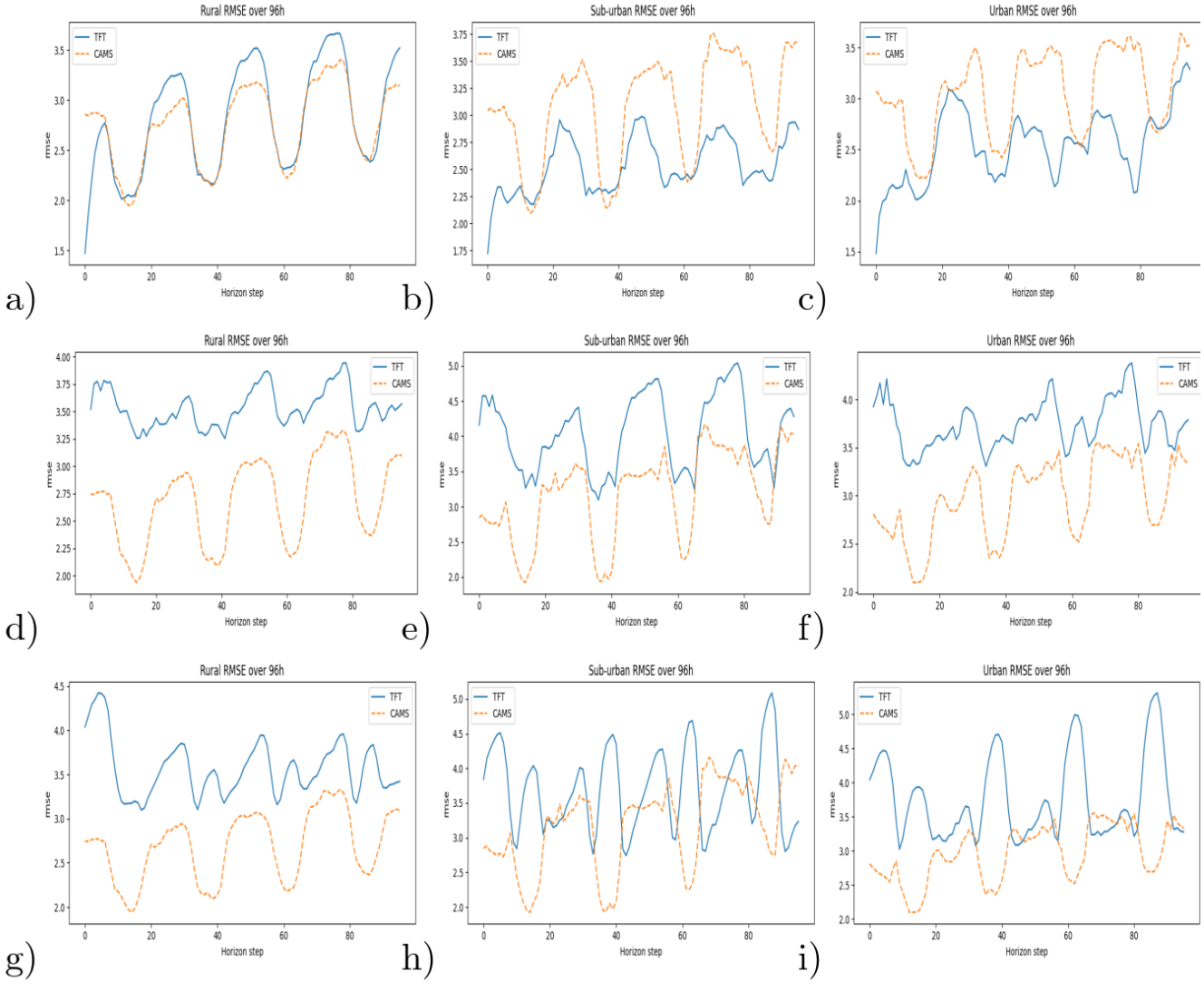

Figure D1. Ablation study of the Temporal Fusion Transformer forecast skill relative to CAMS during summer 2023 (25 June–25 September), stratified by station type. Rows correspond to different ablation settings: (plots a-c) reduced encoder context window of 168 h (one week), (plots d-f) removal of known future meteorological covariates, and (plots g-i) removal of static anthropogenic metadata. Columns show averaged RMSE for rural, suburban, and urban stations, respectively.

In these configurations, TFT exhibits negative skill relative to CAMS, validating the importance of future meteorological conditioning and station-level metadata for hourly ozone forecasting.

