

APPLICATION PAPER  

Short-term forecasting of ozone air pollution across Europe with transformers

Sebastian H. M. Hickman^{1,2} , Paul T. Griffiths^{1,3} , Peer J. Nowack⁴ and Alexander T. Archibald^{1,3}

¹Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, United Kingdom

²The Alan Turing Institute, London, United Kingdom

³National Centre for Atmospheric Science, University of Cambridge, Cambridge, United Kingdom

⁴Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

Corresponding author: Sebastian H. M. Hickman; Email: shmh4@cam.ac.uk

Received: 28 February 2023; **Revised:** 22 August 2023; **Accepted:** 24 October 2023

Keywords: Air pollution; forecasting; machine learning; transferability; transformers

Abstract



Surface ozone is an air pollutant that contributes to hundreds of thousands of premature deaths annually. Accurate short-term ozone forecasts may allow improved policy actions to reduce the risk to human health. However, forecasting surface ozone is a difficult problem as its concentrations are controlled by a number of physical and chemical processes that act on varying timescales. We implement a state-of-the-art transformer-based model, the temporal fusion transformer, trained on observational data from three European countries. In four-day forecasts of daily maximum 8-hour ozone (DMA8), our novel approach is highly skillful (MAE = 4.9 ppb, coefficient of determination $R^2 = 0.81$) and generalizes well to data from 13 other European countries unseen during training (MAE = 5.0 ppb, $R^2 = 0.78$). The model outperforms other machine learning models on our data (ridge regression, random forests, and long short-term memory networks) and compares favorably to the performance of other published deep learning architectures tested on different data. Furthermore, we illustrate that the model pays attention to physical variables known to control ozone concentrations and that the attention mechanism allows the model to use the most relevant days of past ozone concentrations to make accurate forecasts on test data. The skillful performance of the model, particularly in generalizing to unseen European countries, suggests that machine learning methods may provide a computationally cheap approach for accurate air quality forecasting across Europe.

Impact Statement

Ozone is a harmful air pollutant that contributes to hundreds of thousands of deaths every year. Making accurate short-term forecasts of ozone is necessary to provide the public with timely and accurate air quality warnings. We propose a forecasting system for ozone air pollution using a transformer model, a machine learning architecture that allows accurate and computationally cheap forecasts of ozone using meteorological variables as inputs. The model performs skillfully in countries across Europe, highlighting its transferability.

1. Introduction

Ozone is a secondary pollutant that is not directly emitted by anthropogenic activities but formed in the troposphere via a series of photochemical reactions (Finlayson-Pitts and Pitts, 1997). Once formed, ozone

  This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

can be transported over long distances and across continents (Lin et al., 2015). There are a number of relationships between ozone and other variables, such as other pollutants and temperature (Laña et al., 2016), which are, for example, relevant for severe ozone pollution during major heat waves (Archibald et al., 2020). Ozone at the surface is estimated to contribute to between 365,000 and 1,100,000 premature deaths worldwide annually (Anenberg et al., 2010; Silva et al., 2013; Malley et al., 2017; Murray et al., 2020), primarily by causing cardiovascular and respiratory diseases (Filippidou and Koukoulia, 2011; Kim et al., 2020; Sun et al., 2022). The impacts of ozone pollution have been linked to both long- and short-term exposure (Bell et al., 2004; Nuvolone et al., 2018).

Ozone air pollution is both a global and a local issue. Background levels of ozone in remote areas often exceed guidelines set by the World Health Organization (WHO), while local ozone concentrations in urban and peri-urban areas can far exceed these guidelines. The WHO estimates that 99% of the world's population lives in areas where pollutant concentrations routinely exceed guidelines (WHO and ECE, 2021).

Due to the phytotoxicity of ozone, the negative effects of ozone air pollution on vegetation, ecosystems, and crop yields are also significant (Emberson et al., 2001; Fowler et al., 2009; Emberson, 2020). This damage leads to both considerable economic losses from reduced crop yields (Burney and Ramanathan, 2014), and the potential for increased climate change as damaged vegetation has a reduced capacity to sequester carbon dioxide from the atmosphere (Sitch et al., 2007; Ainsworth et al., 2012; Lombardozi et al., 2015; Wang et al., 2016; Oliver et al., 2018).

There is strong evidence that ozone levels increase with increasing temperature (Porter et al., 2015), leading to the suggestion that under climate change, some regions will become more polluted with ozone (Bloomer et al., 2009; Rasmussen et al., 2013; Schnell et al., 2016; Brown et al., 2022), in turn leading to increased risk to human health. This effect is known as the ozone “climate penalty” (Rasmussen et al., 2013). The risks of increased ozone are compounded as extreme ozone episodes are often accompanied by high temperatures, leading to a combination of risks that further increase mortality (Dear et al., 2005; Filleul et al., 2006; Lei et al., 2012). These compound events are also associated with increased levels of other pollutants, such as PM_{2.5}, and therefore, understanding and predicting their impact is a fruitful avenue to mitigate risks to health (Galindo et al., 2011; Schnell and Prather, 2017).

In addition to chemical effects under climate change, the lifetime of ozone, which is of the order of weeks in the free troposphere (Jacob et al., 1996; Fiore et al., 2002), means that transport and meteorology are also important when determining local ozone levels (Thompson et al., 1996; Zhang et al., 2008). For example, transport between North America and Europe has been purported to lead to increased ozone concentrations in Europe (Li et al., 2002; Derwent et al., 2004).

Regional scale features such as blocking result in extended periods of stagnant conditions, which may also lead to increased ozone (Garrido-Perez et al., 2019; Otero et al., 2021). These meteorological effects typically cause changes in other environmental variables, leading to compound effects, such as the combined effect of heat waves and pollution (Li et al., 2020).

In order to mitigate the impacts of ozone pollution on human health, skillful short-term forecasts of ozone concentrations, particularly at extrema, would allow preventative government policy, such as providing air quality warnings (Kelly et al., 2012; Iordache et al., 2015). To reduce ozone concentrations, a better quantitative understanding of the causes and drivers of ozone would provide a basis for governmental interventions to reduce ozone and hence risk to human health, and provide better understanding of how ozone may evolve under changing climates (Archibald et al., 2020).

Due to the transport of ozone precursors and the strong diurnal cycle of ozone concentrations, it was recognized that severe ozone episodes can last for up to 8 hours. In 1997, the United States Environmental Protection Agency updated their guideline ozone metric to the daily maximum 8-hour mean concentration (DMA8) (Chameides et al., 1997; EPA U, 1997). Since this decision, the DMA8 metric has typically been used to evaluate the risk of ozone pollution to human health and is used by the WHO to set target ozone concentrations (WHO and ECE, 2021). This metric has been found to be more strongly associated with adverse health outcomes such as respiratory and cardiovascular diseases than other metrics (Bell et al.,

2005; Yang et al., 2012; Li et al., 2018). As the most widely used daily metric for ozone, including by the WHO (WHO and ECE, 2021), it is DMA8 that we focus on.

1.1. Air pollution forecasting

Traditionally, numerical chemical transport models (CTMs) have been used for air pollution forecasting. However, these models are typically highly computationally expensive, and resolution is often an issue requiring parameterizations, which can introduce inconsistencies in model predictions (Weng et al., 2023). Machine learning (ML) may provide a complement to existing numerical CTMs and simple statistical approaches to modeling air pollution, and climate phenomena in general, as ML allows automatic learning of the behavior of a complex system from data.

In this work, we evaluate the short-term forecasting skill of a transformer-based ML architecture across a range of European countries. The model makes 4-day forecasts of ozone in both urban and rural environments. We build on existing work by applying ML to forecast ozone air pollution, by applying a novel transformer-based method, and by evaluating the skill of the method across a wider range of test data than has been studied previously.

Previous studies focusing on forecasting surface ozone with ML have largely focused on predicting ozone in specific regions, often with relatively short time series of data. A variety of methods have been used, including bias-corrected CTMs (Neal et al., 2014; Ivatt and Evans, 2020), linear regression (Olszyna et al., 1997; Thompson et al., 2001), and feed-forward neural networks (Comrie, 1997; Cobourn et al., 2000). More recently, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been used in an effort to better capture spatial and temporal dependencies (Biancofiore et al., 2015; Eslami et al., 2020; Ma et al., 2020; Sayeed et al., 2020; Kleinert et al., 2021). These studies illustrate the promising advances in short-term ozone forecasting possible with ML methods, in some cases outperforming state-of-the-art CTMs (Table 1).

To our knowledge, this is the first study that uses a purely transformer-based model to make accurate forecasts at a large number of stations across different environments and countries, and furthermore, that

Table 1. The relative performance of different ML and numerical approaches to ozone forecasting

Method (paper)	r (Pearson)	RMSE/ppb	Stations
Chemical transport models			
GEOS-Chem (Ivatt and Evans, 2020)	0.48	16.2	2,200
AQUM (Neal et al., 2014)	0.64	20.9	61
Bias-corrected AQUM (Neal et al., 2014)	0.76	16.4	61
Bias-corrected GEOS-Chem (Ivatt and Evans, 2020)	0.84	7.5	2,200
ML methods			
DRR (Debry and Mallet, 2014)	0.70	6.3	729
CNN (Sayeed et al., 2020)	0.77	8.8	21
CNN (Eslami et al., 2020),	0.79	12.0	25
RNN (Biancofiore et al., 2015)	0.86	12.5	1
CNN-transformer (Chen et al., 2022)	<i>NA</i>	7.8	14
Our dataset			1,012
<i>Persistence</i>	0.67	10.9	
<i>Ridge regression</i>	0.69	10.8	
<i>Random forest</i>	0.80	9.0	
<i>LSTM</i>	0.84	8.3	
TFT	0.91	6.6	

Note. Methods in italics were tested on our dataset, while others used different data. The difficulty of comparing methods tested on different datasets is shown by the varying RMSE values.

evaluates the capacity of an ML model to make forecasts on data drawn from countries outside the training dataset. In addition, as far as we are aware, only one other study has applied an architecture with a transformer-based component to ozone forecasting (Chen et al., 2022).