The reduced-context ablation (168 h) shows performance comparable to CAMS at rural sites and modest improvement at suburban and urban sites. This configuration reflects the attention peak around 168 h shown in Figure 7 and indicates that

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

the model can retain reasonable hourly skill even with shorter historical context. However, identifying an optimal deployable context length was not the focus of this study and is deferred to future work.

It can be seen in case of training without either metadata or future weather the error accumulated by TFT is more and meets the error of CAMS forecasts only in a few hours.

This gives TFT a negative skill in comparison to CAMS as expected, this validates the choice of model architecture, as an operational model like CAMS works with physics and a larger training scope in comparison to a handicapped TFT with a training corpus of 20 years in German region.

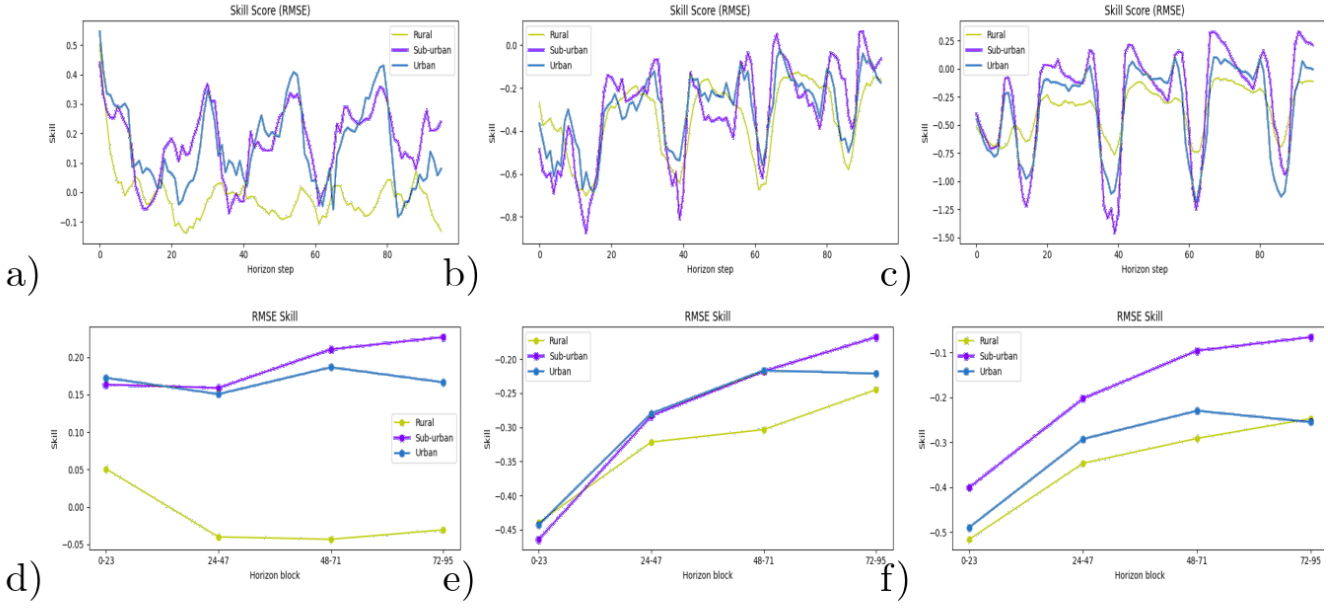

Figure D2. RMSE-based skill of TFT relative to CAMS for the ablation study during summer 2023. Corresponding columns of subplots correspond to ablations: reduced context window (168 h), removal of known future meteorological covariates, and removal of static anthropogenic metadata. plots a-c shows skill aggregated over the full test period, while the plots d-f reports daily-averaged skill. Positive skill indicates improved performance of TFT over CAMS.

On other side a reduced context window of 1 week (chosen based on peak of attention around 168 hours as indicated in Figure 7) shows a comparable performance against CAMS forecasts in rural case and slightly better than CAMS performance in other two types.

This indicates that the model is capable of predicting medium to longer ranges with lesser context, a detailed context horizon estimation for a 4-day forecast is not considered at this time, as the work focuses on establishing the operational readiness of the transformer models for the forecast task and aiming more towards longer forecast ranges.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

D3 Additional Plots and discussions

Below plots are provided in addition the Results and discussion presented in sections 4 and 5 in the main article with an aim to provide a completion to the study.