2. Methods and Data

2.1. Model

Transformers have been shown to be highly effective in sequential domains such as natural language processing (Brown et al., 2020; Ji et al., 2021), in part due to their ability to attend to long-term dependencies in the data. Therefore, a transformer-based model may provide an intrinsic advantage over other ML models (such as random forests) and convolutional and recurrent neural networks that have been previously explored in the ozone forecasting literature (Biancofiore et al., 2015; Eslami et al., 2020; Kleinert et al., 2021).

Therefore, to complement existing numerical CTMs and ML methods for ozone forecasting, we implement a state-of-the-art transformer-based deep learning architecture, temporal fusion transformer (TFT) (Lim et al., 2021). The TFT combines gated residual networks, variable selection networks, a long short-term memory (LSTM) encoder–decoder layer, and multi-head attention and is described in detail in [Appendix A.5](#).

The TFT is able to ingest both static (e.g., local population density, altitude, and landcover) and dynamic (e.g., temperature, wind speed, and cloud cover) features to make forecasts. In order to extract prediction intervals from the TFT, a quantile loss function was implemented ([Appendix A.7](#)), which provides a direct means to estimate forecast uncertainty as part of our methodology (Lim et al., 2021). Where the median quantile is predicted in the quantile loss, the loss is the mean absolute error loss. The quantile loss function is described in more detail in [Appendix A.7](#).

Despite being a relatively computationally expensive ML method, training the TFT on our dataset took less than an hour using 2 Tesla V100 GPUs. Once trained, making consecutive 4-day forecasts for a year of data across 1012 individual stations takes around a minute. This illustrates the substantial time savings compared to CTMs. Hyperparameters were optimized using Bayesian optimization, a method to robustly find optimal hyperparameters (e.g., number of layers and learning rate, as shown in [Table A2](#)) from ranges of possible values (Mockus, 2012). Hyperparameter optimization was carried out on a withheld validation dataset, which was then not used subsequently for model testing.

2.2. Data

The Tropospheric Ozone Assessment Report (TOAR) dataset (Schultz et al., 2017) was selected as a suitable dataset for our forecasting model due to its global coverage and high fidelity and quantity of data, with daily measurements stretching back to the 1980s in some locations. The dataset is hosted by the Jülich Supercomputing Centre and provides around 2.6 billion observations of ozone concentrations in total.

Data from three European countries were collected: the UK, France, and Italy. These were chosen to represent three different domains in order to test whether a single model could be trained to make accurate forecasts across domains. Data from all months of the year and from urban and rural environments were included in our dataset. This dataset therefore provides a larger sample of different environments than have been studied in previous work (Biancofiore et al., 2015; Kleinert et al., 2021), with data from 1997 to 2014, from 1012 individual stations. Our final dataset contains more than 2 million individual days of data.

We processed the data to select features relevant to ozone to use as predictive inputs to the ML model. As the TOAR dataset is designed to include variables relevant for ozone, this streamlined data processing. The data include both static and dynamic features relevant to ozone concentrations. The static features relate to characteristics of a particular station, such as the local population density, while the dynamic features are environmental variables that change through time, such as temperature. The inputs used are

described in full in [Table A1](#). ML algorithms typically require clean datasets without missing values. We therefore removed days of data with missing values and compared summary statistics of the two datasets to ensure this did not introduce bias. Furthermore, if we had missing days of data during a particular 25-day sequence (as used for forecasting by the model), the sequence was dropped. Due to the large size and relative completeness of our dataset, imputing missing values with algorithms such as k-nearest neighbors (Batista et al., 2002) was deemed unnecessary. By simply removing missing data, we removed the risk of bias from data imputation. All dynamic features were scaled with robust scaling ([Appendix A.1](#)), and both planetary boundary layer height and ozone were log-transformed to improve model performance (Jayalakshmi and Santhakumaran, 2011).

To train, validate, and test our models, we first optimized hyperparameters (e.g., number of layers and learning rate) on data from the year 2013, training on all non-test years. The computational expense of Bayesian optimization for hyperparameter tuning meant that we were limited to optimizing the hyperparameters on a single year of validation data. We then trained this fixed architecture on 5 different sets of training data using 1 year at a time as the test data (2008 to 2012) to evaluate the predictive performance of the model across a range of years. For example, when testing on 2008 data, all other years (1997 to 2014, excluding 2008 and 2013) were used as training data, and the model skill was then evaluated on 2008. The same procedure was followed for years 2008 to 2012, and then the model skill metrics were averaged across these five test years to give final model skill metrics.

This temporal splitting of data resulted in an approximately 80%-10%-10% (e.g., testing on 2012, hyperparameter optimization on 2013, and training all other years) split between training, validation, and test data. The previous 21 days of station-specific observations of ozone and dynamic covariates and 4 days of future dynamic covariate data (emulating available weather forecasts at the time) were used to make ozone forecasts up to 4 days ahead at the same station. 21 days was chosen as a suitable trade-off between computational expense and performance, retaining the capacity to account for extended air quality events. The sensitivity of using more previous time-steps of data as inputs to the model was found to be small beyond using the previous 21 days ([Appendix A.2](#)).

3. Results and Discussion

3.1. Forecasting ozone

When forecasting ozone concentrations, the TFT was skillful (MAE = 4.9 ppb, $R^2 = 0.81$, RMSE = 6.6 ppb, $r = 0.91$). These predictions, which use previous ozone observations and ERA5 meteorological reanalysis data as a proxy for both previous and forecasted meteorological data ([Table A1](#)), are therefore suitable for short-term future forecasts with meteorological forecasts as input and also for infilling missing ozone values in historical datasets.

While we cannot make direct comparisons with all similar methods due to differing test datasets, the skill of our method compares favorably to other ML methods and numerical air quality forecasting models such as AQUM (Neal et al., 2014; Im et al., 2015), especially given the size and variety of our test dataset ([Table 1](#)). In addition, a number of other ML algorithms (ridge regression, random forests, and LSTMs) were trained and tested on our data to evaluate the relative skill of the TFT.

A correlation plot of TFT predictions on the test set, against observations, is given in [Figure 1a](#). The model was more accurate than other ML approaches such as random forests and LSTMs and approximately 40% more accurate than a persistence model (which predicts ozone as the same value as the previous day ozone) in terms of RMSE ([Table 1](#)).

We can further visualize the skill of the TFT by looking at predictions and observations at individual stations in our dataset. [Figure 1b](#) shows the previous days that the attention mechanism in the model used to inform the predictions, shown by the gray line denoting attention. The model pays attention to previous high ozone days to make future forecasts of high ozone concentrations. Similarly, when forecasting future low ozone, the model pays attention to previous days of low ozone concentrations ([Figure A.6](#)). [Figure 1b](#) also illustrates the prediction intervals generated by the quantile loss function, which are useful to evaluate

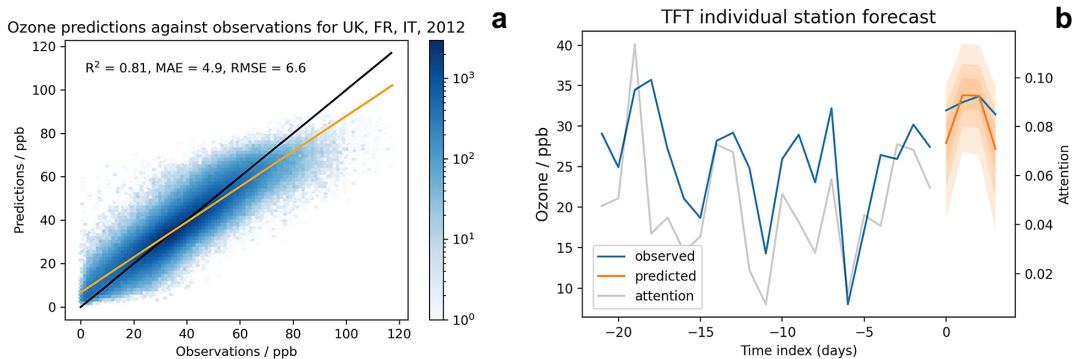


Figure 1. (a) illustrates predictions against observations on 2012 test data for forecasting ozone with the TFT. The number of data points in each bin is shown by the color bar, using a log scale. (b) shows a 4-day forecast on test data at a single station. The gray line shows the attention that the transformer is paying to different days in the time history. The prediction intervals generated with the quantile loss are also shown, with the 7 different quantiles illustrated by orange shading.

trust in the model. As expected, the prediction intervals generally increase for longer lead times, as shown in Figure 4.

The TFT's capacity to make skillful predictions of ozone concentrations at urban and rural stations (as denoted in the TOAR dataset) was also tested. The TFT performed similarly on both urban and rural data (MAE = 5.0, $R^2 = 0.81$ and MAE = 4.8, $R^2 = 0.81$, respectively), which suggests that architectures of this type are able to generalize across these two environments given sufficient training data.

As seen in Table 1, the TFT is the most skillful model on our test data, followed by the LSTM. The TFT and LSTM are designed to deal with sequential data and are able to learn from large quantities of data, which likely contributes to these models performing better than other methods. Ridge regression is a linear model that likely limits its performance; however, it is computationally cheap and provides direct interpretability. The TFT and LSTM also outperform the random forest model, a nonlinear tree-based model. While neural network models are typically not as interpretable as linear models like ridge regression, the TFT does facilitate interpretability studies that provide insight into the model's behavior beyond what is possible in RNNs and LSTMs. Furthermore, the TFT is better adapted to ingesting static features than other machine learning methods tested.

3.2. Forecasting extreme ozone

Ozone concentrations in Europe tend to peak in spring and summer months, typically between April and June (Monks, 2000). Making accurate forecasts of high ozone is important as these high ozone concentrations pose a great threat to health (Bell et al., 2004) and may occur more frequently in some regions in future climates (Doherty et al., 2013; Orru et al., 2019). We therefore evaluated the skill of the TFT during these high ozone periods. Figure 2b illustrates that the TFT was able to make reasonably skillful forecasts on spring and summertime ozone concentrations (MAE = 5.4 ppb, $R^2 = 0.64$). However, the performance was worse than forecasting on data from the rest of the year (MAE = 4.8 ppb, $R^2 = 0.85$). The variability of ozone is significantly higher in spring and summer months, which makes accurate forecasting more difficult, and may require more inputs to the model, such as soil NO_x emissions (Porter and Heald, 2019).

Furthermore, model performance across countries was evaluated during just spring (March, April, and May) and just summer months (June, July, and August). When testing across countries (Figure 3), performance in terms of mean absolute percentage error (MAPE), MAE, and R^2 was significantly better for spring than for summer for most countries. The latitude of the country appeared to have little relationship with R^2 or MAPE; however, unsurprisingly, MAE was higher for the countries where ozone

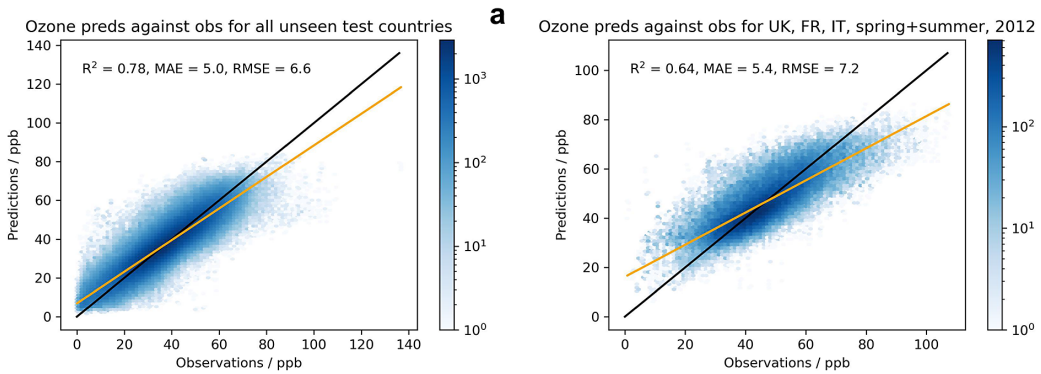


Figure 2. (a) illustrates the performance of the TFT when predicting on 2012 test data from 13 European countries unseen during training. (b) shows that when forecasting on spring and summertime test data, the performance of the TFT was poorer (MAE = 5.4 ppb, $R^2 = 0.64$) than predicting on the whole year.

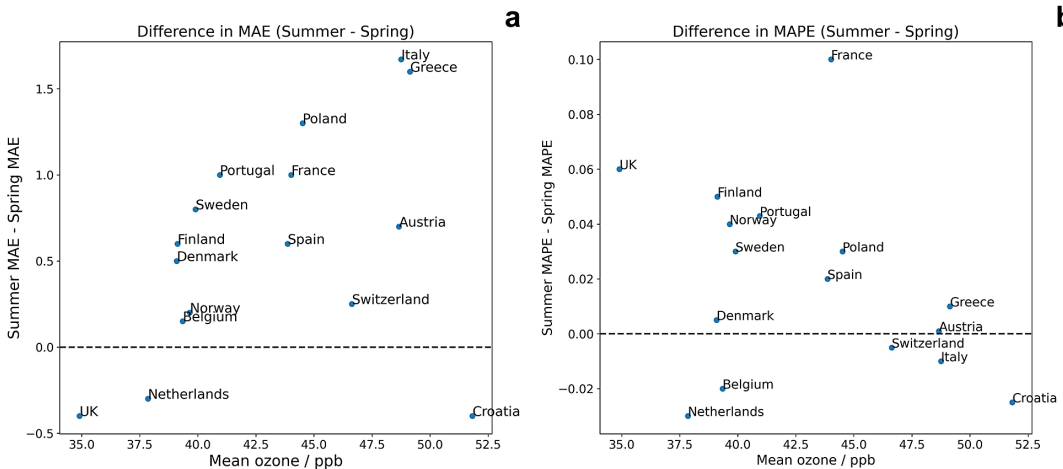


Figure 3. (a) illustrates the difference in performance of the model in different countries, in terms of MAE, between spring and summer, while (b) illustrates the same for MAPE. The countries are plotted by the mean ozone during spring and summer in the country.

levels are greater on average in summer. The poorer performance of the model in summer points to variables affecting ozone in summer months that are unaccounted for in the model. This could be caused by a number of factors, including the effect of biogenic VOCs (Porter and Heald, 2019; Cao et al., 2022), or the calculation of planetary boundary layer height in the reanalysis data, as planetary boundary layer height is driven by convection, a parameterized process. For example, it has recently been shown that the European Centre for Medium-Range Weather Forecasts’ Integrated Forecast System overestimates planetary boundary layer height over the Eastern Mediterranean in summertime compared to ceilometer observations (Uzan et al., 2020).

3.3. TFT generalizes better than other ML approaches

While the improved skill of the TFT in the domain of the data it was trained on (UK, France, and Italy) is impressive, it is important for production ML models to perform accurately on data outside the country

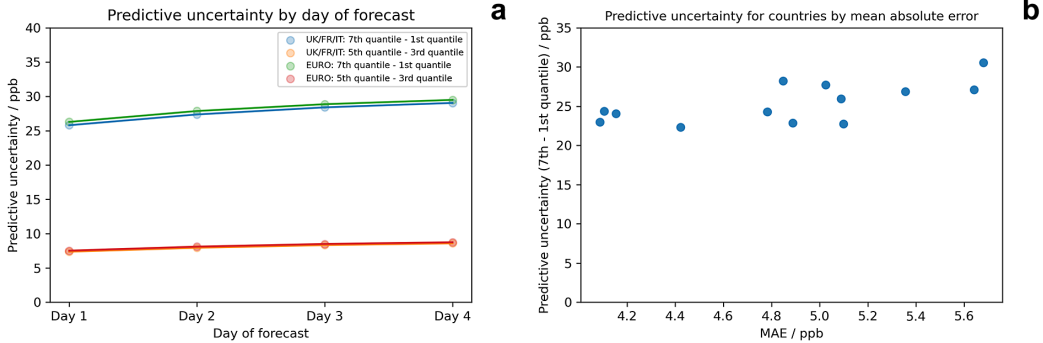


Figure 4. (a) shows the width of prediction intervals generated by the model on the test data for the countries used for training (UK, France, Italy) and for test data from the unseen European countries. The prediction intervals are marginally wider for the unseen countries, in line lower model performance in these countries. (b) illustrates that the prediction intervals increase as the MAE of the predictions increases, consistent with well-calibrated prediction intervals.

Table 2. The relative performance of different ML approaches to ozone forecasting for different domains within our data

Performance on UK, IT, FR	R ²	RMSE / ppb	MAE/ppb
Persistence	0.53	10.9	8.8
Ridge regression	0.54	10.8	8.6
Random forest	0.66	9.0	6.9
LSTM	0.70	8.3	6.1
TFT	0.81	6.6	4.9
Rest of Europe			
Persistence	0.55	9.3	7.3
Ridge regression	0.21	12.5	9.9
Random forest	0.14	12.8	10.1
LSTM	0.68	8.5	6.2
TFT	0.78	6.6	5.0

Note. The values show the ability of the LSTM and the TFT to generalize well to new countries across Europe, while other methods fail to generalize as effectively. Detailed results for each testing country are available in [Appendix A.3](#).

or domain that they are trained on. The TFT was not only better able to perform within the domain of its training data, it also better generalized to new domains than the other ML methods, which suggests that the model is better able to capture the underlying dynamics of the system controlling ozone concentrations (Table 2). To evaluate the skill of our model in generalizing to unseen data, we deployed the model, trained solely on data from the UK, France, and Italy, to make forecasts in a separate test set comprising 13 other European countries with the same covariates. The model was able to generalize impressively on data from these unseen countries when evaluated on a single year of test data ($R^2 = 0.78$, MAE = 5.0 ppb, RMSE = 6.6 ppb), as shown in Figure 2a. This suggests that the model could act as a Europe-wide predictive model, without requiring retraining. The predictive uncertainty for the model on the unseen countries was marginally wider but, similar to the training countries, was in line with the skill of the predictions. Furthermore, the predictive uncertainty was reasonably well calibrated with the accuracy of the model—in countries where the model performed poorly, the predictive uncertainty was larger (Figure 4).

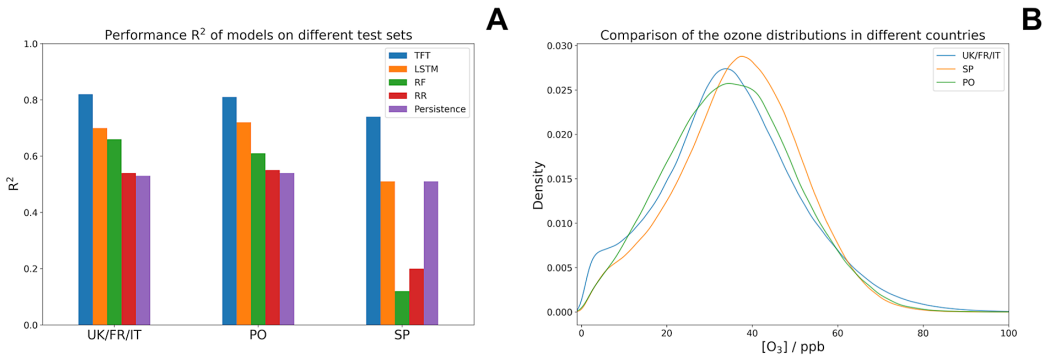


Figure 5. (a) shows the accuracy of different ML methods on the test data for the countries used for training (UK, France, Italy) and for a pair of unseen countries (Spain and Poland). The performance of the LSTM and the TFT is relatively stable when forecasting in new countries, while the random forest and ridge regression models perform poorly. (b) shows a density plot of ozone concentrations observed in the test data for the countries used for training (UK, France, Italy) and for 2 of the 13 unseen test countries (Spain and Poland).