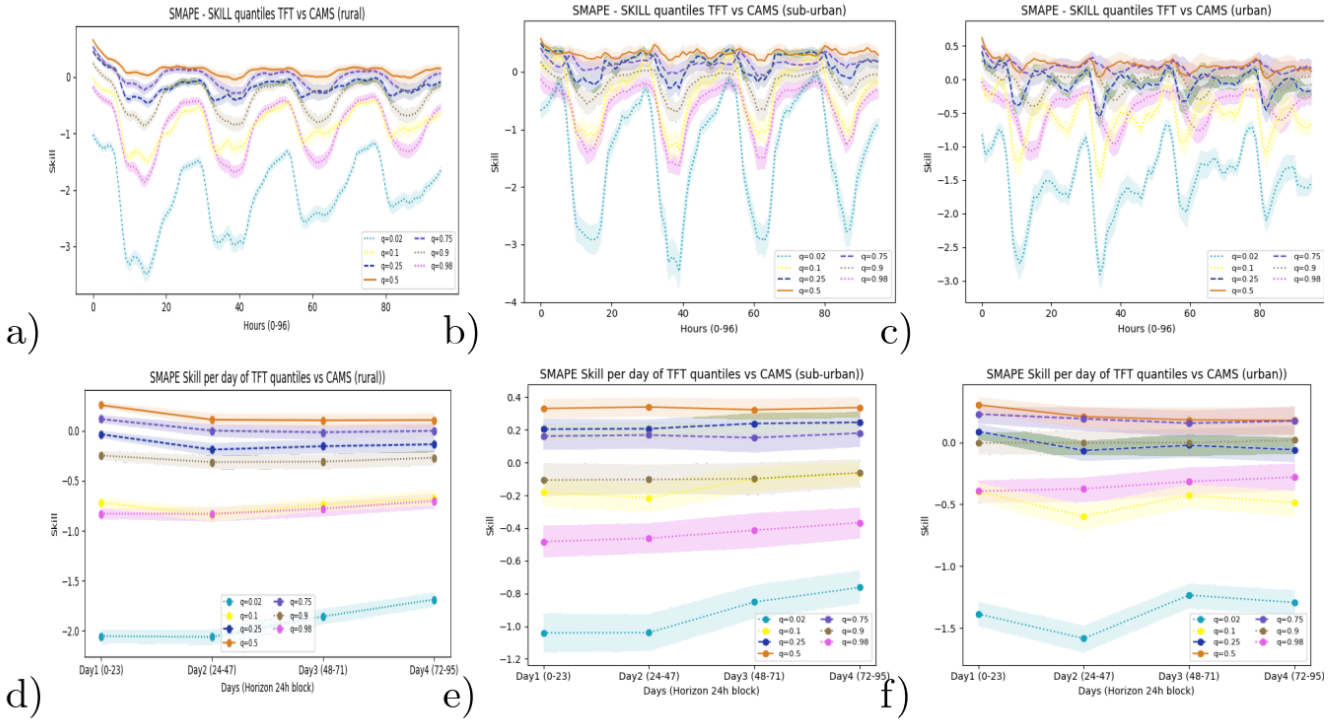

Figure D3. SMAPE based Skill scores of TFT predictions for all 7 quantiles against CAMS baseline for rural, sub-urban and urban station categories are presented in subplots (a-c) respectively. Corresponding per day averaged skill scores are presented in subplots (d-f) with the most broadest confidence intervals of 95% and 98% indicated as dotted lines and closest CI of 50% presented as dashed lines

Figure D3 presents the SMAPE-based skill scores for all TFT quantile predictions relative to CAMS across the forecast horizon and when averaged per forecast day. The central quantiles ($q = 0.25$–$q = 0.75$) generally maintain stable positive skill across station categories, indicating consistent relative error improvements over the baseline. In contrast, the outer quantiles ($q = 0.02$ and $q = 0.98$) exhibit greater variability, reflecting the wider uncertainty bounds associated with extreme prediction intervals. The day-averaged panels show a similar structure, where the central quantiles remain comparatively stable while the spread among quantiles increases toward longer forecast days.