In previous work, when testing in a new domain, the similarity of the domain to the training domain appeared to significantly affect performance. This is likely due to correlations learned from the training set that do not hold in the new domain, or country. Such inconsistencies in the relationships point to an issue related to the models being able to identify the generic causal drivers of ozone, rather than spurious correlations specific to individual locations (Runge et al., 2019). At least, we empirically find that the TFT performance is more robust to being deployed away from its spatial training domain than, for example, random forest models, which often poorly generalize to out-of-distribution data. We highlight this important result for two of the unseen European countries, Spain and Poland, to illustrate the relative skill in terms of R^2 different methods in Figure 5a.

4. What is the TFT Paying Attention To?

We extracted feature importances, derived from the weights of attention mechanism in our ozone forecasting model, to examine which dynamic features are the most important for the model when making forecasts with our data (Figure 6). These importances correspond well with expected physical drivers: Both temperature and planetary boundary layer height are key variables (Porter and Heald, 2019). The skillful performance of the TFT suggests that good forecasting skill can be achieved with only the meteorological variables used by the model, which may simplify implementation of this method operationally, similar to recent results using random forests (Weng et al., 2022). Static feature importances are shown in Figure A.7. When training and testing the model without static features, we found only slight degradation of model performance ($R^2=0.81$, MAE = 5.0 ppb, RMSE = 6.7 ppb), suggesting that the inclusion of these variables is not essential for high model skill, further simplifying operational implementation.

5. Conclusions

Despite decades-long model development, ozone remains challenging to forecast with existing numerical methods. Our ML model, the TFT, makes skillful forecasts of ozone concentrations at stations across Europe. The model requires only static features and dynamic variables available from weather forecasts and could therefore feasibly be used operationally to forecast ozone at observational stations. The model is able to make accurate forecasts across environments and performs reasonably well when

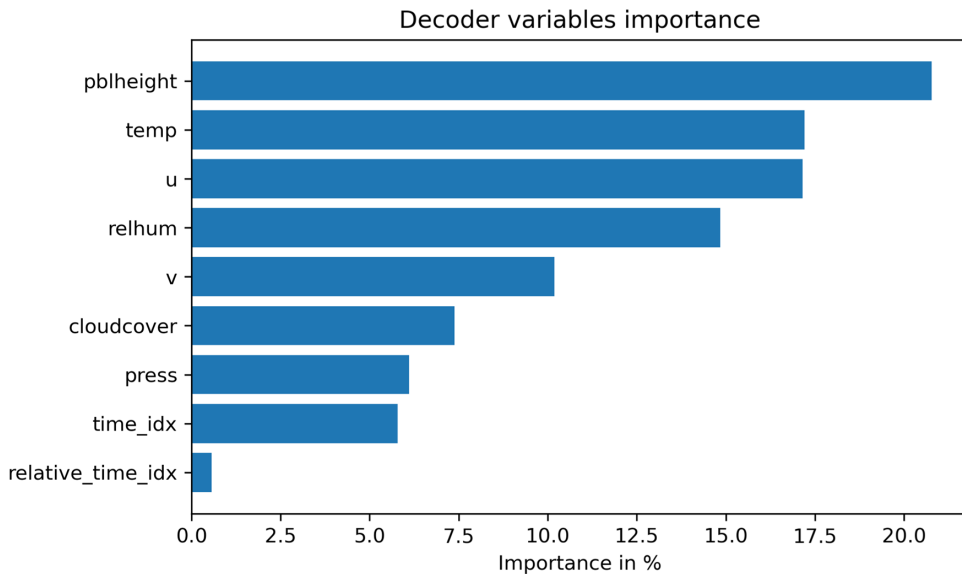


Figure 6. The variable importances of the TFT when making forecasts, derived from the weights of the attention mechanism. These are largely in line with expected physical relationships.

predicting extrema. The model is able to generalize well to data from 13 European countries unseen in training, outperforming other ML methods in these new domains. This suggests, in combination with feature importances, that the model is learning sound physical relationships in the data. The model provides a promising, computationally cheap method to make accurate forecasts of ozone across Europe.

However, further work is required to ensure the model can make accurate predictions of extrema, such as explicitly encoding known physical relationships in the model, and furthermore that operational forecasts at short lead times can be used instead of reanalysis data. In addition, while we have demonstrated that the model generalizes well to other European countries, it would be a worthwhile question to explore how far the model can be taken across to countries and world regions with even more pronounced differences in meteorology and emission regulations. However, an initial extension of this work could be to evaluate the skill of the model in locations with similar ozone levels and meteorological conditions, such as other regions in the Northern Hemisphere midlatitudes.

Acknowledgments. The authors thank the NERC Earth Observation Data Acquisition and Analysis Service (NEODAAS) for access to compute resources and staff support for this study and the TOAR database maintainers.

Author contribution. Formal analysis: S.H.M.H.; Investigation: S.H.M.H. and P.T.G.; Methodology: S.H.M.H., P.T.G., P.J.N., and A.T.A.; Software: S.H.M.H.; Visualization: S.H.M.H., P.T.G., P.J.N., and A.T.A.; Writing—original draft: S.H.M.H.; and Writing—review and editing: S.H.M.H., P.T.G., P.J.N., and A.T.A.

Competing interest. The authors declare none.

Data availability statement. The TOAR dataset is publicly available online (<https://join.fz-juelich.de/services/rest/surfacedata/>, last accessed 22/08/2023) (Schultz et al., 2017), and the code of this work is available at <https://github.com/shmh40/forecasto3> (last accessed 22/08/2023).

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. S.H.M.H. acknowledges funding from EPSRC via the AI4ER CDT at the University of Cambridge (EP/S022961/1) and support from The Alan Turing Institute. P.T.G. and A.T.A. were financially supported by NERC through NCAS (R8/H12/83/003). P.N. was supported by the Natural Environment Research Council (grant no. NE/V012045/1).

References

- Ainsworth EA, Yendrek CR, Sitch S, Collins WJ, Emberson LD, et al. (2012) The effects of tropospheric ozone on net primary productivity and implications for climate change. *Annual Review of Plant Biology* 63(1), 637–661.
- Anenberg SC, Horowitz LW, Tong DQ and West JJ (2010) An estimate of the global burden of anthropogenic ozone and fine particulate matter on premature human mortality using atmospheric modeling. *Environmental Health Perspectives* 118(9), 1189–1195.
- Archibald AT, Turnock ST, Griffiths PT, Cox T, Derwent RG, Knote C and Shin M (2020) On the changes in surface ozone over the twenty-first century: Sensitivity to changes in surface temperature and chemical mechanisms. *Philosophical Transactions of the Royal Society A* 378(2183), 20190329.
- Batista GE, Monard MC, et al. (2002) A study of k-nearest neighbour as an imputation method. *His* 87(251–260), 48.
- Bell ML, Hobbs BF and Ellis H (2005) Metrics matter: Conflicting air quality rankings from different indices of air pollution. *Journal of the Air & Waste Management Association* 55(1), 97–106.
- Bell ML, McDermott A, Zeger SL, Samet JM and Dominici F (2004) Ozone and short-term mortality in 95 US urban communities, 1987–2000. *JAMA* 292(19), 2372–2378.
- Biancofiore F, Verdecchia M, Di Carlo P, Tomassetti B, Aruffo E, Busilacchio M, Bianco S, Di Tommaso S and Colangeli C (2015) Analysis of surface ozone using a recurrent neural network. *Science of the Total Environment* 514, 379–387.
- Biewald L (2020) Experiment tracking with weights and biases. Available at <https://www.wandb.com/> (accessed 13 November 2023).
- Bloomer BJ, Stehr JW, Piety CA, Salawitch RJ and Dickerson RR (2009) Observed relationships of ozone pollution with temperature and emissions. *Geophysical Research Letters* 36(9).
- Brown F, Folberth GA, Sitch S, Bauer S, Bauters M, Boeckx P, Cheesman AW, Deushi M, Dos Santos I, Galy-Lacaux C, et al. (2022). The ozone–climate penalty over South America and Africa by 2100. *EGUsphere* 1–33.
- Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. (2020) Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33, 1877–1901.
- Burney J and Ramanathan V (2014) Recent climate and air pollution impacts on Indian agriculture. *Proceedings of the National Academy of Sciences* 111(46), 16319–16324.
- Cao J, Situ S, Hao Y, Xie S and Li L (2022) Enhanced summertime ozone and SOA from biogenic volatile organic compound (BVOC) emissions due to vegetation biomass variability during 1981–2018 in China. *Atmospheric Chemistry and Physics* 22(4), 2351–2364.
- Chameides W, Saylor R and Cowling E (1997) Ozone pollution in the rural United States and the new NAAQS. *Science* 276(5314), 916–916.
- Chen Y, Chen X, Xu A, Sun Q and Peng X (2022) A hybrid CNN-transformer model for ozone concentration prediction. *Air Quality, Atmosphere & Health* 15, 1–14.
- Cobourn WG, Dolcine L, French M and Hubbard MC (2000) A comparison of nonlinear regression and neural network models for ground-level ozone forecasting. *Journal of the Air & Waste Management Association* 50(11), 1999–2009.
- Comrie AC (1997) Comparing neural networks and regression models for ozone forecasting. *Journal of the Air & Waste Management Association* 47(6), 653–663.
- Dear K, Ranmuthugala G, Kjellström T, Skinner C and Hanigan I (2005) Effects of temperature and ozone on daily mortality during the August 2003 heat wave in France. *Archives of Environmental & Occupational Health* 60(4), 205–212.
- Debry E and Mallet V (2014) Ensemble forecasting with machine learning algorithms for ozone, nitrogen dioxide and pm10 on the Prev'Air platform. *Atmospheric Environment* 91, 71–84.
- Derwent R, Stevenson D, Collins W and Johnson C (2004) Intercontinental transport and the origins of the ozone observed at surface sites in Europe. *Atmospheric Environment* 38(13), 1891–1901.
- Doherty R, Wild O, Shindell D, Zeng G, MacKenzie I, Collins W, Fiore AM, Stevenson D, Dentener F, Schultz M, et al. (2013) Impacts of climate change on surface ozone and intercontinental ozone pollution: A multi-model study. *Journal of Geophysical Research: Atmospheres* 118(9), 3744–3763.
- Emberson L, Ashmore M, Simpson D, Tuovinen J-P and Cambridge H (2001) Modelling and mapping ozone deposition in Europe. *Water, Air, and Soil Pollution* 130(1), 577–582.
- Emberson L (2020) Effects of ozone on agriculture, forests and grasslands. *Philosophical Transactions of the Royal Society A* 378(2183), 20190327.
- EPA U (1997) National ambient air quality standards for ozone: Proposed decision. *Federal Register* 61(241), 65717–65750.
- Eslami E, Choi Y, Lops Y and Sayeed A (2020) A real-time hourly ozone prediction system using deep convolutional neural network. *Neural Computing and Applications* 32(13), 8783–8797.
- Filippidou E and Koukoulia A (2011) Ozone effects on the respiratory system. *Progress in Health Sciences* 1(2).
- Filleul L, Cassadou S, Médina S, Fabres P, Lefranc A, Eilstein D, Le Tertre A, Pascal L, Chardon B, Blanchard M, et al. (2006) The relation between temperature, ozone, and mortality in nine French cities during the heat wave of 2003. *Environmental Health Perspectives* 114(9), 1344–1347.
- Finlayson-Pitts BJ and Pitts JN (1997) Tropospheric air pollution: Ozone, airborne toxics, polycyclic aromatic hydrocarbons, and particles. *Science* 276(5315), 1045–1051.

- Fiore AM, Jacob DJ, Bey I, Yantosca RM, Field BD, Fusco AC and Wilkinson JG (2002) Background ozone over the United States in summer: Origin, trend, and contribution to pollution episodes. *Journal of Geophysical Research: Atmospheres* 107 (D15), ACH-11.
- Fowler D, Pilegaard K, Sutton M, Ambus P, Raivonen M, Duyzer J, Simpson D, Fagerli H, Fuzzi S, Schjoerring J, et al. (2009) Atmospheric composition change: Ecosystems–atmosphere interactions. *Atmospheric Environment* 43(33), 5193–5267.
- Galindo N, Varea M, Gil-Moltó J, Yubero E and Nicolás J (2011) The influence of meteorology on particulate matter concentrations at an urban mediterranean location. *Water, Air, & Soil Pollution* 215(1), 365–372.
- Garrido-Perez JM, Ordóñez C, Garcia-Herrera R and Schnell JL (2019) The differing impact of air stagnation on summer ozone across Europe. *Atmospheric Environment* 219, 117062.
- Im U, Bianconi R, Solazzo E, Kioutsioukis I, Badia A, Balzarini A, Baró R, Bellasio R, Brunner D, Chemel C, et al. (2015) Evaluation of operational on-line-coupled regional air quality models over Europe and North America in the context of AQMEII phase 2. Part i: Ozone. *Atmospheric Environment* 115, 404–420.
- Iordache S, Dunea D, Lungu E, Predescu L, Dumitru D, Ianache C and Ianache R (2015) A cyberinfrastructure for air quality monitoring and early warnings to protect children with respiratory disorders. In *2015 20th International Conference on Control Systems and Computer Science*. Bucharest, Romania: IEEE, pp. 789–796.
- Ivatt PD and Evans MJ (2020) Improving the prediction of an atmospheric chemistry transport model using gradient-boosted regression trees. *Atmospheric Chemistry and Physics* 20(13), 8063–8082.
- Jacob DJ, Heikes E, Fan S-M, Logan JA, Mauzerall D, Bradshaw J, Singh H, Gregory G, Talbot R, Blake D, et al. (1996) Origin of ozone and NO_x in the tropical troposphere: A photochemical analysis of aircraft observations over the South Atlantic basin. *Journal of Geophysical Research: Atmospheres* 101, 24235–24250.
- Jayalakshmi T and Santhakumaran A (2011) Statistical normalization and back propagation for classification. *International Journal of Computer Theory and Engineering* 3(1), 1793–8201.
- Ji Y, Zhou Z, Liu H and Davuluri RV (2021) Dnabert: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 37(15), 2112–2120.
- Kelly FJ, Fuller GW, Walton HA and Fussell JC (2012) Monitoring air pollution: Use of early warning systems for public health. *Respirology* 17(1), 7–19.
- Kim S-Y, Kim E and Kim WJ (2020) Health effects of ozone on respiratory diseases. *Tuberculosis and Respiratory Diseases* 83 (Supple 1), S6.
- Kleinert F, Leufen LH and Schultz MG (2021) Intellio3-ts v1. 0: A neural network approach to predict near-surface ozone concentrations in Germany. *Geoscientific Model Development* 14(1), 1–25.
- Laña I, Del Ser J, Padró A, Vélez M and Casanova-Mateo C (2016) The role of local urban traffic and meteorological conditions in air pollution: A data-based case study in Madrid, Spain. *Atmospheric Environment* 145, 424–438.
- Lei H, Wuebbles DJ and Liang X-Z (2012) Projected risk of high ozone episodes in 2050. *Atmospheric Environment* 59, 567–577.
- Li H, Wu S, Pan L, Xu J, Shan J, Yang X, Dong W, Deng F, Chen Y, Shima M, et al. (2018) Short-term effects of various ozone metrics on cardiopulmonary function in chronic obstructive pulmonary disease patients: Results from a panel study in Beijing, China. *Environmental Pollution* 232, 358–366.
- Li M, Yao Y, Simmonds I, Luo D, Zhong L and Chen X (2020) Collaborative impact of the NAO and atmospheric blocking on European heatwaves, with a focus on the hot summer of 2018. *Environmental Research Letters* 15(11), 114003.
- Li Q, Jacob DJ, Bey I, Palmer PI, Duncan BN, Field BD, Martín RV, Fiore AM, Yantosca RM, Parrish DD, et al. (2002) Transatlantic transport of pollution and its effects on surface ozone in Europe and North America. *Journal of Geophysical Research: Atmospheres* 107(D13), ACH-4.
- Lim B, Arik SÖ, Loeff N and Pfister T (2021) Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* 37(4), 1748–1764.
- Lin M, Fiore AM, Horowitz LW, Langford AO, Oltmans SJ, Tarasick D and Rieder HE (2015) Climate variability modulates western us ozone air quality in spring via deep stratospheric intrusions. *Nature Communications* 6(1), 7105.
- Lombardozi D, Levis S, Bonan G, Hess P and Sparks J (2015) The influence of chronic ozone exposure on global carbon and water cycles. *Journal of Climate* 28(1), 292–305.
- Ma J, Li Z, Cheng JC, Ding Y, Lin C and Xu Z (2020) Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network. *Science of the Total Environment* 705, 135771.
- Malley CS, Henze DK, Kuylenstierna JC, Vallack HW, Davila Y, Anenberg SC, Turner MC and Ashmore MR (2017) Updated global estimates of respiratory mortality in adults 30 years of age attributable to long-term ozone exposure. *Environmental Health Perspectives* 125(8), 087021.
- Mockus J (2012) *Bayesian Approach to Global Optimization: Theory and Applications*, Vol. 37. Berlin: Springer Science & Business Media.
- Monks PS (2000) A review of the observations and origins of the spring ozone maximum. *Atmospheric Environment* 34(21), 3545–3561.
- Murray CJ, Aravkin AY, Zheng P, Abbafati C, Abbas KM, Abbasi-Kangevari M, Abd-Allah F, Abdelalim A, Abdollahi M, Abdollahpour I, et al. (2020) Global burden of 87 risk factors in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019. *The Lancet* 396(10258), 1223–1249.
- Neal L, Agnew P, Moseley S, Ordóñez C, Savage N and Tilbee M (2014) Application of a statistical post-processing technique to a gridded, operational, air quality forecast. *Atmospheric Environment* 98, 385–393.