As mentioned earlier for Korean region, although CAMS surface ozone was not available in the extracted product.A proxy of CAMS total-column ozone ($gtco3, kgm^{-3}$) was explored as auxiliary diagnostic of large-scale ozone variability and results presented in D4. Total-column ozone is not directly comparable to near-surface observations ($nmolmol^{-1}$), because it is

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

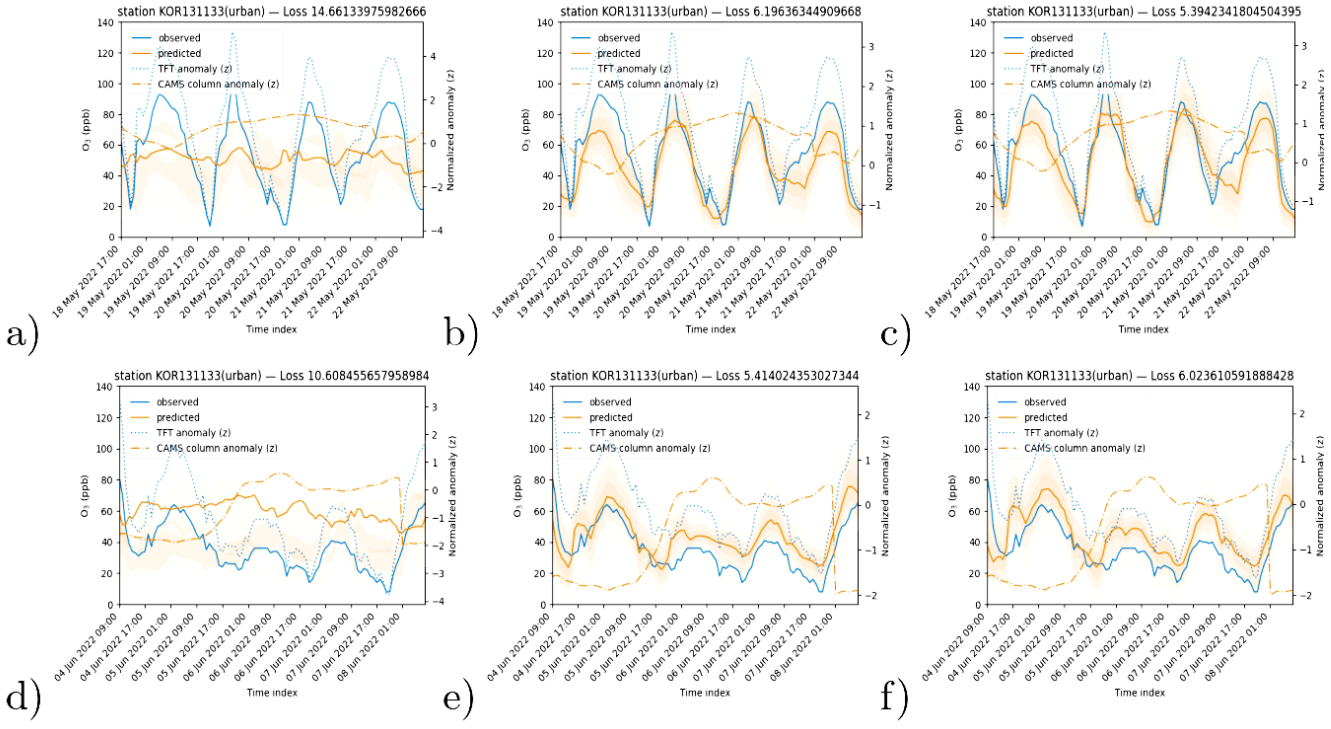

Figure D4. Transfer learning ablation results for the Korea domain adaptation experiment. Plots show representative actual versus predicted time series for two unseen validation samples under three fine-tuning strategies: (left plots a,d) retraining static anthropogenic and spatial metadata embeddings only, (centre plots b,e) retraining metadata embeddings and output layers with attention layers frozen, and (right plots c,f) full fine-tuning including metadata embeddings, attention mechanisms, and output layers.

dominated by stratospheric ozone. Consequently, we do not interpret $gtco3$-based scores as surface forecast skill, but rather as a consistency check on synoptic-scale ozone evolution.

The station-wise time series in D4 compare observed surface ozone concentrations, model predictions, and normalized anomaly signals derived from CAMS total-column ozone. The anomaly series are obtained by standardizing each signal relative to its local mean and variance, which highlights temporal variability while removing differences in absolute magnitude. Across the stations shown for the Korean region, the model-derived anomalies follow the observed surface ozone variability reasonably well, capturing the timing of the main peaks and troughs associated with the diurnal cycle and multi-day variability. In contrast, the normalized anomalies derived from CAMS total-column ozone appear smoother and exhibit smaller relative fluctuations, reflecting the fact that column-integrated ozone is primarily influenced by large-scale atmospheric processes rather than near-surface photochemical variability.