- Nuvolone D, Petri D and Voller F** (2018) The effects of ozone on human health. *Environmental Science and Pollution Research* 25(9), 8074–8088.
- Oliver RJ, Mercado LM, Sitch S, Simpson D, Medlyn BE, Lin Y-S and Folberth GA** (2018) Large but decreasing effect of ozone on the European carbon sink. *Biogeosciences* 15(13), 4245–4269.
- Olaszyna K, Luria M and Meagher J** (1997) The correlation of temperature and rural ozone levels in Southeastern USA. *Atmospheric Environment* 31(18), 3011–3022.
- Orru H, Åström C, Andersson C, Tamm T, Ebi KL and Forsberg B** (2019) Ozone and heat-related mortality in Europe in 2050 significantly affected by changes in climate, population and greenhouse gas emission. *Environmental Research Letters* 14(7), 074013.
- Otero N, Jurado O, Butler T and Rust HW** (2022) The impact of atmospheric blocking on the compounding effect of ozone pollution and temperature: A copula-based approach. *Atmospheric Chemistry and Physics Discussions* 22, 1905–1919.
- Porter WC and Heald CL** (2019) The mechanisms and meteorological drivers of the summertime ozone–temperature relationship. *Atmospheric Chemistry and Physics* 19(21), 13367–13381.
- Porter WC, Heald CL, Cooley D and Russell B** (2015) Investigating the observed sensitivities of air-quality extremes to meteorological drivers via quantile regression. *Atmospheric Chemistry and Physics* 15(18), 10349–10366.
- Rasmussen D, Hu J, Mahmud A and Kleeman MJ** (2013) The ozone–climate penalty: Past, present, and future. *Environmental Science & Technology* 47(24), 14258–14266.
- Runge J, Nowack P, Kretschmer M, Flaxman S and Sejdinovic D** (2019) Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances* 5(11), eaau4996.
- Sayed A, Choi Y, Esлами E, Lops Y, Roy A and Jung J** (2020) Using a deep convolutional neural network to predict 2017 ozone concentrations, 24 hours in advance. *Neural Networks* 121, 396–408.
- Schnell JL and Prather MJ** (2017) Co-occurrence of extremes in surface ozone, particulate matter, and temperature over eastern North America. *Proceedings of the National Academy of Sciences* 114(11), 2854–2859.
- Schnell JL, Prather MJ, Josse B, Naik V, Horowitz LW, Zeng G, Shindell DT and Faluvegi G** (2016) Effect of climate change on surface ozone over North America, Europe, and East Asia. *Geophysical Research Letters* 43(7), 3509–3518.
- Schultz MG, Schröder S, Lyapina O, Cooper OR, Galbally I, Petropavlovskikh I, Von Schneidmesser E, Tanimoto H, Elshorbany Y, Naja M, et al.** (2017) Tropospheric ozone assessment report: Database and metrics data of global surface ozone observations. *Elementa: Science of the Anthropocene* 5, 58.
- Silva RA, West JJ, Zhang Y, Anenber SC, Lamarque J-F, Shindell DT, Collins WJ, Dalsoren S, Faluvegi G, Folberth G, et al.** (2013) Global premature mortality due to anthropogenic outdoor air pollution and the contribution of past climate change. *Environmental Research Letters* 8(3), 034005.
- Sitch S, Cox P, Collins W and Huntingford C** (2007) Indirect radiative forcing of climate change through ozone effects on the land-carbon sink. *Nature* 448(7155), 791–794.
- Sun HZ, Yu P, Lan C, Wan MW, Hickman S, Murulitharan J, Shen H, Yuan L, Guo Y and Archibald AT** (2022) Cohort-based long-term ozone exposure-associated mortality risks with adjusted metrics: A systematic review and meta-analysis. *The Innovation* 3, 100246.
- Thompson A, Pickering K, McNamara D, Schoeberl M, Hudson R, Kim J, Browell E, Kirchhoff V and Nganga D** (1996) Where did tropospheric ozone over Southern Africa and the tropical Atlantic come from in October 1992? Insights from TOMS, GTE TRACE A, and SAFARI 1992. *Journal of Geophysical Research: Atmospheres* 101(D19), 24251–24278.
- Thompson ML, Reynolds J, Cox LH, Guttorp P and Sampson PD** (2001) A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment* 35(3), 617–630.
- Uzan L, Egert S, Khain P, Levi Y, Vadislavsky E and Alpert P** (2020) Ceilometers as planetary boundary layer height detectors and a corrective tool for COSMO and IFS models. *Atmospheric Chemistry and Physics* 20(20), 12177–12192.
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł and Polosukhin I** (2017) Attention is all you need. *Advances in Neural Information Processing Systems* 30.
- Wang B, Shugart HH, Shuman JK and Lerdau MT** (2016) Forests and ozone: Productivity, carbon storage and feedbacks. *Scientific Reports* 6(1), 22133.
- Wen R, Torkkola K, Narayanaswamy B and Madeka D** (2017) A multi-horizon quantile recurrent forecaster. Preprint, arXiv: 1711.11053.
- Weng X, Forster G and Nowack P** (2022) A machine learning approach to quantify meteorological drivers of recent ozone pollution in China from 2015 to 2019. *Atmospheric Chemistry and Physics* 22, 8385–8402.
- Weng X, Li J, Forster GL and Nowack P** (2023) Large modeling uncertainty in projecting decadal surface ozone changes over city clusters of China. *Geophysical Research Letters* 50(9), e2023GL103241.
- WHO, & ECE** (2021) *Who Global Air Quality Guidelines: Particulate Matter (PM_{2.5} and PM₁₀), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. World Health Organization. Available at <https://www.who.int/publications/i/item/9789240034228> (accessed 13 November 2023).
- Yang C, Yang H, Guo S, Wang Z, Xu X, Duan X and Kan H** (2012) Alternative ozone metrics and daily mortality in Suzhou: The China air pollution and health effects study (CAPES). *Science of the Total Environment* 426, 83–89.
- Zhang L, Jacob DJ, Boersma K, Jaffe D, Olson J, Bowman K, Worden J, Thompson A, Avery M, Cohen RC, et al.** (2008) Transpacific transport of ozone pollution and the effect of recent Asian emission increases on air quality in North America: An integrated analysis using satellite, aircraft, ozonesonde, and surface observations. *Atmospheric Chemistry and Physics* 8(20), 6117–6136.

A. Appendix

A.1. Features from the TOAR dataset

Table A1 describes the data used as features for the machine learning model. The features are split into static and dynamic features. Static features describe the characteristics of a particular station, while dynamic features vary through time. Due to the large size and relative completeness of our dataset, imputing missing values was deemed unnecessary, and rows with missing data were dropped. The way these features are ingested by the TFT are described by Figure A.4. The static features are used as shown, and the dynamic features (apart from ozone) are treated as known future inputs. The reanalysis data are taken in this case to be a proxy for a meteorological forecast, which would be used operationally. All features were scaled with robust scaling (Equation A.1), following testing with standard and min-max scaling on the validation data. Ozone was log-transformed, as was planetary boundary layer height, which improved model performance.

$$\text{Scaled} = \frac{\text{original} - \text{median}}{\text{IQR}} \quad (\text{A.1})$$

Table A1. Relevant data extracted from the TOAR database

Variable name	Description
<i>Static</i>	
station type	Characterization of site, e.g., “background,” “industrial,” and “traffic”
landcover	The dominant IGBP landcover classification at the station location extracted from the MODIS MCD12C1 dataset (original resolution: 0.05 degrees)
toar category	A station classification for the Tropospheric Ozone Assessment Report based on the station proxy data that are stored in the database. One of unclassified, low elevation rural, high elevation rural, or urban
pop density	Year 2010 human population per square km from CIESIN GPW v3 (original horizontal resolution: 2.5 arc minutes)
max 5km pop density	Maximum population density in a radius of 5 km around the station location
max 25km pop density	Maximum population density in a radius of 25 km around the station location
nightlight 1km	Year 2013 Nighttime lights brightness values from NOAA DMSP (original horizontal resolution: 0.925 km)
nightlight max 25km	Year 2013 Nighttime lights brightness values (original horizontal resolution: 5 km)
alt	Altitude of station (in m above sea level). Best estimate of the station altitude, which frequently uses the elevation from Google Earth
station etopo alt	Terrain elevation at the station location from the 1 km resolution ETOPO1 dataset
nox emi	Year 2010 NOx emissions from EDGAR HTAP inventory V2 in units of $\text{gm}^{-2}\text{yr}^{-1}$ (original resolution: 0.1 degrees)
omi nox	Average 2011-2015 tropospheric NO ₂ columns from OMI at 0.1 degree resolution (Env. Canada) in units of 10^{15} molecules cm^{-2}
<i>Dynamic</i>	
o3 (forecasted variable)	Ozone concentration, daily maximum 8-hour average statistics according to the EU definition of the daily 8-hour window starting from 17 h of the previous day. Measured at the station, with UV absorption
cloudcover	Daily average cloud cover from ERA5 reanalysis for the grid cell containing a particular station

Continued

Table A1. Continued

Variable name	Description
relhum	Daily average relative humidity from ERA5 reanalysis for the grid cell containing a particular station
press	Daily average pressure from ERA5 reanalysis for the grid cell containing a particular station
temp	Daily average temperature from ERA5 reanalysis for the grid cell containing a particular station
v	Daily average meridional wind speed from ERA5 reanalysis for the grid cell containing a particular station
u	Daily average zonal wind speed from ERA5 reanalysis for the grid cell containing a particular station
pblheight	Daily average planetary boundary layer height from ERA5 reanalysis for the grid cell containing a particular station

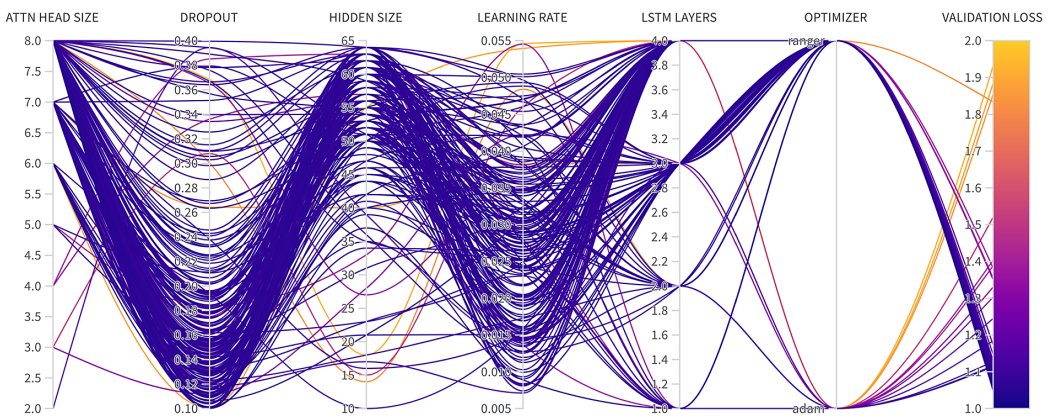


Figure A.1. Plot illustrating the hyperparameter optimization and the skill, in terms of the loss on the validation data, of various hyperparameter combinations for the TFT.

A.2. Model hyperparameter optimization

Table A2 details the hyperparameters used for the TFT model. These hyperparameters were determined with Bayesian optimization, implemented with weights and biases (Biewald, 2020). Bayesian optimization is a method used to determine the optimal hyperparameters of a model, by defining a Gaussian process that describes the function controlling the performance of the model with respect to the hyperparameters. This function is evaluated and updated as the hyperparameter space is explored, as shown in Figure A.1. We found that the most influential hyperparameters were the learning rate and the choice of optimizer. Furthermore, we also evaluated the optimal number of past time-steps necessary for skillful prediction. While increasing the number of past time-steps past 21 days did improve performance marginally, there was little improvement relative to the increased computational burden. Additionally, decreasing the number of time-steps below 21 days did hinder performance considerably, especially at very low numbers (fewer than 5 days). The hyperparameters of the ridge regression, random forest, and LSTM models were optimized with grid search.

Table A2. Hyperparameters for the final TFT and LSTM models used for model evaluation

Model	Hyperparameter value
<i>TFT</i>	
attention head size	8
dropout	0.110
hidden continuous size	14
hidden size	50
learning rate	0.0099
lstm layers	3
optimizer	Ranger
<i>LSTM</i>	
dropout	0.100
hidden size	100
learning rate	0.023
layers	5
optimizer	Ranger

Note. We carried out 120 runs of Bayesian optimization for the TFT.

A.3. Performance across countries

The performance of the optimized model in forecasting O₃ in different countries is given in Figure A.2 below. While some European countries had very little data available, others provided a good quantity of data. Performance is relatively good across countries, with some poorer skill in Norway (which has little data) and Portugal.

A.4. Comparison of model performance in spring and summer across countries

To evaluate the processes that might be poorly represented by our model, we tested the trained model across all countries, using just spring and then just summer data. We found that in general, the model performance was poorer in summer than in spring. Furthermore, we found that there was little correlation between the latitude of the country and the difference in performance between spring and summer. The findings are described in Figure 3; however, an analysis of the spring–summer difference in just the UK, France, and Italy dataset is provided in Figure A.3. In terms of RMSE and MAE, we found the performance of the TFT was better in spring but poorer in terms of R².

A.5. Architecture of the temporal fusion transformer

The temporal fusion transformer (Lim et al., 2021) is designed to carry out multi-horizon (multiple time-step) forecasting of a target variable using both static and dynamic covariates. The fundamental setup of this forecasting task is described in Figure A.4. Both dynamic and static inputs are passed to the model, which then makes forecasts.

The TFT is an attention-based ML model that uses the attention mechanism (Vaswani et al., 2017) to learn long-term dependencies in data, combined with recurrent layers to learn short-term dependencies. The TFT also deploys gating layers to minimize the effect of less relevant predictors. The architecture of the model is given in Figure A.5.

As described in Lim et al. (2021), the TFT offers a number of key features to improve upon existing models for multi-horizon forecasting.

The first is the use of gated residual networks (GRNs) to suppress nonlinear relationships in the model. Variable selection networks are deployed to identify and reduce the effect of poor predictors. Inputs are passed first through a GRN, followed by a Softmax layer, a function that converts the output vector to a probability distribution, to yield weights describing how relevant a feature is for predicting an output. The original features are then weighted by the variable selection weights to yield appropriately weighted features. Static covariates (the covariates which do not vary through time, such as landcover), are encoded with GRNs to produce 4 context vectors, which are then combined with the dynamic features. The TFT then uses a long short-term memory network (LSTM) encoder–decoder to provide an alternative to the positional embeddings used in traditional transformer architectures. The context vectors from the static covariates are used to initialize the cell and hidden state (a representation of the previous data in the sequence seen by the model) in the first LSTM in the layer.

The TFT then uses multi-head attention, as detailed in Vaswani et al. (2017), which allows the model to learn long-term dependencies in data. The weights of the attention mechanism allow us to interpret that features and previous time-steps that are the

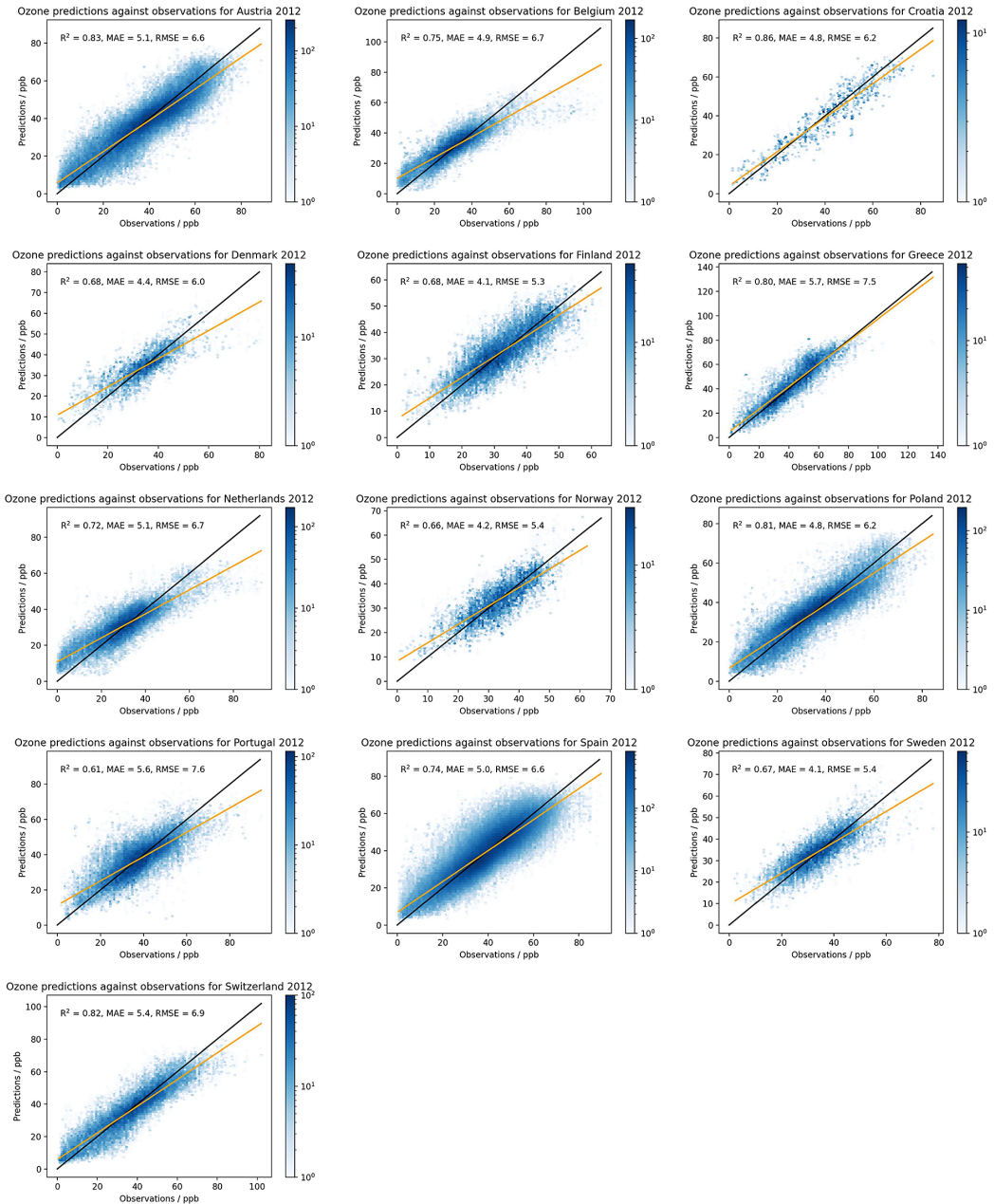


Figure A.2. Plots illustrating the skill of the model in predicting ozone in different countries across Europe.

most important for prediction, aiding model interpretability. The decoder deploys masking to ensure that it can only attend to features preceding it. The outputs from the attention layer are then passed to a feed forward GRN, which also takes inputs via a gated residual connection that skips the attention layer providing a simpler model if the attention layer is unnecessary for accurate predictions (this behavior is learned during training), as shown in Figure A.5. Finally, prediction intervals from the outputs of the final GRN are generated by making predictions at different quantiles, using a linear transformation of the output.

The TFT is trained by minimizing the sum of the quantile loss across quantile outputs.

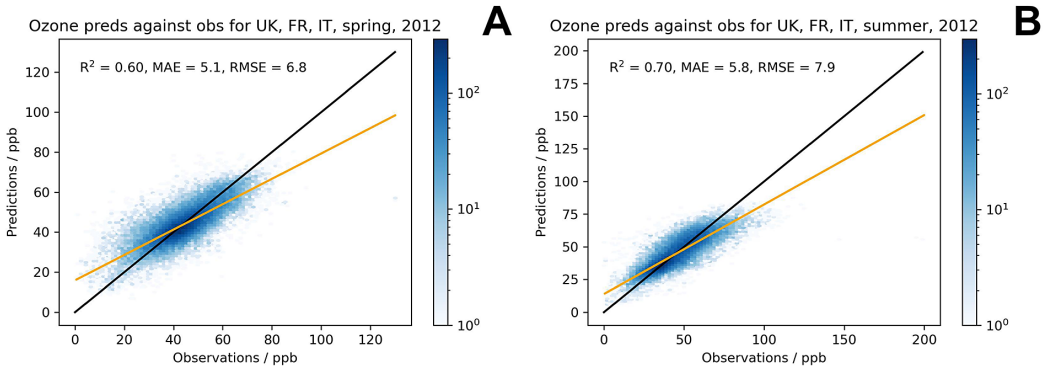


Figure A.3. (a) Performance of the TFT when predicting on 2012 spring test data from the UK, France, and Italy. (b) Performance of the TFT when predicting on 2012 summer test data from the UK, France, and Italy.