Despite this limitation, the column ozone anomalies provide a useful qualitative indicator of broader synoptic-scale ozone variability affecting the region. Periods of enhanced or reduced column anomalies often coincide with larger-scale atmospheric

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

changes that can also influence surface ozone conditions. The proxy comparison therefore serves as a supplementary diagnostic to verify that the model responds consistently to large-scale variability, while avoiding a direct interpretation as surface forecast skill.

The similar behaviour across the three ablation configurations further suggests that the differences between model variants mainly affect near-surface ozone dynamics rather than the large-scale ozone variability represented by the CAMS total-column field.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

Appendix E: Temporal Fusion Transformer Hyperparameters

Category	Setting
Data / windowing (PyTorch Forecasting TimeSeriesDataSet)	
Time index / target	`time_idx` / O_3
Group identifiers	`[station_code, latitude, longitude, altitude]`
Encoder length	336 timesteps (hours)
Prediction length	96 timesteps (hours)
Missing timesteps	`allow_missing_timesteps = True`
Relative time index	`add_relative_time_idx = True`
Target scaling features	`add_target_scales = True`
Encoder length feature	`add_encoder_length = True`
Static categorical	`[station_type, station_code, station_area]`
Static real	`[mean_emission, mean_pop_den, max_pop_den, mean_nlights, max_nlights, mean_topalt, min_topaltrel, max_topaltrel, stddev_topaltrel, roaddist, encoder_length, O3_center, O3_scale]`
Known real	`[cloudcover, pblheight, press, relhum, temp, u, v, relative_time_idx]`
Unknown real	`[O3, NO, NO2]`
Model architecture (TFT)	
Causal attention	`True`
Attention head size	4
Hidden size	16
LSTM layers	1
Dropout	0.1
Output size	7 (quantiles)
Embeddings	`station_code: (387, 45), station_type: (3, 3), station_area: (3, 3)`
Optimization / training	
Optimizer	`ranger`
Learning rate	3.7153523e-3
ReduceLROnPlateau	`min_lr = 1e-5, patience = 4, factor = 2.0`
Logging interval	10 / 10 (train / val)
Loss and reported metrics (from model printout)	
Loss	Quantile loss with $\tau \in \{0.02, 0.10, 0.25, 0.50, 0.75, 0.90, 0.98\}$
Metrics	sMAPE, MAE, RMSE, MAPE

Table E1. Key hyperparameters for the Temporal Fusion Transformer (TFT) ozone forecasting model.

https://doi.org/10.5194/egusphere-2026-1562
Preprint. Discussion started: 21 April 2026

E1 Korea station network

Category	Setting
Data / windowing (TimeSeriesDataSet configuration)	
Time index / target	`time_idx` / O_3
Group identifiers	`[station_code, latitude, longitude, altitude]`
Encoder length	336 timesteps (hours)
Prediction length	96 timesteps (hours)
Missing timesteps	`allow_missing_timesteps = True`
Relative time index	`add_relative_time_idx = True`
Target scaling features	`add_target_scales = True`
Encoder length feature	`add_encoder_length = True`
Static categorical covariates	`[station_type, station_code, station_area]`
Static real covariates	`[mean_emission, mean_pop_den, max_pop_den, mean_nlights, max_nlights, mean_topalt, min_topaltrel, max_topaltrel, stddev_topaltrel, roaddist]`
Known real covariates	`[cloudcover, pblheight, press, relhum, temp, u, v]`
Unknown real covariates	[O_3, NO_2]
Model architecture (TFT)	
Causal attention	`True`
Attention head size	4
Hidden size	16
LSTM layers	1
Dropout	0.1
Output size	7 (quantile outputs)
Embeddings	`station_code: (59, 16), station_type: (1, 1), station_area: (1, 1)`
Optimization / training	
Optimizer	ranger
Learning rate	1.0e-3
Weight decay	0.0
ReduceLROnPlateau	`min_lr = 1e-5, patience = 1000, factor = 2.0`
Logging interval (train / val)	`-1 / -1` (disabled)
Loss and reported metrics	
Loss	Quantile loss with $\tau \in \{0.02, 0.10, 0.25, 0.50, 0.75, 0.90, 0.98\}$
Metrics	sMAPE, MAE, RMSE, MAPE

Table E2. Key hyperparameters for the Temporal Fusion Transformer (TFT) configuration used for the Korea station network experiment.