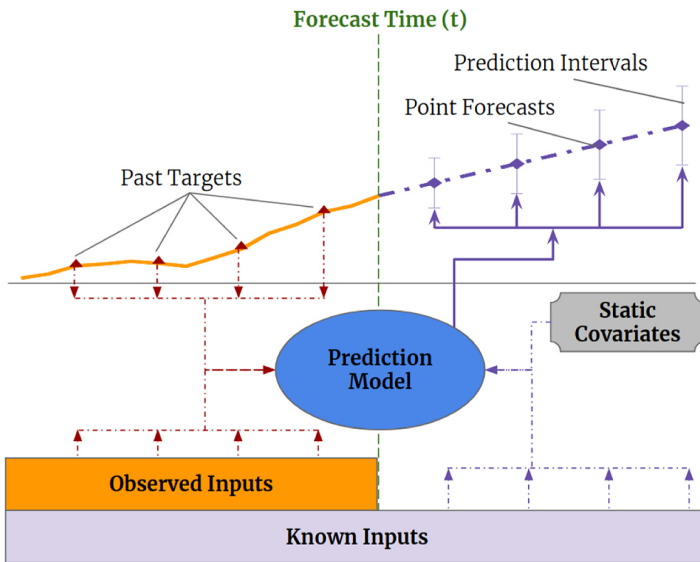


Figure A.4. Setup of a typical multi-horizon forecasting problem. Source: 'Temporal Fusion Transformers for interpretable multi-horizon time series forecasting', Lim et al. (2021), licensed under CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>).

A.6. Attention in the temporal fusion transformer

The attention values in the TFT are calculated as in Lim et al. (2021). In the TFT architecture, standard multi-head attention is additively aggregated to interpretable multi-head attention. In this formulation, each attention head learns different temporal relationships from the same set of input features, and then the aggregation of these attention weights can be viewed as an ensemble, allowing interpretability studies. We refer the reader to Lim et al. (2021) for full details.

In the context of this study, we generally see that the model, as expected, pays more attention to the more recent past days when making future forecasts and that the model is able to attend to recent low days of ozone when making forecasts of low ozone and, similarly, to recent days of high ozone when forecasting future high ozone.

Since the attention values for the dynamic and the static features are calculated separately, we include the feature importances for the static variables here.

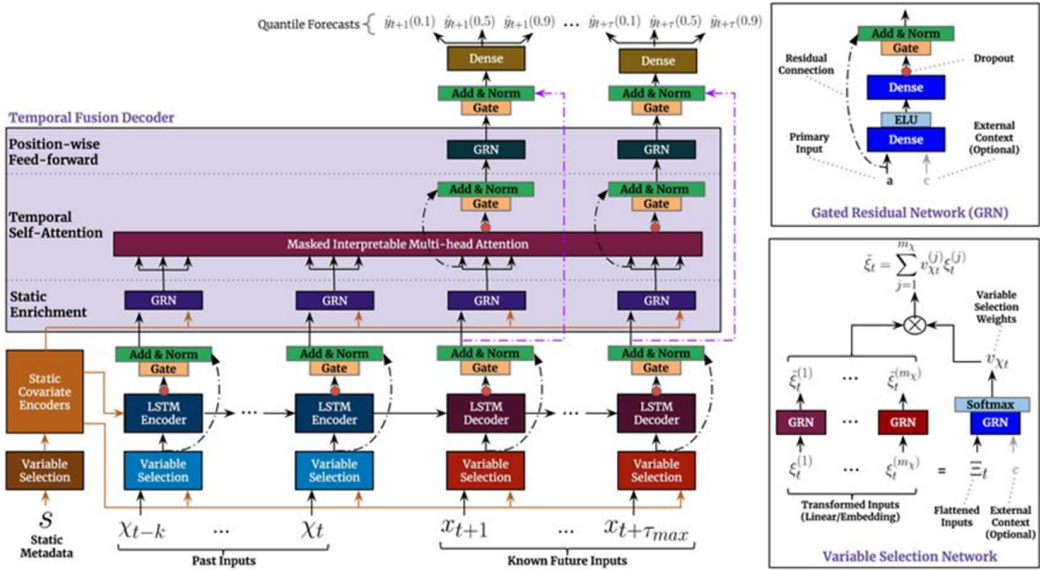


Figure A.5. Architecture of the temporal fusion transformer model. The model consists of a combination of RNN encoders, followed by an attention layer, and then a fully connected decoder layer. Source: 'Temporal Fusion Transformers for interpretable multi-horizon time series forecasting', Lim et al. (2021), licensed under CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>).

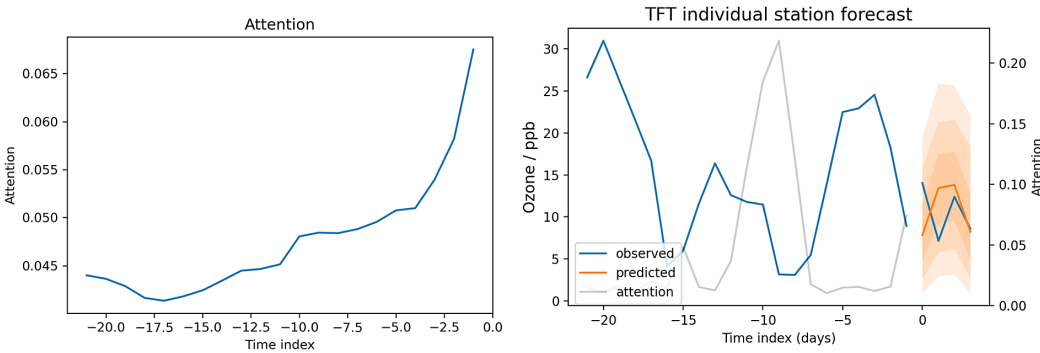


Figure A.6. Left-hand plot shows the attention paid to different days in the past, averaged over the whole test set. The right-hand plot illustrates an example of the model attending to previous low ozone days when making forecasts of future low ozone. This example is from a station in one of the unseen countries.

A.7. Quantile loss

The quantile loss function is used in this work to extract prediction intervals from the TFT model. The quantile loss function is given as follows:

$$L_q(y, \hat{y}) = q(y - \hat{y})_+ + (1 - q)(\hat{y} - y)_+ \tag{A.2}$$

where $(\cdot)_+$ is equal to $\max(0, \cdot)$ (Wen et al., 2017). Note that when $q = 0.5$, the quantile loss is the same as using the mean absolute error loss function. The model is trained to find the weights of the model that minimize the total loss of the quantile loss over various values of q for various prediction horizons.

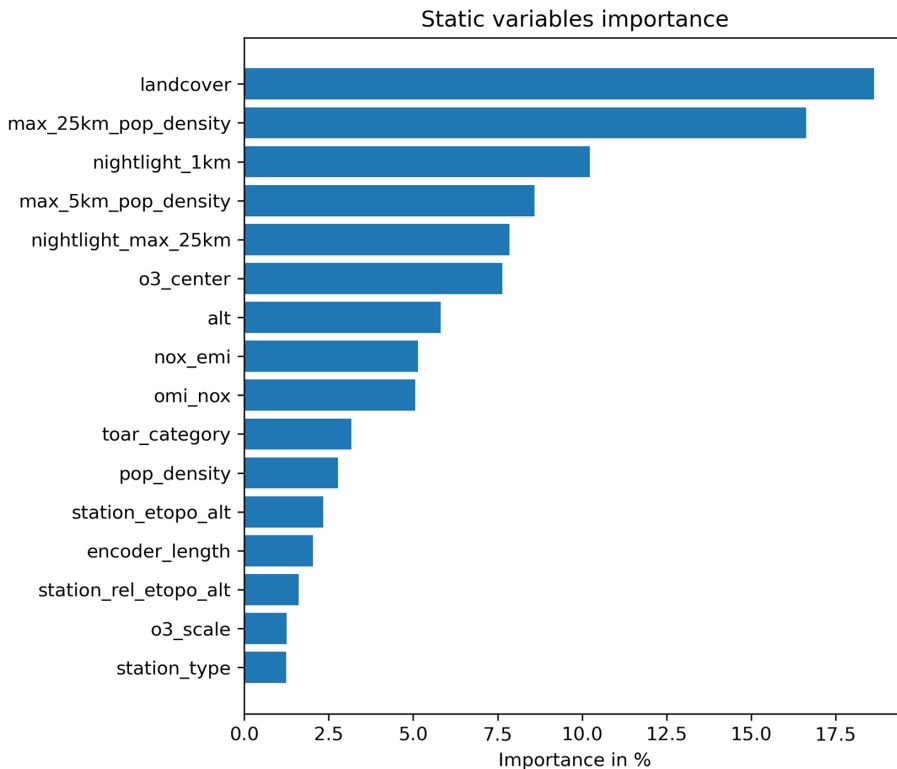


Figure A.7. Plot illustrating the feature importances for the static features.

A.8. Benchmark models

We used 4 other approaches to compare with the TFT. The first of these is a persistence model, which simply predicts the next timestep of ozone to be equal to the previous time-step. This provides our baseline model.

We also evaluated a ridge regression model. Ridge regression (Tikhonov regularization) is often deployed in multiple regression tasks where covariates are correlated.

$$(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta \quad (3)$$

Ridge regression uses the L_2 penalty on the coefficients of the model, similar to LASSO regression, which uses the L_1 penalty. Ridge regression is a standard benchmark used in the environmental sciences but is a linear method, which limits its flexibility.

Random forests are a widely used nonlinear ensemble machine learning technique, which construct multiple decision trees during training, using bootstrap aggregating (bagging) to reduce instability during training. Bagging generates new training datasets from the original training data by sampling from the original data, with replacement. Decision trees are then trained on these new training sets, and the output of the model is determined by aggregating the predictions of the individual trees on the new training datasets. Random forests are prone to overfitting training data, which can lead to poor out-of-distribution performance.

While random forests are a nonlinear model and therefore may be able to better fit data with nonlinear relationships, they do not necessarily exploit temporal dependencies in data well.

Long short-term memory networks are a form of recurrent neural network that have been shown to perform very well on time series data, aided by the use of memory cells which improve on vanilla recurrent neural networks. However, recent work in fields such as natural language processing has illustrated that transformers perform better on sequential data, in part due to their capacity to model long-term dependencies in data. Furthermore, the temporal fusion transformer is adapted to ingest both static and dynamic features, unlike the other models, which provides a small improvement in model skill.